Biost/Stat 533
Spring 2008

Homework Assignment #8

1. Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where ϵ_i are *iid* $N(0, \sigma^2)$. Assume $\overline{x} = 0$. Derive an F-statistic for testing H: $\beta_0 = \beta_1$.

Solution: The linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is given by

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

We assume that some of x_i is not zero so that fitting to this linear model has a practical sense. Then by the constraint $\overline{x} = 0$ implies that the design matrix $\boldsymbol{X}$ is of full rank. We compute

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \left(\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \right)^{-1} = \left(\begin{bmatrix} n & 0 \\ 0 & \sum x_i^2 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum x_i^2 \end{bmatrix}.$$

Thus, the least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}
(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} &= \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum x_i^2 \end{bmatrix} \left(\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^T \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\
&= \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum x_i^2 \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix} \\
&= \begin{bmatrix} \overline{Y}_n \\ \frac{\sum x_i Y_i}{\sum x_i^2} \end{bmatrix}
\end{aligned}$$

where $\overline{Y}_n \equiv \sum Y_i / n$.

Let $\boldsymbol{A} = (1, -1)$. Note that the null hypothesis H is equivalent to the hypothesis $H : \boldsymbol{A}\boldsymbol{\beta} = 0$ and that rank of $\boldsymbol{A}$ is 1. It is easy to see that this hypothesis is testable (why?). We have

$$\begin{aligned}
RSS_H - RSS &= (\boldsymbol{A}\hat{\boldsymbol{\beta}})'(\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A}')^{-1})^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}) \\
&= \left((1,-1) \begin{bmatrix} \overline{Y}_n \\ \frac{\sum x_i Y_i}{\sum x_i^2} \end{bmatrix} \right)^T \left((1,-1) \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum x_i^2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right)^{-1} \left((1,-1) \begin{bmatrix} \overline{Y}_n \\ \frac{\sum x_i Y_i}{\sum x_i^2} \end{bmatrix} \right) \\
&= \left(\overline{Y}_n - \frac{\sum x_i Y_i}{\sum x_i^2} \right) \left(\frac{1}{n} + \frac{1}{\sum x_i^2} \right)^{-1} \left(\overline{Y}_n - \frac{\sum x_i Y_i}{\sum x_i^2} \right) \\
&= \left(\overline{Y}_n - \frac{\sum x_i Y_i}{\sum x_i^2} \right)^2 \left(\frac{1}{n} + \frac{1}{\sum x_i^2} \right)^{-1}
\end{aligned}$$

and

$$
\begin{aligned}
RSS &= (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum (Y_i - \overline{Y}_n - \frac{\sum x_i Y_i}{\sum x_i^2} x_i)^2 \\
&= \sum (Y_i - \overline{Y}_n)^2 - 2\sum (Y_i - \overline{Y}_n)\frac{\sum x_j Y_j}{\sum x_j^2} x_i + \sum \left(\frac{\sum x_i Y_i}{\sum x_i^2}\right)^2 x_i^2 \\
&= \sum (Y_i - \overline{Y}_n)^2 - 2\frac{\sum x_j Y_j}{\sum x_j^2}\sum x_i Y_i - 2\frac{\overline{Y}_n \sum x_j Y_j}{\sum x_j^2}\sum x_i + \left(\frac{\sum x_j Y_j}{\sum x_j^2}\right)^2 \sum x_i^2 \\
&= \sum (Y_i - \overline{Y}_n)^2 - 2\frac{(\sum x_j Y_j)^2}{\sum x_j^2} + 0 + \frac{(\sum x_j Y_j)^2}{\sum x_j^2} \\
&= \sum (Y_i - \overline{Y}_n)^2 - \frac{(\sum x_j Y_j)^2}{\sum x_j^2}.
\end{aligned}
$$

Finally, the F-statistic is

$$
\begin{aligned}
F &= \frac{(RSS_H - RSS)/1}{RSS/(n-2)} \\
&= \frac{(n-2)\left(\overline{Y}_n - \frac{\sum x_i Y_i}{\sum x_i^2}\right)^2}{\left(\frac{1}{n} + \frac{1}{\sum x_i^2}\right)\left(\sum (Y_i - \overline{Y}_n)^2 - \frac{(\sum x_j Y_j)^2}{\sum x_j^2}\right)}
\end{aligned}
$$

which is distributed as F distribution with degrees of freedom 1 and $n-2$. When we say “without loss of generality $\overline{x} = 0$, we typically mean to reparametrize the linear model as $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i = (\beta_0 + \beta_1 \overline{x}) + \beta_1 (x_i - \overline{x}) + \epsilon_i \equiv \gamma_0 + \gamma_1 x_i + \epsilon_i$ where $\gamma_0 = \beta_0 + \beta_1 \overline{x}$ and $\gamma_1 = \beta_1$. In this case the hypothesis $\beta_0 = \beta_1$ is different from the hypothesis $\gamma_0 = \gamma_1$.

2. Suppose the postulated regression model is

$$E(Y) = \beta_0 + \beta_1 x$$

when, in fact, the true model is

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

(a) If we have observations at $x = -3, -2, -1, 0, 1, 2, 3$ and fit the postulated model, what bias will be introduced to those estimates?
(b) Answer the same question if the true and postulated models are reversed.

Solution: (a) The true model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\eta} + \boldsymbol{\epsilon}$ is given by

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 9 & -27 \\ 4 & -8 \\ 1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 4 & 8 \\ 9 & 27 \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}.$$

Thus, the bias is given by

$$\begin{aligned}
(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z\eta} &= \begin{bmatrix} 1/7 & 0 \\ 0 & 1/\sum x_i^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 9 & -27 \\ 4 & -8 \\ 1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 4 & 8 \\ 9 & 27 \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} \\
&= \begin{bmatrix} 1/7 & 0 \\ 0 & 1/28 \end{bmatrix} \begin{bmatrix} 28 & 0 \\ 0 & 196 \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} \\
&= \begin{bmatrix} 4 & 0 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} \\
&= \begin{bmatrix} 4\beta_2 \\ 7\beta_3 \end{bmatrix}.
\end{aligned}$$

(b) Now suppose the true model is

$$E[Y] = X\beta$$

yet the postulated model is

$$E[Y] = X\beta + Z\eta$$

Note that $\mathcal{R}(X) \subset \mathcal{R}([X, Z])$, therefore $\hat{\beta}$ is unbiased.

3. Consider a randomized clinical trial for the effect of a treatment on some positive continuous trait (blood pressure, cholesterol level, body weight...). z is the pre-treatment value of the trait and y is the post-treatment value. x denote assignment to the active treatment ($x = 1$) or placebo ($x = 0$).

Suppose that in truth the effect of treatment is linear on the relative change in the trait. The true model is

$$(y_i - z_i)/z_i = \alpha_0 + \alpha_1 x_i + \epsilon_i$$

(a) Write the true model in matrix notation (assume subjects are randomized equally to treatment and placebo).

Suppose that the data are modeled using absolute change:

$$y_i - z_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

(b) Write the model for absolute change in matrix notation and then show the least squares estimator of β_1 has expectation $E[\hat{\beta}_1] \approx \overline{z}\alpha_1$.
(c) Would testing the hypothesis $H : \beta_1 = 0$ be a valid test for a treatment effect? Explain.
(d) Suppose that one tested $H : \beta_1 = 0$ with a Wald test, i.e., one use the statistic $T = \hat{\beta}_1 / \sqrt{v\hat{a}r(\hat{\beta}_1)}$. Would the test be conservative? Anticonservative? Explain.

Solution: (a) In matrix notation

$$D_z^{-1}(Y - Z) = \begin{bmatrix} 1_n & X \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} + \epsilon$$

where $D_z \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the ith elements equal to z_i, $Y \in \mathbb{R}^{n \times 1} = (y_1, \ldots, y_n)^T$, $Z \in \mathbb{R}^{n \times 1} = (z_1, \ldots, z_n)^T$, $1_n \in \mathbb{R}^{n \times 1} = (1, \ldots, 1)^T$, $X \in \mathbb{R}^{n \times 1} = (x_1, \ldots, x_n)^T$, and $\epsilon \in \mathbb{R}^{n \times 1} = (\epsilon_1, \ldots, \epsilon_n)^T$. Note X denotes the vector of randomization assignments to treatment (1) or placebo (0) so that half of the entries are 0 and the other half are 1. Thus,

$$X'X \approx \frac{n}{2}.$$

(b) In matrix notation,

$$(Y - Z) = \begin{bmatrix} 1_n & Z & X \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{bmatrix} + \epsilon$$

Let $W = \begin{bmatrix} 1_n & Z \end{bmatrix}$ so that we partition the design matrix to $\begin{bmatrix} W & X \end{bmatrix}$. It can be shown that the least squares estimator for $\hat{\beta}_1$ is

$$\hat{\beta}_1 = (X'MX)^{-1} X'M(Y - Z)$$

where M is the projection onto the orthogonal compliment of the column space of W, i.e.,

$$M = I - W\,(W'W)^{-1}\,W'.$$

To simplify the algebra, we construct a vector S so that $s_i = x_i$ if $x_i = 1$, and $s_i = -1$ if $x_i = 0$.

$$X = \frac{1}{2}\,[1 + S]$$

S is simply another representation of treatment assignment, yet it has some nice properties because of the equal randomization scheme:

$$S'1 \approx 0$$
$$S'Z \approx 0$$
$$S'W \approx 0$$
$$S'M \approx 0$$

and

$$S'MS \approx S'S \approx n$$

We calculate the estimator of $\hat{\beta}_1$,

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[(X'MX)^{-1} X'M(Y-X)\right] \\ &= \left[\frac{1}{4}(1+S)' M (1+S)\right]^{-1} \left[\frac{1}{2}(1+S)' M\right] E[Y-Z] \\ &= 2\left[1'M1 + 1'MS + S'M1 + S'M1 + S'MS\right]^{-1} \left[1'M + S'M\right](\alpha_0 Z + \alpha_1 D_z X) \end{aligned}$$

Notice that

$$1'M = 0$$

since M is the projection onto the orthogonal space of $\mathcal{R}([1, Z])$. Hence

$$\begin{aligned} E[\hat{\beta}_1] &\approx 2\left[S'S\right]^{-1}\left[S'\right](\alpha_0 Z + \alpha_1 D_z X) \\ &\approx \frac{2}{n}\frac{\alpha_1 \sum_{i=1}^n z_i}{2} \\ &= \alpha_1 \bar{z} \end{aligned}$$

(c) Under $H_1 : \alpha_1 = 0$ (true model), there is no treatment effect. Under this null hypothesis,

$$E[\hat{\beta}_1] \approx \alpha_1 \bar{z} = 0$$

Therefore testing H: $\beta_1 = 0$ is equivalent to testing H: $\alpha_1 = 0$ if we use $\hat{\beta}_1$. This is a valid test for testing H: $\alpha_1 = 0$.

(d) Let V be the design matrix fitting postulated model, i.e.

$$V = \begin{bmatrix} 1^{n\times 1} & Z^{n\times 1} & X^{n\times 1} \end{bmatrix}$$

Therefore

$$\begin{aligned} (V'V)^{-1} &= \left[X'MX\right]^{-1} \\ &= 4\left[(1+S)'M(1+S)\right]^{-1} \\ &= 4\left[S'S\right]^{-1} \\ &= \frac{4}{n} \end{aligned}$$

and

$$
\begin{aligned}
\widehat{var}(\hat{\beta}_1) &= (V'V)^{-1}\,\widehat{var}(Y-Z) \\
&= (V'V)^{-1}\,\frac{RSS}{n-p} \\
&\approx \frac{4}{n}\frac{RSS}{n-p}
\end{aligned}
$$

Here RSS is the residual sum of squares from the postulated model. We can further calculate its expectation. P is the projection matrix fitting the postulated model, i.e., $P = V(V'V)^{-1}V'$

$$
\begin{aligned}
E[RSS] &= E[(Y-Z)'(I-P)(Y-Z)] \\
&= tr\{(I-P)\sigma^2 D_z^2\} + (E[Y-Z])'(I-P)(E[Y-Z]) \\
&= \sigma^2 \sum_{i=1}^{n} z_i^2 (1 - v_i(V'V)^{-1}v_i') + (\alpha_0 Z + \alpha_1 D_z X)'(I-P)(\alpha_0 Z + \alpha_1 D_z X) \\
&= \sigma^2 \sum_{i=1}^{n} z_i^2 (1 - v_i(V'V)^{-1}v_i')
\end{aligned}
$$

Here $\sigma^2 = var(\epsilon_i)$. Note $v_i(V'V)^{-1}v_i'$ is the diagonal term of projection matrix P when fitting the postulated model. Denote $P_{ii} = v_i(V'V)^{-1}v_i'$. We have

$$
\sum_{i=1}^{n} P_{ii} = \sum_{i=1}^{n} v_i(V'V)^{-1}v_i' = n-p
$$

Hence

$$
E[\widehat{var}(\hat{\beta}_1)] \approx \frac{4\sigma^2}{n}\frac{\sum_{i=1} z_i^2 P_{ii}}{n-p}
$$

On the other hand, if we fit the true model, we will get the true variance of $\hat{\alpha}_1$

$$
\begin{aligned}
var(\hat{\alpha}_0, \hat{\alpha}_1) &= (\begin{bmatrix} 1 & X \end{bmatrix}' \begin{bmatrix} 1 & X \end{bmatrix})^{-1}\sigma_2 \\
&= \begin{bmatrix} n & n/2 \\ n/2 & n/2 \end{bmatrix}^{-1} \sigma^2
\end{aligned}
$$

so

$$
var(\hat{\alpha}_1) = \frac{4\sigma^2}{n}
$$

So the variance of $\hat{\beta}_1$ is inflated by $\frac{\sum_{i=1}^{n} z_i^2 P_{ii}}{n-p}$ compared to the variance of $\hat{\alpha}_1$. And from part (b), the expectation of $\hat{\beta}_1$ is $\alpha_1 \bar{z}$. Combining these two results, if $\sqrt{\frac{\sum_{i=1}^{n} z_i^2 P_{ii}}{n-p}} > \bar{z}$, T will be smaller than it should be. Hence the test based on T will be conservative.
If $\sqrt{\frac{\sum_{i=1}^{n} z_i^2 P_{ii}}{n-p}} < \bar{z}$, T will be larger than it should be. Hence the test based on T will be anti-conservative.

4. Let the true and fitted model be reversed from the question 4. Find $E[\hat{\alpha}_1]$.

Solution: Suppose the true model is "absolute change," yet the postulated model is "relative change." Fitting the OLS by postulated model gives

$$\hat{\alpha_1} = \left(X'M_1X\right)^{-1} X'M_1D_z^{-1}(Y - Z)$$

where $M_1 = I - 1(1'1)^{-1}1'$.

$$X'M_1X \approx \frac{n}{4}$$

and

$$X'M_1 = X' - \frac{1}{2}1' = \frac{1}{2}S$$

where S is the row vector with i^{th} element 1 if $x_i = 1$, - 1 if $x_i = 0$. The estimator has expected value

$$\begin{aligned} E[\hat{\alpha}_1] &= (X'M_1X)^{-1}X'M_1D_z^{-1}(1\beta_0 + X\beta_1 + Z\beta_2) \\ &= \left(\frac{n}{4}\right)^{-1}(1/2)SD_z^{-1}(1\beta_0 + X\beta_1 + Z\beta_2) \\ &\approx \frac{2}{n}SD_z^{-1}X\beta_1 \end{aligned}$$

Since by randomization we have

$$SD_z^{-1}1 = 0$$

and

$$SD_z^{-1}Z = 0$$

Moreover,

$$SD_z^{-1}X \approx \frac{\sum_{i=1}^n z_i^{-1}}{2}$$

Hence

$$E[\hat{\alpha}_1] \approx \frac{\beta_1}{n}\sum_{i=1}^n z_i^{-1} \approx \frac{\beta_1}{\bar{z}}$$