Balance Assessment of Matched Data with Multiple Treatment Levels

Yiling Huang

Department of Statistics

University of Michigan

Supervisor

Mark M. Fredrickson Ph.D.

In partial fulfillment of the requirements for the degree of

Bachelor of Science in Statistics

April 2022

Acknowledgements

I would like to acknowledge many people who, directly or indirectly, contributed to this thesis.

First, I would like to express my deepest gratitude to Dr. Mark Fredrickson who has been a brilliant mentor. This thesis would not have been possible without his consistent help and insightful suggestions. He set me on the path of statistics research and introduced to me the field of causal inference, which I am very passionate about. Being his student and mentee has been one of the greatest pleasures in my college experience.

I am very thankful to the Department of Statistics at the University of Michigan, for organizing the Undergraduate Research Program in Statistics (URPS). Getting involved in a real research is such a rare experience for statistics undergraduate students, and I am very fortunate that our department offers this opportunity.

Beyond those who contributed directly to the project, I also want to thank many of my peers at the university. This includes my friends who, through the meaningful discussions and companionships, contributed to a huge portion of my happiness. I especially want to thank my girlfriend, who has always been by my side, for her unconditional support and encouragement throughout the whole project, and for being like a family member.

Finally, I would like to sincerely acknowledge my parents for raising me to be who I am, educating me to be curious and skeptical about the world, and encouraging me to be confident enough to take up any challenges in my life. Anything I have achieved would not have been possible without their efforts, and I constantly feel blessed for being their child.

Abstract

Identifying and estimating causal effects of treatments is of significant research interest. In doing so, similar data are oftentimes matched into one stratum, and subsequent inferences of causality are carried out based on these strata. In particular, when the data are from observational studies, properly matching observations by their treatment assignment probabilities are especially important for removing potential selection bias induced by selecting observations that receive specific treatments in a non-randomized fashion. Therefore, it is an important task to evaluate whether matching was done properly, that is, whether the covariates are equally distributed in different treatment groups given the matching information. Traditional methods of matching evaluation involve visually investigating summary statistics, such as the standardized mean difference, by covariate, but lack uncertainty quantification of the conclusion and are less convenient compared to an omnibus test that checks matching validity for all covariates one-shot. We propose a hypothesis test that expresses treatment assignment probabilities by an adjacent category logistic regression model and provides an omnibus test of matching for all covariates by testing the global null $\beta = \mathbf{0}$ in the language of regression models. In this thesis, we adopt a χ^2 approximation of the asymptotic distribution of the test statistic, inspired by the Rao score test. An application of the test indicates the matching results produced by a matching algorithm can be further improved.

Contents

Chapter 1

Introduction

Valid identification and estimation of causal effects of treatments is a critical task in scientific research (Holland 1986). In modern scientific research areas, when investigators intend to discover causal relationships between events, the ideal and widely accepted solution is to conduct randomized controlled trials (RCTs), where research investigators have the discretion to randomly assign treatments to participants of the trial so that all confounding factors are controlled and they may thus attribute differences in the response variable entirely to the treatment assignment. However, we do not always have the luxury of setting up a randomized controlled trial and collecting data from it. Due to factors such as costs and ethical concerns, we must rely on observational data instead (Rosenbaum 2010). On the positive side, using observational data to infer causality allows us to observe objects and events in their natural environments, which is more realistic compared to RCTs that are set up artificially. However, one caveat that might arise from using such observational data to make causal claims is that in a real-world scenario, treatment assignments and the response variables may be systematically related, and we may not have information about the mechanism behind the treatment assignments. Consequently, without being able to control the confounders that relate to both the treatment assignment and the response variable of interest, we may be subject to biases in our reasoning.

For this reason, it is important to use observational data to resemble controlled trials, where the pretreatment covariates in different treatment groups are roughly equally distributed, i.e., balanced (Hansen and Bowers 2008). Rosenbaum and Rubin (1983) showed, when the treatment is binary, the propensity score, i.e. the probability of receiving treatment, can serve as a *balancing score*, which equivalently means that the treatment assignment and covariates are conditionally independent given the propensity score. Moreover, under the *strong ignorability assumption*, matching on propensity scores may yield unbiased estimates of treatment effects (Rosenbaum and Rubin 1983). Some existing literature had generalized the notion of the propensity score to problems where the treatment is other than binary, such as continuous, nominal, or ordinal; and under some mild conditions, treatment effect estimates remain unbiased (Yang et al. 2016; Imai and Van Dyk 2004).

In this work, we are primarily concerned with matched data with multilevel (nominal/ordinal) treatments. Examples of these types of treatments include drugs of different types or manufacturers, and drug doses with different amounts, such as low/medium/high. With proper tools to achieve matching in the covariates, it then becomes relevant to evaluate the balance within these matched results, i.e., whether the covariates are equally distributed in different treatment groups given the matching. When the treatment is binary, the classical textbook method to evaluate matching incorporates procedures such as investigating the marginal distribution of each covariate in the treated and controlled sets, before and after matching (Rosenbaum 2010). However, it is still preferable to give a more quantifiable and statistically meaningful assessment of the matching quality. Instead of using descriptive statistics to assess the covariate balance, Hansen and Bowers (2008) called for an omnibus statistical test that performs balance assessment one-shot. Although they were primarily focused on matching balance assessment in randomized trials, we can easily generalize their notion to observational study because under the null hypothesis that a good matching was done, an observational study should be no different from an experimental study.

To evaluate covariate balance, we motivate our approach using an adjacent category logistic regression model for treatment assignments as a function of a set of properly chosen background covariates. With presumably correctly matched sets of observations, the null hypothesis that the treatment assignment is independent of any covariates could be expressed as the global null $\beta = \mathbf{0}$ in the context of logistic regression. We hereby propose the usage of a statistic inspired by the Rao score test as an omnibus method to assess matching, using likelihood functions conditioning on the observed strata-specific treatment counts to account for stratification. We approximate the asymptotic behavior of the test statistic using a χ^2 distribution to complement the omnibus test pipeline, and prove the limiting distribution holds under mild assumptions. Then, we apply the method to an Ohio observational health study where matching was done for multilevel-treated observations and conclude the method was insufficient to induce balance on all covariates simultaneously. It is to be noted that although the work is motivated by matching for observational data, it is generally applicable to any type of matched data with multiple treatment levels, including matched data from randomized controlled trials, because the null hypothesis of balanced covariates, or valid matching, is generally applicable to data collected by any methods.

Finally, simulation studies show that this method has proper convergence behavior and thus obtains valid Type I error rate under the null hypothesis, and it achieves desirable power when the sample size is relatively large.

Chapter 2

Methods

This chapter contains our main method for covariate balance assessment under the aforementioned scenario, i.e., one with a multilevel treatment variable and matched (grouped) observations. The method is rather straightforward. First assume we were given N observations that are grouped into s different strata based on some similarity metrics, according to any valid matching procedure. We now state our null hypothesis in words, i.e. that the matching is valid. Then this implies that observations in the same strata have the same treatment assignment probabilities. Because the treatment assignment $Y = 1, 2, \dots J$ should be independent of the background covariates $\mathbf{x} \in \mathbb{R}^p$ conditioning on propensity score-based matching information, when modeling the treatment assignment by the background covariates using a regression model, the slope parameters β of the model should satisfy $\beta = \mathbf{0}$.

Therefore the problem gets reduced to testing $\beta = \mathbf{0}$ in a regression model that can account for the fact that the observations have a grouping structure, and within groups, the observations are homogeneous in terms of treatment assignment probabilities. It can be seen later that a slightly modified version of an adjacent category logistic regression model will be sufficient. To derive the Rao score test statistics requires working on the likelihood function yielded by the model, and we further condition on several strata-specific treatment counts to account for the hypothetical setup of a stratified treatment assignment regime. It can be seen in Section 2.5 that we can express the test statistic in a relatively simple form that is easily computable. To perform testing, some existing results are employed to approximate the limiting behavior of the statistic.

We hereby give the outline of the chapter. Section 2.1 introduce the adjacent category logistic regression model. Section 2.2 extends the original regression model to allow for a stratified/grouped design in observations, as is the case in our problem. We construct conditional likelihood of the stratified adjacent category logistic regression model in Section 2.3. Section 2.4 introduces the (generalized) Rao score test as an omnibus testing method to infer global null effects in the model. And finally, we derive the test statistic in Section 2.5, with proofs postponed to Appendix A.

2.1 Adjacent Category Logistic Regression

In this model, let Y be the treatment variable with possible outcome labels $1, 2, 3, ..., J$, serving as aliases of the original treatments levels, and denote $\pi_j(\mathbf{x}) = P(Y = j)$, with the probability implicitly conditioning on the covariates $\mathbf{x}$, which are background variables that we deem as relevant to the natural mechanism that influences treatment assignment in a non-experimental setting. The model is given as

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + {\beta_j}^T \mathbf{x}, \quad j = 1, 2, \ldots, J-1$$

With some simple algebra, we observe that

$$\pi_j(\mathbf{x}) = \begin{cases} \frac{\exp\{\sum_{k=j}^{J-1}(\alpha_k + \beta_k^T \mathbf{x})\}}{1+\sum_{i=1}^{J-1}\exp\{\sum_{k=i}^{J-1}(\alpha_k+\beta_k^T\mathbf{x})\}}, & j < J \\ \frac{1}{1+\sum_{i=1}^{J-1}\exp\{\sum_{k=i}^{J-1}(\alpha_k+\beta_k^T\mathbf{x})\}}, & j = J \end{cases} \tag{2.1}$$

Then, to get the likelihood function, we denote the observed sample by

$$X = [x_{ij}]_{n\times p}\ , \quad Y = [Y_i]_{n\times 1}, \quad n : \text{sample size}, \quad p : \text{number of covariates}$$

We let $t_i = [t_{i,j}]_{1\times J} = \begin{bmatrix} t_{i,1} & t_{i,2} & \ldots & t_{i,J} \end{bmatrix}$ be the set of indicator variables representing the category that y_i falls in, where the j-th element of t_i is the individual indicator of whether the i-th response falls into the category j; it follows that $\sum_{j=1}^{J} t_{i,j} = 1,\ \forall i = 1, \ldots, n$. Consequently, we have the equivalent expression for the likelihood of observing a certain result at the ith observation

$$P(Y_i = y_i) = \prod_{j=1}^{J} \pi_j(\mathbf{x_i})^{t_{i,j}}, \quad t_{i,j} = \begin{cases} 1, & j = y_i \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

2.2 Adjacent Category Logistic Regression with Stratification

In our working scenario, the data were grouped into strata based on information characterizing their probability distributions of receiving each treatment. Hence, it is of our central interest to extend our previous work on adjacent category logistic model to a situation where the data is stratified into s groups. Here, we similarly let Y be the response with possible outcome labels $1, 2, 3, ..., J$. Moreover, let $b = 1, ..., s$ be the indices of the strata. Now, for any realization $\mathbf{x}$ that belongs to the b-th strata, we denote its assignment probability as $\pi_j(\mathbf{x}) = P(Y = j)$, with the probability implicitly conditioning on the background covariates $\mathbf{x}$ and the strata b. The model is given as

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_{jb} + {\beta_j}^T \mathbf{x}, \quad j = 1, 2, \ldots, J-1$$

Here, we introduced a different intercept for each stratum to take into account the fact that the data is stratified based on how their assignment probabilities differ and that data points within the same stratum have the same assignment probability. And it is to be noted that it is sufficient to only introduce a strata-specific design for the intercepts, but not the slopes. To adopt a common slope over strata is equivalent to assuming that the effects of the background covariates on the treatment assignments behave uniformly across strata. However, it is exactly the case in our null hypothesis that the background covariates have null effect $\beta = \mathbf{0}$ in each stratum, so assuming a uniform slope to characterize the effect of the background variables is sufficient, convenient, and more logical in our situation. Given this model, we analogously observe that

$$\pi_j(\mathbf{x}) = \begin{cases} \frac{\exp\{\sum_{k=j}^{J-1}(\alpha_{kb}+\beta_k^T\mathbf{x})\}}{1+\sum_{i=1}^{J-1}\exp\{\sum_{k=i}^{J-1}(\alpha_{kb}+\beta_k^T\mathbf{x})\}}, & j < J \\ \frac{1}{1+\sum_{i=1}^{J-1}\exp\{\sum_{k=i}^{J-1}(\alpha_{kb}+\beta_k^T\mathbf{x})\}}, & j = J \end{cases} \tag{2.3}$$

As in the non-stratified case, we denote the observed sample by

$$X = [x_{ij}]_{n\times p}\ , \quad Y = [Y_i]_{n\times 1}, \quad n : \text{sample size}, \quad p : \text{number of covariates}$$

And we defined $t_i = [t_{i,j}]_{1\times J} = \begin{bmatrix} t_{i,1} & t_{i,2} & \dots & t_{i,J} \end{bmatrix}$ to be the set of indicator functions representing the category that y_i falls in, where the j-th element of t_i is the individual indicator of whether the i-th response falls into the category j; it follows that $\sum_{j=1}^{J} t_{i,j} = 1,\ \forall i = 1, \dots, n$. Also, since we have s different strata, each owns some observations, we use $I(b)$ to denote the set of indices of observations belonging to the b-th stratum, $b = 1, 2, ..., s$. Therefore, as an extension of (2.2), we have the following expression for the likelihood of the data:

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{b=1}^{s}\prod_{i\in I(b)}\prod_{j=1}^{J}\pi_j(\mathbf{x_i})^{t_{ij}}$$

Remark 2.2.1. *Equations* (2.1) *and* (2.3) *gives the assignment probabilities under the ordinary adjacent-category logistic model. However, in the actual computation of the likelihood, we will consider a slight variant of the model, i.e., by conditioning on the sufficient statistics of the intercepts, so as to account for stratification within data.*

2.3 Conditional Likelihood Functions for Adjacent Category Logistic Models

Given the models expressing likelihoods of the adjacent category logistic model, we further condition on several sufficient statistics. As can be seen in the stratified case, conditioning on these statistics brings into our model information from the matching (stratification).

Proposition 2.3.1 (Non-stratified Conditional Likelihood)**.** *Conditioning on the sufficient statistics c_j, defined as the realizations of $C_j := \sum_{k=1}^{j}\sum_{i=1}^{n} t_{ik}$, for all $j = 1, 2, \ldots, J$, the conditional likelihood function of the data is expressed as*

$$\begin{aligned} &P(Y_1 = y_1, \ldots, Y_n = y_n \mid \sum_{k=1}^{j}\sum_{i=1}^{n} t_{ik} = c_j,\ \forall j = 1, 2, \ldots, J) \\ =&\frac{\exp\Big\{\sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}(\sum_{k=1}^{j} t_{ik}\mathbf{x_i})\Big]\Big\}}{\sum\limits_{y^*\in S(t)}\exp\Big\{\sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}(\sum_{k=1}^{j} t^*_{ik}\mathbf{x_i})\Big]\Big\}} \end{aligned}$$

where $S(t) = \big\{(y_1^*, y_2^*, \ldots, y_n^*) : \sum_{k=1}^{j}\sum_{i=1}^{n} t_{ik}^* = c_j\ \forall j = 1, \ldots, J\big\}$ denotes the set of all possible categorical outcomes Y, under the constraint that the cumulative sum of observations under each category being the observed value c_j.

Proof. A.1 □

As mentioned previously, we extend the conditioning regime to the stratified model, so that the matching is taken into account in our likelihood.

Proposition 2.3.2 (Stratified Conditional Likelihood)**.** *For a data with strata $b = 1, \ldots, s$, let $I(b)$ denotes the collection of indices of observations belonging to the b-th stratum. Conditioning on the sufficient statistics c_{jb}'s, defined as $c_{jb} = \sum_{k=1}^{j}\sum_{i\in I(b)} t_{ik}$, $j = 1, 2, ..., J;\ b = 1, 2, ..., s$, the conditional likelihood function of the data is expressed as*

$$\begin{aligned} &P(Y_1 = y_1, \ldots, Y_n = y_n \mid \sum_{k=1}^{j}\sum_{i\in I(b)} t_{ik} = c_{jb},\ \forall j = 1, 2, ..., J;\ b = 1, 2, ..., s) \\ =&\frac{\exp\Big\{\sum_{j=1}^{J-1}\beta_j^T\big[\sum_{i=1}^{n}\mathbf{x_i}\big(\sum_{k=1}^{j} t_{ik}\big)\big]\Big\}}{\sum\limits_{y^*\in S(t)}\exp\Big\{\sum_{j=1}^{J-1}\beta_j^T\big[\sum_{i=1}^{n}\mathbf{x_i}\big(\sum_{k=1}^{j} t^*_{ik}\big)\big]\Big\}} \end{aligned}$$

where $S(t) = \big\{(y_1^*, y_2^*, \ldots, y_n^*) : \sum_{k=1}^{j}\sum_{i\in I(b)} t_{ij}^* = c_{jb}\ \forall j = 1, \ldots, J; b = 1, \ldots, s\big\}$ denotes the set of all possible (categorical) outcomes, under the constraint that the strata-wise cumulative sum of observations under each category being the observed value c_{jb}.

Proof. A.4 □

2.4 The Rao Score Test

The Rao score test is paired with our model to statistically quantify the quality of matching (Rao 2005). We first present a general version of the test, which motivates our method for balance testing. The Rao score test states the following

Theorem 2.4.1 (Rao score test)**.** *Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n from the density function $p(x, \theta)$ where θ is a p-vector parameter, and denote the joint density by $P(X, \theta) = p(x_1, \theta) \ldots p(x_n, \theta)$ and the log likelihood by $l(\theta \mid X) = \log P(X, \theta)$. The score vector of p components is defined as*

$$s(\theta) = \frac{\partial}{\partial \theta} l(\theta \mid X) = \left[\frac{\partial}{\partial \theta_1} l(\theta \mid X) \quad \frac{\partial}{\partial \theta_2} l(\theta \mid X) \quad \ldots \quad \frac{\partial}{\partial \theta_p} l(\theta \mid X) \right]^T$$

And the Fisher information matrix is defined as

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} l(\theta \mid X) \right]$$

where $\frac{\partial^2}{\partial \theta^2} l(\theta \mid X)$ is the Hessian of the log likelihood (Lehmann and Casella 2006). Then under $H_0 : \theta = \theta_0$, where θ_0 is a specified p-dimensional real vector,

$$[s(\theta_0)]^T [I(\theta_0)]^{-1} [s(\theta_0)] \xrightarrow{d} \chi^2_p$$

Remark 2.4.1. *Note that the Rao score test statistic behaves in a χ^2 fashion, which directly motivates our approximation.*

Proposition 2.4.1. *Under regularity conditions (B.2.1),*

$$T^2 := [s(\mathbf{0})]^T [I(\mathbf{0})]^- [s(\mathbf{0})] \xrightarrow{d} \chi^2_r$$

where $[I(\mathbf{0})]^-$ is a generalized inverse of $I(\mathbf{0}) = -\mathbb{E}[H(\mathbf{0})] = -H(\mathbf{0})$, $r = rank(I(\theta))$, and $s(\mathbf{0})$, $H(\mathbf{0})$ are score vector and Hessian matrix of the log-likelihood evaluated at $\beta = \mathbf{0}$, respectively. Details of these quantities are given in the next section.

Proof. We postpone the proof to Appendix B. □

Remark 2.4.2. *Proposition 2.4.1 naturally introduces a hypothesis test for the null hypothesis $H_0 : \beta = \mathbf{0}$ for the stratified adjacent category logistic regression model. Chapter 4 is devoted to study the operating characteristics of this test.*

2.5 The Rao Score Test Statistic for Balance Testing

Our main contribution in this work is the derivation of the exact score vector and Hessian matrix of the log-likelihood under stratified and non-stratified designs, where the result in the latter case, as an extension of the former case, will be applied to balance testing. We hereby state our main propositions.

In order to test the null hypothesis $H_0 : \beta_j = 0, \quad \forall j = 1, ..., J-1$ in a quadratic form using Rao's notation, we stack the slope vectors together as a $[p(J-1)] \times 1$ vector β, and the following development of the test statistic relies on this vectorization

$$\beta = \begin{bmatrix} \beta_1^T & | & \beta_2^T & | & \cdots & | & \beta_{J-1}^T \end{bmatrix}^T$$

Proposition 2.5.1. *The score function of the unstratified conditional log-likelihood taken with respect to the j-th block of β and evaluated at $\beta_j = \mathbf{0}$ is*

$$s_j(\mathbf{0}) = \frac{\partial l}{\partial \beta_j}\Big|_{\beta_j=\mathbf{0}} = \left(\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{1} & \mathbf{1} & \cdots & \mathbf{1} \\ | & | & & | \end{bmatrix} \begin{bmatrix} \overline{x_1} & 0 & \cdots & 0 \\ 0 & \overline{x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \overline{x_p} \end{bmatrix} \right)^T \mathbf{z}_j$$

where $\mathbf{x}_j$ is the data vector for the j-th covariate, $\mathbf{1}$ is the n-dimensional all-one vector, $\mathbf{z}_j = \begin{bmatrix} \sum_{k=1}^j t_{1,k} & \sum_{k=1}^j t_{2,k} & \cdots & \sum_{k=1}^j t_{n,k} \end{bmatrix}^T = \begin{bmatrix} \mathbb{1}\{Y_1 \le j\} & \mathbb{1}\{Y_2 \le j\} & \cdots & \mathbb{1}\{Y_n \le j\} \end{bmatrix}^T$, and $\overline{x_c} = \frac{1}{n}\sum_{i=1}^n x_{ic}$.

Proof. A.2 □

Proposition 2.5.2. *The score function of the stratified conditional log-likelihood taken with respect to the j-th block of β and evaluated at $\beta_j = \mathbf{0}$ is*

$$s_j(\mathbf{0}) = \frac{\partial l}{\partial \beta_j}\Big|_{\beta_j=\mathbf{0}} = \left(\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{1}_1 & \mathbf{1}_2 & \cdots & \mathbf{1}_s \\ | & | & & | \end{bmatrix} \begin{bmatrix} \overline{x_1}^{(1)} & \overline{x_2}^{(1)} & \cdots & \overline{x_p}^{(1)} \\ \overline{x_1}^{(2)} & \overline{x_2}^{(2)} & \cdots & \overline{x_p}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x_1}^{(s)} & \overline{x_2}^{(s)} & \cdots & \overline{x_p}^{(s)} \end{bmatrix} \right)^T \mathbf{z}_j$$

where $\mathbf{z}_j = \begin{bmatrix} \sum_{k=1}^j t_{1k} & \sum_{k=1}^j t_{2k} & \cdots & \sum_{k=1}^j t_{nk} \end{bmatrix}^T = \begin{bmatrix} \mathbb{1}\{Y_1 \le j\} & \mathbb{1}\{Y_2 \le j\} & \cdots & \mathbb{1}\{Y_n \le j\} \end{bmatrix}^T$, $\mathbf{1_b} = \begin{bmatrix} \mathbb{1}\{1 \in I(b)\} & \mathbb{1}\{2 \in I(b)\} & \cdots & \mathbb{1}\{n \in I(b)\} \end{bmatrix}^T$ is the indicator vector of strata membership, $\mathbf{x}_j$ is the data vector for the j-th covariate, n_b is the size of the b-th stratum, and $\overline{x_c}^{(b)} = \frac{1}{n_b}\sum_{i \in I(b)} x_{ic}$ is the local average in stratum b of the c-th covariate.

Proof. A.5 □

Proposition 2.5.3. *The Hessian of the unstratified conditional log-likelihood taken with respect to β evaluated at $\beta = \mathbf{0}$ is*

$$H(\mathbf{0}) = \frac{\partial^2 l}{\partial \beta^2}\Big|_{\beta=\mathbf{0}} = A \otimes \widehat{cov}(X) \quad \in \mathbb{R}^{q\times q}$$

where $A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,(J-1)} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,(J-1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(J-1),1} & a_{(J-1),2} & \cdots & a_{(J-1),(J-1)} \end{bmatrix}$ *is a symmetric matrix of constant multipliers, with* $a_{j,l} = \begin{cases} \frac{c_j(c_l-n)}{n}, & \text{if } j < l \\ \frac{c_j(c_j-n)}{n}, & \text{if } j = l \\ \frac{c_l(c_j-n)}{n}, & \text{if } j > l \end{cases}$.
The $\otimes$ operator denotes the Kronecker product, and $q = p(J-1)$ denotes the dimension of the parameter space.

Proof. A.3 □

Proposition 2.5.4. *The Hessian of the stratified conditional log-likelihood taken with respect to β evaluated at $\beta = \mathbf{0}$ is*

$$H(\mathbf{0}) = \frac{\partial^2 l}{\partial \beta^2}\Big|_{\beta=\mathbf{0}} = \sum_{b=1}^{s} A^{(b)} \otimes \widehat{cov}^{(b)}(X) \quad \in \mathbb{R}^{q\times q}$$

where $A^{(b)} = \begin{bmatrix} a^{(b)}_{1,1} & a^{(b)}_{1,2} & \cdots & a^{(b)}_{1,(J-1)} \\ a^{(b)}_{2,1} & a^{(b)}_{2,2} & \cdots & a^{(b)}_{2,(J-1)} \\ \vdots & \vdots & \ddots & \vdots \\ a^{(b)}_{(J-1),1} & a^{(b)}_{(J-1),2} & \cdots & a^{(b)}_{(J-1),(J-1)} \end{bmatrix}$ *is a symmetric matrix of constant multipliers specifically defined for each stratum, with the (j,l)-th entry defined as*

$$a^{(b)}_{j,l} = \begin{cases} \frac{c_{jb}(c_{lb}-n_b)}{n_b}, & \text{if } j < l \\ \frac{c_{jb}(c_{jb}-n_b)}{n_b}, & \text{if } j = l \\ \frac{c_{lb}(c_{jb}-n_b)}{n_b}, & \text{if } j > l \end{cases}$$

The $\otimes$ operator denotes the Kronecker product, and $\widehat{cov}^{(b)}(X)$ is defined as the sample covariance matrix of stratum b, with the (c,d)-th entry defined as $\widehat{cov}^{(b)}(x_c, x_d) = \frac{1}{n_b-1}\Big[\sum_{i\in I(b)} x_{ic}x_{id} - \frac{1}{n_b}(\sum_{i\in I(b)} x_{ic})(\sum_{i\in I(b)} x_{id})\Big]$, and n_b is the size of the b-th stratum. and $q = p(J-1)$ denotes the dimension of the parameter space.

Proof. A.6 □

Chapter 3

Application to Ohio Medicaid Data

Next, we apply our method to an observational dataset with multiple treatment levels, with data being matched into strata of size three using a propensity score-based method (Nattino, Lu, et al. 2021). For data with three treatment levels, called 1, 2, and 3, the authors proposed an algorithm that produces 1:1:1 matched sets, i.e., within each matched set, each observation has a unique treatment. They defined a distance measurement by first calculating a three-way distance of each matched triplet, and then summing over the three-way distances over all triplets, where the three-way distance is the sum of the pairwise distance between all possible pairs within a triplet. Here, the propensity score information might be used as the pairwise distance. And the match is created based on the following algorithm that aims to minimize the total three-way distance within the matched data:

1. Choose two starting treatment levels, say 1-2, and create a 1:1 match between these treatments;

2. For each of the 1-2 pairs, they optimally match a subject from the third group, and let us call the result *match a*;

3. Then form two alternative matches from *match a* by (i) fixing all 2-3 pairs in *match a*, and optimally match observations with treatment 1, producing *match b*; and (ii) fixing all 1-3 pairs in *match a*, and optimally match observations with treatment 2, producing *match c*;

4. In step 3, if *match b* and *match c* both have greater total distance than *match a*, then conclude the algorithm with *match a*; otherwise, replace *match a* in step 3 by the one from *match b* and *match c* that has lower total distance, and repeat step 3, until termination.

We will focus on evaluating the covariate balance of matchings given by their algorithm. In another paper (Nattino, Song, and Lu 2022), the authors developed an algorithm that matches data with more than three levels of treatment, and they applied their method to the data from two years (2012 & 2015) of the Ohio Medicaid Assessment

Surveys (OMAS), "a survey that is periodically administered by the Ohio Department of Medicaid and aims at assessing access to healthcare system, its utilization and the health status of Ohioans" (Ohio Colleges of Medicine Government Resource Center 2020, as cited in Nattino, Song, and Lu 2022). For context-specific reasons, authors of Nattino, Song, and Lu 2022 defined treatment levels by year of survey (2012 or 2015) and household income level (90%-138% of the Federal Poverty Level (FPL) or 139%-400% of the FPL). To mimic their data example, we considered three levels of treatments: (1) 2012 90-138, (2) 2012 139-400, and (3) 2015 90-138. And as in the case of Nattino et al., we performed matching based on background covariates characterizing the following information: age, sex, race, education level, marital status, number of children in the household, type of county of residence, alcohol use, smoking status, mental health, and disability. It is worth noting that they utilized only categorical (indicator) variables as background covariates, and we therefore also used all covariates in the format of categorical variables.

Applying the triplet matching algorithm in Nattino, Lu, et al. 2021, we obtained 1594 matched sets of three observations, with observations in the same stratum having unique treatment assignments. As their algorithm iteratively performs paired matching, the matching results, therefore, depend on the starting pair of treatments. Thus, we applied the algorithm with all three possible starting pairs: 1-3, 2-3, and 1-2, performing a balance test for each of them. The future analysis in this paper, however, only concerns with the match obtained using the starting pair 1-3. The test results using the χ^2 approximation are summarized in Table 3.1.

	T^2 **statistic**	**df**	**p-value**
1-3	239.679	38	4.293×10^{-31}
2-3	282.218	38	4.588×10^{-39}
1-2	241.362	38	2.096×10^{-31}

Table 3.1: Balance test result for matched data using the triplet matching algorithm, with varying starting treatment pairs; p-values are calculated from the χ^2 approximation

It is observable that we have strong evidence to reject the null hypothesis that the algorithm produces a sufficient matching, no matter how we change the starting pair. To investigate why this is the case, we calculated the standardized mean difference between each treatment group and the rest two groups, before and after matching, as an examination of covariate balance. The results are given in Table 3.2.

From inspecting the standardized mean differences, we observe that despite the general trend that the algorithm gave decreased mean differences after matching, it failed to induce balance, or even inflated imbalance, for some covariates and treatment groups. And we suspect that although this issue is not uncommon for many matching algorithms, it could have contributed to the observation that our test rejected the null hypothesis of covariate balance.

Variable	2012 90-138		2012 139-400		2015 90-138	
	Pre	Post	Pre	Post	Pre	Post
Age (25-34 vs. 19-24)	-0.022	-0.003	-0.124	0.024	0.148	-0.022
Age (35-44 vs. 19-24)	-0.025	0.059	0.043	-0.014	-0.033	-0.046
Age (45-54 vs. 19-24)	-0.012	0.005	0.081	-0.019	-0.083	0.014
Age (55-64 vs. 19-24)	0.044	-0.039	0.094	0.004	-0.132	0.035
Sex (Female vs. Male)	0.055	0.026	-0.036	-0.049	0.004	0.024
Race (Black vs. White)	0.128	-0.005	-0.167	-0.015	0.107	0.020
Race (Hispanic vs. White)	0.099	0.044	-0.084	-0.012	0.030	-0.033
Race (Other vs. White)	-0.031	0.094	-0.193	0.016	0.220	-0.118
Education (High school diploma vs. No high school diploma)	0.078	0.010	-0.151	-0.019	0.122	0.010
Education (College degree or higher vs. No high school diploma)	-0.168	-0.038	0.304	0.037	-0.239	0.001
Marital status (Married vs. Not married)	-0.131	0.032	0.606	0.022	-0.603	-0.055
Number of children (1 vs. 0)	0.026	0.054	0.028	-0.030	-0.048	-0.025
Number of children (>1 vs. 0)	-0.039	0.036	0.053	-0.009	-0.037	-0.027
County type (Suburban vs. Metropolitan)	-0.058	0.016	0.023	-0.002	0.007	-0.014
County type (Rural vs. Metropolitan)	-0.063	-0.014	0.069	0.007	-0.041	0.007
Alcohol use past 30 days (Yes vs. No)	-0.084	0.020	0.254	0.010	-0.234	-0.030
Smoking status (Smoked >100 cigarettes vs. No)	-0.069	0.014	0.176	-0.037	-0.156	0.023
Mental health distress (Yes vs. No)	0.221	-0.001	-0.253	0.014	0.141	-0.014
Disability (Yes vs. No)	0.276	-0.036	-0.433	0.031	0.302	0.006

Table 3.2: Standardized Mean Differences (SMD) between the three treatment groups and the other two, where "Pre" and "Post" stand for the pre-matching and post-matching SMD, respectively; the match is obtained using starting treatment pair 1-3

Chapter 4

Simulation Analysis

4.1 Convergence of the Test Statistic

In this section, we aim to visually examine the convergence of T^2 to its limiting distribution under the null hypothesis that the treatment assignment is independent of covariates given the stratification, using the data from Chapter 3. To observe convergence, we randomly subsetted the data into subsets of 100, 500, 1000, and 1594 strata, and investigated how the resampled statistic behaved as we increase strata. In particular, we performed 500 simulation iterations, and in each iteration, we permuted the labels within each stratum and calculated the test statistic based on this permuted set of labels. We performed this resampling for all four datasets with 100, 500, 1000, and 1594 strata. Theoretically, the empirical distribution of the test statistic will be roughly χ^2 (with $df = 38$, given our data) because the treatment assignment regime here conforms with the null hypothesis. The density plot of the resampled T^2, as we increase strata, is shown in Figure 4.1.

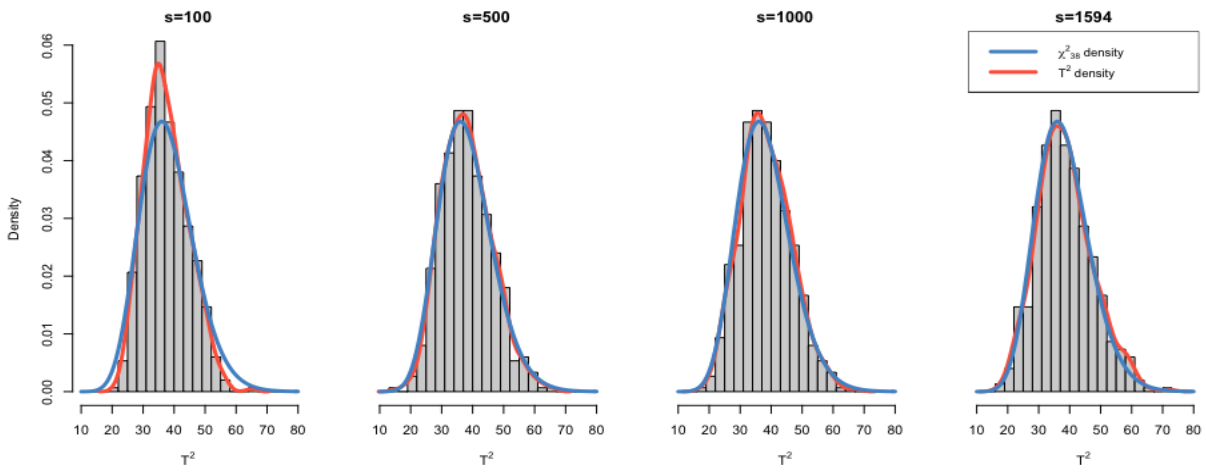


Figure 4.1 Density of the T^2 statistic calculated from labels permuted under the null hypothesis, as we increase the size of the data. Blue curve is the theoretical density of the asymptotic χ^2_{38} distribution; red curve is the estimated density from resamples

From the plot, as we increase the strata number, the empirical density of T^2 converges to the theoretical χ^2 limiting distribution, as expected. This indicates that, under the null hypothesis, the test has a valid size, or Type I error rate, when we use all 1594 strata.

4.2 Power of the Test

In this section, we will study the power of our statistical test under varying sample sizes and signal strengths. We focus on the OMAS data matched using the triplet matching algorithm with 1-3 as the starting pair. The data has $s = 1594$ strata, each having three observations treated uniquely, and thus $n = 4782$.

To evaluate the ability of our method to reject a false null of $\beta = \mathbf{0}$, we generate treatment labels using β vectors drawn from uniform distributions of varying scales in $\{\pm 0.25, \pm 0.5, \pm 0.75\}$. For each β drawn, we randomly sample without replacement a subset of strata from the entirety of 1594 strata, with the size of the subset being chosen from $\{100, 500, 1000, 1594\}$. It is to be noted that although the unconditional treatment assignment probability $P(Y_k = j)$ is modeled by the logistic regression model 2.1 involving beta, we observed each strata of three conditional on the fact that they have distinct treatment levels, and this makes it inappropriate to resample each label naively from their marginal distributions. In order to carry out resampling under a stratified treatment assignment regime, we should draw labels from the conditional law that conditions on the observed treatment counts within each strata. Here, since in every strata i, observations i_1, i_2, i_3 have three different treatments $Y_{i_1}, Y_{i_2}, Y_{i_3}$, we should condition on the event that all observations in any strata has distinct treatment labels, or $|\{Y_{i_1}, Y_{i_2}, Y_{i_3}\}| = 3$.

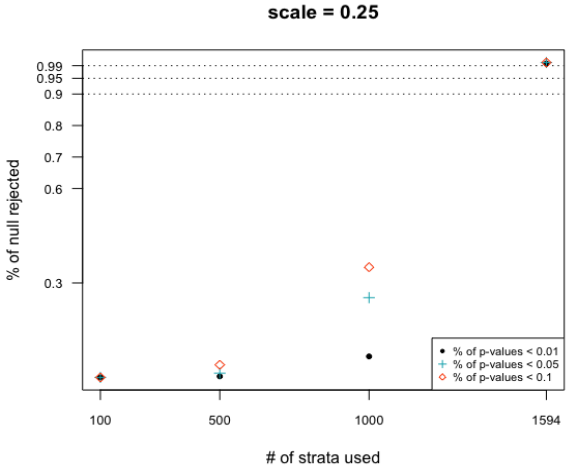
That is, for each pair of β scale and sample size, we resample treatment labels for every stratum, independent of all other strata, under the restriction that all three labels within each stratum are distinct. In particular, our resampling technique works as follows

1. Within a strata, calculate treatment assignment probabilities $\pi_k^j = P(Y_{i_k} = j)$ for all $k = 1, 2, 3$, $j = 1, 2, 3$ using the regression model with known β

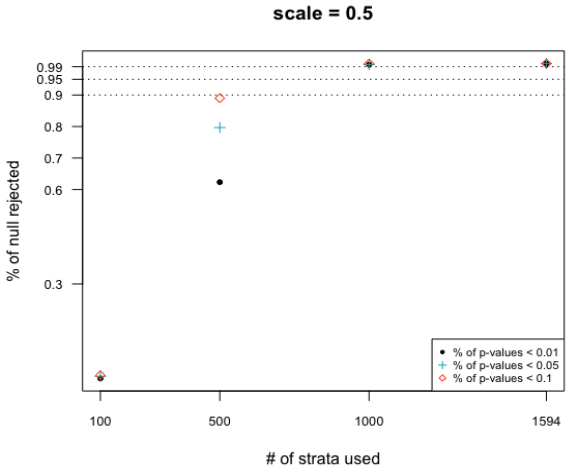
2. For the first observation i_1 in the triplet, calculate $P(Y_{i_1} = j | Y_{i_1} \neq Y_{i_2} \neq Y_{i_3}) = (\sum_{(k,l) \in S(j)} \pi_1^j \pi_2^k \pi_3^l) / (\sum_{(p,q,r) \in S} \pi_1^p \pi_2^q \pi_3^r)$ for all $j = 1, 2, 3$, where $S(j) = \{(k, l) : j \neq k \neq l\}$, and $S = \{(p, q, r) : p \neq q \neq r\}$. Draw Y_{i_1} from this distribution, called y_{i_1}.

3. For the second observation i_2 in the triplet, calculate $P(Y_{i_2} = j | Y_{i_1} \neq Y_{i_2} \neq Y_{i_3}; Y_{i_1} = y_{i_1}) = \pi_2^j \pi_3^k / (\sum_{(p,q) \in S(y_{i_1})} \pi_2^p \pi_3^q)$ for all $j = 1, 2, 3$, where $k \neq j \neq y_{i_1}$, and $S(y_{i_1}) = \{(p, q) : y_{i_1} \neq p \neq q\}$. Draw Y_{i_2} from this distribution, called y_{i_2}.

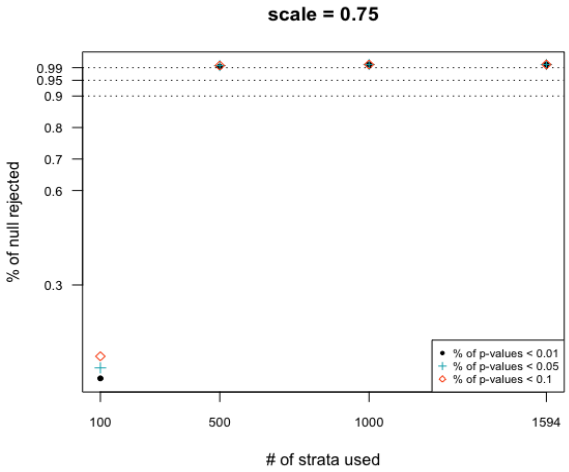
4. Then the realization y_{i_3} of Y_{i_3} will be deterministic, and our triplet drawn is $y_{i_1}, y_{i_2}, y_{i_3}$.



(a)



(b)



(c)

Figure 4.2 Empirical power of the test with levels 0.1, 0.05, and 0.01. Figure 4.2a, 4.2b, and 4.2c corresponds to tests of β drawn from uniform distributions with ranges of ± 0.25, ± 0.5, and ± 0.75. The simulation runs 300 iterations in all cases.

We repeat the previous resampling process in every simulation iterations and calculate the corresponding p-values, over 300 simulation iterations. Results are given in Figure 4.2.

From the simulation results, it is seen that for a sample size as small as 300 (100 strata), even for signals with a range of ± 0.75, the test behaves poorly. However, as we increase the sample size by fivefold to 1500 (500 strata), the test can capture some false null hypotheses. In particular, when β is drawn from a range of ± 0.5, the 0.1 level test has a power of nearly 90%. And for β drawn from a range of ± 0.75, all of the $0.1, 0.05, 0.01$ level tests have powers above 0.99. If we further increase the sample size to 3000 (1000 strata), tests of all levels have powers above 99% for beta with $\pm 0.5, \pm 0.75$ scale. Finally, for a dataset with size 4782 (1594 strata), as is the case of our data, for all signal strengths, all of the $0.1, 0.05, 0.01$ level tests have $> 99\%$ powers.

It is to be remarked that all covariates in the simulation data from the OMAS study are indicator variables, which provides an intuitive understanding of the scale of the covariates and their relative magnitude compared to the β parameter. Empirical observations suggest that in order to recover signs of invalid matching with β that are as weak as ± 0.25 in the context of indicator covariates, a sample size above 3000 will be desirable.

Chapter 5

Discussions and Conclusions

5.1 Discussion and Future Work

We developed a method for balance testing applicable to propensity score-based matching methods, particularly for data with multiple treatments. The methodology of this work is based on expressing the relationship between background covariates and treatment assignment by a conditional multilevel logistic regression model. Although propensity score methods are usually adopted for matching observational data (Rosenbaum 2010), our method does not assume the method by which matching is generated, and can in fact be applied to evaluate matchings generated in any fashion. Hence, one may also use this method to evaluate the matching of data collected from randomized controlled trials (RCTs). Other potential applications include investigating systematic attrition of participants of an RCT, in which the data are initially balanced, but participants may subsequently drop out from the trial, creating a reduced dataset. Our method can be adopted to evaluate the balance of the matching after attrition, to answer whether there is a systematic relation between the attrition event and background variables.

One advantage of our method is that it is an omnibus test. Our test evaluates the balance of all covariates at once, which offers simplicity and statistical quantification of uncertainty when compared to balance assessments using summary statistics such as standardized differences in covariate means between treated/control groups (Rosenbaum 2010). However, to visually examine matching, it is still of our interest to provide a graphical display of these statistics using techniques similar to the Love plot (Love 2002). A potential extension of this work is to provide visualizations of balance assessment statistics (e.g. standardized mean differences) for each level of treatment under the multilevel treatment design, which can be seen as a generalization of the Love plot.

The methods developed in this work are generalizable to other contexts by the nature of a hypothesis test. One could simply accept these methods at the face level and use them to test the global null for a regression model. However, our method utilizes a conditional inference approach, and it might be of potential interest to investigate how these conditional tests compare with their unconditional counterparts in terms of statistical efficiencies, as in Liang 1984.

On the other hand, one weakness of this work is that the limiting distribution of the test statistics relies on several normality conditions. Although these conditions are intuitively practical and comprehensible, it is still of further research interest to study whether relaxations of the conditions are possible.

Numerical simulations exposed some shortcomings of the method. First, in our simulation study, the method obtains poor power when the sample size is as small as 300, even with only roughly 20 covariates. This makes it unattractive to adopt our method in some small sample studies. Also, in our experiment, for weak β signals, the method tends to be not sensitive enough, unless the sample size becomes larger than, e.g., 3000.

5.2 Conclusions

This work provided a statistical test for covariate balance in causal studies with multiple treatments, motivated by propensity score-based matching methods (Yang et al. 2016; Imai and Van Dyk 2004). In particular, our method provides an omnibus test of balance, which offers convenience and uncertainty quantification compared to classical methods of matching evaluation using summary statistics. We used a χ^2 approximation to the limiting distribution of the test statistic, conditioning on the strata-based treatment assignment information to take into account stratification. Based on an observational dataset, we performed simulations to examine the convergence of the statistic and the power of the test. The test has good power when the sample size is large, yet does not obtain a desirable power when either the signal is too weak or the sample size is too small. As a complement to our method, we proposed that a future research direction is to generalize the Love plot (Love 2002) to data with multiple treatment levels, and provide visualizable evaluations thereby.

References

Billingsley, Patrick [1986]. *Probability and Measure*. Second. John Wiley and Sons.

Hansen, Ben B and Jake Bowers [2008]. "Covariate balance in simple, stratified and clustered comparative studies". In: *Statistical Science*, pp. 219–236.

Holland, Paul W [1986]. "Statistics and causal inference". In: *Journal of the American statistical Association* 81.396, pp. 945–960.

Imai, Kosuke and David A Van Dyk [2004]. "Causal inference with general treatment regimes: Generalizing the propensity score". In: *Journal of the American Statistical Association* 99.467, pp. 854–866.

Lehmann, Erich L and George Casella [2006]. *Theory of point estimation*. Springer Science & Business Media.

Li, Xinran and Peng Ding [2016]. *General forms of finite population central limit theorems with applications to causal inference*. arXiv: `1610.04821 [math.ST]`.

Liang, Kung-Yee [1984]. "The asymptotic efficiency of conditional likelihood methods". In: *Biometrika* 71.2, pp. 305–313.

Love, Thomas E [2002]. "Displaying covariate balance after adjustment for selection bias". In: *Joint Statistical Meetings*.

Nattino, Giovanni, Bo Lu, et al. [2021]. "Triplet matching for estimating causal effects with three treatment arms: A comparative study of mortality by trauma center level". In: *Journal of the American Statistical Association* 116.533, pp. 44–53.

Nattino, Giovanni, Chi Song, and Bo Lu [2022]. "Polymatching algorithm in observational studies with multiple treatment groups". In: *Computational Statistics & Data Analysis* 167, p. 107364.

Ohio Colleges of Medicine Government Resource Center [2020]. *Ohio medicaid assessment survey*. URL: `http://grc.osu.edu/OMAS` [visited on November 25, 2021].

Rao, CR [2005]. "Score test: historical review and recent developments". In: *Advances in ranking and selection, multiple comparisons, and reliability*. Springer, pp. 3–20.

Rosenbaum, Paul R [2010]. *Design of observational studies*. Vol. 10. Springer.

Rosenbaum, Paul R and Donald B Rubin [1983]. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55.

Yang, Shu et al. [2016]. "Propensity score matching and subclassification in observational studies with multi-level treatments". In: *Biometrics* 72.4, pp. 1055–1065.

Appendix A

Proof of Propositions

A.1 Proof of Proposition 2.3.1

According to the notation in equation (2.2), we have the equivalent expression for probability of observing a certain result at the ith observation:

$$P(Y_i = k) = \prod_{j=1}^{J} \pi_j(\mathbf{x_i})^{t_{i,j}}, \quad t_{i,j} = \begin{cases} 1, & j = y_i \\ 0, & \text{otherwise} \end{cases}$$

Henceforth, we have the following expression for the likelihood function of the sample:

$$\begin{aligned}
&P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n}\prod_{j=1}^{J} \pi_j(\mathbf{x_i})^{t_{i,j}} \\
=&\prod_{i=1}^{n} \left(\frac{\exp\{\sum_{k=1}^{J-1}(\alpha_k + \beta_k^T \mathbf{x_i})\}}{D_i}\right)^{t_{i,1}} \cdots \left(\frac{\exp\{\sum_{k=J-1}^{J-1}(\alpha_k + \beta_k^T \mathbf{x_i})\}}{D_i}\right)^{t_{i,J-1}} \left(\frac{1}{D_i}\right)^{t_{i,J}} \\
=&\frac{\prod_{i=1}^{n} \exp\{\sum_{k=1}^{J-1}(t_{i,1}\alpha_k + t_{i,1}\beta_k^T \mathbf{x_i})\} \ldots \exp\{\sum_{k=J-1}^{J-1}(t_{i,J-1}\alpha_k + t_{i,J-1}\beta_k^T \mathbf{x_i})\}}{\prod_{i=1}^{n} D_i} \\
=&\frac{\exp\left\{\sum_{i=1}^{n}\left(\sum_{k=1}^{J-1} t_{i,1}\alpha_k + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\alpha_k\right)\right\}}{\prod_{i=1}^{n} D_i} \\
&\times \frac{\exp\left\{\sum_{i=1}^{n}\left(\sum_{k=1}^{J-1} t_{i,1}\beta_k^T \mathbf{x_i} + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\beta_k^T \mathbf{x_i}\right)\right\}}{\prod_{i=1}^{n} D_i}
\end{aligned}$$

where $D_i = 1 + \sum_{t=1}^{J-1} \exp\{\sum_{k=t}^{J-1}(\alpha_k + \beta_k^T \mathbf{x_i})\}$ is the common denominator term for all factors, and this quantity is of no particular interest because it is to be cancelled out eventually.

Now, consider the following equivalency

$$=\Big(\sum_{i=1}^{n}\sum_{k=1}^{J-1} t_{i,1}\alpha_k\Big) + \cdots + \Big(\sum_{i=1}^{n}\sum_{k=J-1}^{J-1} t_{i,J-1}\alpha_k\Big) \tag{A.1}$$

$$=\Big(\sum_{i=1}^{n} t_{i,1}\Big)\sum_{k=1}^{J-1}\alpha_k + \cdots + \Big(\sum_{i=1}^{n} t_{i,J-1}\Big)\sum_{k=J-1}^{J-1}\alpha_k \tag{A.2}$$

$$=n_1\sum_{k=1}^{J-1}\alpha_k + \cdots + n_{J-1}\sum_{k=J-1}^{J-1}\alpha_k \tag{A.3}$$

$$=\sum_{k=1}^{J-1} c_k\alpha_k \tag{A.4}$$

Here, we used $n_j \coloneqq \sum_{i=1}^{n} t_{i,j}$ to denote the total observations with Y falling into category j. Accordingly, we let $c_j \coloneqq \sum_{k=1}^{j}\sum_{i=1}^{n} t_{ik}$, for all $j = 1, 2, \ldots, J$and in future derivations let C_j denote the random variable corresponding to the observed quantity c_j.

Also, consider a similar equivalency

$$\sum_{i=1}^{n}\Big(\sum_{k=1}^{J-1} t_{i,1}\beta_k^T\mathbf{x_i} + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\beta_k^T\mathbf{x_i}\Big) \tag{A.5}$$

$$=\beta_1^T\Big[\sum_{i=1}^{n}\Big(\sum_{k=1}^{1} t_{i,k}\mathbf{x_i}\Big)\Big] + \beta_2^T\Big[\sum_{i=1}^{n}\Big(\sum_{k=1}^{2} t_{i,k}\mathbf{x_i}\Big)\Big] + \cdots + \beta_{J-1}^T\Big[\sum_{i=1}^{n}\Big(\sum_{k=1}^{J-1} t_{i,k}\mathbf{x_i}\Big)\Big] \tag{A.6}$$

$$=\beta_1^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{1} t_{i,k}\Big)\Big] + \beta_2^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{2} t_{i,k}\Big)\Big] + \cdots + \beta_{J-1}^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{J-1} t_{i,k}\Big)\Big] \tag{A.7}$$

$$=\sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}\Big)\Big] \tag{A.8}$$

With (A.4) and (A.8), we rewrite the likelihood $P(Y_1 = y_1, \ldots, Y_n = y_n)$ as

$$\begin{aligned}
&P(Y_1 = y_1, \ldots, Y_n = y_n)\\
=&\frac{\exp\Big\{\sum_{i=1}^{n}\Big(\sum_{k=1}^{J-1} t_{i,1}\alpha_k + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\alpha_k\Big)\Big\}}{\prod_{i=1}^{n} D_i}\\
&\times\frac{\exp\Big\{\sum_{i=1}^{n}\Big(\sum_{k=1}^{J-1} t_{i,1}\beta_k^T\mathbf{x_i} + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\beta_k^T\mathbf{x_i}\Big)\Big\}}{\prod_{i=1}^{n} D_i}\\
=&\frac{\exp\Big\{\sum_{k=1}^{J-1} c_k\alpha_k + \sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}\Big)\Big]\Big\}}{\prod_{i=1}^{n} D_i}
\end{aligned}$$

This means the conjunction probability of $\{Y_1 = y_1, \ldots, Y_n = y_n\}$ with the event $\{C_1 = c_1, \ldots, C_J = c_J\}$, i.e. the joint likelihood, can be expressed as

$$\begin{aligned}
&P(Y_1 = y_1, \ldots, Y_n = y_n, C_1 = c_1, \ldots, C_J = c_J) \\
=&\frac{\exp\Big\{\sum_{k=1}^{J-1} c_k\alpha_k + \sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}\Big)\Big]\Big\}}{\prod_{i=1}^{n} D_i}
\end{aligned}$$

Let $S(t) = \big\{(y_1, y_2, \ldots, y_n) : \sum_{k=1}^{j}\sum_{i=1}^{n} t_{i,k} = c_j \ \forall j = 1, \ldots, J\big\}$ denote the set of all possible outcomes Y satisfying the constraint that the cumulative sum of observations lower than or equal to each category j equals c_j. Here, we keep the notation of $t_{i,k}$ to denote the indicator of $\mathbb{1}\{y_i = k\}$.

Summing the joint likelihood over this set of permutation $S(t)$ gives the probability of the event $\{C_1 = c_1, \ldots, C_J = c_J\}$, i.e.,

$$\begin{aligned}
&P(C_1 = c_1, \ldots, C_J = c_J) \\
=&\sum_{y^*\in S(t)} P(Y_1 = y_1^*, \ldots, Y_n = y_n^*, C_1 = c_1, \ldots, C_J = c_J) \\
=&\sum_{y^*\in S(t)} \frac{\exp\Big\{\sum_{k=1}^{J-1} c_k\alpha_k + \sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}^*\Big)\Big]\Big\}}{\prod_{i=1}^{n} D_i}
\end{aligned}$$

where each $t_{i,k}^*$ is the deterministic indicator analogously defined for y^*.

With these results, it directly follows that the conditional likelihood on the sufficient statistic $\{C_1 = c_1, \ldots, C_J = c_J\}$ is equal to

$$\begin{aligned}
&P(Y_1 = y_1, \ldots, Y_n = y_n \mid C_1 = c_1, \ldots, C_J = c_J) \\
=&\frac{P(Y_1 = y_1, \ldots, Y_n = y_n, C_1 = c_1, \ldots, C_J = c_J)}{P(C_1 = c_1, \ldots, C_J = c_J)} \\
=&\frac{\exp\Big\{\sum_{k=1}^{J-1} c_k\alpha_k + \sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}\Big)\Big]\Big\}}{\sum_{y^*\in S(t)}\exp\Big\{\sum_{k=1}^{J-1} c_k\alpha_k + \sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}^*\Big)\Big]\Big\}} \\
=&\frac{\exp\Big\{\sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}\Big)\Big]\Big\}}{\sum_{y^*\in S(t)}\exp\Big\{\sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j} t_{i,k}^*\Big)\Big]\Big\}}
\end{aligned}$$

A.2 Proof of Proposition 2.5.1

We first introduce some shorthand notations

$$T(\beta_1, ..., \beta_{J-1}) = \sum_{j=1}^{J-1} \beta_j^T \Big[\sum_{i=1}^{n} \mathbf{x_i} \Big(\sum_{k=1}^{j} t_{i,k} \Big) \Big]$$
$$T^*(\beta_1, ..., \beta_{J-1}) = \sum_{j=1}^{J-1} \beta_j^T \Big[\sum_{i=1}^{n} \mathbf{x_i} \Big(\sum_{k=1}^{j} t^*_{i,k} \Big) \Big]$$

where T^* changes as $y^* \in S(t)$ changes during the iteration.

With these we may re-express the likelihood and log-likelihood in the following fashion

$$\begin{aligned} L(\beta \mid X, Y, C) &:= P(Y_1 = y_1, \ldots, Y_n = y_n \mid C_1 = c_1, \ldots, C_J = c_J) \\ &= \frac{\exp\Big\{T(\beta_1, ..., \beta_{J-1})\Big\}}{\sum\limits_{y^* \in S(t)} \exp\Big\{T^*(\beta_1, ..., \beta_{J-1})\Big\}} \\ l(\beta \mid X, Y, C) &:= \log(L(\beta \mid X, Y, C)) \\ &= T(\beta_1, ..., \beta_{J-1}) - \log\Big(\sum_{y^* \in S(t)} \exp\Big\{T^*(\beta_1, ..., \beta_{J-1})\Big\}\Big) \end{aligned}$$

Now, we differentiate $T(\beta_1, ..., \beta_{J-1})$ and $T^*(\beta_1, ..., \beta_{J-1})$ with respect to each entry c (the slope corresponding to the c-th covariate) of each β_j (vector of slope coefficients for the j-th category)

$$\frac{\partial}{\partial \beta_j^{(c)}} T(\beta_1, ..., \beta_{J-1}) = \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big)$$
$$\frac{\partial}{\partial \beta_j^{(c)}} T^*(\beta_1, ..., \beta_{J-1}) = \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t^*_{i,k} \Big)$$

With the above expressions, we may take the derivative of the log-likelihood function with respect to each entry c (the slope corresponding to the c-th covariate) of each β_j (vector of slope coefficients for the j-th category)

$$\frac{\partial}{\partial \beta_j^{(c)}} l(\beta \mid X, Y, C) = \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \frac{\sum\limits_{y^* \in S(t)} \exp(T^*) \Big[\sum_{i=1}^{n} x_{ic} (\sum_{k=1}^{j} t^*_{i,k}) \Big]}{\sum\limits_{y^* \in S(t)} \exp(T^*)}$$

But this, evaluated at $\beta_j = 0$ under the null hypothesis, can be expressed as

$$\frac{\partial}{\partial \beta_j^{(c)}} l(\beta \mid X, Y, C)\Big|_{\beta_j = \mathbf{0}} = \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \frac{\sum\limits_{y^* \in S(t)} \sum_{i=1}^{n} x_{ic} (\sum_{k=1}^{j} t^*_{i,k})}{\sum\limits_{y^* \in S(t)} 1} \tag{A.9}$$

$$= \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \frac{\sum_{i=1}^{n} x_{ic} \sum\limits_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{i,k}}{\sum\limits_{y^* \in S(t)} 1} \tag{A.10}$$

Now, letting $n_j = \sum_{i=1}^{n} t_{i,j}$ denote the observed category frequencies that can be directly calculated once we condition on $c_1, \ldots, c_J$, we may notice the following equivalencies

$$\sum_{y^* \in S(t)} 1 = \binom{n}{n_1, \ldots, n_J} \tag{A.11}$$

$$\sum_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{i,k} = \frac{c_j}{n} \binom{n}{n_1, \ldots, n_J} \tag{A.12}$$

where equation (A.11) holds by its nature as a multinomial combinatorial problem. Equation (A.12) may not be easily seen, but can be shown by considering

$$\sum_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{i,k} = \sum_{y^* \in S(t)} \mathbb{1}\{y_i^* \leq j\} \tag{A.13}$$

$$= \binom{n-1}{n_1 - 1, n_2, \ldots, n_j, \ldots n_J} + \binom{n-1}{n_1, n_2 - 1, n_3, \ldots, n_j, \ldots, n_J} + \cdots + \tag{A.14}$$

$$\binom{n-1}{n_1, n_2, \ldots, n_j - 1, \ldots, n_J} \tag{A.15}$$

$$= \binom{n}{n_1, \ldots, n_J} \Big(\frac{n_1}{n} + \frac{n_2}{n} + \cdots + \frac{n_j}{n} \Big) = \binom{n}{n_1, \ldots, n_J} \frac{c_j}{n} \tag{A.16}$$

where

1. $\mathbb{1}\{y_i^* \leq j\}$ is the indicator of whether the observation y_i^* falls in a category of order less than or equal to j;

2. equation (A.15) follows from the possible permutation of labels under the constraint that $y_i^* \leq j$ and that we have $n_1, \ldots, n_J$ observations in categories $1, \ldots, J$.

With these conclusions, we see that equation (A.10) transforms to

$$\begin{aligned}
&\frac{\partial}{\partial \beta_j^{(c)}} l(\beta \mid X, Y, C)\Big|_{\beta_j = \mathbf{0}} \\
=&\sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \frac{c_j}{n} \sum_{i=1}^{n} x_{ic} \\
=&\sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \frac{1}{n} \Big(\sum_{i=1}^{n} x_{ic} \Big) \Big(\sum_{i=1}^{n} \sum_{k=1}^{j} t_{i,k} \Big) \\
=&(\mathbf{x}_c - \overline{x_c}\mathbf{1})^T \mathbf{z}_j
\end{aligned}$$

where where $\mathbf{x}_c$ is the data vector for the c-th covariate, $\mathbf{1}$ is the n-dimensional all-one vector, $\mathbf{z}_j = \begin{bmatrix} \sum_{k=1}^{j} t_{1,k} & \sum_{k=1}^{j} t_{2,k} & \cdots & \sum_{k=1}^{j} t_{n,k} \end{bmatrix}^T = \begin{bmatrix} \mathbb{1}\{y_1 \le j\} & \mathbb{1}\{y_2 \le j\} & \cdots & \mathbb{1}\{y_n \le j\} \end{bmatrix}^T$, and $\overline{x_c} = \frac{1}{n} \sum_{i=1}^{n} x_{ic}$.

Noticing $\beta_j = \begin{bmatrix} \beta_j^{(1)} & \cdots & \beta_j^{(p)} \end{bmatrix}^T$, we have

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} l(\beta \mid X, Y, C)\Big|_{\beta_j = \mathbf{0}} &= \begin{bmatrix} — & (\mathbf{x}_1 - \overline{x_1}\mathbf{1})^T & — \\ — & (\mathbf{x}_2 - \overline{x_2}\mathbf{1})^T & — \\ & \vdots & \\ — & (\mathbf{x}_p - \overline{x_p}\mathbf{1})^T & — \end{bmatrix} \mathbf{z}_j \\
&= \left(\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{1} & \mathbf{1} & \cdots & \mathbf{1} \\ | & | & & | \end{bmatrix} \begin{bmatrix} \overline{x_1} & 0 & \cdots & 0 \\ 0 & \overline{x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \overline{x_p} \end{bmatrix} \right)^T \mathbf{z}_j
\end{aligned}$$

as desired.

A.3 Proof of Proposition 2.5.3

Recall that we stacked the slope vectors together as a $[p(J-1)] \times 1$ vector

$$\beta = \Big[\beta_1^T \quad | \quad \beta_2^T \quad | \quad \cdots \quad | \quad \beta_{J-1}^T\Big]^T$$

And we focus on every entry of the Hessian of the log-likelihood, denoted by $\frac{\partial^2 l(\cdot)}{\partial \beta_j^{(c)} \partial \beta_l^{(d)}}$, where $l(\cdot)$ is the log-likelihood function, and this quantity lies at the $\Big(p(j-1)+c\ ,\ p(l-1)+d\Big)$-th and (by symmetry of Hessian) the $\Big(p(l-1)+d\ ,\ p(j-1)+c\Big)$-th entry of the Hessian matrix.

These second derivative matrices, however, have simplified solution that depends on the relative values of j and l. Before getting to the derivatives, we first introduce a new notation c_{ij} that helps us get rid of one layer of summation in the derivation:

$$c_{ij} := \sum_{k=1}^{j} t_{i,k} = \mathbb{1}\{y_i \leq j\}$$

where i stands for the i-th observation and j stands for the j-th order category. And it is to be noted that we should discriminate c_{ij} from the sufficient statistics of form c_j, where the former has two indexing quantities and the latter has one. Also, as we did in (A.2), let $n_j = \sum_{i=1}^n t_{i,j}$ denote the observed category frequencies.

Now we similarly take the second derivatives based on our results of first derivatives:

$$\frac{\partial^2}{\partial \beta_j^{(c)} \partial \beta_l^{(d)}} l(\beta \mid X, Y, C) = \frac{\partial^2}{\partial \beta_l^{(d)} \partial \beta_j^{(c)}} l(\beta \mid X, Y, C) = \frac{\partial}{\partial \beta_l^{(d)}} \left[\frac{\partial}{\partial \beta_j^{(c)}} l(\beta \mid X, Y, C) \right]$$

$$= \frac{\partial}{\partial \beta_l^{(d)}} \left\{ \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \frac{\sum\limits_{y^* \in S(t)} \exp(T^*) \Big[\sum_{i=1}^n x_{ic} (\sum_{k=1}^j t_{i,k}^*) \Big]}{\sum\limits_{y^* \in S(t)} \exp(T^*)} \right\}$$

$$= - \frac{\partial}{\partial \beta_l^{(d)}} \frac{\sum\limits_{y^* \in S(t)} \exp(T^*) \Big[\sum_{i=1}^n x_{ic} (\sum_{k=1}^j t_{i,k}^*) \Big]}{\sum\limits_{y^* \in S(t)} \exp(T^*)}$$

$$=\frac{\Big\{\sum\limits_{y^*\in S(t)}\exp(T^*)\Big[\sum_{i=1}^n x_{ic}(\sum_{k=1}^j t^*_{i,k})\Big]\Big\}\Big\{\sum\limits_{y^*\in S(t)}\exp(T^*)\Big[\sum_{m=1}^n x_{md}(\sum_{k=1}^l t^*_{m,k})\Big]\Big\}}{\left[\sum\limits_{y^*\in S(t)}\exp(T^*)\right]^2}$$

$$-\frac{\Big\{\sum\limits_{y^*\in S(t)}\exp(T^*)\Big[\sum_{i=1}^n x_{ic}(\sum_{k=1}^j t^*_{i,k})\Big]\Big[\sum_{m=1}^n x_{md}(\sum_{k=1}^l t^*_{m,k})\Big]\Big\}\Big\{\sum\limits_{y^*\in S(t)}\exp(T^*)\Big\}}{\left[\sum\limits_{y^*\in S(t)}\exp(T^*)\right]^2}$$

Evaluating the second derivative at the null hypothesis that $\beta = 0$, we observe

$$\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}}l(\beta \mid X,Y,C)\Big|_{\beta=\mathbf{0}} \tag{A.17}$$

$$=\frac{\Big\{\sum\limits_{y^*\in S(t)}\sum_{i=1}^n x_{ic}(\sum_{k=1}^j t^*_{i,k})\Big\}\Big\{\sum\limits_{y^*\in S(t)}\sum_{m=1}^n x_{md}(\sum_{k=1}^l t^*_{m,k})\Big\}}{\left[\sum\limits_{y^*\in S(t)}1\right]^2} \tag{A.18}$$

$$-\frac{\sum\limits_{y^*\in S(t)}\Big[\sum_{i=1}^n x_{ic}(\sum_{k=1}^j t^*_{i,k})\Big]\Big[\sum_{m=1}^n x_{md}(\sum_{k=1}^l t^*_{m,k})\Big]}{\sum\limits_{y^*\in S(t)}1} \tag{A.19}$$

$$=\frac{c_j}{n}\frac{c_l}{n}\Big(\sum_{i=1}^n x_{ic}\Big)\Big(\sum_{m=1}^n x_{md}\Big)-\frac{\sum\limits_{y^*\in S(t)}\Big[\sum_{i=1}^n x_{ic}c^*_{ij}\Big]\Big[\sum_{m=1}^n x_{md}c^*_{ml}\Big]}{\sum\limits_{y^*\in S(t)}1} \tag{A.20}$$

$$=\frac{c_jc_l}{n^2}\Big(\sum_{i=1}^n x_{ic}\Big)\Big(\sum_{m=1}^n x_{md}\Big)-\frac{\sum\limits_{y^*\in S(t)}\sum_{i=1}^n\sum_{m=1}^n x_{ic}x_{md}c^*_{ij}c^*_{ml}}{\sum\limits_{y^*\in S(t)}1} \tag{A.21}$$

$$=\frac{c_jc_l}{n^2}\Big(\sum_{i=1}^n x_{ic}\Big)\Big(\sum_{m=1}^n x_{md}\Big)-\frac{\Big[\sum_{i=1}^n\sum_{m=1}^n x_{ic}x_{md}\Big(\sum\limits_{y^*\in S(t)}c^*_{ij}c^*_{ml}\Big)\Big]}{\sum\limits_{y^*\in S(t)}1} \tag{A.22}$$

$$=\frac{c_jc_l}{n^2}\Big(\sum_{i=1}^n x_{ic}\Big)\Big(\sum_{m=1}^n x_{md}\Big)-\frac{\Big[\sum\limits_{i\neq m}x_{ic}x_{md}\Big(\sum\limits_{y^*\in S(t)}c^*_{ij}c^*_{ml}\Big)\Big]}{\sum\limits_{y^*\in S(t)}1}-\frac{\Big[\sum_{i=1}^n x_{ic}x_{id}\Big(\sum\limits_{y^*\in S(t)}c^*_{ij}c^*_{il}\Big)\Big]}{\sum\limits_{y^*\in S(t)}1} \tag{A.23}$$

Now, the derivation reduces to evaluating $\sum_{y^* \in S(t)} c^*_{ij} c^*_{ml}$ $(i \neq m)$ and $\sum_{y^* \in S(t)} c^*_{ij} c^*_{il}$. And it can be seen that these quantities can only be expressed in closed form if we discuss by cases:

1. $j = l$
2. $j < l$
3. $j > l$

Case 1: $j = l$

The evaluation of $\sum_{y^* \in S(t)} c^*_{ij} c^*_{ml}$ reduces to the evaluation of the combinatorial problem of picking two slots out of a total of n, under the constraint that both slots has to be among the top j category of slots, where each category k of slots has a total n_k of slots inside.

$$\sum_{y^* \in S(t)} c^*_{ij} c^*_{ml} = \sum_{y^* \in S(t)} c^*_{ij} c^*_{mj} = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \leq j, y^*_m \leq j\} \tag{A.24}$$

$$= \binom{n-2}{n_1 - 2, n_2, \ldots, n_j, \ldots n_J} + \binom{n-2}{n_1, n_2 - 2, n_3, \ldots, n_j, \ldots, n_J} + \cdots + \binom{n-2}{n_1, n_2, \ldots, n_j - 2, \ldots, n_J} \tag{A.25}$$

$$+ \binom{n-2}{n_1 - 1, n_2 - 1, n_3, \ldots, n_j, \ldots n_J} + \cdots + \binom{n-2}{n_1, n_2, \ldots, n_{j-1} - 1, n_j - 1, \ldots, n_J} \tag{A.26}$$

$$= \binom{n}{n_1, \ldots, n_J} \left[\frac{n_1(n_1-1)}{n(n-1)} + \frac{n_2(n_2-1)}{n(n-1)} + \cdots + \frac{n_j(n_j-1)}{n(n-1)} + \frac{n_1 n_2}{n(n-1)} + \cdots + \frac{n_{j-1} n_j}{n(n-1)} \right] \tag{A.27}$$

$$= \binom{n}{n_1, \ldots, n_J} \frac{1}{n(n-1)} \Big[-c_j + \sum_{p=1}^{j} \sum_{q=1}^{j} n_p n_q \Big] \tag{A.28}$$

$$= \binom{n}{n_1, \ldots, n_J} \frac{c_j(c_j - 1)}{n(n-1)} \tag{A.29}$$

Here, equation (A.25) is enumerating all permutation of Y labels in which y_i and y_m have the same category, and equation (A.26) is enumerating cases in which their categories differ. To obtain the $[\cdot]$ quantity in equation (A.27), one simply divide every multinomial coefficient from equation (A.25) and (A.26) by $\binom{n}{n_1, \ldots, n_J}$.

The evaluation of $\sum_{y^* \in S(t)} c^*_{ij} c^*_{il}$ is equivalent to that of $\sum_{y^* \in S(t)} c^*_{ij} c^*_{ij} = \sum_{y^* \in S(t)} c^*_{ij}$ since in the first case we have $j = l$. So this reduces to

$$\sum_{y^* \in S(t)} c^*_{ij} c^*_{il} = \sum_{y^* \in S(t)} c^*_{ij} c^*_{ij} = \sum_{y^* \in S(t)} c^*_{ij} = \sum_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{ik} = \binom{n}{n_1, \ldots, n_J} \frac{c_j}{n} \tag{A.30}$$

as we did previously in deriving the first derivatives.
Equation (A.29) and (A.30) allows us to rewrite the second derivatives in equation (A.23) as follows

$$
\begin{aligned}
&\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_j^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} \\
=&\frac{c_j c_j}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{\Big[\sum\limits_{i\neq m} x_{ic}x_{md}\Big(\sum\limits_{y^*\in S(t)} c_{ij}^* c_{mj}^*\Big)\Big]}{\sum\limits_{y^*\in S(t)} 1} - \frac{\Big[\sum_{i=1}^{n} x_{ic}x_{id}\Big(\sum\limits_{y^*\in S(t)} c_{ij}^* c_{ij}^*\Big)\Big]}{\sum\limits_{y^*\in S(t)} 1} \\
=&\frac{c_j c_j}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{\binom{n}{n_1,\ldots,n_J}\frac{c_j(c_j-1)}{n(n-1)}\Big(\sum\limits_{i\neq m} x_{ic}x_{md}\Big)}{\binom{n}{n_1,\ldots,n_J}} - \frac{\binom{n}{n_1,\ldots,n_J}\frac{c_j}{n}\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big)}{\binom{n}{n_1,\ldots,n_J}} \\
=&\frac{c_j c_j}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{c_j(c_j-1)}{n(n-1)}\Big(\sum_{i\neq m} x_{ic}x_{md}\Big) - \frac{c_j}{n}\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big) \\
=&\frac{c_j c_j}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \left\{\frac{c_j(c_j-1)}{n(n-1)}\Big(\sum_{i\neq m} x_{ic}x_{md}\Big) + \Big[\frac{c_j(c_j-1)}{n(n-1)} - \frac{c_j(c_j-n)}{n(n-1)}\Big]\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big)\right\} \\
=&\frac{c_j c_j}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{c_j(c_j-1)}{n(n-1)}\Big(\sum_{i=1}^{n}\sum_{m=1}^{n} x_{ic}x_{md}\Big) + \frac{c_j(c_j-n)}{n(n-1)}\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big) \\
=&\frac{c_j c_j}{n^2}(\mathbf{x}_c^T\mathbf{1})(\mathbf{x}_d^T\mathbf{1}) - \frac{c_j(c_j-1)}{n(n-1)}(\mathbf{x}_c^T\mathbf{1})(\mathbf{x}_d^T\mathbf{1}) + \frac{c_j(c_j-n)}{n(n-1)}(\mathbf{x}_c^T\mathbf{x}_d) \\
=&\frac{c_j(c_j-n)}{n(n-1)}(\mathbf{x}_c^T\mathbf{x}_d) - \frac{c_j(c_j-n)}{n^2(n-1)}(\mathbf{x}_c^T\mathbf{1})(\mathbf{x}_d^T\mathbf{1}) \\
=&\frac{c_j(c_j-n)}{n}\left(\frac{\mathbf{x}_c^T\mathbf{x}_d}{n-1} - \frac{n\bar{x}_c\bar{x}_d}{n-1}\right) \\
=&\frac{c_j(c_j-n)}{n}\widehat{\mathrm{cov}}(x_c, x_d)
\end{aligned}
$$

Case 2: $j < l$

The case when $j \neq l$ is more complicated, and for simplicity of notation we will oftentimes express the multinomial coefficients by their alternative form

$$\binom{n}{n_1, \ldots, n_J} = \frac{n!}{n_1! n_2! \cdots n_J!}$$

The evaluation of $\sum_{y^* \in S(t)} c^*_{ij} c^*_{ml}$ can be regarded as a combinatorial problem again, where we pick two slots out of n, with one of them has category less than j and the other less than l, with $j < l$:

$$\sum_{y^* \in S(t)} c^*_{ij} c^*_{ml} = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \leq j, y^*_m \leq l\} \tag{A.31}$$

$$= \sum_{k=1}^{j} \frac{(n-2)!}{(n_k - 2)! \prod_{p=1,\ldots,J \atop p \neq k} n_p} + \sum_{s=1,\ldots,j;\ t=1,\ldots l;\ s \neq t} \frac{(n-2)!}{(n_s - 1)!(n_t - 1)! \prod_{k=1,\ldots J \atop k \neq s,t} n_k} \tag{A.32}$$

$$= \binom{n}{n_1, \ldots, n_J} \left[\frac{\sum_{k=1}^{j} n_k (n_k - 1)}{n(n-1)} + \frac{\sum_{s=1}^{j} \sum_{t=1}^{l} n_s n_t}{n(n-1)} - \frac{\sum_{k=1}^{j} n_k n_k}{n(n-1)} \right] \tag{A.33}$$

$$= \binom{n}{n_1, \ldots, n_J} \frac{c_j c_l - c_j}{n(n-1)} \tag{A.34}$$

Here, the first quantity in equation (A.32) is enumerating all permutation of Y labels in which y_i and y_m have the same category, and the second is enumerating cases in which their categories differ. To obtain the $[\cdot]$ quantity in equation (A.33), one simply divide every multinomial coefficient from equation (A.32) by $\binom{n}{n_1, \ldots, n_J}$.

The evaluation of $\sum_{y^* \in S(t)} c^*_{ij} c^*_{il}$ is equivalent to that of $\sum_{y^* \in S(t)} c^*_{ij} c^*_{ij} = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \leq j, y^*_i \leq l\}$ and thus only depend on j since we assumed $j < l$. So this reduces to the evaluation of $\sum_{y^* \in S(t)} c^*_{ij}$, which we've done previously:

$$\sum_{y^* \in S(t)} c^*_{ij} c^*_{il} = \sum_{y^* \in S(t)} c^*_{ij} = \binom{n}{n_1, \ldots, n_J} \frac{c_j}{n} \tag{A.35}$$

With equation (A.34) and (A.35), we may rewrite equation (A.23) as:

$$
\begin{aligned}
&\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} \\
=&\frac{c_j c_l}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{\Big[\sum\limits_{i\neq m} x_{ic}x_{md}\Big(\sum\limits_{y^*\in S(t)} c^*_{ij}c^*_{ml}\Big)\Big]}{\sum\limits_{y^*\in S(t)} 1} - \frac{\Big[\sum_{i=1}^{n} x_{ic}x_{id}\Big(\sum\limits_{y^*\in S(t)} c^*_{ij}c^*_{il}\Big)\Big]}{\sum\limits_{y^*\in S(t)} 1} \\
=&\frac{c_j c_l}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{\binom{n}{n_1,\ldots,n_J}\frac{c_j c_l - c_j}{n(n-1)}\Big(\sum\limits_{i\neq m} x_{ic}x_{md}\Big)}{\binom{n}{n_1,\ldots,n_J}} - \frac{\binom{n}{n_1,\ldots,n_J}\frac{c_j}{n}\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big)}{\binom{n}{n_1,\ldots,n_J}} \\
=&\frac{c_j c_l}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{c_j(c_l-1)}{n(n-1)}\Big(\sum_{i\neq m} x_{ic}x_{md}\Big) - \frac{c_j}{n}\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big) \\
=&\frac{c_j c_l}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \left\{\frac{c_j(c_l-1)}{n(n-1)}\Big(\sum_{i\neq m} x_{ic}x_{md}\Big) + \Big[\frac{c_j(c_l-1)}{n(n-1)} - \frac{c_j(c_l-n)}{n(n-1)}\Big]\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big)\right\} \\
=&\frac{c_j c_l}{n^2}\Big(\sum_{i=1}^{n} x_{ic}\Big)\Big(\sum_{m=1}^{n} x_{md}\Big) - \frac{c_j(c_l-1)}{n(n-1)}\Big(\sum_{i=1}^{n}\sum_{m=1}^{n} x_{ic}x_{md}\Big) + \frac{c_j(c_l-n)}{n(n-1)}\Big(\sum_{i=1}^{n} x_{ic}x_{id}\Big) \\
=&\frac{c_j c_l}{n^2}(\mathbf{x}_c^T\mathbf{1})(\mathbf{x}_d^T\mathbf{1}) - \frac{c_j(c_l-1)}{n(n-1)}(\mathbf{x}_c^T\mathbf{1})(\mathbf{x}_d^T\mathbf{1}) + \frac{c_j(c_l-n)}{n(n-1)}(\mathbf{x}_c^T\mathbf{x}_d) \\
=&\frac{c_j(c_l-n)}{n(n-1)}(\mathbf{x}_c^T\mathbf{x}_d) - \frac{c_j(c_l-n)}{n^2(n-1)}(\mathbf{x}_c^T\mathbf{1})(\mathbf{x}_d^T\mathbf{1}) \\
=&\frac{c_j(c_l-n)}{n}\left(\frac{\mathbf{x}_c^T\mathbf{x}_d}{n-1} - \frac{n\bar{x}_c\bar{x}_d}{n-1}\right) \\
=&\frac{c_j(c_l-n)}{n}\widehat{\mathrm{cov}}(x_c, x_d)
\end{aligned}
$$

Case 3: $j > l$

The case when $j > l$ simply follows by symmetry of case 2, swapping the role of j and l in the derivation.

Thus we may observe that

$$
\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} = a_{j,l}\ \widehat{\mathrm{cov}}(x_c, x_d),\ \text{where } a_{j,l} = \begin{cases} \frac{c_j(c_l-n)}{n},\ \text{if } j < l \\ \frac{c_j(c_j-n)}{n},\ \text{if } j = l \\ \frac{c_l(c_j-n)}{n},\ \text{if } j > l \end{cases}
$$

Recall that we partitioned β into $J-1$ segments, and consequently the Hessian were partitioned into $(J-1)^2$ blocks each of size $p \times p$. And it is easy to observe that the (j,l)-th block equals $a_{j,l}\ \widehat{\mathrm{cov}}(X)$, which means the entire Hessian equals $(a_{j,l}) \otimes \widehat{\mathrm{cov}}(X)$. The detailed derivation is omitted here because it highly overlaps with the corresponding part of A.6. In fact, this is a special case of Proposition 2.5.4, which can be seen by setting the total number of strata to 1.

A.4 Proof of Proposition 2.3.2

Let $I(b)$ to denote the set of indices of observations belonging to the b-th stratum, $b = 1, 2, ..., s$. Therefore, as an extension of the non-stratified likelihood, we have the following expression for the likelihood of the data

$$
\begin{aligned}
&P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{b=1}^{s} \prod_{i \in I(b)} \prod_{j=1}^{J} \pi_j(\mathbf{x_i})^{t_{i,j}} \\
=&\prod_{b=1}^{s} \prod_{i \in I(b)} \left(\frac{\exp\{\sum_{k=1}^{J-1}(\alpha_{kb} + \beta_k^T \mathbf{x_i})\}}{D_{ib}} \right)^{t_{i,1}} \cdots \left(\frac{\exp\{\sum_{k=J-1}^{J-1}(\alpha_{kb} + \beta_k^T \mathbf{x_i})\}}{D_{ib}} \right)^{t_{i,J-1}} \left(\frac{1}{D_{ib}} \right)^{t_{i,J}} \\
=&\frac{\prod_{b=1}^{s} \prod_{i \in I(b)} \exp\{\sum_{k=1}^{J-1}(t_{i,1}\alpha_{kb} + t_{i,1}\beta_k^T \mathbf{x_i})\} \ldots \exp\{\sum_{k=J-1}^{J-1}(t_{i,J-1}\alpha_{kb} + t_{i,J-1}\beta_k^T \mathbf{x_i})\}}{\prod_{b=1}^{s} \prod_{i \in I(b)} D_i} \\
=&\frac{\exp\left\{ \sum_{b=1}^{s} \sum_{i \in I(b)} \left(\sum_{k=1}^{J-1} t_{i,1}\alpha_{kb} + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\alpha_{kb} \right) \right\}}{\prod_{b=1}^{s} \prod_{i \in I(b)} D_{ib}} \\
&\times \frac{\exp\left\{ \sum_{b=1}^{s} \sum_{i \in I(b)} \left(\sum_{k=1}^{J-1} t_{i,1}\beta_k^T \mathbf{x_i} + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\beta_k^T \mathbf{x_i} \right) \right\}}{\prod_{b=1}^{s} \prod_{i \in I(b)} D_{ib}}
\end{aligned}
$$

where $D_{ib} = 1 + \sum_{t=1}^{J-1} \exp\{\sum_{k=t}^{J-1}(\alpha_{kb} + \beta_k^T \mathbf{x_i})\}$ is the common denominator term for all factors, and this quantity is of no particular interest because it is to be cancelled out eventually.

Now, consider the following equivalency

$$
\begin{aligned}
&\sum_{b=1}^{s} \sum_{i \in I(b)} \Big(\sum_{k=1}^{J-1} t_{i,1}\alpha_{kb} + \cdots + \sum_{k=J-1}^{J-1} t_{i,J-1}\alpha_{kb} \Big) \\
=&\sum_{b=1}^{s} \sum_{i \in I(b)} \Big(t_{i,1} \sum_{k=1}^{J-1} \alpha_{kb} + \cdots + t_{i,J-1} \sum_{k=J-1}^{J-1} \alpha_{kb} \Big) \\
=&\sum_{b=1}^{s} \Big(\sum_{k=1}^{J-1} \alpha_{kb} \Big) \Big(\sum_{i \in I(b)} t_{i,1} \Big) + \cdots + \Big(\sum_{k=J-1}^{J-1} \alpha_{kb} \Big) \Big(\sum_{i \in I(b)} t_{i,J-1} \Big) \\
=&\sum_{b=1}^{s} n_{1,b} \Big(\sum_{k=1}^{J-1} \alpha_{kb} \Big) + \cdots + n_{J-1,b} \Big(\sum_{k=J-1}^{J-1} \alpha_{kb} \Big) \\
=&\sum_{b=1}^{s} \sum_{j=1}^{J-1} c_{jb}\alpha_{jb}
\end{aligned}
$$

Here, we used $n_{j,b} \coloneqq \sum_{i \in I(b)} t_{i,j}$ to denote the total observations with Y falling into category j. Accordingly, we let $c_{jb} = \sum_{k=1}^{j} \sum_{i \in I(b)} t_{i,k},\ j = 1, 2, ..., J,\ \forall b = 1, 2, ..., s$, and in future derivations let C_{jb} denote the random variable corresponding to the observed quantity c_{jb}.

Also, consider a similar equivalency

$$
\begin{aligned}
&\sum_{b=1}^{s}\sum_{i\in I(b)}\Big(\sum_{k=1}^{J-1}t_{i,1}\beta_k^T\mathbf{x_i}+\cdots+\sum_{k=J-1}^{J-1}t_{i,J-1}\beta_k^T\mathbf{x_i}\Big)\\
=&\sum_{i=1}^{n}\Big(\sum_{k=1}^{J-1}t_{i,1}\beta_k^T\mathbf{x_i}+\cdots+\sum_{k=J-1}^{J-1}t_{i,J-1}\beta_k^T\mathbf{x_i}\Big)\\
=&\beta_1^T\Big[\sum_{i=1}^{n}\Big(\sum_{k=1}^{1}t_{i,k}\mathbf{x_i}\Big)\Big]+\beta_2^T\Big[\sum_{i=1}^{n}\Big(\sum_{k=1}^{2}t_{i,k}\mathbf{x_i}\Big)\Big]+\cdots+\beta_{J-1}^T\Big[\sum_{i=1}^{n}\Big(\sum_{k=1}^{J-1}t_{i,k}\mathbf{x_i}\Big)\Big]\\
=&\beta_1^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{1}t_{i,k}\Big)\Big]+\beta_2^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{2}t_{i,k}\Big)\Big]+\cdots+\beta_{J-1}^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{J-1}t_{i,k}\Big)\Big]\\
=&\sum_{j=1}^{J-1}\beta_j^T\Big[\sum_{i=1}^{n}\mathbf{x_i}\Big(\sum_{k=1}^{j}t_{i,k}\Big)\Big]
\end{aligned}
$$

With these expressions, we have the likelihood being

$$
\begin{aligned}
&P(Y_1=y_1,\ldots,Y_n=y_n)\\
=&\frac{\exp\Big\{\sum_{b=1}^{s}\sum_{i\in I(b)}\big(\sum_{k=1}^{J-1}t_{i,1}\alpha_{kb}+\cdots+\sum_{k=J-1}^{J-1}t_{i,J-1}\alpha_{kb}\big)\Big\}}{\prod_{b=1}^{s}\prod_{i\in I(b)}D_{ib}}\\
&\times\frac{\exp\Big\{\sum_{b=1}^{s}\sum_{i\in I(b)}\big(\sum_{k=1}^{J-1}t_{i,1}\beta_k^T\mathbf{x_i}+\cdots+\sum_{k=J-1}^{J-1}t_{i,J-1}\beta_k^T\mathbf{x_i}\big)\Big\}}{\prod_{b=1}^{s}\prod_{i\in I(b)}D_{ib}}\\
=&\frac{\exp\Big\{\sum_{b=1}^{s}\sum_{j=1}^{J-1}c_{jb}\alpha_{jb}+\sum_{j=1}^{J-1}\beta_j^T\big[\sum_{i=1}^{n}\mathbf{x_i}\big(\sum_{k=1}^{j}t_{i,k}\big)\big]\Big\}}{\prod_{b=1}^{s}\prod_{i\in I(b)}D_{ib}}
\end{aligned}
$$

This means the conjunction probability of $\big\{Y_1=y_1,\ldots,Y_n=y_n\big\}$ with the event $\big\{C_{jb}=c_{jb}\ \forall j=1,\ldots,J,\ \forall b=1,\ldots,s\big\}$, i.e. the joint likelihood, can be expressed as

$$
\begin{aligned}
&P\Big(Y_1=y_1,...,Y_n=y_n,C_{jb}=c_{jb}\ \forall j=1,\ldots,J,\ \forall b=1,\ldots,s\Big)\\
=&\frac{\exp\Big\{\sum_{b=1}^{s}\sum_{j=1}^{J-1}c_{jb}\alpha_{jb}+\sum_{j=1}^{J-1}\beta_j^T\big[\sum_{i=1}^{n}\mathbf{x_i}\big(\sum_{k=1}^{j}t_{i,k}\big)\big]\Big\}}{\prod_{b=1}^{s}\prod_{i\in I(b)}D_{ib}}
\end{aligned}
$$

Now let $S(t)=\big\{(y_1,y_2,\ldots,y_n):\sum_{k=1}^{j}\sum_{i\in I(b)}t_{i,j}=c_{jb}\ \forall j=1,\ldots,J,\ \forall b=1,\ldots,s\big\}$ denote the set of all possible outcomes Y satisfying the constraint that the strata-wise cumulative sum of observations lower than or equal to each category j equals the corresponding sum c_{jb}.

Summing the joint likelihood function over this set of permutation $S(t)$ gives the

probability of the event $\{C_{jb} = c_{jb}\ \forall j = 1, \ldots, J,\ \forall b = 1, \ldots, s\}$, i.e.

$$P(C_{jb} = c_{jb}\ \forall j = 1, \ldots, J,\ \forall b = 1, \ldots, s) = \sum_{y^* \in S(t)} P(Y_1 = y_1^*, \ldots, Y_n = y_n^*)$$

$$= \sum_{y^* \in S(t)} \frac{\exp\Big\{ \sum_{b=1}^{s} \sum_{j=1}^{J-1} c_{jb}\alpha_{jb} + \sum_{j=1}^{J-1} \beta_j^T \big[\sum_{i=1}^{n} \mathbf{x_i} \big(\sum_{k=1}^{j} t_{ik}^* \big) \big] \Big\}}{\prod_{b=1}^{s} \prod_{i \in I(b)} D_{ib}}$$

With these results, it directly follows that the conditional likelihood on the sufficient statistic $\{C_{jb} = c_{jb}\ \forall j = 1, \ldots, J,\ \forall b = 1, \ldots, s\}$ is equal to

$$\begin{aligned} & P\Big(Y_1 = y_1, ..., Y_n = y_n \mid C_{jb} = c_{jb}\ \forall j = 1, \ldots, J,\ \forall b = 1, \ldots, s\Big) \\ =& \frac{P\Big(Y_1 = y_1, ..., Y_n = y_n, C_{jb} = c_{jb}\ \forall j = 1, \ldots, J,\ \forall b = 1, \ldots, s\Big)}{P(C_{jb} = c_{jb}\ \forall j = 1, \ldots, J,\ \forall b = 1, \ldots, s)} \\ =& \frac{\exp\Big\{ \sum_{j=1}^{J-1} \beta_j^T \big[\sum_{i=1}^{n} \mathbf{x_i} \big(\sum_{k=1}^{j} t_{i,k} \big) \big] \Big\}}{\sum_{y^* \in S(t)} \exp\Big\{ \sum_{j=1}^{J-1} \beta_j^T \big[\sum_{i=1}^{n} \mathbf{x_i} \big(\sum_{k=1}^{j} t_{i,k}^* \big) \big] \Big\}} \end{aligned}$$

which is of the same form as the non-stratified version in Proposition 2.3.1, with the only difference being the indexing set $S(t)$.

A.5 Proof of Proposition 2.5.2

Previous results established that the conditional log-likelihood under stratified and unstratified designs are of the same form, up to a difference in indexing sets. Thus we must have the result in (A.10) hold for the stratified situation, i.e.

$$\frac{\partial}{\partial \beta_j^{(c)}} l(\beta \mid X, Y, C)\Big|_{\beta_j=\mathbf{0}} = \sum_{i=1}^{n} x_{ic}\Big(\sum_{k=1}^{j} t_{i,k}\Big) - \frac{\sum_{i=1}^{n} x_{ic} \sum\limits_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{i,k}}{\sum\limits_{y^* \in S(t)} 1} \tag{A.36}$$

$$= \sum_{i=1}^{n} x_{ic}\Big(\sum_{k=1}^{j} t_{i,k}\Big) - \frac{\sum_{b=1}^{s} \sum_{i \in I(b)} x_{ic}\big(\sum\limits_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{i,k}\big)}{\sum\limits_{y^* \in S(t)} 1} \tag{A.37}$$

where $S(t) = \big\{(y_1, y_2, \ldots, y_n) : \sum_{k=1}^{j} \sum_{i \in I(b)} t_{i,j} = c_{jb}\ \forall j = 1, \ldots, J,\ \forall b = 1, \ldots, s\big\}$ as previously defined.

Now the problem becomes the evaluation of the terms $\sum\limits_{y^* \in S(t)}$ and $\sum\limits_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{i,k}$, which can be expressed by the following equivalencies of expressions

$$\sum_{y^* \in S(t)} 1 = \binom{n_1}{n_{1,1}, \ldots, n_{J,1}} \cdots \binom{n_b}{n_{1,b}, \ldots, n_{J,b}} \cdots \binom{n_s}{n_{1,s}, \ldots, n_{J,s}} \tag{A.38}$$

$$\sum_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{ik} = \frac{c_{jb}}{n_b} \binom{n_1}{n_{1,1}, \ldots, n_{J,1}} \cdots \binom{n_b}{n_{1,b}, \ldots, n_{J,b}} \cdots \binom{n_s}{n_{1,s}, \ldots, n_{J,s}} \tag{A.39}$$

where equation (A.38) holds by its nature as a multinomial combinatorial problem. Here, the terms n_{kb} that appears in the multinomial coefficients are defined as usual, i.e., as $n_{j,b} := \sum_{i \in I(b)} t_{i,j}$, that is, the count for each category in each stratum that is observed in the given data, and n_b's denote the size of stratum b, also observed in the data.

Equation (A.39) may not be easily seen, but can be shown by considering the following for the i-th observation in the dataset, belonging to the b-th stratum

$$\sum_{y^* \in S(t)} \sum_{k=1}^{j} t^*_{i,k} = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \le j\} \tag{A.40}$$

$$= \binom{n_1}{n_{1,1}, \ldots, n_{J,1}} \cdots \binom{n_b - 1}{n_{1,b} - 1, n_{2,b}, \ldots, n_{J,b}} \cdots \binom{n_s}{n_{1,s}, \ldots, n_{J,s}} \tag{A.41}$$

$$+ \binom{n_1}{n_{1,1}, \ldots, n_{J,1}} \cdots \binom{n_b - 1}{n_{1,b}, n_{2,b} - 1, \ldots, n_{J,b}} \cdots \binom{n_s}{n_{1,s}, \ldots, n_{J,s}} \tag{A.42}$$

$$+ \cdots + \binom{n_1}{n_{1,1}, \ldots, n_{J,1}} \cdots \binom{n_b - 1}{n_{1,b}, n_{2,b}, \ldots, n_{j,b} - 1, \ldots, n_{J,b}} \cdots \binom{n_s}{n_{1,s}, \ldots, n_{J,s}} \tag{A.43}$$

$$
\begin{aligned}
=&\binom{n_1}{n_{1,1},\ldots,n_{J,1}} \times \cdots \times \left[\binom{n_b-1}{n_{1,b}-1, n_{2,b}, \ldots, n_{J,b}} + \cdots + \binom{n_b-1}{n_{1,b}, n_{2,b}, \ldots, n_{j,b}-1, \ldots, n_{J,b}} \right] \\
&\times \cdots \times \binom{n_s}{n_{1,s},\ldots,n_{J,s}} \\
=&\binom{n_1}{n_{1,1},\ldots,n_{J,1}} \cdots \left[\binom{n_b}{n_{1,b},\ldots,n_{J,b}} \left(\frac{n_{1,b}}{n_b} + \cdots + \frac{n_{j,b}}{n_b} \right) \right] \cdots \binom{n_s}{n_{1,s},\ldots,n_{J,s}} \\
=&\frac{c_{jb}}{n_b} \binom{n_1}{n_{1,1},\ldots,n_{J,1}} \cdots \binom{n_b}{n_{1,b},\ldots,n_{J,b}} \cdots \binom{n_s}{n_{1,s},\ldots,n_{J,s}}
\end{aligned}
$$

where

1. $\mathbb{1}\{y_i^* \leq j\}$ is the indicator of where the observation y_i^* falls in a category of order less than or equal to j;

2. the equation containing the line (A.43) follows from the possible permutation of labels under the constraint that $y_i^* \leq j$ and that we have $n_{1,b}$, ..., $n_{J,b}$ observations in each categories from 1 to J in stratum b.

With (A.38) and (A.39), we may express (A.37) as

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j^{(c)}} l(\beta \mid X, Y, C)\Big|_{\beta_j = \mathbf{0}} &= \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \sum_{b=1}^{s} \sum_{i \in I(b)} x_{ic} \frac{c_{jb}}{n_b} && \text{(A.44)} \\
&= \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \sum_{b=1}^{s} \frac{c_{jb}}{n_b} \sum_{i \in I(b)} x_{ic} && \text{(A.45)} \\
&= \sum_{i=1}^{n} x_{ic} \Big(\sum_{k=1}^{j} t_{i,k} \Big) - \sum_{b=1}^{s} c_{jb} \overline{x_c}^{(b)} && \text{(A.46)} \\
&= \big(\mathbf{x}_c - \sum_{b=1}^{s} \overline{x_c}^{(b)} \mathbf{1_b} \big)^T \mathbf{z}_j && \text{(A.47)}
\end{aligned}
$$

where $\overline{x_c}^{(b)} = \frac{1}{n_b} \sum_{i \in I(b)} x_{ic}$ is the local average in stratum b of the c-th covariate, $\mathbf{z}_j = \begin{bmatrix} \sum_{k=1}^{j} t_{1,k} & \sum_{k=1}^{j} t_{2,k} & \cdots & \sum_{k=1}^{j} t_{n,k} \end{bmatrix}^T = \begin{bmatrix} \mathbb{1}\{y_1 \leq j\} & \mathbb{1}\{y_2 \leq j\} & \cdots & \mathbb{1}\{y_n \leq j\} \end{bmatrix}^T$, and $\mathbf{1_b} = \begin{bmatrix} \mathbb{1}\{1 \in I(b)\} & \mathbb{1}\{2 \in I(b)\} & \cdots & \mathbb{1}\{n \in I(b)\} \end{bmatrix}^T$ is the indicator vector of strata membership.

Thus, following the lines in Appendix Section A.2, we stack the j-th segment of the score vector as

$$
\frac{\partial l}{\partial \beta_j}\Big|_{\beta_j = \mathbf{0}} = \left(\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{1}_1 & \mathbf{1}_2 & \cdots & \mathbf{1}_s \\ | & | & & | \end{bmatrix} \begin{bmatrix} \overline{x_1}^{(1)} & \overline{x_2}^{(1)} & \cdots & \overline{x_p}^{(1)} \\ \overline{x_1}^{(2)} & \overline{x_2}^{(2)} & \cdots & \overline{x_p}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x_1}^{(s)} & \overline{x_2}^{(s)} & \cdots & \overline{x_p}^{(s)} \end{bmatrix} \right)^T \mathbf{z}_j
$$

A.6 Proof of Proposition 2.5.4

As we did in the previous subsection, we may also adopt the results for non-stratified second derivatives to the stratified case, which means we could use the results in (A.18) and (A.19), up to an alternative way to index the observations from 1 to n

$$\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} \tag{A.48}$$

$$= \frac{\Big\{\sum\limits_{y^*\in S(t)} \sum_{b=1}^{s} \sum_{i\in I(b)} x_{ic}(\sum_{k=1}^{j} t^*_{i,k})\Big\}\Big\{\sum\limits_{y^*\in S(t)} \sum_{b'=1}^{s} \sum_{m\in I(b')} x_{md}(\sum_{k=1}^{l} t^*_{m,k})\Big\}}{\Big[\sum\limits_{y^*\in S(t)} 1\Big]^2} \tag{A.49}$$

$$- \frac{\sum\limits_{y^*\in S(t)} \Big[\sum_{b=1}^{s} \sum_{i\in I(b)} x_{ic}(\sum_{k=1}^{j} t^*_{i,k})\Big]\Big[\sum_{b'=1}^{s} \sum_{m\in I(b')} x_{md}(\sum_{k=1}^{l} t^*_{m,k})\Big]}{\sum\limits_{y^*\in S(t)} 1} \tag{A.50}$$

It is obvious that by arguments in (A.39), we may express the term in (A.49) as

$$\frac{\Big\{\sum\limits_{y^*\in S(t)} \sum_{b=1}^{s} \sum_{i\in I(b)} x_{ic}(\sum_{k=1}^{j} t^*_{i,k})\Big\}\Big\{\sum\limits_{y^*\in S(t)} \sum_{b'=1}^{s} \sum_{m\in I(b')} x_{md}(\sum_{k=1}^{l} t^*_{m,k})\Big\}}{\Big[\sum\limits_{y^*\in S(t)} 1\Big]^2} \tag{A.51}$$

$$= \frac{\Big\{\sum\limits_{y^*\in S(t)} \sum_{b=1}^{s} \sum_{i\in I(b)} x_{ic}(\sum_{k=1}^{j} t^*_{i,k})\Big\}}{\sum\limits_{y^*\in S(t)} 1} \cdot \frac{\Big\{\sum\limits_{y^*\in S(t)} \sum_{b'=1}^{s} \sum_{m\in I(b')} x_{md}(\sum_{k=1}^{l} t^*_{m,k})\Big\}}{\sum\limits_{y^*\in S(t)} 1} \tag{A.52}$$

$$= \Big\{\sum_{b=1}^{s} \frac{c_{jb}}{n_b} \sum_{i\in I(b)} x_{ic}\Big\} \cdot \Big\{\sum_{b'=1}^{s} \frac{c_{lb}}{n_{b'}} \sum_{m\in I(b')} x_{md}\Big\} \tag{A.53}$$

$$= \Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\} \cdot \Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} \tag{A.54}$$

Now let us focus on the quantity in (A.50). With some algebra we may simplify it to a more manageable form

$$\frac{\sum\limits_{y^*\in S(t)}\Big[\sum_{b=1}^{s}\sum_{i\in I(b)}x_{ic}(\sum_{k=1}^{j}t^*_{i,k})\Big]\Big[\sum_{b'=1}^{s}\sum_{m\in I(b')}x_{md}(\sum_{k=1}^{l}t^*_{m,k})\Big]}{\sum\limits_{y^*\in S(t)}1} \tag{A.55}$$

$$=\frac{\sum\limits_{y^*\in S(t)}\sum_{b=1}^{s}\sum_{i\in I(b)}\sum_{b'=1}^{s}\sum_{m\in I(b')}x_{ic}x_{md}(\sum_{k=1}^{j}t^*_{i,k})(\sum_{k=1}^{l}t^*_{m,k})}{\sum\limits_{y^*\in S(t)}1} \tag{A.56}$$

$$=\frac{\sum_{b=1}^{s}\sum_{b'=1}^{s}\sum_{i\in I(b)}\sum_{m\in I(b')}x_{ic}x_{md}\sum\limits_{y^*\in S(t)}(\sum_{k=1}^{j}t^*_{i,k})(\sum_{k=1}^{l}t^*_{m,k})}{\sum\limits_{y^*\in S(t)}1} \tag{A.57}$$

Hence we turn our focus to the evaluation of $\sum\limits_{y^*\in S(t)}(\sum_{k=1}^{j}t^*_{i,k})(\sum_{k=1}^{l}t^*_{m,k})$, and it is always to be noted that in this expression $i\in I(b)$ and $j\in I(b')$ for some strata b, b'. To fully transform this into a combinatorial problem, we note that $\sum\limits_{y^*\in S(t)}(\sum_{k=1}^{j}t^*_{i,k})(\sum_{k=1}^{l}t^*_{m,k}) = \sum\limits_{y^*\in S(t)}\mathbb{1}\{Y^*_i\le j\}\mathbb{1}\{Y^*_m\le j\}$, and evaluate it individually for the following three cases

1. $b\neq b'$
2. $b=b', i=m$
3. $b=b', i\neq m$

Case 1: $b\neq b'$

In this case, we first introduce the following quantity to simplify notation

$$F(b,j):=\binom{n_b-1}{n_{1,b},n_{2,b},\ldots,n_{j,b}-1,\ldots,n_{J,b}}=\frac{n_{j,b}}{n_b}\binom{n_b}{n_{1,b},n_{2,b},\ldots,n_{J,b}},\quad \forall b=1,...,s; j=1,...,J$$

With this, it is easy to observe, when $b \neq b'$, the following equivalency

$$\sum_{y^* \in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k}) = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \leq j\}\mathbb{1}\{y^*_m \leq l\} \tag{A.58}$$

$$= \sum_{p=1}^{j}\sum_{q=1}^{l}\left[F(b,p)\cdot F(b',q)\prod_{B\neq b,b'}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}}\right] \tag{A.59}$$

$$= \left[\sum_{p=1}^{j}\sum_{q=1}^{l} F(b,p)\cdot F(b',q)\right]\prod_{B\neq b,b'}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.60}$$

$$= \left[\sum_{p=1}^{j}\sum_{q=1}^{l}\frac{n_{p,b}}{n_b}\binom{n_b}{n_{1,b}, n_{2,b}, \ldots, n_{J,b}}\frac{n_{q,b'}}{n'_b}\binom{n'_b}{n_{1,b'}, n_{2,b'}, \ldots, n_{J,b'}}\right]\prod_{B\neq b,b'}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.61}$$

$$= \left[\sum_{p=1}^{j}\sum_{q=1}^{l}\frac{n_{p,b}}{n_b}\frac{n_{q,b'}}{n'_b}\right]\prod_{B=1}^{s}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.62}$$

$$= \frac{c_{jb}}{n_b}\frac{c_{lb'}}{n'_b}\prod_{B=1}^{s}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.63}$$

Case 2: $b = b', i = m$

In this case, we will find it useful to define $r \coloneqq \min(j,l)$. Along with the shorthand notation $F(b,j)$ defined previously, we may easily observe the following equivalency

$$\sum_{y^* \in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k}) = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \leq j\}\mathbb{1}\{y^*_m \leq l\} = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \leq r\} \tag{A.64}$$

$$= \sum_{p=1}^{r}\left[F(b,p)\prod_{B\neq b}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}}\right] \tag{A.65}$$

$$= \left[\sum_{p=1}^{r} F(b,p)\right]\prod_{B\neq b}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.66}$$

$$= \left[\sum_{p=1}^{r}\frac{n_{p,b}}{n_b}\binom{n_b}{n_{1,b}, n_{2,b}, \ldots, n_{J,b}}\right]\prod_{B\neq b,b'}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.67}$$

$$= \left[\sum_{p=1}^{r}\frac{n_{p,b}}{n_b}\right]\prod_{B=1}^{s}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.68}$$

$$= \frac{c_{r,b}}{n_b}\prod_{B=1}^{s}\binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.69}$$

Case 3a: $b = b', i \neq m, j = l$

In this case, we need to split our derivation into even more cases, depending on the relative values of j and l. W.L.O.G., aside from the case that $j = l$ we will only consider the case of $j < l$, which is case 3b, simply because the result for $j > l$ can simply be obtained by swapping the role of j and l in the case of $j < l$. Now, back to the case of $j = l$, the following equivalency can be shown similarly to what we did in (A.29)

$$\sum_{y^* \in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k}) = \sum_{y^* \in S(t)} \mathbb{1}\{y^*_i \leq j\}\mathbb{1}\{y^*_m \leq l\} \tag{A.70}$$

$$= \Bigg[\binom{n_b - 2}{n_{1,b} - 2, n_{2,b}, \ldots, n_{j,b}, \ldots n_{J,b}} + \binom{n_b - 2}{n_{1,b}, n_{2,b} - 2, n_{3,b}, \ldots, n_{j,b}, \ldots, n_{J,b}} \tag{A.71}$$

$$+ \cdots + \binom{n_b - 2}{n_{1,b}, n_{2,b}, \ldots, n_{j,b} - 2, \ldots, n_{J,b}} + \binom{n_b - 2}{n_{1,b} - 1, n_{2,b} - 1, n_{3,b}, \ldots, n_{j,b}, \ldots n_{J,b}} \tag{A.72}$$

$$+ \cdots + \binom{n_b - 2}{n_{1,b}, n_{2,b}, \ldots, n_{j-1,b} - 1, n_{j,b} - 1, \ldots, n_{J,b}} \Bigg] \tag{A.73}$$

$$\times \prod_{B \neq b} \binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.74}$$

$$= \binom{n_b}{n_{1,b}, \ldots, n_{J,b}} \left[\frac{n_{1,b}(n_{1,b} - 1)}{n_b(n_b - 1)} + \frac{n_{2,b}(n_{2,b} - 1)}{n_b(n_b - 1)} + \cdots + \frac{n_{j,b}(n_{j,b} - 1)}{n_b(n_b - 1)} + \frac{n_{1,b}n_{2,b}}{n_b(n_b - 1)} + \cdots + \frac{n_{j-1,b}n_{j,b}}{n_b(n_b - 1)} \right] \tag{A.75}$$

$$= \frac{1}{n_b(n_b - 1)} \Big[-\sum_{k=1}^{j} n_{jb} + \sum_{p=1}^{j}\sum_{q=1}^{j} n_{p,b}n_{q,b} \Big] \prod_{B=1}^{s} \binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.76}$$

$$= \frac{-c_{jb} + c_{jb}c_{jb}}{n_b(n_b - 1)} \prod_{B=1}^{s} \binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} = \frac{c_{jb}(c_{jb} - 1)}{n_b(n_b - 1)} \prod_{B=1}^{s} \binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.77}$$

Here, the first three terms in equation (A.71) and (A.72) are enumerating all permutation of Y labels in which y_i and y_m have the same category, and the following terms are enumerating cases in which their categories differ. To obtain the $[\cdot]$ quantity in equation (A.75), one simply divide every multinomial coefficient from equation (A.71), (A.72), and (A.73) by $\binom{n_b}{n_{1b}, \ldots, n_{Jb}}$.

Case 3b: $b = b', i \neq m, j < l$

Here, as mentioned previously, we will investigate the case of $j < l$. The evaluation of $\sum_{y^* \in S(t)} c^*_{ij} c^*_{ml}$ can be regarded as a combinatorial problem again, analogous to (A.34)

$$\sum_{y^* \in S(t)} c^*_{ij} c^*_{ml} = \sum_{y^* \in S(t)} \mathbb{1}\{Y^*_i \leq j, Y^*_m \leq l\} \tag{A.78}$$

$$= \left[\sum_{k=1}^{j} \frac{(n_b - 2)!}{(n_{k,b} - 2)! \prod_{p \neq k}^{p=1,\ldots,J} n_{p,b}} + \sum_{s \neq t}^{s=1,\ldots,j;\ t=1,\ldots l} \frac{(n_b - 2)!}{(n_{s,b} - 1)!(n_{t,b} - 1)! \prod_{k \neq s,t}^{k=1,\ldots J} n_{k,b}} \right] \times \prod_{B \neq b} \binom{n_B}{n_{1,B}, \ldots, n_{J,B}} \tag{A.79}$$

$$= \binom{n_b}{n_{1,b}, \ldots, n_{J,b}} \left[\frac{\sum_{k=1}^{j} n_{k,b}(n_{k,b} - 1)}{n_b(n_b - 1)} + \frac{\sum_{s=1}^{j} \sum_{t=1}^{l} n_{s,b} n_{t,b}}{n_b(n_b - 1)} - \frac{\sum_{k=1}^{j} n_{k,b} n_{k,b}}{n_b(n_b - 1)} \right] \times \prod_{B \neq b} \binom{n_B}{n_{1,B}, \ldots, n_{J,B}} \tag{A.80}$$

$$= \frac{c_{jb} c_{lb} - c_{jb}}{n_b(n_b - 1)} \times \prod_{B=1}^{s} \binom{n_B}{n_{1,B}, n_{2,B}, \ldots, n_{J,B}} \tag{A.81}$$

Here, the first quantity in the $[\cdot]$ term in equation (A.79) is enumerating all permutation of Y labels in which Y_i and Y_m have the same category, and the second is enumerating cases in which their categories differ, subject to the condition that $i, m \in I(b)$. To obtain the $[\cdot]$ quantity in equation (A.80), one simply divide every multinomial coefficient from equation (A.79) by $\binom{n_b}{n_{1b}, \ldots, n_{Jb}}$.

As we have computed the values of $\sum_{y^* \in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})$ with $i \in I(b)$, $m \in I(b')$ under cases where $b \neq b'$, $b = b', i = m$, and $b = b', i \neq m$, we may evaluate the second derivative at $\beta = 0$, using the results we obtained. In order to take advantage of these results, we will split the second derivative accordingly, as we will see in a moment.

So, according to (A.54) and (A.57) we have

$$\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} \tag{A.82}$$

$$= \Big\{ \sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)} \Big\} \cdot \Big\{ \sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')} \Big\} \tag{A.83}$$

$$- \frac{\sum_{b=1}^{s}\sum_{b'=1}^{s}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \sum\limits_{y^*\in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})}{\sum\limits_{y^*\in S(t)} 1} \tag{A.84}$$

$$= \Big\{ \sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)} \Big\} \cdot \Big\{ \sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')} \Big\} \tag{A.85}$$

$$- \frac{\sum_{b=1}^{s} \sum\limits_{b'\neq b}^{b'=1,\ldots,s} \sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \sum\limits_{y^*\in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})}{\sum\limits_{y^*\in S(t)} 1} \tag{A.86}$$

$$- \frac{\sum_{b=1}^{s}\sum_{i\in I(b)} \sum\limits_{m\neq i}^{m\in I(b)} x_{ic}x_{md} \sum\limits_{y^*\in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})}{\sum\limits_{y^*\in S(t)} 1} \tag{A.87}$$

$$- \frac{\sum_{b=1}^{s}\sum_{i\in I(b)} x_{ic}x_{id} \sum\limits_{y^*\in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{i,k})}{\sum\limits_{y^*\in S(t)} 1} \tag{A.88}$$

Now we can observe that lines (A.86), (A.87), (A.88) corresponds to the cases of $\{b \neq b'\}$, $\{b = b', i \neq m\}$, and $\{b = b', i = m\}$, respectively. Much like how we dealt with the non-stratified case, we will hereby discuss by cases

1. $j = l$
2. $j < l$
3. $j > l$

A.6.1 Final Expressions of the Stratified Second-order Derivatives

In this subsection, we combine results from the previous subsection together, to derive an analytical form of the second-order derivatives. First, let us consider the case when $j = l$.

Case 1: $j = l$

$$\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} \tag{A.89}$$

$$= \Big\{ \sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)} \Big\} \cdot \Big\{ \sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')} \Big\} \tag{A.90}$$

$$- \frac{\sum_{b=1}^{s} \sum_{b'\neq b}^{b'=1,\ldots,s} \sum_{i\in I(b)} \sum_{m\in I(b')} x_{ic}x_{md} \sum_{y^*\in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})}{\sum_{y^*\in S(t)} 1} \tag{A.91}$$

$$- \frac{\sum_{b=1}^{s} \sum_{i\in I(b)} \sum_{m\neq i}^{m\in I(b)} x_{ic}x_{md} \sum_{y^*\in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})}{\sum_{y^*\in S(t)} 1} \tag{A.92}$$

$$- \frac{\sum_{b=1}^{s} \sum_{i\in I(b)} x_{ic}x_{id} \sum_{y^*\in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{i,k})}{\sum_{y^*\in S(t)} 1} \tag{A.93}$$

$$= \Big\{ \sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)} \Big\} \cdot \Big\{ \sum_{b'=1}^{s} c_{jb'}\overline{x_d}^{(b')} \Big\} \tag{A.94}$$

$$- \sum_{b=1}^{s} \sum_{b'\neq b}^{b'=1,\ldots,s} \sum_{i\in I(b)} \sum_{m\in I(b')} x_{ic}x_{md} \frac{c_{jb}}{n_b} \frac{c_{jb'}}{n'_b} \tag{A.95}$$

$$- \sum_{b=1}^{s} \sum_{i\in I(b)} \sum_{m\neq i}^{m\in I(b)} x_{ic}x_{md} \frac{c_{jb}(c_{jb}-1)}{n_b(n_b-1)} \tag{A.96}$$

$$- \sum_{b=1}^{s} \sum_{i\in I(b)} x_{ic}x_{id} \frac{c_{jb}}{n_b} \tag{A.97}$$

Note that (A.95), (A.96), (A.97) follows respectively from applying (A.63), (A.77), and (A.69). Now with the following equations

$$\frac{c_{jb}}{n_b} = \frac{c_{jb}(c_{jb}-1)}{n_b(n_b-1)} - \frac{c_{jb}(c_{jb}-n_b)}{n_b(n_b-1)}$$
$$\frac{c_{jb}(c_{jb}-1)}{n_b(n_b-1)} = \frac{c_{jb}^2}{n_b^2} + \frac{c_{jb}(c_{jb}-n_b)}{n_b^2(n_b-1)}$$

we may continue the derivation as follows

$$\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} \tag{A.98}$$

$$= \Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} \tag{A.99}$$

$$- \sum_{b=1}^{s}\sum_{b'\neq b}^{b'=1,\ldots,s} \frac{c_{jb}}{n_b}\frac{c_{jb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \tag{A.100}$$

$$- \sum_{b=1}^{s}\frac{c_{jb}(c_{jb}-1)}{n_b(n_b-1)}\sum_{i\in I(b)}\sum_{m\neq i}^{m\in I(b')} x_{ic}x_{md} \tag{A.101}$$

$$- \Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{jb}-1)}{n_b(n_b-1)}\sum_{i\in I(b)} x_{ic}x_{id}\Big] + \Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{jb}-n_b)}{n_b(n_b-1)}\sum_{i\in I(b)} x_{ic}x_{id}\Big] \tag{A.102}$$

$$= \Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} \tag{A.103}$$

$$- \sum_{b=1}^{s}\sum_{b'\neq b}^{b'=1,\ldots,s} \frac{c_{jb}}{n_b}\frac{c_{jb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \tag{A.104}$$

$$- \sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{jb}-1)}{n_b(n_b-1)}\sum_{m\in I(b)} x_{ic}x_{md} \tag{A.105}$$

$$+ \sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{jb}-n_b)}{n_b(n_b-1)} x_{ic}x_{id} \tag{A.106}$$

$$= \Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} \tag{A.107}$$

$$- \sum_{b=1}^{s}\sum_{b'\neq b}^{b'=1,\ldots,s} \frac{c_{jb}}{n_b}\frac{c_{jb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \tag{A.108}$$

$$- \Big[\sum_{b=1}^{s}\frac{c_{jb}^2}{n_b^2}\sum_{i\in I(b)}\sum_{m\in I(b)} x_{ic}x_{md}\Big] - \Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{jb}-n_b)}{n_b^2(n_b-1)}\sum_{i\in I(b)}\sum_{m\in I(b)} x_{ic}x_{md}\Big] \tag{A.109}$$

$$+ \sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{jb}-n_b)}{n_b(n_b-1)} x_{ic}x_{id} \tag{A.110}$$

$$= \Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} - \sum_{b=1}^{s}\sum_{b'=1}^{s}\frac{c_{jb}}{n_b}\frac{c_{jb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \tag{A.111}$$

$$- \Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{jb}-n_b)}{n_b^2(n_b-1)}\sum_{i\in I(b)}\sum_{m\in I(b)} x_{ic}x_{md}\Big] + \sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{jb}-n_b)}{n_b(n_b-1)} x_{ic}x_{id} \tag{A.112}$$

$$= \sum_{b=1}^{s}\frac{c_{jb}(c_{jb}-n_b)}{n_b}\widehat{\mathrm{COV}}^{(b)}(x_c, x_d) \tag{A.113}$$

Note that (A.111) equals zero, and this can be seen after changing the order of summation signs. And (A.112) equals $\sum_{b=1}^{s} \frac{c_{jb}(c_{jb}-n_b)}{n_b} \widehat{\mathrm{cov}}^{(b)}(x_c, x_d)$, where $\widehat{\mathrm{cov}}^{(b)}(x_c, x_d)$ is defined as the sample covariance of the c-th and d-th covariates in stratum b:

$$\widehat{\mathrm{cov}}^{(b)}(x_c, x_d) = \frac{1}{n_b - 1}\Big[\sum_{i \in I(b)} x_{ic}x_{id} - \frac{1}{n_b}(\sum_{i \in I(b)} x_{ic})(\sum_{i \in I(b)} x_{id})\Big]$$

And that concludes the discussion of the case when $j = l$.

Case 2: $j < l$

W.L.O.G. we consider the other case, i.e. $j < l$, in the same spirit as the previous case

$$\frac{\partial^2}{\partial \beta_j^{(c)} \partial \beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} \tag{A.114}$$

$$= \Big\{ \sum_{b=1}^{s} c_{jb} \overline{x_c}^{(b)} \Big\} \cdot \Big\{ \sum_{b'=1}^{s} c_{lb'} \overline{x_d}^{(b')} \Big\} \tag{A.115}$$

$$- \frac{\sum_{b=1}^{s} \sum_{b' \neq b}^{b'=1,\ldots,s} \sum_{i \in I(b)} \sum_{m \in I(b')} x_{ic}x_{md} \sum_{y^* \in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})}{\sum_{y^* \in S(t)} 1} \tag{A.116}$$

$$- \frac{\sum_{b=1}^{s} \sum_{i \in I(b)} \sum_{m \neq i}^{m \in I(b)} x_{ic}x_{md} \sum_{y^* \in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{m,k})}{\sum_{y^* \in S(t)} 1} \tag{A.117}$$

$$- \frac{\sum_{b=1}^{s} \sum_{i \in I(b)} x_{ic}x_{id} \sum_{y^* \in S(t)} (\sum_{k=1}^{j} t^*_{i,k})(\sum_{k=1}^{l} t^*_{i,k})}{\sum_{y^* \in S(t)} 1} \tag{A.118}$$

$$= \Big\{ \sum_{b=1}^{s} c_{jb} \overline{x_c}^{(b)} \Big\} \cdot \Big\{ \sum_{b'=1}^{s} c_{lb'} \overline{x_d}^{(b')} \Big\} \tag{A.119}$$

$$- \sum_{b=1}^{s} \sum_{b' \neq b}^{b'=1,\ldots,s} \sum_{i \in I(b)} \sum_{m \in I(b')} x_{ic}x_{md} \frac{c_{jb}}{n_b} \frac{c_{lb'}}{n'_b} \tag{A.120}$$

$$- \sum_{b=1}^{s} \sum_{i \in I(b)} \sum_{m \neq i}^{m \in I(b)} x_{ic}x_{md} \frac{c_{jb}(c_{lb} - 1)}{n_b(n_b - 1)} \tag{A.121}$$

$$- \sum_{b=1}^{s} \sum_{i \in I(b)} x_{ic}x_{id} \frac{c_{jb}}{n_b} \tag{A.122}$$

Note that (A.120), (A.121), (A.122) follows respectively from applying (A.63), (A.81), and (A.69). Now with the following equations

$$\frac{c_{jb}}{n_b} = \frac{c_{jb}(c_{lb}-1)}{n_b(n_b-1)} - \frac{c_{jb}(c_{lb}-n_b)}{n_b(n_b-1)}$$
$$\frac{c_{jb}(c_{lb}-1)}{n_b(n_b-1)} = \frac{c_{jb}c_{lb}}{n_b^2} + \frac{c_{jb}(c_{lb}-n_b)}{n_b^2(n_b-1)}$$

we may continue the derivation as follows

$$\begin{aligned}
&\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X,Y,C)\Big|_{\beta=\mathbf{0}} \\
=&\Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} \\
&-\sum_{b=1}^{s}\sum_{b'\neq b}^{b'=1,\ldots,s}\frac{c_{jb}}{n_b}\frac{c_{lb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \\
&-\sum_{b=1}^{s}\frac{c_{jb}(c_{lb}-1)}{n_b(n_b-1)}\sum_{i\in I(b)}\sum_{m\neq i}^{m\in I(b')} x_{ic}x_{md} \\
&-\Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{lb}-1)}{n_b(n_b-1)}\sum_{i\in I(b)} x_{ic}x_{id}\Big] + \Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{lb}-n_b)}{n_b(n_b-1)}\sum_{i\in I(b)} x_{ic}x_{id}\Big] \\
=&\Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} \\
&-\sum_{b=1}^{s}\sum_{b'\neq b}^{b'=1,\ldots,s}\frac{c_{jb}}{n_b}\frac{c_{lb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \\
&-\sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{lb}-1)}{n_b(n_b-1)}\sum_{m\in I(b)} x_{ic}x_{md} \\
&+\sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{lb}-n_b)}{n_b(n_b-1)} x_{ic}x_{id}
\end{aligned}$$

$$=\Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\} \tag{A.123}$$

$$-\sum_{b=1}^{s}\sum_{b'\neq b}^{b'=1,\ldots,s}\frac{c_{jb}}{n_b}\frac{c_{lb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \tag{A.124}$$

$$-\Big[\sum_{b=1}^{s}\frac{c_{jb}c_{lb}}{n_b^2}\sum_{i\in I(b)}\sum_{m\in I(b)} x_{ic}x_{md}\Big]-\Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{lb}-n_b)}{n_b^2(n_b-1)}\sum_{i\in I(b)}\sum_{m\in I(b)} x_{ic}x_{md}\Big] \tag{A.125}$$

$$+\sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{lb}-n_b)}{n_b(n_b-1)}x_{ic}x_{id} \tag{A.126}$$

$$=\Big\{\sum_{b=1}^{s} c_{jb}\overline{x_c}^{(b)}\Big\}\cdot\Big\{\sum_{b'=1}^{s} c_{lb'}\overline{x_d}^{(b')}\Big\}-\sum_{b=1}^{s}\sum_{b'=1}^{s}\frac{c_{jb}}{n_b}\frac{c_{lb'}}{n'_b}\sum_{i\in I(b)}\sum_{m\in I(b')} x_{ic}x_{md} \tag{A.127}$$

$$-\Big[\sum_{b=1}^{s}\frac{c_{jb}(c_{lb}-n_b)}{n_b^2(n_b-1)}\sum_{i\in I(b)}\sum_{m\in I(b)} x_{ic}x_{md}\Big]+\sum_{b=1}^{s}\sum_{i\in I(b)}\frac{c_{jb}(c_{lb}-n_b)}{n_b(n_b-1)}x_{ic}x_{id} \tag{A.128}$$

$$=\sum_{b=1}^{s}\frac{c_{jb}(c_{lb}-n_b)}{n_b}\widehat{\mathrm{cov}}^{(b)}(x_c,x_d) \tag{A.129}$$

Note that (A.127) equals zero, and this can be seen after changing the order of summation signs.

A.6.2 Vectorization of the Second Derivatives, with Stratification

By (A.113) and (A.129), we see that

$$\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}}l(\beta\mid X,Y,C)\Big|_{\beta=\mathbf{0}}=\begin{cases}\sum_{b=1}^{s}\frac{c_{jb}(c_{lb}-n_b)}{n_b}\widehat{\mathrm{cov}}^{(b)}(x_c,x_d), & \text{if } j<l;\\ \sum_{b=1}^{s}\frac{c_{jb}(c_{jb}-n_b)}{n_b}\widehat{\mathrm{cov}}^{(b)}(x_c,x_d), & \text{if } j=l;\\ \sum_{b=1}^{s}\frac{c_{lb}(c_{jb}-n_b)}{n_b}\widehat{\mathrm{cov}}^{(b)}(x_c,x_d), & \text{if } j>l.\end{cases}$$

Nevertheless, if we define the coefficient

$$a_{j,l}^{(b)}=\begin{cases}\frac{c_{jb}(c_{lb}-n_b)}{n_b}, & \text{if } j<l;\\ \frac{c_{jb}(c_{jb}-n_b)}{n_b}, & \text{if } j=l;\\ \frac{c_{lb}(c_{jb}-n_b)}{n_b}, & \text{if } j>l.\end{cases}$$

then it is easily seen that

$$\frac{\partial^2}{\partial\beta_j^{(c)}\partial\beta_l^{(d)}} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} = \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_c, x_d)$$

Then it follows that if we again stack $\beta_j^{(1)}, \ldots, \beta_j^{(p)}$ and $\beta_l^{(1)}, \ldots, \beta_l^{(p)}$, respectively, then we have the second derivative matrices as

$$\frac{\partial^2}{\partial\beta_j\partial\beta_l} l(\beta \mid X, Y, C)\Big|_{\beta=\mathbf{0}} = \begin{bmatrix} \frac{\partial^2 l}{\partial\beta_j^{(1)}\partial\beta_l^{(1)}} & \frac{\partial^2 l}{\partial\beta_j^{(1)}\partial\beta_l^{(2)}} & \cdots & \frac{\partial^2 l}{\partial\beta_j^{(1)}\partial\beta_l^{(p)}} \\ \frac{\partial^2 l}{\partial\beta_j^{(2)}\partial\beta_l^{(1)}} & \frac{\partial^2 l}{\partial\beta_j^{(2)}\partial\beta_l^{(2)}} & \cdots & \frac{\partial^2 l}{\partial\beta_j^{(2)}\partial\beta_l^{(p)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial\beta_j^{(p)}\partial\beta_l^{(1)}} & \frac{\partial^2 l}{\partial\beta_j^{(p)}\partial\beta_l^{(2)}} & \cdots & \frac{\partial^2 l}{\partial\beta_j^{(p)}\partial\beta_l^{(p)}} \end{bmatrix}_{\beta=\mathbf{0}} \tag{A.130}$$

$$= \begin{bmatrix} \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_1, x_1) & \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_1, x_2) & \cdots & \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_1, x_p) \\ \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_2, x_1) & \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_2, x_2) & \cdots & \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_p, x_1) & \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_p, x_2) & \cdots & \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(x_p, x_p) \end{bmatrix} \tag{A.131}$$

$$= \sum_{b=1}^{s} a_{j,l}^{(b)} \begin{bmatrix} \widehat{\mathrm{cov}}^{(b)}(x_1, x_1) & \widehat{\mathrm{cov}}^{(b)}(x_1, x_2) & \cdots & \widehat{\mathrm{cov}}^{(b)}(x_1, x_p) \\ \widehat{\mathrm{cov}}^{(b)}(x_2, x_1) & \widehat{\mathrm{cov}}^{(b)}(x_2, x_2) & \cdots & \widehat{\mathrm{cov}}^{(b)}(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\mathrm{cov}}^{(b)}(x_p, x_1) & \widehat{\mathrm{cov}}^{(b)}(x_p, x_2) & \cdots & \widehat{\mathrm{cov}}^{(b)}(x_p, x_p) \end{bmatrix} \tag{A.132}$$

$$= \sum_{b=1}^{s} a_{j,l}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) \tag{A.133}$$

where $\widehat{\mathrm{cov}}^{(b)}(X)$ is the sample covariance matrix estimated from the b-th strata.

Recall we stacked the slope vectors together as a $[p(J-1)] \times 1$ vector

$$\beta = \begin{bmatrix} \beta_1^T & | & \beta_2^T & | & \cdots & | & \beta_{J-1}^T \end{bmatrix}^T$$

And it is notable that (A.133) gives the (j, l)-th block of the Hessian matrix of β. With this observation, we can express the Hessian matrix of the entire β vector as the following

square block matrix with $(J-1)^2$ blocks

$$\frac{\partial^2 l}{\partial \beta^2}\Big|_{\beta=\mathbf{0}} = H(\beta)\Big|_{\beta=\mathbf{0}} = \begin{bmatrix} \frac{\partial^2 l}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_{J-1}} \\ \frac{\partial^2 l}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 l}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial^2 l}{\partial \beta_2 \partial \beta_{J-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \beta_{J-1} \partial \beta_1} & \frac{\partial^2 l}{\partial \beta_{J-1} \partial \beta_2} & \cdots & \frac{\partial^2 l}{\partial \beta_{J-1} \partial \beta_{J-1}} \end{bmatrix}_{\beta=\mathbf{0}}$$

$$= \begin{bmatrix} \sum_{b=1}^{s} a_{1,1}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \sum_{b=1}^{s} a_{1,2}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \cdots & \sum_{b=1}^{s} a_{1,(J-1)}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) \\ \sum_{b=1}^{s} a_{2,1}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \sum_{b=1}^{s} a_{2,2}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \cdots & \sum_{b=1}^{s} a_{2,(J-1)}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{b=1}^{s} a_{(J-1),1}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \sum_{b=1}^{s} a_{(J-1),2}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \cdots & \sum_{b=1}^{s} a_{(J-1),(J-1)}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) \end{bmatrix}$$

$$= \sum_{b=1}^{s} \begin{bmatrix} a_{1,1}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & a_{1,2}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \cdots & a_{1,(J-1)}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) \\ a_{2,1}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & a_{2,2}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \cdots & a_{2,(J-1)}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) \\ \vdots & \vdots & \ddots & \vdots \\ a_{(J-1),1}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & a_{(J-1),2}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) & \cdots & a_{(J-1),(J-1)}^{(b)} \widehat{\mathrm{cov}}^{(b)}(X) \end{bmatrix}$$

$$= \sum_{b=1}^{s} A^{(b)} \otimes \widehat{\mathrm{cov}}^{(b)}(X) \qquad \in \mathbb{R}^{q \times q}$$

where $A^{(b)} = \begin{bmatrix} a_{1,1}^{(b)} & a_{1,2}^{(b)} & \cdots & a_{1,(J-1)}^{(b)} \\ a_{2,1}^{(b)} & a_{2,2}^{(b)} & \cdots & a_{2,(J-1)}^{(b)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(J-1),1}^{(b)} & a_{(J-1),2}^{(b)} & \cdots & a_{(J-1),(J-1)}^{(b)} \end{bmatrix}$ is a symmetric matrix of constant multipliers specifically defined for each stratum, the $\otimes$ operator denotes the Kronecker product, and $q = p(J-1)$ denotes the dimension of the parameter space.

Appendix B

Technical Development of the Asymptotic Distribution of T^2

B.1 Decomposition of the Hessian

Lemma B.1.1. *Let $\tilde{H}$ denote $-H(\mathbf{0})$ and let $\tilde{H} = UDU^T$ be an orthogonal eigen decomposition of $\tilde{H} \in \mathbb{R}^{q\times q}$, where U is a $q \times q$ orthogonal matrix and D is a rank-r positive semi-definite diagonal matrix. Let $H^- = UD^-U^T$ denote the generalized inverse of $\tilde{H}$, where*

$$(D^-)_{ii} = \begin{cases} (D_{ii})^{-1}, & D_{ii} > 0 \\ 0, & D_{ii} = 0 \end{cases}$$

Then H^- can be decomposed into $H^- = (\tilde{D}^{-\frac{1}{2}}\tilde{U}^T)^T(\tilde{D}^{-\frac{1}{2}}\tilde{U}^T)$ for some full row rank $\tilde{D}^{-\frac{1}{2}}\tilde{U}^T$.

Proof. Let $\tilde{H} = UDU^T$ be an eigen decomposition of $\tilde{H}$ such that the r nonzero entries of D are at the top-left r diagonal entries of D. That is, $\tilde{H}^T\tilde{H} = \text{diag}(d_{11}^2, d_{22}^2, ..., d_{rr}^2, 0, ..., 0)$, where $d_{11}, ..., d_{rr} > 0$. Now consider the block representations of U and D

$$U = \begin{bmatrix} \tilde{U} & \bar{U} \\ {\scriptstyle q\times r} & {\scriptstyle q\times(q-r)} \end{bmatrix}, D = \begin{bmatrix} \tilde{D} & 0 \\ {\scriptstyle r\times r} & {\scriptstyle r\times(q-r)} \\ 0 & 0 \\ {\scriptstyle (q-r)\times r} & {\scriptstyle (q-r)\times(q-r)} \end{bmatrix}$$

Obviously, $\tilde{H} = UDU^T = \tilde{U}\tilde{D}\tilde{U}^T$. And by the same reasoning, $H^- = \tilde{U}\tilde{D}^{-1}\tilde{U}^T$.

Therefore, consider $\tilde{D}^{-\frac{1}{2}}\tilde{U}^T$, and it can be seen from the definition that this is a full row rank matrix satisfying

$$(\tilde{D}^{-\frac{1}{2}}\tilde{U}^T)^T(\tilde{D}^{-\frac{1}{2}}\tilde{U}^T) = \tilde{U}\tilde{D}^{-\frac{1}{2}}\tilde{D}^{-\frac{1}{2}}\tilde{U}^T = \tilde{U}\tilde{D}^{-1}\tilde{U}^T = H^-$$

□

Next we prove a useful lemma for the derivation of the limiting distribution of T^2. The lemma allows us to express the dot product between $\tilde{D}^{-\frac{1}{2}}\tilde{U}^T\tilde{C}\mathbf{z}$ and any constant vector as a sum over terms indexed by strata, and in particular, independent across strata, thus admitting limiting behaviors as the strata size $s \to \infty$.

Lemma B.1.2. *Let*

$$s(\mathbf{0}) = \begin{bmatrix} s_1(\mathbf{0})^T & | & s_2(\mathbf{0})^T & | & \cdots & | & s_{J-1}(\mathbf{0})^T \end{bmatrix}^T$$

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1^T & | & \mathbf{z}_2^T & | & \cdots & | & \mathbf{z}_{J-1}^T \end{bmatrix}^T$$

be the stacked representations.

Let $C \coloneqq \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{1}_1 & \mathbf{1}_2 & \cdots & \mathbf{1}_s \\ | & | & & | \end{bmatrix} \begin{bmatrix} \overline{x_1}^{(1)} & \overline{x_2}^{(1)} & \cdots & \overline{x_p}^{(1)} \\ \overline{x_1}^{(2)} & \overline{x_2}^{(2)} & \cdots & \overline{x_p}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x_1}^{(s)} & \overline{x_2}^{(s)} & \cdots & \overline{x_p}^{(s)} \end{bmatrix}.$

Equivalently, we write $\mathbf{s}(\mathbf{0})$ *as* $s(\mathbf{0}) = \tilde{C}\mathbf{z}$*, where* $\tilde{C} = \begin{bmatrix} C^T & 0 & \cdots & 0 \\ 0 & C^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C^T \end{bmatrix}.$

Let $a \in \mathbb{R}^r$ *be any vector. Then*

1. $\exists$ *constant vectors* $k^{(1)}, k^{(2)}, \ldots, k^{(s)}$ *and* $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(s)}$ *as functions of* $\mathbf{z}$*, s.t.* $a^T \tilde{D}^{-\frac{1}{2}} \tilde{U}^T \tilde{C}\mathbf{z} = \sum_{b=1}^{s} (k^{(b)})^T \mathbf{z}^{(b)}$*, and furthermore,*
2. $var(a^T \tilde{D}^{-\frac{1}{2}} \tilde{U}^T \tilde{C}\mathbf{z}) = a^T a$.

Proof. We now prove the first claim. Let $a = \begin{bmatrix} a_1 & a_2 & \cdots & a_r \end{bmatrix}^T \in \mathbb{R}^r$ be arbitrary. Define $Q \coloneqq \tilde{D}^{-\frac{1}{2}} \tilde{U}^T \tilde{C}$, with row representation

$$Q = \begin{bmatrix} —q_1^T— \\ —q_2^T— \\ \vdots \\ —q_r^T— \end{bmatrix}$$

Now by reordering rows of $\mathbf{z}$ into

$$\tilde{\mathbf{z}} = \begin{bmatrix} \mathbf{z}^{(1)} \\ \mathbf{z}^{(2)} \\ \vdots \\ \mathbf{z}^{(s)} \end{bmatrix}$$

such that $\mathbf{z}^{(b)}$, the b-th block of $\tilde{\mathbf{z}}$, represents entries in $\mathbf{z}$ that belongs to the b-th stratum.

Reorder the columns of Q accordingly, we obtain

$$\tilde{Q} = \begin{bmatrix} (q_1^{(1)})^T & (q_1^{(2)})^T & \cdots & (q_1^{(s)})^T \\ (q_2^{(1)})^T & (q_2^{(2)})^T & \cdots & (q_2^{(s)})^T \\ \vdots & \vdots & \ddots & \vdots \\ (q_r^{(1)})^T & (q_r^{(2)})^T & \cdots & (q_r^{(s)})^T \end{bmatrix}$$

And note that by matrix multiplication rules $\tilde{Q}\tilde{\mathbf{z}} = Q\mathbf{z}$.
We define $k^{(1)}, k^{(2)}, \ldots, k^{(s)}$ by $k^{(b)} := \sum_{i=1}^{r} a_i (q_i^{(b)})^T$ and observe

$$a^T\tilde{Q} = \begin{bmatrix} (k^{(1)})^T & (k^{(2)})^T & \cdots & (k^{(s)})^T \end{bmatrix}$$

Therefore, it follows that $a^T\tilde{D}^{-\frac{1}{2}}\tilde{U}^T\tilde{C}\mathbf{z} = a^TQ\mathbf{z} = a^T\tilde{Q}\tilde{\mathbf{z}} = \sum_{b=1}^{s}(k^{(b)})^T\mathbf{z}^{(b)}$.
The second claim follows from the following calculations

$$\begin{aligned} \text{var}(a^T\tilde{D}^{-\frac{1}{2}}\tilde{U}^T\tilde{C}\mathbf{z}) &= a^T\tilde{D}^{-\frac{1}{2}}\tilde{U}^T\text{cov}(\tilde{C}\mathbf{z})(\tilde{D}^{-\frac{1}{2}}\tilde{U}^T)^Ta \\ &= a^T\tilde{D}^{-\frac{1}{2}}\tilde{U}^T\text{cov}(\mathbf{s}(\mathbf{0}))\tilde{U}\tilde{D}^{-\frac{1}{2}}a \\ &= a^T\tilde{D}^{-\frac{1}{2}}\tilde{U}^T(-\mathbf{H}(\mathbf{0}))\tilde{U}\tilde{D}^{-\frac{1}{2}}a \\ &= a^T\tilde{D}^{-\frac{1}{2}}\tilde{U}^T\tilde{H}\tilde{U}\tilde{D}^{-\frac{1}{2}}a \\ &= a^T\tilde{D}^{-\frac{1}{2}}\tilde{U}^T(\tilde{U}\tilde{D}\tilde{U}^T)\tilde{U}\tilde{D}^{-\frac{1}{2}}a \\ &= a^T\tilde{D}^{-\frac{1}{2}}\tilde{D}\tilde{D}^{-\frac{1}{2}}a \\ &= a^Ta \end{aligned}$$

where the third inequality follows from a standard result in mathematical statistics. □

We cite the following lemma from Billingsley 1986 as an important tool for our proof.

Lemma B.1.3 (Lindeberg's Theorem)**.** *Let $X_1, X_2, ..., X_s$ be independent random variables with finite variances. Denote $\mathbb{E}[X_b] = \mu_b$, var$(X_b) = \sigma_b^2$, and $\Sigma_s^2 = \sum_{b=1}^{s}\sigma_b^2$. Suppose for $\epsilon > 0$,*

$$\lim_{s\to\infty}\frac{1}{\Sigma_s^2}\sum_{b=1}^{s}\mathbb{E}\Big[(X_b-\mu_b)^2\cdot\mathbf{1}\{|\mathbf{X_b}-\mu_\mathbf{b}| > \epsilon\mathbf{\Sigma_s}\}\Big] = 0$$

then

$$\frac{\sum_{b=1}^{s}X_b-\mu_b}{\Sigma_s} \xrightarrow{d} N(0,1)$$

B.2 Regularity Conditions and the Main Proof

We hereby state our regularity conditions.

Definition B.2.1 (Regularity Conditions)**.** *Define the following regularity conditions:*

1. *(Asymptotic in-stratum Similarity)* $\forall \epsilon > 0$, $\exists S \in \mathbb{N}$ *s.t.* $\forall s \geq S$, *we have*

$$\forall b = 1, ..., s, \max_{i \in I(b)} ||\tilde{\mathbf{x}}_i|| < \epsilon$$

   where $\tilde{\mathbf{x}}_i = \begin{bmatrix} x_{i1} & x_{i2} & ... & x_{ip} \end{bmatrix}^T - \begin{bmatrix} \overline{x_1}^{(b)} & \overline{x_2}^{(b)} & ... & \overline{x_p}^{(b)} \end{bmatrix}^T$ *is the strata-mean centered data vector for the* i*-th subject.*

2. *(Rank Stability)* $\exists S \in \mathbb{N}, \exists r \in \mathbb{N}$ *s.t.* $\forall s \geq S$, $rank(H(\mathbf{0})) = r$.

3. *(Hessian Stability)* $\exists M > 0$ *s.t.* $\forall s \in \mathbb{N}$, $\forall u \in \mathbb{R}^r$ *s.t.* $||u|| = 1$, *we have*

$$||u^T D_s^{-\frac{1}{2}} \tilde{U}_s^T|| < M$$

   where the subscript s *in* $D_s^{-\frac{1}{2}}$, $\tilde{U}_s^T$ *is meant to refer to these quantities for each particular* s.

4. *(Bounded Strata Sizes)* $\exists N \in \mathbb{N}$ *s.t.* $\forall s \in \mathbb{N}$, $\max_{b=1,...,s} n_b < N$

Remark B.2.1. *The conditions 1 can be interpreted as follows under a matching setting: as the total number of strata (sample) increase, we are matching over a larger pool of subjects, which makes it easier to find groups of high similarity.*

Remark B.2.2. *The conditions 2 & 3, although do not appear practically meaningful, are easily satisfied if* $H(\mathbf{0}) \to H_0$ *for some* H_0*, which is in turn a statistically meaningful yet stronger assumption.*

Proposition B.2.1. *Suppose the regularity conditions in Definition B.2.1 hold. Then for $a \in \mathbb{R}^r$,*

$$a^T Q\mathbf{z} \xrightarrow{d} a^T X, \text{ where } X \sim N_r(0, I_r)$$

Proof. For any strata size $s \in \mathbb{N}$, we define the familiar notations that

$$C_s \coloneqq \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{1}_1 & \mathbf{1}_2 & \cdots & \mathbf{1}_s \\ | & | & & | \end{bmatrix} \begin{bmatrix} \overline{x_1}^{(1)} & \overline{x_2}^{(1)} & \cdots & \overline{x_p}^{(1)} \\ \overline{x_1}^{(2)} & \overline{x_2}^{(2)} & \cdots & \overline{x_p}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x_1}^{(s)} & \overline{x_2}^{(s)} & \cdots & \overline{x_p}^{(s)} \end{bmatrix}, \tilde{C}_s \coloneqq \begin{bmatrix} C_s^T & 0 & \cdots & 0 \\ 0 & C_s^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_s^T \end{bmatrix}$$

We may consider the following column representation of C_s^T:

$$C_s^T = \begin{bmatrix} | & & | \\ \tilde{\mathbf{x}}_1 & \cdots & \tilde{\mathbf{x}}_n \\ | & & | \end{bmatrix}, \tilde{\mathbf{x}}_i = \begin{bmatrix} x_{i1} - \overline{x_1}^{(b)} \\ x_{i2} - \overline{x_2}^{(b)} \\ \vdots \\ x_{ip} - \overline{x_p}^{(b)} \end{bmatrix} \text{ where } i \in I(b)$$

That is, the transpose of the strata-mean centered data matrix can be represented column-wise where each column is the strata-mean centered data vector for an observation.

Now, consider the strata-based reordering of the columns of $\tilde{C}_s$ into $\bar{C}_s$ and the corresponding reordering of rows of $\mathbf{z}$ into $\tilde{\mathbf{z}}$:

$$\bar{C}_s = \begin{bmatrix} C_s^{(1)} & \cdots & C_s^{(b)} & \cdots & C_s^{(s)} \end{bmatrix}, \text{ where } C_s^{(b)} = \begin{bmatrix} \tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)} & 0 & \cdots & 0 \\ 0 & \tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)} \end{bmatrix}$$

and

$$\tilde{\mathbf{z}} = \begin{bmatrix} \mathbf{z}^{(1)} \\ \mathbf{z}^{(2)} \\ \vdots \\ \mathbf{z}^{(s)} \end{bmatrix}, \text{ where } \mathbf{z}^{(b)} = \begin{bmatrix} \mathbf{z}_1^{(b)} \\ \mathbf{z}_2^{(b)} \\ \vdots \\ \mathbf{z}_{J-1}^{(b)} \end{bmatrix}, \text{ where } \mathbf{z}_j^{(b)} = \begin{bmatrix} \mathbb{1}\{Y_1^{(b)} \leq j\} \\ \mathbb{1}\{Y_2^{(b)} \leq j\} \\ \vdots \\ \mathbb{1}\{Y_{n_b}^{(b)} \leq j\} \end{bmatrix}$$

where we introduced the notation $(\cdot)_i^{(b)}$, $i = 1, ..., n_b$ to denote a quantity that belongs to the i-th observation of the b-th stratum. For example, $\tilde{\mathbf{x}}_i^{(b)}$ is the centered data vector for the i-th observation in stratum b, and $\mathbb{1}\{Y_i^{(b)} \leq j\}$ is the treatment indicator for the i-th observation in stratum b.

Then we make the following observations:

$$\tilde{C}_s\mathbf{z} = \bar{C}_s\tilde{\mathbf{z}} = \sum_{b=1}^{s} C_s^{(b)}\mathbf{z}^{(b)}, \text{ where } C_s^{(b)}\mathbf{z}^{(b)} = \begin{bmatrix} \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \mathbf{z}_1^{(b)} \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \mathbf{z}_2^{(b)} \\ \vdots \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \mathbf{z}_{J-1}^{(b)} \end{bmatrix}$$

whereas

$$\mathbb{E}[C_s^{(b)}\mathbf{z}^{(b)}] = \begin{bmatrix} \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \mathbb{E}[\mathbf{z}_1^{(b)}] \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \mathbb{E}[\mathbf{z}_2^{(b)}] \\ \vdots \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \mathbb{E}[\mathbf{z}_{J-1}^{(b)}] \end{bmatrix} = \begin{bmatrix} \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \frac{c_{1b}}{n_b}\mathbf{1} \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \frac{c_{2b}}{n_b}\mathbf{1} \\ \vdots \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix} \frac{c_{(J-1)b}}{n_b}\mathbf{1} \end{bmatrix} = \mathbf{0}$$

and the last equality is because the columns of $\begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}$ are mean-centered.

And due to conditioning on $\{C_{jb} = c_{jb}\}, j = 1, ..., J-1, b = 1, ..., s$,

$$\underset{1\times n_b}{\mathbf{1}^T}\mathbf{z}_j^{(b)} = c_{jb}$$

Then, let $a \in \mathbb{R}^r$ be arbitrary. To show

$$a^T Qz \xrightarrow{d} a^T X, \text{ where } X \sim N_r(0, I_r)$$

it suffices to check the Lindeberg's condition.

Now, fix arbitrary $\epsilon > 0$, and according to Hessian Stability in Definition B.2.1, fix $M > 0$ s.t. $\forall k \in \mathbb{N}$, $\forall u \in \mathbb{R}^r$ s.t. $||u|| = 1$, M satisfies

$$||u^T D_k^{-\frac{1}{2}} \tilde{U}_k^T|| < M$$

Now, this implies for $u = \frac{a}{||a||} \in \mathbb{R}^r$, M satisfies

$$||u^T D_s^{-\frac{1}{2}} \tilde{U}_s^T|| < M$$

Furthermore, by the Asymptotic in-stratum Similarity and the boundedness of strata sizes in Definition B.2.1, we may fix $S \in \mathbb{N}$ s.t. $\forall s \geq S$, we have

$$\forall b = 1, ..., s, \max_{i \in I(b)} ||\tilde{\mathbf{x}}_i|| < \epsilon \cdot (M(J-1)N)^{-1} < \epsilon \cdot (M(J-1)\max_{j,b} c_{jb})^{-1}$$

Then by $C_s^{(b)}\mathbf{z}^{(b)} = \begin{bmatrix} \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}\mathbf{z}_1^{(b)} \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}\mathbf{z}_2^{(b)} \\ \vdots \\ \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}\mathbf{z}_{J-1}^{(b)} \end{bmatrix}$ and $\begin{matrix} || \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}\mathbf{z}_1^{(b)}|| < (\max_{i\in I(b)} ||\tilde{\mathbf{x}}_i||)c_{1b} \\ || \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}\mathbf{z}_2^{(b)}|| < (\max_{i\in I(b)} ||\tilde{\mathbf{x}}_i||)c_{2b} \\ \vdots \\ || \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}\mathbf{z}_{J-1}^{(b)}|| < (\max_{i\in I(b)} ||\tilde{\mathbf{x}}_i||)c_{(J-1)b} \end{matrix}$,
we conclude that $\forall b = 1, ..., s$,

$$||C_s^{(b)}\mathbf{z}^{(b)}|| \leq \sum_{j=1}^{J-1} || \begin{bmatrix}\tilde{\mathbf{x}}_1^{(b)} \cdots \tilde{\mathbf{x}}_{n_b}^{(b)}\end{bmatrix}\mathbf{z}_j^{(b)}|| < \sum_{j=1}^{J-1}(\max_{i\in I(b)} ||\tilde{\mathbf{x}}_i||)c_{jb} < \sum_{j=1}^{J-1} \epsilon \cdot [M(J-1)]^{-1} = \epsilon/M$$

Therefore, $\forall s \geq S, \forall b = 1, ..., s$,

$$\begin{aligned}
&\mathbb{P}[||C_s^{(b)}\mathbf{z}^{(b)}|| > \epsilon/M] = 0 \\
\implies &\mathbb{P}[M \cdot ||C_s^{(b)}\mathbf{z}^{(b)}|| > \epsilon] = 0 \\
\implies &\mathbb{P}[||u^T D_s^{-\frac{1}{2}} \tilde{U}_s^T|| \cdot ||C_s^{(b)}\mathbf{z}^{(b)}|| > \epsilon] = 0, \text{ where } u = \frac{a}{||a||} \\
\implies &\mathbb{P}[||a^T D_s^{-\frac{1}{2}} \tilde{U}_s^T|| \cdot ||C_s^{(b)}\mathbf{z}^{(b)}|| > ||a|| \cdot \epsilon] = 0 \\
\implies &\mathbb{P}[||a^T D_s^{-\frac{1}{2}} \tilde{U}_s^T C_s^{(b)}\mathbf{z}^{(b)}|| > ||a|| \cdot \epsilon] = 0 \\
\implies &\mathbb{P}[||(k^{(b)})^T\mathbf{z}^{(b)}|| > ||a|| \cdot \epsilon] = 0
\end{aligned}$$

where the last implication holds if we admit the following decomposition from Lemma B.1.2:

$$a^T \tilde{D}_s^{-\frac{1}{2}} \tilde{U}_s^T \bar{C}_s \tilde{\mathbf{z}} = a^T \tilde{D}_s^{-\frac{1}{2}} \tilde{U}_s^T \tilde{C}_s \mathbf{z} = \sum_{b=1}^{s} (k^{(b)})^T \mathbf{z}^{(b)}$$

And therefore,

$$\begin{aligned}
&\forall s \geq S, \ \sum_{b=1}^{s} \mathbb{E}\Big[[(k^{(b)})^T\mathbf{z}^{(b)}]^2 \cdot \mathbf{1}\{|(k^{(b)})^T\mathbf{z}^{(b)}| > ||a|| \cdot \epsilon\}\Big] = 0 \\
\implies &\lim_{s\to\infty} \sum_{b=1}^{s} \mathbb{E}\Big[[(k^{(b)})^T\mathbf{z}^{(b)}]^2 \cdot \mathbf{1}\{|(k^{(b)})^T\mathbf{z}^{(b)}| > ||a|| \cdot \epsilon\}\Big] = 0
\end{aligned}$$

But $\sum_{b=1}^{s} \text{var}((k^{(b)})^T\mathbf{z}^{(b)}) = a^T a = ||a||^2$, and $\mathbb{E}[(k^{(b)})^T\mathbf{z}^{(b)})] = 0$, and the Lindeberg Theorem's implies

$$\begin{aligned}
&\frac{1}{||a||} \sum_{b=1}^{s} (k^{(b)})^T\mathbf{z}^{(b)} \xrightarrow{d} N(0,1) \\
\implies &a^T \tilde{D}_s^{-\frac{1}{2}} \tilde{U}_s^T \tilde{C}_s \mathbf{z} = \sum_{b=1}^{s} (k^{(b)})^T\mathbf{z}^{(b)} \xrightarrow{d} N(0, a^T a)
\end{aligned}$$

□

To finalize the proof of the χ^2_r distribution of T^2, we introduce the following lemma (Billingsley 1986).

Lemma B.2.1 (Cramér-Wold Device)**.** *For random vectors $X_n = (X_{n1}, ..., X_{nk})$ and $Y = (Y_1, ..., Y_k)$, a necessary and sufficient condition for $X_n \xrightarrow{d} Y$ is that $t^T X_n \xrightarrow{d} t^T Y$ for each $t = (t_1, ..., t_k)$ in $\mathbb{R}^k$.*

This allows us to state the following conclusions as final pieces of the proof.

Corollary B.2.1. *Suppose the regularity conditions in Definition B.2.1 hold, then*

$$Q\mathbf{z} \xrightarrow{d} N_r(0, I_r)$$

Proof. By B.2.1 and B.2.1 □

Corollary B.2.2. *Suppose the regularity conditions in Definition B.2.1 hold, then*

$$T^2 \xrightarrow{d} \chi^2_r$$

Proof.

$$\begin{aligned}
T^2 &= s(\mathbf{0})^T(-H(\mathbf{0}))s(\mathbf{0}) \\
&= s(\mathbf{0})^T\tilde{H}s(\mathbf{0}) \\
&= s(\mathbf{0})^T(UDU^T)s(\mathbf{0}) \\
&= s(\mathbf{0})^T(\tilde{U}\tilde{D}^{-\frac{1}{2}}\tilde{D}^{-\frac{1}{2}}\tilde{U}^T)s(\mathbf{0}) \\
&= \mathbf{z}^T\tilde{C}^T(\tilde{U}\tilde{D}^{-\frac{1}{2}}\tilde{D}^{-\frac{1}{2}}\tilde{U}^T)\tilde{C}\mathbf{z} \\
&= (\mathbf{z}^T\tilde{C}^T\tilde{U}\tilde{D}^{-\frac{1}{2}})(\tilde{D}^{-\frac{1}{2}}\tilde{U}^T\tilde{C}\mathbf{z}) \\
&= (Q\mathbf{z})^T(Q\mathbf{z}) \xrightarrow{d} \chi^2_r
\end{aligned}$$

□