Statistics 600 Problem Set 1
Due in class on Wednesday, October 7th.

1. Suppose we have a least squares problem with more variables than observations. That is, we observe a response vector $Y \in \mathbb{R}^n$, and a design matrix $X \in \mathbb{R}^{n \times p}$ where $p \geq n$. You may assume that $Y \in \text{col}(X)$.

(a) Derive an expression for the vector $\hat{\beta}$ that minimizes $\|\beta\|^2$ subject to $X\beta = Y$.

Solution: Using the QR decomposition, write $X' = QR$, so the equation $X\beta = Y$ becomes $Q'\beta = G$, where $G = R^{-T}Y$. Next we will show that $\hat{\beta} \in \text{col}(Q)$. We can write $\beta = \theta + \gamma$, where $\theta \in \text{col}(Q)$ and $\gamma \in \text{col}(Q)^\perp$. Note that $Q'\beta = Q'\theta$, and $\|\beta\|^2 = \|\theta\|^2 + \|\gamma\|^2$. Thus for any choice of $\theta$ satisfying $Q'\theta = G$, $\|\beta\|^2$ will always be minimized by setting $\gamma = 0$. Since $\theta \in \text{col}(Q)$, we can write $\theta = Q\eta$ for some $\eta \in \mathbb{R}^n$, and we have $Q'Q\eta = \eta = G$, and $\theta = QG$. Thus the solution is $\hat{\beta} = QR^{-T}Y$.

(b) Under what conditions is $\hat{\beta}$ unbiased? You may assume the usual generating model $Y = X\beta + \epsilon$ with $E(\epsilon|X) = 0$.

Solution: We can write

$$\hat{\beta} = QR^{-T}Y$$
$$= QR^{-T}(X\beta + \epsilon)$$
$$= QR^{-T}(R'Q'\beta + \epsilon)$$
$$= QQ'\beta + QR^{-T}\epsilon.$$

Thus $E(\hat{\beta}|X) = QQ'\beta$, which is equal to $\beta$ under the condition that $\beta \in \text{col}(Q) = \text{col}(X')$.

(c) Derive an expression for $\text{cov}(\hat{\beta}|X)$, under the generating model $Y = X\beta + \epsilon$ with $E(\epsilon|X) = 0$ and $\text{cov}(\epsilon|X) = \sigma^2 I$.

Solution:

$$\text{cov}(\hat{\beta}|X) = \text{cov}(QR^{-T}Y|X)$$
$$= \text{cov}(QR^{-T}(X\beta + \epsilon)|X)$$
$$= \text{cov}(QR^{-T}(R'Q'\beta + \epsilon)|X)$$
\[
= \operatorname{cov}(QQ'\beta + QR^{-T}\epsilon|X) \\
= \operatorname{cov}(QR^{-T}\epsilon|X) \\
= \sigma^2 QR^{-T} R^{-1}Q'.
\]

(d) Let \( \hat{Y} = X\hat{\beta} \) be the usual fitted values. What is the value of \( E\|\hat{Y} - Y\|^2? \)

**Solution:**

\[
\hat{Y} = XQR^{-T}Y = R'Q'R^{-T}Y = Y.
\]

Thus \( E\|\hat{Y} - Y\|^2 = 0 \) – or \( \hat{Y} \) is always equal to \( Y \).

(e) What is the value of \( E\|\hat{Y} - EY\|^2/n? \)

**Solution:**

\[
E\|\hat{Y} - EY\|^2/n = E\|Y - EY\|^2/n = \sigma^2.
\]

(f) Suppose we observe a random vector \( Y^* \in \mathcal{R}^n \) that has the same distribution as \( Y \), but is independent of \( Y \). What is the value of \( E\|Y^* - \hat{Y}\|^2/n? \)

**Solution:** Write \( Y^* = X\beta + \epsilon^* \),

\[
E\|X\beta + \epsilon^* - (X\beta + \epsilon)\|^2/n = E\|\epsilon - \epsilon^*\|^2/n = 2\sigma^2.
\]

2. Suppose we observe data from a simple linear model \( Y = \alpha + \beta X + \epsilon \) where \( X, Y \in \mathcal{R}^n \), \( E(\epsilon|X) = 0 \) and \( \operatorname{cov}(\epsilon|X) = \sigma^2 I \). Suppose \( X \) and \( Y \) are partitioned as

\[
Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},
\]

where \( Y_1 \) and \( Y_2 \) each have half the length of \( Y \), and \( X_1 \) and \( X_2 \) each have half the length of \( X \). Let \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) denote the least squares estimates obtained by regressing \( Y_1 \) on \( X_1 \) and \( Y_2 \) on \( X_2 \), respectively, and let \( \hat{\beta} = (\hat{\beta}_1 + \hat{\beta}_2)/2 \).

(a) If \( \bar{X}_1 = \bar{X}_2 = \bar{X} \), state a condition such that \( \hat{\beta} \) has the same variance as the least squares estimate \( \beta \) obtained by regressing \( Y \) on \( X \). Then state whether when this condition holds, \( \hat{\beta} \) is the least squares estimate, or is a different estimate with the same variance.
Solution: Let \( T_1 = \sum_{i=1}^{n/2} \epsilon_i (X_i - \bar{X}_1) \) and \( T_2 = \sum_{i=n/2+1}^{n} \epsilon_i (X_i - \bar{X}_2) \), and let \( S_1 = \sum_{i=1}^{n/2} (X_i - \bar{X}_1)^2 \) and \( S_2 = \sum_{i=n/2+1}^{n} (X_i - \bar{X}_2)^2 \). Then

\[
\hat{\beta}_1 = \beta + \frac{T_1}{S_1},
\]

\[
\hat{\beta}_2 = \beta + \frac{T_2}{S_2},
\]

and

\[
\hat{\beta} = \beta + \frac{T_1}{2S_1} + \frac{T_2}{2S_2}.
\]

Since \( \text{var}(T_j) = \sigma^2 S_j \) for \( j = 1, 2 \), it follows that

\[
\text{var}\hat{\beta} = \frac{\sigma^2}{4S_1} + \frac{\sigma^2}{4S_2}.
\]

The variance of the least squares estimate using all the data is

\[
\sigma^2 / \sum_i (X_i - \bar{X})^2 = \sigma^2 / (S_1 + S_2).
\]

The two variances are equal if only if

\[
(S_1 + S_2)^2 = 4S_1 S_2,
\]

which is easily seen to hold if and only if \( S_1 = S_2 \). This is the condition required for the variance of \( \hat{\beta} \) to equal the variance of \( \hat{\beta} \), and it is easy to see that when \( S_1 = S_2 \), \( \hat{\beta} = \hat{\beta} \).

(b) Now consider the more general case where \( \bar{X}_1 \) and \( \bar{X}_2 \) may differ. Show that in this case \( \text{var}\hat{\beta} \) is always greater than \( \text{var}\hat{\beta} \), and derive a concise expression for the difference.

Solution: By the Gauss-Markov theorem, since \( \hat{\beta} \) is linear and unbiased, if \( \hat{\beta} \neq \hat{\beta} \), then \( \text{var}(\hat{\beta}) \) must be greater than \( \text{var}(\hat{\beta}) \).

We can show this directly as follows.

\[
\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n/2}(X_i - \bar{X}_1 + \bar{X}_1 - \bar{X}_2 + \bar{X}_2 - \bar{X})^2 + \]

3
\[ \sum_{i=n/2+1}^{n} (X_i - \bar{X}_1 + \bar{X}_1 - \bar{X}_2 + \bar{X}_2 - \bar{X})^2 \]

Taking the first term,

\[ \sum_{i=1}^{n/2} (X_i - \bar{X}_1 + \bar{X}_1 - \bar{X}_2 + \bar{X}_2 - \bar{X})^2 \]

\[ = \sum_{i=1}^{n/2} (X_i - \bar{X}_1)^2 + (\bar{X}_1 - \bar{X}_2)^2 + (\bar{X}_2 - \bar{X})^2 + (X_i - \bar{X}_1)(\bar{X}_1 - \bar{X}_2) + (X_i - \bar{X}_1)(\bar{X}_2 - \bar{X}) + (\bar{X}_1 - \bar{X}_2)(\bar{X}_2 - \bar{X}) \]

\[ = \sum_{i=1}^{n/2} (X_i - \bar{X}_1)^2 + (\bar{X}_1 - \bar{X}_2)^2 + (\bar{X}_2 - \bar{X})^2 + (X_i - \bar{X}_1)(\bar{X}_1 - \bar{X}_2) \]

\[ = S_1 + n(\bar{X}_1 - \bar{X}_2)^2/2 + n(\bar{X}_2 - \bar{X})^2/2 + n(\bar{X}_1 - \bar{X}_2)(\bar{X}_2 - \bar{X})/2. \]

We can apply a similar calculation to obtain

\[ \sum_{i=n/2+1}^{n} (X_i - \bar{X}_1 + \bar{X}_1 - \bar{X}_2 + \bar{X}_2 - \bar{X})^2 \]

\[ = S_2 + n(\bar{X}_1 - \bar{X}_2)^2/2 + n(\bar{X}_1 - \bar{X})^2/2 + n(\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X})/2. \]

Since

\[ (\bar{X}_1 - \bar{X}_2)(\bar{X}_2 - \bar{X}) + (\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}) = 0, \]

we have

\[ \sum_{i=1}^{n} (X_i - \bar{X})^2 = S_1 + S_2 + n(\bar{X}_1 - \bar{X}_2)^2 + n(\bar{X}_1 - \bar{X})^2/2 + n(\bar{X}_2 - \bar{X})^2/2. \]

Thus the difference in variances is

\[ \text{var}(\beta) - \text{var}(\beta) = \sigma^2/4S_1 + \sigma^2/4S_2 - 1/(S_1 + S_2 + D) \]

where \( D = n(\bar{X}_1 - \bar{X})^2/n + n(\bar{X}_2 - \bar{X})^2/2. \)

The difference in variances simplifies to

\[ \frac{(S_1 - S_2)^2 + D(S_1 + S_2)}{4S_1S_2(S_1 + S_2 + D)}. \]
3. Prove that the “horizontal residuals” in simple linear regression sum to zero in a least squares fit of $Y$ (the dependent variable) on $X$ (the independent variable). The horizontal residuals are the line segments connecting each data point $X_i, Y_i$ to the fitted line $\hat{\alpha} + \hat{\beta}X$.

**Solution:** To get the $i^{th}$ horizontal residual, solve

$$\hat{\alpha} + \hat{\beta}X = Y_i$$

to get $\hat{X}_i = (Y_i - \hat{\alpha})/\hat{\beta}$, so the residual becomes $H_i \equiv X_i - (Y_i - \hat{\alpha})/\hat{\beta}$. Now if we sum these values we get

$$\sum_i H_i = \sum_i X_i - (Y_i - \bar{Y} + \hat{\beta}(X_i - \bar{X}))/\hat{\beta}$$

$$= \sum_i (\bar{Y} - Y_i)/\hat{\beta} + \sum_i (X_i - \bar{X})$$

$$= 0.$$  

4. (a) Suppose that $F \in \mathcal{R}^d$ is a vector, and $I$ is the $d \times d$ identity matrix. Derive explicit expressions for $(I + FF')^{-1}$ and $(I - FF')^{-1}$. Hint: the answers have the form $I + \lambda FF'$, for $\lambda \in \mathcal{R}$.

**Solution:** To determine the inverse of $I + FF'$, set

$$I = (I + FF')(I + \lambda FF')$$

$$= I + \lambda FF' + FF' + \lambda \|F\|^2 FF'$$

$$= I + (\lambda + 1 + \lambda \|F\|^2)FF'.$$

We must have $1 + \lambda(1 + \|F\|^2) = 0$, so $\lambda = -1/(1 + \|F\|^2)$. To determine the inverse of $I - FF'$, set

$$I = (I - FF')(I + \lambda FF')$$

$$= I + \lambda FF' - FF' - \lambda \|F\|^2 FF'$$

$$= I + (\lambda - 1 - \lambda \|F\|^2)FF'.$$
We must have $-1 + \lambda(1 - \|F\|^2) = 0$, so $\lambda = 1/(1 - \|F\|^2)$.

(b) Suppose we have an orthogonal design matrix $X \in \mathbb{R}^{n \times p+1}$, and we are able to add one additional observation to the data set (i.e. add one row to $X$). This row, denoted $x$, must satisfy the constraint $\|x\|^2 = 1$. Describe how $x$ should be chosen so as to minimize the maximum of the variances of $\hat{\beta}_0, \ldots, \hat{\beta}_p$.

**Solution:** Let $\hat{\beta}$ denote the slope estimates based on all $n+1$ cases. Then $X'X = I + xx'$, so $\text{cov}(\hat{\beta}) = I - xx'/2$. Thus the variance of $\hat{\beta}_j$ is $\sigma^2(1-x_j^2/2)$. The maximum of these variances is determined by the smallest of the $x_j^2$. Thus we want to maximize $\min_j x_j^2$ subject to $\sum_j x_j^2 = 1$. The solution is to have $x_j = 1/\sqrt{p+1}$ for all $j$.

5. (a) Derive an expression for $\text{cov}(Y, \hat{Y})$, i.e. the $n \times n$ matrix containing all population covariances between elements of $Y$ and elements of $\hat{Y}$.

(b) Derive an expression for the sample covariance between the observed and fitted values, $E\text{cov}(\hat{Y}, Y)$ – note that this is a scalar. Consider whether this covariance can or cannot be positive, negative, or zero.

**Solution:** Let $P$ be the projection matrix onto $\text{col}(X)$. Then,

$$\text{cov}(\hat{Y}, Y) = (PY)'(Y - \bar{Y})/(n - 1)$$

$$= (Y - \bar{Y} + Y'P(Y - \bar{Y})/n - 1)$$

$$= (Y - \bar{Y})'P(Y - \bar{Y})/n - 1 + \bar{Y}'P(Y - \bar{Y})/n - 1.$$ 

Here, $\bar{Y}$ is interpreted as an $n$-vector in which all values are equal to the sample mean of the $Y_i$. This can be written $\bar{Y} = n^{-1}11'Y$, where $1$ is an $n$-vector of 1's. Since there is an intercept in the model, $P1 = 1$, so the second summand above is equal to

$$n^{-1}11'(Y - \bar{Y})/n - 1,$$

which is zero since $1'(Y - \bar{Y}) = 0$. Thus

$$\text{cov}(\hat{Y}, Y) = (Y - \bar{Y})'P(Y - \bar{Y})/n - 1 \geq 0.$$
The covariance cannot be negative. It can only be zero if \( Y \) is a constant vector.

6. “Total least squares” (TLS) for one covariate aims to identify a line \( \ell \) that minimizes

\[
\sum_i d((X_i, Y_i), \ell)^2,
\]

where \( d(Q, \ell) \) is the minimum distance in \( \mathbb{R}^2 \) between the point \( Q \) and any point on the line \( \ell \).

(a) Parameterize \( \ell \) in the form \( \{ (X, \alpha + \beta X) | X \in \mathbb{R} \} \), for scalars \( \alpha \) and \( \beta \). Write down expressions for \( d(Q, \ell) \) and a loss function that can be minimized to identify \( \alpha \) and \( \beta \). Both expressions should be explicit functions of \( \alpha \) and \( \beta \).

**Solution:** To identify the point on \( \ell \) that is closest to \( X_i, Y_i \), we minimize

\[
(X - X_i)^2 + (\alpha + \beta X - Y_i)^2
\]

as a function of \( X \). Setting the first derivative to zero yields

\[
X = \frac{X_i - \alpha \beta + Y_i \beta}{1 + \beta^2},
\]

and the second derivative is \( 2(1 + \beta^2) \), so this is a global minimizer. The loss function is

\[
(1 + \beta^2)^{-1} \sum_i R_i^2,
\]

where \( R_i = Y_i - \alpha - \beta X_i \) is the usual OLS residual.

(b) Parameterize \( \ell \) in the form \( \{ Z \in \mathbb{R}^2 | B'(Z - W) = 0 \} \), for 2-vectors \( B \) and \( W \) with \( ||B|| = 1 \). Write down expressions for \( d(Q, \ell) \) and a loss function that can be minimized to identify \( B \) and \( W \) (\( W \) can be any point on \( \ell \) and is therefore not uniquely identified). Both expressions should be explicit functions of \( B \) and \( W \).

**Solution:** Let \( Q_i = (X_i, Y_i) \) be a data point. Let \( P_i \) be the point on \( \ell \) that is closest to \( Q_i \). Then \( Q_i - P_i \) is parallel to \( B \), so we can write \( P_i = Q_i - \lambda B \) for some \( \lambda \in \mathbb{R} \), and since \( P_i \) is on \( \ell \) we must have \( B'(P_i - W) = 0 \). Combining these two equations we can identify \( \lambda = B'(Q_i - W) \). Therefore the \( d(Q_i, \ell)^2 \) is

7
\[ B'(Q_i - W)(Q_i - W)'B \]

so the loss function is

\[ B' \left( \sum_i (Q_i - W)(Q_i - W)' \right) B. \]

(c) Based on your expression in part (b), show that the TLS solution passes through the center of the data \((\bar{X}, \bar{Y})\), and use this to define a minimizing value for \(W\).

**Solution:**

\[
\sum_i (Q_i - W)(Q_i - W)' = \sum_i (Q_i - \bar{Q} + \bar{Q} - W)(Q_i - \bar{Q} + \bar{Q} - W)'
\]
\[
= \sum_i (Q_i - \bar{Q})(Q_i - \bar{Q})' + \sum_i (Q_i - \bar{Q})(\bar{Q} - W)' + \sum_i (\bar{Q} - W)(Q_i - \bar{Q})' + n(\bar{Q} - W)(\bar{Q} - W)'
\]
\[
= \sum_i (Q_i - \bar{Q})(Q_i - \bar{Q})' + n(\bar{Q} - W)(\bar{Q} - W)'.
\]

therefore the value of the loss function will either stay constant or be reduced if we set \(W = \bar{Q}\), which guarantees that \(\ell\) contains \(\bar{Q}\).

(d) Building on (b) and (c), construct a quadratic form whose minimizing value subject to \(\|B\| = 1\) solves the TLS problem for \(B\).

**Solution:** The quadratic form is

\[ B' \left( \sum_i (Q_i - \bar{Q})(Q_i - \bar{Q})' \right) B. \]

7. (a) Suppose we are fitting a simple linear regression model to a data set of size \(n\). Let \(V_n = \text{var}(X_1, \ldots, X_n)\). Determine the fastest rate at which \(V_n \to 0\) for which we still have \(\text{var}(\beta_n) \to 0\).
Solution: Since
\[
\text{var}(\hat{\beta}) = \frac{\sigma^2}{(n-1)V_n}
\]
we need \(nV_n \to \infty\) (or \(V_n \to 0\) “slower than \(1/n\”).

(b) Suppose we are fitting a regression model of the form \(\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2\), and the covariates are asymptotically standardized so that \(\bar{X}_1, \bar{X}_2 \to 0\), and \(\text{var}(X_1), \text{var}(X_2) \to 1\). Let \(r_n = \text{cov}(X_1, X_2)\). What is the fastest rate at which \(r_n \to 1\) such that we will still have \(\text{var}(\hat{\beta}_1), \text{var}(\hat{\beta}_2) \to 0\)?

Solution: The variance of \(\hat{\beta}_1\) (which is the same as the variance of \(\hat{\beta}_2\) is
\[
\frac{1}{n(1-r_n^2)} = \frac{1}{n(1-r_n)(1+r_n)}.
\]
So we need \(n(1-r_n) \to \infty\), or \(1-r_n\) goes to zero “slower than rate \(1/n\”).

8. This exercise aims to illustrate the effect of outliers in least squares fitting. Suppose we observe data that follows a linear model with \(p = 1\) covariate: \(Y = \alpha + \beta X + \epsilon\). Specifically, consider a triangular array of data \(Y_{in}, X_{in}\), where \(i = 1, \ldots, n\). There is also a random indicator \(\delta_{in}\), that we do not observe, such that \(\text{var}(\epsilon_{in}|X, \delta_{in} = 1) = k_n\sigma^2\), and \(\text{var}(\epsilon_{in}|X, \delta_{in} = 0) = \sigma^2\) (the errors are centered, so that \(E(\epsilon|X, \delta) \equiv 0\)). Suppose \(X\) is sampled from a population with variance \(\sigma_X^2\), and \(P(\delta_{in} = 1) = p_n\). Note that \(n \cdot \text{var}(\hat{\beta})\) has a finite limit when \(k_n \equiv 1\). Derive conditions on \(k_n\) and \(p_n\) such that (i) \(n \cdot \text{var}(\hat{\beta})\) has a finite limit, and (ii) \(n \cdot \text{var}(\hat{\beta})\) has the same limit that would occur if \(k_n \equiv 1\).

Solution: The least squares estimator can be written
\[
\hat{\beta}_n = \beta + \sum_i \epsilon_{in}(X_{in} - \bar{X}_n) / \sum_i (X_{in} - \bar{X}_n)^2.
\]
Since the variance of the error term can be expressed
\[
\text{var}(\epsilon_{in}) = \text{var}E(\epsilon_{in}|\delta_{in}) + E\text{var}(\epsilon_{in}|\delta_{in}) = \sigma^2(p_n k_n + 1 - p_n),
\]
the variance of the estimator is

$$\text{var}\hat{\beta}_n = \sigma^2(p_nk_n + 1 - p_n)\sum_{i}(X_{in} - \bar{X}_n)^2.$$ 

Scaling by $n$,

$$n \times \text{var}\hat{\beta}_n = \sigma^2(p_nk_n + 1 - p_n)/n^{-1}\sum_{i}(X_{in} - \bar{X}_n)^2 \sim \sigma^2(p_nk_n + 1 - p_n)/\sigma_x^2.$$ 

Thus for (i), we need $p_nk_n + 1 - p_n$ to have a limit, and for (ii), we need $p_n(k_n - 1) \to 0$. A reasonable interpretation of this is that the outliers will not prevent the variance of $\hat{\beta}$ from going to zero at the usual rate as long as $p_nk_n$ stays bounded. For example, if fraction $p_n = 0.1$ of the errors have $k_n = 10$ times greater variance ($\sqrt{10} \approx 3.2$ times greater standard deviation), then the variance of $\hat{\beta}_n$ will decrease at the usual rate. But if we want the limiting variance ($\lim_{n\to\infty} n\text{var}\hat{\beta}_n$) to be the same as when no outliers are present, we would need $p_n$ to be much smaller, say $p_n = 0.01$. 