

Written problems to be handed in Wednesday, April 29.

1. Consider a linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ where the first column of the design matrix is filled with 1's (so we are fitting an intercept). Assume that $\text{var}(\vec{\epsilon}) = \sigma^2 \mathbf{I}_n$.

- a. Suppose $E[\epsilon_i] = \theta$ for $i = 1, \dots, n$. What is the distribution for the OLS estimator $\hat{\vec{\beta}}$? In particular, how does the assumption of nonzero mean for the errors alter the interpretation and distribution of the slope parameters $\beta_1, \dots, \beta_{p-1}$?

Ans: Under the above assumptions, we have that $E[\vec{Y}|\mathbf{X}] = \mathbf{X}\vec{\beta} + \theta\vec{1}_n$, where $\vec{1}_n$ is an n-vector containing all 1's, and $\text{var}(\vec{Y}|\mathbf{X}) = \sigma^2 \mathbf{I}_n$. To make this whole problem easier, we note that

$$\theta\vec{1}_n = \mathbf{X}\vec{\Delta} = \mathbf{X} \begin{pmatrix} \theta \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The OLSE $\hat{\vec{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{Y}$ thus has expectation

$$\begin{aligned} E[\hat{\vec{\beta}}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\vec{Y}] \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\vec{\beta} + \mathbf{X}\vec{\Delta}) \\ &= \vec{\beta} + \vec{\Delta} \end{aligned}$$

Hence, we find that errors that do not have zero mean do not affect the expectation of $\hat{\beta}_1, \dots, \hat{\beta}_{p-1}$ as the 2nd - p th elements of $\vec{\Delta}$ are 0. Furthermore, because the variance of a random variable is unaffected by adding a constant, the variance of the estimates is similarly unchanged.

The interpretation of the slope parameter estimates is unchanged by such nonzero means for the error distribution.

- b. Let \mathbf{X}^* be a design matrix derived from \mathbf{X} by subtracting the corresponding column means from the elements in columns 2 through p . That is $X_{ij}^* = X_{ij} - \sum_{i=1}^n X_{ij}/n$. If we fit the regression model $\vec{Y} = \mathbf{X}^* \vec{\beta}^* + \vec{\epsilon}$, how does the OLS estimator $\hat{\vec{\beta}}^*$ relate to $\hat{\vec{\beta}}$ from the original problem. In particular, how does the interpretation and distribution of each of the regression parameters change?

Ans: This problem is easiest in matrix notation using partitioned matrices, though it can also be done by brute force. Consider the partitioning of \mathbf{X} and \mathbf{X}^* each into n by 1 and n by $p-1$ matrices

$$\mathbf{X} = \begin{pmatrix} \vec{1}_n & \mathbf{W} \end{pmatrix} \quad \mathbf{X}^* = \begin{pmatrix} \vec{1}_n & \mathbf{W}^* \end{pmatrix}$$

where $\mathbf{W}^* = \mathbf{W} - \frac{1}{n} \vec{1}_n \vec{1}_n^T \mathbf{W}$ has subtracted the means of the column of \mathbf{W} from the elements in each corresponding column. We thus find

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} \vec{1}_n^T \\ \mathbf{W}^T \end{pmatrix} \begin{pmatrix} \vec{1}_n & \mathbf{W} \end{pmatrix} = \begin{pmatrix} n & \vec{1}_n^T \mathbf{W} \\ \mathbf{W}^T \vec{1}_n & \mathbf{W}^T \mathbf{W} \end{pmatrix}$$

and

$$\mathbf{X}^{*T} \mathbf{X}^* = \begin{pmatrix} (\mathbf{I}_n - \frac{1}{n} \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T) \mathbf{W}^T \\ \bar{\mathbf{1}}_n^T (\mathbf{I}_n - \frac{1}{n} \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T) \mathbf{W} \end{pmatrix} = \begin{pmatrix} n & \mathbf{0}_n^T \\ \mathbf{0}_n & \mathbf{W}^{*T} \mathbf{W}^* \end{pmatrix}$$

Now the inverse of a partitioned symmetric matrix can be given by

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F} \mathbf{E}^{-1} \mathbf{F}^T & -\mathbf{F} \mathbf{E}^{-1} \\ -\mathbf{E}^{-1} \mathbf{F}^T & \mathbf{E}^{-1} \end{pmatrix}$$

where $\mathbf{E} = \mathbf{D} - \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}$ and $\mathbf{F} = \mathbf{A}^{-1} \mathbf{B}$ (this is one of the forms given by Seber and Lee, page 466). Now we merely need to compare $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \bar{\mathbf{Y}}$ to $(\mathbf{X}^{*T} \mathbf{X}^*)^{-1} \mathbf{X}^{*T} \bar{\mathbf{Y}}$. By straightforward matrix application of the above formula for the inverse and then matrix multiplication we find

$$\begin{aligned} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T &= \begin{pmatrix} \frac{1}{n} \bar{\mathbf{1}}_n^T - \bar{\mathbf{1}}_n^T \mathbf{W} (\mathbf{W}^T \mathbf{W} - \frac{1}{n} \mathbf{W}^T \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{I}_n - \frac{1}{n} \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T) \\ (\mathbf{W}^T \mathbf{W} - \frac{1}{n} \mathbf{W}^T \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{I}_n - \frac{1}{n} \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T) \end{pmatrix} \\ (\mathbf{X}^{*T} \mathbf{X}^*)^{-1} \mathbf{X}^{*T} &= \begin{pmatrix} \frac{1}{n} \bar{\mathbf{1}}_n^T \\ (\mathbf{W}^T \mathbf{W} - \frac{1}{n} \mathbf{W}^T \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{I}_n - \frac{1}{n} \bar{\mathbf{1}}_n \bar{\mathbf{1}}_n^T) \end{pmatrix} \end{aligned}$$

Note that in the above, the only difference is in the 1 by n upper matrix in the partition. Thus, while $\hat{\beta}_0$ will not in general equal $\hat{\beta}_0^*$, we will have $\hat{\beta}_j = \hat{\beta}_j^*$ for $j = 1, \dots, p-1$.

Now $var(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ and $var(\hat{\beta}^*) = \sigma^2 (\mathbf{X}^{*T} \mathbf{X}^*)^{-1}$. Note that because the lower right matrix in the partition of $(\mathbf{X}^T \mathbf{X})^{-1}$

$$(\mathbf{W}^T \mathbf{W} - \frac{1}{n} \bar{\mathbf{1}}_n^T \mathbf{W} \mathbf{W}^T \bar{\mathbf{1}}_n)^{-1}$$

equals the lower right matrix in the partition of $(\mathbf{X}^{*T} \mathbf{X}^*)^{-1}$

$$(\mathbf{W}^{*T} \mathbf{W}^*)^{-1},$$

the covariance matrix for $(\hat{\beta}_1, \dots, \hat{\beta}_{p-1})$ is equal to the covariance matrix for $(\hat{\beta}_1^*, \dots, \hat{\beta}_{p-1}^*)$. So the estimated variances will also be the same for the slope estimates for the uncentered and the centered models.

The interpretation of the slope estimates is unchanged by centering, because the j th slope parameter continues to model the difference in means between two subjects who differ by one unit in their values of X_j but are alike with respect to all other modelled covariates.

2. Consider the simple linear regression model $Y_i = \beta_0 + x_i \beta_1 + \epsilon_i$ for $i = 1, \dots, n$, with x_i known predictors, $\vec{\beta} = (\beta_0, \beta_1)^T$ an unknown parameter vector to be estimated and/or tested, and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. Without loss of generality, we will assume that $\sum_{i=1}^n x_i = 0$. Let $\sigma_i^2 = \alpha_i + x_i \gamma > 0$ with γ and $\vec{\alpha}$ unknown nuisance parameters subject to $\vec{\alpha}^T \vec{x} = 0$. Let $\hat{\vec{\beta}}$ be the ordinary least squares estimate of $\vec{\beta}$.

- a. What is the mean and variance of $\hat{\beta}_1$?

Ans: In this problem with $\bar{\mathbf{Y}} = \sum_{i=1}^n Y_i/n$, $S_{xx} = \sum_{i=1}^n x_i^2$, and $S_{xY} = \sum_{i=1}^n x_i Y_i$

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix} \quad (\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix} \quad \mathbf{X}^T \bar{\mathbf{Y}} = \begin{pmatrix} n \bar{\mathbf{Y}} \\ S_{xY} \end{pmatrix}$$

so

$$\hat{\vec{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \bar{\mathbf{Y}} = \begin{pmatrix} \bar{\mathbf{Y}} \\ \frac{S_{xY}}{S_{xx}} \end{pmatrix}$$

$E[\widehat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\vec{Y}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \vec{\beta} = \vec{\beta}$ and the variance of $\widehat{\beta}$ is found by

$$\begin{aligned} \mathbf{X}^T \mathbf{V} \mathbf{X} &= \begin{pmatrix} \sum \alpha_i + \gamma \sum x_i & \sum \alpha_i x_i + \gamma \sum x_i^2 \\ \sum \alpha_i x_i + \gamma \sum x_i^2 & \sum \alpha_i x_i^2 + \gamma \sum x_i^3 \end{pmatrix} = \begin{pmatrix} \sum \alpha_i & \gamma \sum x_i^2 \\ \gamma \sum x_i^2 & \sum \alpha_i x_i^2 + \gamma \sum x_i^3 \end{pmatrix} \\ \text{Var}(\widehat{\beta}) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix} \begin{pmatrix} \sum \alpha_i & \gamma \sum x_i^2 \\ \gamma \sum x_i^2 & \sum \alpha_i x_i^2 + \gamma \sum x_i^3 \end{pmatrix} \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix} \\ &= \begin{pmatrix} \sum \alpha_i / n^2 & \gamma / n \\ \gamma / n & \sum \alpha_i x_i^2 / S_{xx}^2 + \gamma \sum x_i^3 / S_{xx}^2 \end{pmatrix} \end{aligned}$$

From this we find that $E[\widehat{\beta}_1] = \beta_1$ and $\text{Var}(\widehat{\beta}_1) = (\sum \alpha_i x_i^2 + \gamma \sum x_i^3) / S_{xx}^2$.

- b. Under what conditions will the estimated variance of $\widehat{\beta}_1$ based on the ordinary least squares regression analysis be consistent for the true variance of $\widehat{\beta}_1$.

Ans: Under OLS, we assume that the errors have constant variance, and our estimate

$$\hat{\sigma}^2 = \frac{1}{n-p} (\vec{Y} - \mathbf{X} \widehat{\beta})^T (\vec{Y} - \mathbf{X} \widehat{\beta})$$

consistently estimates the limit of $\sum_{i=1}^n (\alpha_i + \gamma x_i) / n = \sum_{i=1}^n \alpha_i / n = \bar{\alpha}$, assuming such a limit exists. Let α be that limit. Then, the OLSE variance estimate of $\widehat{\beta}$ is

$$\widehat{\text{Var}}(\widehat{\beta}) = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1} = \bar{\alpha} \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix}$$

and the variance estimate for $\widehat{\beta}_1$ is therefore just

$$\widehat{\text{Var}}(\widehat{\beta}_1) = \frac{\bar{\alpha}}{\sum_{i=1}^n x_i^2} = \frac{\bar{\alpha}}{n V_x}$$

where $V_x = S_{xx} / n$ is the variance of the x_i 's.

From part (a), we find

$$\text{Var}(\widehat{\beta}_1) = \frac{\sum \alpha_i x_i^2}{n^2 V_x^2} + \frac{\gamma \sum x_i^3}{n^2 V_x^2}$$

From this we see that $\widehat{\text{Var}}(\widehat{\beta}_1)$ will tend toward $\text{Var}(\widehat{\beta}_1)$ certainly when $\alpha_i \equiv \alpha$ for all i and either $\gamma = 0$ or the distribution of the x_i 's is unskewed.

Now it is stipulated that $\sum \alpha_i x_i = 0$. If this were to be strengthened to the case that the α_i 's are also sampled independently of the x_i 's (and note that $\sum \alpha_i x_i = 0$ merely suggests that they are uncorrelated, not necessarily independent), then $\sum \alpha_i x_i^2 / n = (\sum \alpha_i / n) (\sum x_i^2 / n)$. Thus

$$\text{Var}(\widehat{\beta}_1) = \frac{\bar{\alpha}}{n V_x} + \frac{\gamma \sum x_i^3}{n^2 V_x^2}$$

and $\widehat{\text{Var}}(\widehat{\beta}_1)$ will tend toward $\text{Var}(\widehat{\beta}_1)$ when either $\gamma = 0$ or the distribution of the x_i 's is unskewed.

- c. What restrictions on the problem would be necessary for $\widehat{\beta}$ to be asymptotically normally distributed? (You need not rigorously derive an asymptotic distribution, instead just briefly discuss the ways that this setting differs from the assumptions under which we derived the asymptotic distribution in class, and what general requirements might address those problems.)

Ans: When the errors are uncorrelated and identically distributed, in order to derive the asymptotic normal results for the OLSE we had to place restrictions on the sampling of the x_i 's to ensure that the contribution of any particular x_i to the total variance of the x_i 's was negligible as $n \rightarrow \infty$ in simple linear regression. This restriction translates into a requirement that the smallest eigenvalue of $\mathbf{X}^T \mathbf{X}$ tend to infinity as $n \rightarrow \infty$, along with some requirements that none of the cases sampled be too influential (see class notes). In this more general case we must be concerned in the way that we sample the x_i 's and the ϵ_i 's. This will translate into a requirement that the smallest eigenvalue of $\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}$ approach infinity as $n \rightarrow \infty$, along with some requirements that the most influential cases not also tend to have the errors with the largest variance.

- d. What would be the effect of using the asymptotic results for ordinary least squares regression analysis on tests of $H_0 : \beta_1 = 0$? Consider the effect that the variance of the α_i s and the value of γ has on your answer.

Ans: Assuming that asymptotic normality holds, we need only worry about when the variance estimate under OLSE over- or underestimates the true variance. When the variance estimate underestimates the true variance, tests of H_0 will be anti-conservative in that the true type I error will be larger than desired. When the variance estimate overestimates the true variance, tests of H_0 will tend to have a smaller type I error than desired.

When the α_i 's are independent of the x_i 's (which would include the case when $Var(\alpha_i) = 0$), then the variance estimate $\widehat{Var}(\hat{\beta}_1)$ will underestimate the true variance when γ and the skewness of the x_i 's are of the same sign, and it will overestimate the true variance when γ and the skewness of the x_i 's are of opposite sign. For instance, if the distribution of the x_i 's is positively skewed, then a tendency for larger variance with larger values of x_i will lead to anti-conservative tests, while a tendency for smaller variance with larger values of x_i will lead to conservative tests (and loss of statistical power).

When $\gamma = 0$ or the distribution of the x_i 's is unskewed, then the variance estimate $\widehat{Var}(\hat{\beta}_1)$ will underestimate the true variance when the weighted average of the α_i 's based on weights x_i^2 is greater than $\bar{\alpha}$ and will overestimate the true variance when the weighted average is less than $\bar{\alpha}$. The weighted average will tend to be greater than $\bar{\alpha}$ when the more extreme values of x_i are associated with larger α_i . Consider for example the simple example where $n = 5$ and $\bar{x} = (-2, -1, 0, 1, 2)^T$ (so $\sum x_i = 0$). Now if $\bar{\alpha} = (3, 1, 2, 1, 3)^T$ (so $\sum \alpha_i x_i = 0$, but the α_i 's are not independent of the x_i 's), then $\bar{\alpha} = 2$, but the weighted average $(\sum \alpha_i x_i^2) / (\sum x_i^2) = 2.6$ and inference based on the OLSE estimate of the variance is anti-conservative. On the other hand, if $\bar{\alpha} = (1, 3, 2, 3, 1)^T$ (so $\sum \alpha_i x_i = 0$, but the α_i 's are not independent of the x_i 's), then $\bar{\alpha} = 2$, but the weighted average $(\sum \alpha_i x_i^2) / (\sum x_i^2) = 1.4$ and inference based on the OLSE estimate of the variance is conservative.

Clearly, as both the distribution of the α_i 's relative to the x_i 's and the value of γ and/or the skewness of the x_i 's are allowed to vary, the tendency for the OLS variance estimate to over- or underestimate the true variance will reflect the combination of those effects.

- e. What would be the effect of using the asymptotic results for ordinary least squares regression analysis on confidence intervals for β_1 ? Consider the effect that the variance of the α_i s and the value of γ has on your answer.

Ans: Assuming that asymptotic normality holds, we need only worry about when the variance estimate under OLSE over- or underestimates the true variance. When the variance estimate underestimates the true variance, confidence intervals for β_1 will tend to be too narrow, and thus will have a coverage probability that is less than the desired level. When the variance estimate overestimates the true variance, confidence intervals for β_1 will tend to be too wide, and thus will have a coverage probability that is greater than the desired level. Discussion of the cases that such over- or underestimation occurs is exactly the same as for part (d).

- f. How do the above results compare to results for the t tests as considered in homework # 1?

Ans: These results exactly predict the results in homework # 1 for the t test that presumes equal variances: With only two values of the predictor, any heteroscedasticity is necessarily linear in X . Furthermore, inequality of sample sizes results in skewness of the predictor distribution.

3. Let $Y_i \sim \text{Bernoulli}(p_i), i = 1, \dots, n$ 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 $i = 1, \dots, n$ (as it would be under the null hypothesis $H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$), 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 2 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 2 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 any 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 $\hat{\vec{\beta}}^{(0)}$ and $\hat{p}_i^{(0)} = \vec{x}_i^T \hat{\vec{\beta}}^{(0)}$. 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 $\hat{\vec{\beta}}_G^{(1)}$. These estimates are then used to find $\hat{p}_i^{(1)}$ and $\mathbf{V}^{(1)}$. The process is then repeated with GLSE $\hat{\vec{\beta}}_G^{(k)}$ estimated using $\mathbf{V}^{(k-1)}$ until $(\hat{\vec{\beta}}_G^{(k)} - \hat{\vec{\beta}}_G^{(k-1)})^T (\hat{\vec{\beta}}_G^{(k)} - \hat{\vec{\beta}}_G^{(k-1)})$ 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 $\mathbf{X} = (\vec{1}_n \ \vec{W} \ \vec{Z})$ has $X_{i1} \equiv 1$ for $i = 1, \dots, n$ 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 $\text{var}(\vec{\epsilon}) = \sigma^2 \mathbf{I}_n$.

- 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 $\mathbf{X} = (\vec{1}_n \ \mathbf{U})$ similar to problem 1. By problem 1, 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 j th and k th parameter estimates will be 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 $j = 2, \dots, p - 1$. We thus further partition $\mathbf{U}^* = (\vec{W}^* \ \mathbf{V}^*)$ 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

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

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

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

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

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

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 $S_{WW} = (\bar{W} - E[\bar{W}])^T(\bar{W} - E[\bar{W}])$, S_{ZZ} , and σ^2 constant, but we may freely vary $S_{WZ} = (\bar{W} - E[\bar{W}])^T(\bar{Z} - E[\bar{Z}])$. 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 \bar{W} and \bar{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 Problem 4 in which we will assume the true model is

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

but we want to also consider fitting a model

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

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

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

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{Y} \\ \hat{\gamma} &= (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \vec{Y} \end{aligned}$$

and we want to find when $\begin{pmatrix} 0 & 1 \end{pmatrix} \hat{\gamma} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \hat{\beta}$. Following the approaches used above with $r = r_{WZ} = S_{WZ} / \sqrt{S_{WW} S_{ZZ}}$ we find

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

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

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

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

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

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

$$\begin{aligned} \text{var}(\hat{\beta}) &= \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}, \text{ and} \\ \text{var}(\hat{\gamma}) &= \tau^2(\mathbf{W}^T \mathbf{W})^{-1} \end{aligned}$$

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

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

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

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

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

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

Ans: We would typically estimate

$$\begin{aligned} \widehat{\text{Var}}(\hat{\beta}_1) &= \hat{\sigma}^2 \frac{1}{(1-r^2)S_{WW}} \\ \widehat{\text{Var}}(\hat{\gamma}_1) &= \hat{\tau}^2 \frac{1}{S_{WW}} \end{aligned},$$

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

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-3} (\vec{Y} - \mathbf{X}\hat{\beta})^T (\vec{Y} - \mathbf{X}\hat{\beta}) = \frac{1}{n-3} \vec{Y}^T (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \vec{Y} \\ \hat{\tau}^2 &= \frac{1}{n-2} (\vec{Y} - \mathbf{W}\hat{\gamma})^T (\vec{Y} - \mathbf{W}\hat{\gamma}) = \frac{1}{n-2} \vec{Y}^T (\mathbf{I}_n - \mathbf{W}(\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T) \vec{Y} \end{aligned}$$

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:

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

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, $\hat{\beta}$ 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 interested student can watch my take on these issues as given in lectures 6, 7, and 8 from a short course on applied regression analysis. These lectures can be obtained via streaming video at

<http://www.uwv.org/programs/displayseries.aspx?fid=746>.

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 negligible with a decent sample size.

6. Consider a linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ where the first column of the design matrix is filled with 1's (so we are fitting an intercept). Consider adding an additional predictor \vec{Z} to the model where, for some fixed j , $Z_i = 1$ if $i = j$ and $Z_i = 0$ otherwise. Let \mathbf{X}^* be the augmented matrix in which the $(p + 1)$ th column is \vec{Z} , and consider fitting the regression model $\vec{Y} = \mathbf{X}^*\vec{\gamma} + \vec{\epsilon}^*$

a. How do the parameter estimates $\hat{\gamma}_0, \dots, \hat{\gamma}_{p-1}$ differ from $\hat{\beta}$?

Ans: Without loss of generality, I consider the case of deleting the first case. To find $\hat{\gamma}$ I consider the partitioned matrix $\mathbf{X}^* = (\mathbf{X} \quad \vec{Z})$. Then letting $\vec{x}_1 = (\vec{Z}^T \mathbf{X})$ be the covariate vector for the first case, we have

$$\mathbf{X}^{*T} \mathbf{X}^* = \begin{pmatrix} \mathbf{X}^T \mathbf{X} & \vec{x}_1 \\ \vec{x}_1^T & 1 \end{pmatrix}$$

and using the formula for the inverse of a partitioned matrix given above, we find

$$(\mathbf{X}^{*T} \mathbf{X}^*)^{-1} = \begin{pmatrix} (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \vec{x}_1 \vec{x}_1^T (\mathbf{X}^T \mathbf{X})^{-1}}{(1 - h_{11})} & -\frac{(\mathbf{X}^T \mathbf{X})^{-1} \vec{x}_1}{(1 - h_{11})} \\ -\frac{\vec{x}_1^T (\mathbf{X}^T \mathbf{X})^{-1}}{(1 - h_{11})} & \frac{1}{(1 - h_{11})} \end{pmatrix}$$

where $h_{11} = \vec{x}_1^T (\mathbf{X}^T \mathbf{X})^{-1} \vec{x}_1$ is the first element on the diagonal of the hat matrix. Thus

$$\begin{aligned} \hat{\gamma} &= (\mathbf{X}^{*T} \mathbf{X}^*)^{-1} \mathbf{X}^{*T} \vec{Y} \\ &= \begin{pmatrix} \hat{\beta} - \frac{(\mathbf{X}^T \mathbf{X})^{-1} \vec{x}_1 (\vec{Y} - \vec{x}_1^T \hat{\beta})}{(1 - h_{11})} \\ \frac{(\vec{Y} - \vec{x}_1^T \hat{\beta})}{(1 - h_{11})} \end{pmatrix} \end{aligned}$$

where $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{Y}$.

By inspection, then, adding the covariate \vec{Z} to the model has $\hat{\beta}_j - \hat{\gamma}_j$ equal to the j th element of $(\mathbf{X}^T \mathbf{X})^{-1} \vec{x}_1 (\vec{Y} - \vec{x}_1^T \hat{\beta}) / (1 - h_{11})$.

It should be noted that $(\vec{x}_1^T \quad 1) \hat{\gamma} = Y_1$. Thus adding the covariate indicating a single case results in a model which predicts that case exactly.

b. How do the parameter estimates $\hat{\gamma}_0, \dots, \hat{\gamma}_{p-1}$ differ from the estimates obtained by fitting the first model with the j th case deleted?

Ans: Note first that if we partition

$$\mathbf{X} = \begin{pmatrix} \vec{x}_1^T \\ \mathbf{W} \end{pmatrix} \quad \vec{Y} = \begin{pmatrix} Y_1 \\ \vec{U} \end{pmatrix}$$

(so \mathbf{W} contains rows 2 through n of \mathbf{X} and $\vec{U} = (Y_2, \dots, Y_n)^T$), then

$$\begin{aligned} \mathbf{X}^T \mathbf{X} &= \mathbf{W}^T \mathbf{W} + \vec{x}_1 \vec{x}_1^T, \text{ so} \\ \mathbf{W}^T \mathbf{W} &= \mathbf{X}^T \mathbf{X} - \vec{x}_1 \vec{x}_1^T, \text{ and} \\ \mathbf{X}^T \vec{Y} &= \vec{x}_1 Y_1 + \mathbf{W}^T \vec{U}. \end{aligned}$$

Now an alternative formula for the inverse of a partitioned symmetric matrix is given by

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{G}^{-1} & -\mathbf{G}^{-1}\mathbf{J} \\ -\mathbf{J}^T\mathbf{G}^{-1} & \mathbf{D}^{-1} + \mathbf{J}^T\mathbf{G}^{-1}\mathbf{J} \end{pmatrix}$$

where $\mathbf{G} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T$ and $\mathbf{F} = \mathbf{B}\mathbf{D}^{-1}$ (again, see Seber and Lee, page 466). Using this formula and the above relations between $\mathbf{X}^T\mathbf{X}$ and $\mathbf{W}^T\mathbf{W}$, we find that $(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}$, $\mathbf{X}^{*T}\vec{Y}$, and $\hat{\vec{\gamma}}$ from part (a) can be written as

$$(\mathbf{X}^{*T}\mathbf{X}^*)^{-1} = \begin{pmatrix} (\mathbf{W}^T\mathbf{W})^{-1} & -(\mathbf{W}^T\mathbf{W})^{-1}\vec{x}_1 \\ -\vec{x}_1^T(\mathbf{W}^T\mathbf{W})^{-1} & 1 + \vec{x}_1^T(\mathbf{W}^T\mathbf{W})^{-1}\vec{x}_1 \end{pmatrix} \quad \mathbf{X}^{*T}\vec{Y} = \begin{pmatrix} \mathbf{W}^T\vec{U} + \vec{x}_1 Y_1 \\ Y_1 \end{pmatrix}$$

and

$$\hat{\vec{\gamma}} = \begin{pmatrix} (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\vec{U} \\ Y_1 - \vec{x}_1^T(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\vec{U} \end{pmatrix}$$

Thus we see that $(\hat{\gamma}_0, \dots, \hat{\gamma}_{p-1})$ are exactly the OLS estimates that we would have obtained if the first case had been deleted from the dataset.

This result gives us computationally useful ways to compute the influence of individual cases: We can compute the change in the parameter estimates using the estimates from the full data case and the design matrix. We do not really have to fit separate regressions for every case deletion. I note, however, that we will not have such a result for other forms of regression. Furthermore, computing the difference in the P values is a little more difficult without actually fitting all the regressions.