Biost/Stat 533
Emerson, Spr 14
Homework #4 Key
April 30, 2014

Written problems to be handed in Monday, May 5.

Under no circumstances may you refer to a homework key from this or other classes. While you may work with other students to derive a solution, when you write up your final solution, you may not refer to any other source. You must be able to develop your answer as if it were being done in a closed book, closed notes examination. You must provide a signed pledge to that effect:

On my honor I have neither given nor received unauthorized aid on the completion of this homework.

1. Consider again the setting in which $Y_i \sim (\mu_0, \sigma^2)$ for $i = 1, \ldots, n_0$ and $Y_i \sim (\mu_1, \sigma^2)$ for $i = n_0 + 1, \ldots, n = n_0 + n_1 = 2n_0$, except observations within each group are correlated. That is, we have $Cov(Y_i, Y_j) = \rho\sigma^2$ for $i, j = 1, \ldots, n_0; i \neq j$, $Cov(Y_i, Y_j) = \rho\sigma^2$ for $i, j = n_0 + 1, \ldots, n; i \neq j$, and $Cov(Y_i, Y_j) = 0$ for $i = 1, \ldots, n_0; j = n_0 + 1, \ldots, n$. For notational convenience, let $\vec{w}$ be an n-vector such that $w_i = 1$ for $1 \leq i \leq n_0$ and $w_i = 0$ otherwise, and let $\vec{z} = \vec{1}_n - \vec{w}$. Consider linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with $\mathbf{X} = (\vec{w} \quad \vec{z})$. We are interested in estimating $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$.

   a. Show that the ordinary least squares estimator $\widehat{\vec{\beta}}$ is equal to the generalized least squares estimator $\widehat{\vec{\beta}}_G$. What is the mean and variance of these estimators?

Ans: The OLSE $\widehat{\vec{\beta}}$ is found from

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \vec{w}^T\vec{w} & \vec{w}^T\vec{z} \\ \vec{z}^T\vec{w} & \vec{z}^T\vec{z} \end{pmatrix} = \begin{pmatrix} n_0 & 0 \\ 0 & n_1 \end{pmatrix} \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_0} & 0 \\ 0 & \frac{1}{n_1} \end{pmatrix}$$

$$\mathbf{X}^T\vec{Y} = \begin{pmatrix} n_0\overline{Y}_0 \\ n_1\overline{Y}_1 \end{pmatrix} \qquad \widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y} = \begin{pmatrix} \overline{Y}_0 \\ \overline{Y}_1 \end{pmatrix}$$

Taking the expectation of $\widehat{\vec{\beta}}$ we find

$$E[\widehat{\vec{\beta}}] = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}$$

The variance of $\widehat{\vec{\beta}}$ is found by

$$Var(\widehat{\vec{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\vec{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

$$= \begin{pmatrix} \frac{1}{n_0} & 0 \\ 0 & \frac{1}{n_1} \end{pmatrix} \sigma^2 \begin{pmatrix} n_0(1 + (n_0 - 1)\rho) & 0 \\ 0 & n_1(1 + (n_1 - 1)\rho) \end{pmatrix} \begin{pmatrix} \frac{1}{n_0} & 0 \\ 0 & \frac{1}{n_1} \end{pmatrix}$$

$$= \sigma^2 \begin{pmatrix} \frac{1+(n_0-1)\rho}{n_0} & 0 \\ 0 & \frac{1+(n_1-1)\rho}{n_1} \end{pmatrix}$$

To find the GLSE $\widehat{\vec{\beta}}_G$, we first consider the form of $\mathbf{V}^{-1}$. Let $\mathbf{R}_m$ be a $m \times m$ matrix with 1's on the

diagonal ane ρ elsewhere, and $\mathbf{O}$ be a conformable matrix full of 0's. Then

$$\mathbf{V} = \sigma^2 \begin{pmatrix} \mathbf{R}_{n_0} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_{n_1} \end{pmatrix} \quad \text{and} \quad \mathbf{V}^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} \mathbf{R}_{n_0}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_{n_1}^{-1} \end{pmatrix}$$

where $\mathbf{R}_m^{-1}$ has the same symmetrical structure as $\mathbf{R}_m$. Let the diagonal elements of $\mathbf{R}_m^{-1}$ be equal to r and the off diagonal elements be equal to s. Then because $\mathbf{R}_m\mathbf{R}_m^{-1} = \mathbf{I}_m$ we have the simultaneous equations

$$\begin{aligned} 1 &= r + (m-1)s\rho \\ 0 &= r\rho + s + (m-2)s\rho \end{aligned}$$

which can be solved to yield

$$\begin{aligned} r &= \frac{1 + (m-2)\rho}{1 + (m-2)\rho - (m-1)\rho^2} \\ s &= -\frac{\rho}{1 + (m-2)\rho - (m-1)\rho^2} \end{aligned}$$

Let r_0 and s_0 be the values of r and s when $m = n_0$, and r_1 and s_1 be the values of r and s when $m = n_1$. From this we can then find

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} = \frac{1}{\sigma^2} \begin{pmatrix} n_0(r_0 + (n_0-1)s_0) & 0 \\ 0 & n_1(r_1 + (n_1-1)s_1) \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n_0(r_0+(n_0-1)s_0)} & 0 \\ 0 & \frac{1}{n_1(r_1+(n_1-1)s_1)} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{V}^{-1}\vec{Y} = \begin{pmatrix} n_0(r_0 + (n_0-1)s_0)\overline{Y}_0 \\ n_1(r_1 + (n_1-1)s_1)\overline{Y}_1 \end{pmatrix}$$

$$\widehat{\vec{\beta}}_G = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\vec{Y} = \begin{pmatrix} \overline{Y}_0 \\ \overline{Y}_1 \end{pmatrix}$$

which is the same as the OLSE, and thus has the same expectation and variance (you can check that $\sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ gives the same answer as found above– it does).

Note that this agreement between the OLSE and GLSE in this case is specific to the particular design matrix. In general the OLSE and GLSE will not be equal. However, when they are equal, it only stands to reason that their standard errors must also be equal. This does not mean, however, that standard statistical software for OLSE and standard statistical software for GLSE will provide the same inference. That will have to do with the estimates of σ^2 as noted in the part b.

b. Provide an estimate of the variance of $\widehat{\vec{\beta}}_G$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ assuming that ρ is known.

Ans: The variance of $\widehat{\vec{\beta}}_G$ is given above. In order to estimate $\mu_1 - \mu_0$, we are interested in estimating $\vec{a}^T\vec{\beta}$, where $\vec{a} = (-1\ 1)^T$. The variance of the GLSE for that estimable function is thus

$$Var(\vec{a}^T\widehat{\vec{\beta}}_G) = \vec{a}^T Var(\widehat{\vec{\beta}}_G)\vec{a} = \sigma^2 \left(\frac{1 + (n_0-1)\rho}{n_0} + \frac{1 + (n_1-1)\rho}{n_1} \right)$$

Now we know n_0, n_1, and (by assumption) ρ. Hence to estimate the variance, we only need estimate σ^2. It would seem logical to consider the residuals $\vec{e} = \vec{Y} - \mathbf{X}\widehat{\vec{\beta}}_G$, which owing to the unbiasedness of the GLSE would have distribution

$$\vec{e} \sim \left(\vec{0}, \mathbf{V} = \sigma^2 \begin{pmatrix} \mathbf{R}_{n_0} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_{n_1} \end{pmatrix} \right).$$

Owing to the correlation among the residuals, the sample variance of the residuals will not estimate σ^2 directly. But we can transform the residuals to independence, and then take the sample variance of those transformed residuals. To do this we find a transformation matrix $\mathbf{A}$ such that $\mathbf{AVA}^T = \sigma^2\mathbf{I}_n$. We can find such a matrix by considering the linear algebra result that says that every symmetric positive definite matrix $\mathbf{V}$ can be expressed as a product involving an invertible symmetric matrix $\mathbf{V} = \mathbf{V}^{1/2}\mathbf{V}^{1/2}$ (where the notation is obviously mnemonic). For our purposes, then, we would want to find matrices $\mathbf{R}_{n_0}^{1/2}$ and $\mathbf{R}_{n_1}^{1/2}$, and then define our "whitening" transformation (terminology out of signal process, where independent errors simulate white noise) as

$$\mathbf{A} = \begin{pmatrix} \mathbf{R}_{n_0}^{-1/2} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_{n_1}^{-1/2} \end{pmatrix},$$

where, again, we use the obvious notation that $\mathbf{R}_{n_0}^{-1/2}$ is the inverse of $\mathbf{R}_{n_0}^{1/2}$.

On approach to finding $\mathbf{R}_{n_0}^{1/2}$ is to guess that it will have a structure similar to $\mathbf{R}$ with some constant a on the diagonal and another constant b on the off-diagonals. Then we would have that

$$\begin{aligned} 1 &= a^2 + (n_0 - 1)b^2 \\ \rho &= 2ab + (n_0 - 1)b^2 \end{aligned}$$

which can be solved to find

$$\begin{aligned} a &= \frac{(n_0 - 1)\sqrt{1-\rho} \pm \sqrt{1 + (n_0 - 1)\rho}}{n} \\ b &= \frac{-\sqrt{1-\rho} \pm \sqrt{1 + (n_0 - 1)\rho}}{n}. \end{aligned}$$

(Note that either the plus or the minus will work.) We would then have

$$\mathbf{A}\vec{e} \sim \left(\vec{0}, \sigma^2\mathbf{I}_n\right),$$

and could use

$$\hat{\sigma}^2 = \frac{1}{n}\vec{e}^T\mathbf{A}^T\mathbf{A}\vec{e}$$

as a consistent estimate (where consistency comes from WLLN). (Note that our true usual practice would be to divide by $n-2$ in this problem, due to the 2 dimensional $\widehat{\vec{\beta}}_G$. This would give an unbiased estimate of σ^2.)

c. Provide an estimate of the variance of $\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}$ under the assumption that the observations are independent. How do they compare to the answers in b?

Ans: When we assume $\rho = 0$, we obtain

$$Var(\widehat{\vec{\beta}}) = \sigma^2 \begin{pmatrix} \frac{1}{n_0} & 0 \\ 0 & \frac{1}{n_1} \end{pmatrix}$$

$$Var(\vec{a}^T\widehat{\vec{\beta}}) = \sigma^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right)$$

Note that for positive ρ, the true variance is greater than that which would be estimated when we assume $\rho = 0$. Thus in this case where the data within groups defined by predictors are positively correlated, inference based on the assumption of independence with the true value of σ^2 would be anti-conservative. Of course, if we presume independence of the observations, we would not transform the residuals to estimate σ^2. For the same vector of residuals, we can show that for $\rho > 0$ (and this is a

limit to which the correlation can be negative within the "exchangeable" structure for the correlation within a group that we are considering here)

$$\vec{e}^T\vec{e} - \vec{e}^T\mathbf{A}^T\mathbf{A}\vec{e} < 0,$$

thus by incorrectly assuming independence, when having to estimate our nuisance parameter σ^2, we will also underestimate σ^2, thereby making our inference even more anti-conservative.

Note that the degree of error we make depends both on ρ and the sample size n_0 and n_1 within "clusters": The increase in variability over what might be obtained with independent observations depends on the product of sample size and correlation. Hence, even very small correlation in large clusters (e.g., hospitals, schools, cities) causes a problem and must be accounted for.

2. Now consider the setting in which $Y_i \sim (\mu_0, \sigma^2)$ for $i = 1, \ldots, n_0$ and $Y_i \sim (\mu_1, \sigma^2)$ for $i = n_0 + 1, \ldots, n = n_0 + n_1 = 2n_0$, except observations are paired across groups. That is, we have $Cov(Y_i, Y_i) = \sigma^2$ for $i = 1, \ldots, n$, $Cov(Y_i, Y_{n_0+i}) = \rho\sigma^2$ for $i = 1, \ldots, n_0$, and $Cov(Y_i, Y_j) = 0$ otherwise. For notational convenience, let $\vec{w}$ be an n-vector such that $w_i = 1$ for $1 \leq i \leq n_0$ and $w_i = 0$ otherwise, and let $\vec{z} = \vec{1}_n - \vec{w}$. Consider linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with $\mathbf{X} = (\vec{w} \quad \vec{z})$. We are interested in estimating $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$.

   a. Show that the ordinary least squares estimator $\widehat{\vec{\beta}}$ is equal to the generalized least squares estimator $\widehat{\vec{\beta}}_G$. What is the mean and variance of these estimators?

Ans: The OLSE $\widehat{\vec{\beta}}$ is the same as given in problem 1a, and the expectation is the same as was given in that answer. The variance of $\widehat{\vec{\beta}}$ is found from the results for $(\mathbf{X}^T\mathbf{X})^{-1}$ with $n_0 = n_1$

$$\mathbf{V} = Var(\vec{Y}) = \sigma^2 \begin{pmatrix} \mathbf{I}_{n_0} & \rho\mathbf{I}_{n_0} \\ \rho\mathbf{I}_{n_0} & \mathbf{I}_{n_0} \end{pmatrix}$$

$$\begin{aligned} Var(\widehat{\vec{\beta}}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\vec{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \begin{pmatrix} \frac{1}{n_0} & 0 \\ 0 & \frac{1}{n_0} \end{pmatrix} \sigma^2 \begin{pmatrix} n_0 & n_0\rho \\ n_0\rho & n_0 \end{pmatrix} \begin{pmatrix} \frac{1}{n_0} & 0 \\ 0 & \frac{1}{n_0} \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} \frac{1}{n_0} & \frac{\rho}{n_0} \\ \frac{\rho}{n_0} & \frac{1}{n_0} \end{pmatrix} \end{aligned}$$

To find the GLSE $\widehat{\vec{\beta}}_G$, we use the result for inverse of a symmetric partitioned matrix to find

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} \frac{1}{1-\rho^2}\mathbf{I}_{n_0} & \frac{-\rho}{1-\rho^2}\mathbf{I}_{n_0} \\ \frac{-\rho}{1-\rho^2}\mathbf{I}_{n_0} & \frac{1}{1-\rho^2}\mathbf{I}_{n_0} \end{pmatrix}$$

. From this we can then find

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} = \frac{1}{\sigma^2} \begin{pmatrix} \frac{n_0}{1-\rho^2} & -\frac{n_0\rho}{1-\rho^2} \\ -\frac{n_0\rho}{1-\rho^2} & \frac{n_0}{1-\rho^2} \end{pmatrix} \qquad (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n_0} & \frac{\rho}{n_0} \\ \frac{\rho}{n_0} & \frac{1}{n_0} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{V}^{-1}\vec{Y} = \frac{1}{\sigma^2} \begin{pmatrix} \frac{n_0}{1-\rho^2}\overline{Y}_0 - \frac{n_0\rho}{1-\rho^2}\overline{Y}_1 \\ \frac{n_0}{1-\rho^2}\overline{Y}_1 - \frac{n_0\rho}{1-\rho^2}\overline{Y}_0 \end{pmatrix} \qquad \widehat{\vec{\beta}}_G = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\vec{Y} = \begin{pmatrix} \overline{Y}_0 \\ \overline{Y}_1 \end{pmatrix}$$

which is the same as the OLSE, and thus has the same expectation and variance (you can check that $\sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ gives the same answer as found above– it does).

Note that this agreement between the OLSE and GLSE in this case is specific to the particular design matrix. In general the OLSE and GLSE will not be equal. However, when they are equal, it only stands to reason that their standard errors must also be equal. This does not mean, however, that standard statistical software for OLSE and standard statistical software for GLSE will provide the same inference. That will have to do with the estimates of σ^2 as noted in part b.

b. Provide an estimate of the variance of $\widehat{\vec{\beta}}_G$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ assuming that ρ is known.

Ans: The variance of $\widehat{\vec{\beta}}_G$ is given above. In order to estimate $\mu_1 - \mu_0$, we are interested in estimating $\vec{a}^T\vec{\beta}$, where $\vec{a} = (-1\ 1)^T$. The variance of the GLSE for that estimable function is thus

$$Var(\vec{a}^T\widehat{\vec{\beta}}_G) = \vec{a}^T Var(\widehat{\vec{\beta}}_G)\vec{a} = \sigma^2\frac{2(1-\rho)}{n_0}$$

We again have to estimate σ^2, which we can effect by methods similar to those used for problem 1. We can also think about it quite simply in this case: The paired observations would allow us to note that $Y_i - Y_{n_0+i} \sim \left(\mu_0 - \mu_1, \sigma^2(2-2\rho)\right)$, so we could take the sample variance of the paired differences and obtain an unbiased estimate of $\sigma^2(2-2\rho)$, and then use the known value of ρ to solve for an unbiased estimate of σ^2.

c. Provide an estimate of the variance of $\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}$ under the assumption that the observations are independent. How do they compare to the answers in b?

Ans: When we assume $\rho = 0$, we obtain

$$Var(\widehat{\vec{\beta}}) = \sigma^2\begin{pmatrix}\frac{1}{n_0} & 0\\ 0 & \frac{1}{n_1}\end{pmatrix}$$

$$Var(\vec{a}^T\widehat{\vec{\beta}}) = \sigma^2\left(\frac{2}{n_0}\right)$$

Note that for positive ρ, the true variance is less than that which would be estimated when we assume $\rho = 0$. Furthermore, when making inference using an estimate of σ^2, incorrectly assuming independence rather than a true positive correlation would overestimate σ^2. Thus in this case when the correlated observations are sampled at different values of the covariate, inference based on the assumption of independence would be conservative, resulting in a substantial loss of statistical power.

d. How does the effect of correlated observations affect an ordinary least squares analysis differ when the correlated observations are within groups sharing the same predictor values versus when the correlated observations have different predictor values?

Ans: As noted above, when we consider a cluster of correlated observations of response, if the correlation among the predictors is of the same sign as the correlation among the errors within that cluster, the true variance tends to be greater than the variance estimated under independence, and tests and confidence intervals will be anti-conservative. On the other hand, if the correlation among the predictors within a cluster is of opposite sign of the correlation among the errors, then the true variance tends to be smaller than the variance estimated under independence.

So, for instance, in problem 1 the predictors in a cluster were positively correlated in the sense that the cluster had all the same values for the predictor. In that problem, when $\rho > 0$, the estimated variance was too small. However, if $\rho < 0$ in that problem, the variance estimated under independence is too large. But as noted in problem 1, there is a lower bound on how negative a common correlation may be for a specific sample size within clusters: For a cluster size of 2, any correlation is possible, for a cluster size of n, we must have $\rho > -1/(n-1)$.

In problem 2, the predictors in a cluster were negatively correlated in the sense that repeated observations within a cluster were for different values of the predictor. In that problem, when $\rho > 0$, the variance

estimated under independence was too large. On the other hand, if $\rho < 0$ the variance estimated under independence was too small, thereby leading to anti-conservative testing.

3. Let $Y_i \sim Bernoulli(p_i), i = 1, \ldots, n$ be independent random variables with $p_i = \vec{x}_i^T \vec{\beta}$ for known predictor vector $\vec{x}_i$.

   a. Is inference about $\vec{\beta}$ using ordinary least squares regression analysis asymptotically valid for this problem? If so, provide justification. If not, are there any situations in which it might be approximately valid?

Ans: In the Bernoulli model, we have regression model $Y_i = \vec{x}_i^T \vec{\beta} + \epsilon_i$ with ϵ_i's being independently distributed with mean 0 and variance $p_i(1-p_i)$. Hence, unless $\vec{x}_i^T \vec{\beta}$ is constant for all $i = 1, \ldots, n$ (as it would be under the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$), there is heteroscedasticity, with a relationship between the predictors and the magnitude of the error variance. Hence, inference based on OLS will in general not be correct. Note, however, that if $.4 \leq p_i \leq .6$, then $.24 \leq p_i(1-p_i) \leq .25$, and the heteroscedasticity is not too severe. In fact, a commonly quoted rule of thumb is that if $.3 \leq p_i \leq .7$, then $.21 \leq p_i(1-p_i) \leq .25$, and OLS based inference in this setting will generally be okay. Some others will further relax this to $.2 \leq p_i \leq .8$. I note that if the x_i's are sampled such that the distribution of p_i's is fairly symmetric about .5, then we can use the results of problem 6 of homework 3 about the α_i's (because there will be no linear trend in the error variances with the x_i's) to infer that the OLS inference will tend to be conservative, because the more extreme values of x_i will tend to be associated with the smaller error variance. On the other hand, if the x_i's are sampled such that the distribution of the p_i's is skewed about .5, then the results of problem 6 of homework 3 about γ can be used, because there will tend to be a trend in the error variance with the values of the x_i's.

All of the above caveats have to do with trying to make quantitative inference about β_1. With respect to qualitative inference about an association between Y and X, there is no possibility of invalid inference under the strongest null hypothesis of distributions that are exactly equal for all values of X, because in that case there can be no heteroscedasticity and tests of the strong null hypothesis are statistically valid. It is the estimation under the alternative (as required for frequentist CI and <u>any</u> Bayesian inference) that would ever be problematic.

Of course, if we were merely trying to test for no linear trend in p by X (a very weak null hypothesis), we could have the wrong distribution for β_1 due to the heteroscedasticity possible under this weak null. As a rule, the conservatism or anti-conservatism of this inference in "attributable risk regression" will depend upon the individual group proportions.

   b. Describe an iterative approach in which weighted least squares might be used to address this problem. What undesirable small sample behavior with respect to the range of estimates $\hat{p}_i$ might persist under this analysis scheme?

Ans: If we knew the error variances, we could use weighted least squares (generalized least squares with a diagonal covariance matrix for $\vec{\epsilon}$) to obtain valid inference. One approach would be to first use OLS to estimate $\widehat{\vec{\beta}}^{(0)}$ and $\hat{p}_i^{(0)} = \vec{x}_i^T \widehat{\vec{\beta}}^{(0)}$. Then use $\mathbf{V}^{(0)}$ with $V_{ii}^{(0)} = \hat{p}_i^{(0)}(1-\hat{p}_i^{(0)})$ and $V_{ij}^{(0)} = 0$ for $i \neq j$ to find GLSE $\widehat{\vec{\beta}}_G^{(1)}$. These estimates are then used to find $\hat{p}_i^{(1)}$ and $\mathbf{V}^{(1)}$. The process is then repeated with GLSE $\widehat{\vec{\beta}}_G^{(k)}$ estimated using $\mathbf{V}^{(k-1)}$ until $(\widehat{\vec{\beta}}_G^{(k)} - \widehat{\vec{\beta}}_G(k-1))^T (\widehat{\vec{\beta}}_G^{(k)} - \widehat{\vec{\beta}}_G(k-1))$ is sufficiently small. Inference is then based on estimates of the variance of the regression parameter vector derived under weighted (generalized) least squares theory.

This is an asymptotically valid procedure under the correct model. However, in the setting of small samples, it may well happen that estimates $\hat{p}_i$ would be less than 0 or greater than 1. This is often regarded as undesirable, because such estimates would lead to nonsensible weights in the GLSE. This problem does not occur in logistic regression, where instead of the mean p_i, the log odds $\log(p_i/(1-p_i))$ (which can range from $-\infty$ to ∞) is modeled by $\vec{x}_i^T \vec{\beta}$. Finding estimates for logistic regression uses an iteratively reweighted least squares approach very similar to that described above, except transformations of the observations are used. (This will be discussed later in the quarter.)

4. Consider a linear regression model relating response $\vec{Y}$ to an intercept and two predictor vectors $\vec{W}$ and $\vec{Z}$ (so design matrix $\mathbf{X} = (\vec{1}_n \quad \vec{W} \quad \vec{Z})$ has $X_{i1} \equiv 1$ for $i = 1, \ldots, n$ and $X_{i2} = W_i$ and $X_{i3} = Z_i$ and $\vec{\beta} = (\beta_0, \beta_1, \beta_2)^T$). Assume $E[\vec{\epsilon}] = \vec{0}$ and $var(\vec{\epsilon}) = \sigma^2 \mathbf{I}_n$.

   a. Show that the correlation between OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ is opposite in sign to the sample correlation between $\vec{W}$ and $\vec{Z}$ and that the two slope estimates are uncorrelated if the sample correlation between $\vec{W}$ and $\vec{Z}$ is zero.

Ans: I will work the first part of this problem in more generality, assuming $(p-1)$ covariates. Partition $\mathbf{X} = (\vec{1}_n \quad \mathbf{U})$ similar to problem 1 of homework 3. By problem 1 of homework 3, we can without loss of generality center the covariates to obtain $\mathbf{U}^* = (\mathbf{I}_n - \frac{1}{n}\vec{1}_n\vec{1}_n^T))\mathbf{U}$. Hence, the covariance between the jth and kth parameter estimates will the (j,k)th element of $\sigma^2(\mathbf{U}^{*T}\mathbf{U}^*)^{-1}$. Without loss of generality, assume $\sigma^2 = 1$. Due to the symmetry of the problem, it will be sufficient to consider the covariance between $\hat{\beta}_1$ and $\hat{\beta}_j$ for $j = 2, \ldots, p-1$. We thus further partition $\mathbf{U}^* = (\vec{W}^* \quad \mathbf{V}^*)$ into an n by 1 matrix and an n by $(p-2)$ matrix. Hence

$$\mathbf{U}^{*T}\mathbf{U}^* = \begin{pmatrix} \vec{W}^{*T}\vec{W} & \vec{W}^{*T}\mathbf{V}^* \\ \mathbf{V}^{*T}\vec{W}^* & \mathbf{V}^{*T}\mathbf{V}^* \end{pmatrix}$$

Using the formula for the inverse of a partitioned matrix we find that the upper left matrix in the partition of $(\mathbf{U}^{*T}\mathbf{U}^*)^{-1}$ is

$$\begin{aligned} var(\hat{\beta}_1) &= (\vec{W}^{*T}\vec{W}^*)^{-1} \\ &\quad + (\vec{W}^{*T}\vec{W}^*)^{-1}\vec{W}^{*T}\mathbf{V}^*(\mathbf{V}^{*T}\mathbf{V}^* - \mathbf{V}^{*T}\vec{W}^*(\vec{W}^{*T}\vec{W}^*)^{-1}\vec{W}^{*T}\mathbf{V}^*)^{-1}\mathbf{V}^{*T}\vec{W}^*(\vec{W}^{*T}\vec{W}^*)^{-1} \end{aligned}$$

and the upper right matrix in the partition of $(\mathbf{U}^{*T}\mathbf{U}^*)^{-1}$ is

$$cov(\hat{\beta}_1, (\hat{\beta}_2, \ldots, \hat{\beta}_{p-1})) = -(\vec{W}^{*T}\vec{W}^*)^{-1}\vec{W}^{*T}\mathbf{V}^*(\mathbf{V}^{*T}\mathbf{V}^* - \mathbf{V}^{*T}\vec{W}^*(\vec{W}^{*T}\vec{W}^*)^{-1}\vec{W}^{*T}\mathbf{V}^*)^{-1}$$

Now suppose $p = 3$ and $\mathbf{V}^* = \vec{Z}^*$. Then

$$\begin{aligned} \vec{W}^{*T}\vec{W}^* &= S_{WW} \\ \vec{W}^{*T}\mathbf{V}^* = \vec{W}^{*T}\vec{Z}^* &= S_{WZ} \\ \mathbf{V}^{*T}\mathbf{V}^* &= S_{ZZ} \end{aligned}$$

Letting $r_{WZ} = S_{WZ}/\sqrt{S_{WW}S_{ZZ}}$ be the sample correlation between $\vec{W}$ and $\vec{Z}$, we thus have

$$\begin{aligned} var(\hat{\beta}_1) &= \frac{1}{S_{WW}}\left(\frac{1}{1-r_{WZ}^2}\right) \\ cov(\hat{\beta}_1, \hat{\beta}_2) &= \frac{-r_{WZ}}{(1-r_{WZ}^2)\sqrt{S_{WW}S_{ZZ}}} \end{aligned}$$

By inspection, the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ is opposite in sign to the sample correlation r_{WZ}, and it will only be zero if $\vec{W}$ and $\vec{Z}$ are uncorrelated.

   b. Suppose we hold $S_{WW} = (\vec{W} - E[\vec{W}])^T(\vec{W} - E[\vec{W}])$, S_{ZZ}, and σ^2 constant, but we may freely vary $S_{WZ} = (\vec{W} - E[\vec{W}])^T(\vec{Z} - E[\vec{Z}])$. For what value of S_{WZ} do we minimize the variance of $\hat{\beta}_1$ and $\hat{\beta}_2$? What does this suggest about our ability to test for an association between Y and W adjusting for Z when W and Z are correlated?

Ans: From the results given above, it can be seen that the variance of $\hat{\beta}_1$ increases as the absolute value of the sample correlation between $\vec{W}$ and $\vec{Z}$ increases. Hence, we will have the greatest power to detect an

association between Y and W when the sample correlation between W and Z is 0. This would be true on average if we sample in such a way that W and Z are independent (e.g., a completely randomized design), but we can obtain more efficient studies if we guarantee that W and Z are uncorrelated in our sample through experimental design.

This tendency for the standard error of a slope estimate to be increased by the modelling of a correlated variable is termed 'variance inflation'. Note that this effect exists even when the correlated variable does not predict the response. This in turn argues that adjusting for truly unimportant variables decreases the statistical power to detect associations between the response and other predictors.

5. Consider again the linear regression model in which we will assume the true model is

$$\vec{Y} = \beta_0 + \vec{W}\beta_1 + \vec{Z}\beta_2 + \vec{\epsilon}$$

but we want to also consider fitting a model

$$\vec{Y} = \gamma_0 + \vec{W}\gamma_1 + \vec{\epsilon}^*$$

.

a. Under what conditions is the OLS estimate $\hat{\beta}_1$ equal to the OLS estimate $\hat{\gamma}_1$?

Ans: Suppose that $\vec{1}_n^T\vec{W} = \vec{1}_n^T\vec{Z} = 0$. (This can be achieved by centering the covariate vectors, and by problem 1 of homework 3, this does not affect the slope parameter estimates or distributions. Later we will consider the general case.) Define $\mathbf{W} = (\vec{1}_n \quad \vec{W})$ and $\mathbf{X} = (\mathbf{W} \quad \vec{Z})$. Then

$$\widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$$
$$\widehat{\vec{\gamma}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\vec{Y}$$

and we want to find when $(0 \quad 1)\widehat{\vec{\gamma}} = (0 \quad 1 \quad 0)\widehat{\vec{\beta}}$. Following the approaches used above with $r = r_{WZ} = S_{WZ}/\sqrt{S_{WW}S_{ZZ}}$ we find

$$(\mathbf{W}^T\mathbf{W})^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{WW}} \end{pmatrix}$$
$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n} & 0 & 0 \\ 0 & \frac{1}{S_{WW}(1-r^2)} & -\frac{r}{(1-r^2)\sqrt{S_{WW}S_{ZZ}}} \\ 0 & -\frac{r}{(1-r^2)\sqrt{S_{WW}S_{ZZ}}} & \frac{1}{S_{ZZ}(1-r^2)} \end{pmatrix}$$

Thus $\hat{\beta}_1 = \hat{\gamma}_1$ when

$$\frac{\vec{W}^T\vec{Y}}{(1-r^2)S_{WW}} - \frac{r\vec{Z}^T\vec{Y}}{(1-r^2)\sqrt{S_{WW}S_{ZZ}}} = \frac{\vec{W}^T\vec{Y}}{S_{WW}}$$

which in turn is satisfied if $r = 0$ or if

$$r = \sqrt{\frac{S_{WW}}{S_{ZZ}}}\frac{\vec{Z}^T\vec{Y}}{\vec{W}^T\vec{Y}}$$

Obviously, $\vec{Y}$ is random, and thus the second condition cannot be set by experimental design. We can set $r_{WZ} = 0$ by experimental design.

For arbitrary $\vec{W}$ and $\vec{Z}$, the above results obtain so long as the centered vectors have correlation 0. Of course, adding constants to vectors does not change their correlation, so for arbitrary $\vec{W}$ and $\vec{Z}$, $\hat{\gamma}_1 = \hat{\beta}_1$ so long as $S_{WZ}/\sqrt{S_{WW}S_{ZZ}} = 0$.

b. Under what conditions is the standard error of $\hat{\beta}_1$ equal to the standard error of $\hat{\gamma}_1$.

Ans: Now

$$var(\widehat{\vec{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}, \text{ and}$$
$$var(\widehat{\vec{\gamma}}) = \tau^2(\mathbf{W}^T\mathbf{W})^{-1}$$

where $\sigma^2 = var(Y|W,Z)$ and

$$\tau^2 = var(Y|W) = E_Z[var(Y|W,Z)] + var_Z(E[Y|W,Z]) = \sigma^2 + \beta_2^2 var(Z|W).$$

For the standard errors of $\hat{\beta}_2$ and $\hat{\gamma}_2$ to be equal, we must have

$$\sigma^2 \frac{1}{(1-r^2)S_{WW}} = (\sigma^2 + \beta_2^2 var(Z|W))\frac{1}{S_{WW}}$$

This will be satisfied if $r = 0$ and $\beta_2 = 0$ or if $r = 0$ and $var(Z|W) = 0$. The above equation can also be satisfied by putting suitable restrictions on $var(Z|W) = \frac{r^2\sigma^2}{(\beta_2^2(1-r^2))}$ for nonzero β_2, but this is difficult to do by experimental design when β_2 is unknown.

c. Under what conditions is the <u>estimated</u> standard error of $\hat{\beta}_1$ equal to the <u>estimated</u> standard error of $\hat{\gamma}_1$.

Ans: We would typically estimate

$$\widehat{Var}(\hat{\beta}_1) = \hat{\sigma}^2 \frac{1}{(1-r^2)S_{WW}}$$
$$\widehat{Var}(\hat{\gamma}_1) = \hat{\tau}^2 \frac{1}{S_{WW}}$$

,

where (using the fact that $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a projection matrix)

$$\hat{\sigma}^2 = \frac{1}{n-3}(\vec{Y} - \mathbf{X}\widehat{\vec{\beta}})^T(\vec{Y} - \mathbf{X}\widehat{\vec{\beta}}) = \frac{1}{n-3}\vec{Y}\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\vec{Y}$$
$$\hat{\tau}^2 = \frac{1}{n-2}(\vec{Y} - \mathbf{W}\widehat{\vec{\gamma}})^T(\vec{Y} - \mathbf{W}\widehat{\vec{\gamma}}) = \frac{1}{n-2}\vec{Y}\left(\mathbf{I}_n - \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\right)\vec{Y}$$

.

Now, the condition that the projection matrices be the same for the two models would demand that $\mathbf{X}$ and $\mathbf{W}$ have the same range. This condition is not of interest. The condition that $\mathbf{X}$ and $\mathbf{W}$ have different ranges, but the same projection of $\vec{Y}$ would lead to $\hat{\beta}_2 = 0$, something that we cannot control by experimental design, nor can we expect it to happen regularly even when $\beta_2 = 0$.

Hence, we will get equality in the estimated standard errors only when the sum of squared errors from the unadjusted model is very little more than the sum of squared errors from the adjusted model, so as to allow the larger denominator of $(n-2)$ versus $(n-3)(1-r^2)$ to make up the difference. This is very hard to control by experimental design. I also note that in common practice, we use the t distribution with $n-2$ degrees of freedom to find critical values in the unadjusted model, while we use the t distribution with $n-3$ degrees of freedom in the adjusted model. This can also lead to slight differences when making inference.

d. Under what conditions is $\hat{\gamma}_1$ unbiased for β_1?

Ans:

$$\begin{aligned}
E[\widehat{\vec{\gamma}}] &= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T E[\vec{Y}] \\
&= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}\vec{\beta} \\
&= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T(\mathbf{W}(\beta_0 \quad \beta_1)^T + \vec{Z}\beta_2) \\
&= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{W}(\beta_0 \quad \beta_1)^T + (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\vec{Z}\beta_2) \\
&= (\beta_0 \quad \beta_1)^T + (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\vec{Z}\beta_2)
\end{aligned}$$

Hence, using our above results for the structure of $(\mathbf{W}^T\mathbf{W})^{-1}$, $\hat{\gamma}_1$ is unbiased for β_1 if only if $r_{WZ} = 0$ or $\beta_2 = 0$.

e. Under what conditions is $\hat{\gamma}_1$ BLUE for β_1?

Ans: By Gauss-Markov theorem, $\widehat{\vec{\beta}}$ is BLUE for $\vec{\beta}$. Hence the only time that $\hat{\gamma}_1$ will be BLUE is when $\hat{\gamma}_1 = \hat{\beta}_1$ under the conditions of part (a.).

f. Suppose in particular that $\beta_1 = 0$ and $\beta_2 \neq 0$. What is the impact of this situation on the distribution of $\hat{\gamma}_1$, and how would $\hat{\gamma}_1$ compare to $\hat{\beta}_1$ from the full model? Compare this situation to the setting in which $\beta_2 = 0$ and $\beta_1 \neq 0$.

Ans: If $\beta_1 = 0$, $\beta_2 \neq 0$, and $r_{WZ} \neq 0$, the estimate $\hat{\gamma}_1$ will be biased towards finding an association between Y and W when there is truly none after conditioning on Z. $\hat{\beta}_1$ will tend to be close to zero, but $\hat{\gamma}_1$ will tend to be too large or too small depending upon the sign of β_2 and the sign of the correlation between W and Z.

If $\beta_1 = 0$, $\beta_2 \neq 0$, and $r_{WZ} = 0$, the estimate $\hat{\gamma}_1$ will be unbiased for β_1. If $r_{WZ} = 0$ by design, the estimated standard error of $\hat{\gamma}_1$ will tend to be too large leading to confidence intervals that are too wide.

On the other hand, if $\beta_1 \neq 0$ and $\beta_2 = 0$, this is the situation where the smaller model provides regression estimates that are BLUE.

Bottom line: The first question in deciding whether you want to fit the adjusted or unadjusted model is whether you are more interested in β_1 or γ_1. There are many scientific issues that must be considered in that decision.

The remainder of this discussion presumes that we are truly interested in β_1, which may or may not be equal to γ_1. The relative advantages of fitting the adjusted or unadjusted model depend upon the values of β_2 and r_{WX}.

- If $\beta_2 \neq 0$, then, when possible, we would like to choose our experimental design matrix such that $r_{WX} = 0$. In that setting, both $\hat{\beta}_1$ and $\hat{\gamma}_1$ are unbiased for β_1. Furthermore, $\hat{\beta}_1 = \hat{\gamma}_1$, so they would have the same standard error. However, using the usual statistical software with the unadjusted model will overestimate the true standard error of $\hat{\beta}_1$, because it will use $\hat{\tau}^2$ instead of $\hat{\sigma}^2$. In this case, we definitely want to use the adjusted model.
- If $\beta_2 \neq 0$ and we are stuck with $r_{WX} \neq 0$, we want to use the adjusted model in order to obtain an unbiased estimate of β_1. In that setting, the standard error of $\hat{\beta}_1$ may be either larger or smaller than the standard error of $\hat{\gamma}_1$, depending upon the relative sizes of β_2 and r_{WX}. Large β_2 will tend to decrease the value of σ^2 relative to τ^2, but if $|r_{WX}|$ is large, then the $(1 - r_{WX}^2)$ term in the denominator will tend to increase the standard error of $\hat{\beta}_1$.
- If $\beta_2 = 0$ and we are stuck with $r_{WX} \neq 0$, then we certainly do not want to use the adjusted model, despite the fact that $\hat{\gamma}_1$ will be unbiased for β_1. In this setting, $\hat{\sigma}^2$ will tend to be close to $\hat{\tau}^2$, because $\sigma^2 = \tau^2$. However, the $(1 - r_{WX}^2)$ term in the denominator will tend to increase the standard error of $\hat{\beta}_1$, without any concomitant gain in precision.
- If $\beta_2 = 0$ and $r_{WX} = 0$, then we also do not want to use the adjusted model, although there is relatively less harm if we use it anyway. We will have that $\hat{\gamma}_1$ will be unbiased for β_1, and we do not have to worry about any variance inflation from the $(1 - r_{WX}^2)$ term in the denominator. We can guess that $\hat{\sigma}^2$ will tend to be close to $\hat{\tau}^2$, because $\sigma^2 = \tau^2$, but insofar as the sums of squared errors are close to each other, in the adjusted model we are dividing by $n-3$ instead of $n-2$ in the unadjusted model. By habit, we will also use the critical values from a t distribution with $n-3$ degrees of freedom, which critical values are larger than the corresponding critical values for a t distribution with $n-2$ degrees of freedom. This last point will tend to make our CI wider and our tests less powerful, though this is fairly negigible with a decent sample size.