Lecture 3

13. The case of random regressors
   a. Remember that above treatment presumes explanatory variables chosen by design.
   b. In random case, condition on explanatory variables.

14. Graphical approach in the bivariate normal case
   a. Expectations $\mu_x$, $\mu_y$, standard deviations $\sigma_x$, $\sigma_y$, correlation $\rho$.
   b. Rescale each of explanatory variable and response variable by
      i. subtracting off mean and
      ii. dividing by standard deviation
   c. Bivariate normality implies an elliptical point cloud.
   d. lines with slope $\pm 1$ divide point cloud symmetrically
   e. Suppose $+1$ gives major axis
   f. Then this line gives best description of how variables move together
      i. Freedman, Pisani, and Purves call this the sd line
      ii. But it is too high for the best explain the various response values associated with a fixed explanatory variable
      iii. Better answer: flatten this line by multiplying by correlation.
      iv. Phenomenon is called regression to the mean
         a. On original scale, $\beta_1 = \rho \sigma_y/\sigma_x$, $\beta_0 = \mu_y - \mu_x \beta_1$.
         b. Estimate of correlation is
            
            $r = \hat{\rho} = \frac{n}{\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y})}/\sqrt{\sum_{j=1}^{n} (X_j - \bar{X})^2 \sum_{j=1}^{n} (Y_j - \bar{Y})^2}$
            
            c. Cf. $\hat{\beta}_1 = \frac{\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$.
            d. Let $\hat{\sigma}_X^2 = \sum_{j=1}^{n} (X_j - \bar{X})^2/(n-1)$,
               $\hat{\sigma}_Y^2 = \sum_{j=1}^{n} (Y_j - \bar{Y})^2/(n-1)$.
            e. Then $\hat{\beta}_1 = \rho \hat{\sigma}_Y/\hat{\sigma}_X$.
               i. That is, use correlation and rescale to represent ratio of response and explanatory variables.
   e. Test of $H_0 : \rho = 0$ is equivalent to test of $H_0 : \beta_1 = 0$.
      i. Recall $T = (\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\sigma}/\sum_{j=1}^{n} (X_j - \bar{X})^2} \sim T_{n-2}$.
      ii. $(n-1)\hat{\sigma} = S_y y - S_{xy}/S_{xx}$
         a. Because
            $(n-1)\hat{\sigma}^2 = \sum_{j=1}^{n} \left( y_j - \bar{y} - \hat{\beta}_1(x_j - \bar{x}) \right)^2$
            $= S_{yy} - \hat{\beta}_1^2 S_{xy} + \hat{\beta}_1^2 S_{xx}$
            $= S_{yy} - 2\hat{\beta}_1 S_{xy} + S_{xy}^2 + S_{xx}$
            $= S_{yy} - S_{xy}^2/S_{xx}$
         b. For $S_{yy} = \sum_{j=1}^{n} (y_j - \bar{y})^2$ and $S_{xy} = \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})$.
         iii. Then $\sqrt{n-2}/\hat{\rho}/\sqrt{1-\rho^2} \sim T_{n-2}$

15. Correlation and Regression

   - because $T = \sqrt{n-2}(S_{xy}/S_{xx})/(\sqrt{S_{yy} - S_{xy}^2/S_{xx}}/\sqrt{S_{xx}}) = \sqrt{n-2}(S_{xy}/\sqrt{S_{xy}S_{yy}})/\sqrt{1-S_{xy}^2/(S_{xx}S_{yy})} = \sqrt{n-2}\hat{\rho}/\sqrt{1-\rho^2}$
   - Outside of the $\rho^2 = 0$ case:
      i. $\beta_0^2 \neq \rho^2 \sigma_y/\sigma_x$ unless both null values are zero.
      ii. Recall inference is conditional on $X$ but not on $Y$.
      iii. Sampling distribution of $\hat{\rho}$ is more complicated than that of $\hat{\beta}_1$, because of division by $\hat{\sigma}_Y$.
      iv. Sampling distribution is skewed.
      v. A Taylor series approach to removing skewness leads to the transformation to $Z = \frac{1}{2} \log \left( \frac{1+\hat{\rho}}{1-\hat{\rho}} \right)$: Fisher’s $Z$ transform (Wikipedia credits this to Fisher, 1915. SAS credits it to Fisher, 1921.)
         a. $Z \sim N \left( \frac{1}{2} \log \left( \frac{1+\hat{\rho}}{1-\hat{\rho}} \right), 1/(n-3) \right)$
   - Define interval $I = \{ \rho \mid \sqrt{n-3} \left\{ \frac{1}{2} \log \left( \frac{1+\hat{\rho}}{1-\hat{\rho}} \right) - \frac{1}{2} \log \left( \frac{1+\rho}{1-\rho} \right) \right\} \leq z_{\alpha/2} \}$.
      a. A small bias correction of size $O(1/n)$ might or might not be used.
      b. It does not impact accuracy of the CIs.
      ii. Interval is
         $\left( \frac{1-\rho}{1+\rho} \exp(z_{\alpha/2}/\sqrt{n-3} - \hat{\rho}/2 - 1) \frac{1}{1-\rho} \right) \left( \frac{1-\rho}{1+\rho} \exp(z_{\alpha/2}/\sqrt{n-3} - \hat{\rho}/2 + 1) \frac{1}{1-\rho} \right) \left( \frac{1}{1+\rho} \exp(z_{\alpha/2}/\sqrt{n-3} + \hat{\rho}/2 - 1) \right) \left( \frac{1}{1+\rho} \exp(z_{\alpha/2}/\sqrt{n-3} + \hat{\rho}/2 + 1) \right) \exp(z_{\alpha/2}/\sqrt{n-3})
      iii. Can replace this as
         $tanh(tanh^{-1}(\hat{\rho}) - z_{\alpha/2}/\sqrt{n-3}) \leq z_{\alpha/2}/\sqrt{n-3}$
   - for $tanh(u) = \frac{exp(u)-exp(-u)}{exp(u)+exp(-u)}$,
     and its inverse is $tanh^{-1}(x) = \frac{1}{2} \log \left( \frac{1+x}{1-x} \right)$.

MPV: 2.12

D. Influence and Outliers

1. Sometimes an observation is strongly dissimilar to the remainder of the observations in a data set.
   a. Such an observation is called an outlier
   b. Can be a rare event.
   c. Can be an experimental error.
   d. Can be a data transcription error.

2. Points farther from the center of the explanatory variable values have more influence on the regression parameter estimate
   a. Recall $\hat{\beta}_1 = \sum_{j=1}^{n} c_j y_j$ for $c_j = (X_j - \bar{X})/S_{xx}$
   b. Hence $\hat{\beta}_1$ changes most for a given change in $Y_j$ when $|X_j - \bar{X}|$ largest
   c. If $X_j = \bar{X}$, then $Y_j$ has no influence on $\hat{\beta}_1$.
   d. Hence a subject whose explanatory variable is an outlier is influential.
   e. An outlier among the $(X_j, Y_j)$ pairs is one that fails to follow the relationship between the response and explanatory variable.

MPV: 3.0-3.1

IV. Multiple Regression

A. The Multiple Regression Model

1. Adding an additional variable

MPV: 2.9
B. Estimation of Coefficients

1. Multivariate Normal Equations

a. Suppose we want to add a second (for now) additional variable.
   i. Giving a new symbol for each new thing we want in the model will become unwieldy.

b. \( Y_i = \beta_0 + x_i \beta_1 + x_{i2} \beta_2 + \epsilon_i \)
   i. Second index on \( x \) refers to which variable
   ii. First index refers to which subject.

c. \( Y_i = \beta_0 + x_{i1} \beta_1 + x_{i2} \beta_2 + \epsilon_i \)
   i. Recall that \( x_{i0} \) are all 1.

2. Least Squares

a. \( S = \sum(Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \)

b. Strategy with one regressor was to
   i. Find best intercept term in terms of the unknown slope
   ii. Plug this best intercept term into equation for best slope
   iii. Solve this one equation with one unknown.

c. As before, best intercept is \( \hat{\beta}_0 = \bar{Y} - \bar{x}_1 \beta_1 - \bar{x}_2 \beta_2 \)
   i. Plugging this into the normal equations gives
      • two equations (one for \( \frac{\partial}{\partial \beta_1} S = 0 \) and one for \( \frac{\partial}{\partial \beta_2} S = 0 \))
      • Two unknowns \( \beta_1 \) and \( \beta_2 \)
   ii. Solving these is trivial
   iii. We could write down a formula, but it would be complicated.

d. Juxtaposition of \( Y \) and \( \beta \) represents matrix multiplication
   i. That is, if \( Y = X \beta \), then \( \beta \) is the vector
      • with as many components as there are observations
      • Value in slot \( i \) is \( \beta_0 + \sum_{j=1}^{k} x_{ij} \beta_j = \sum_{j=0}^{k} x_{ij} \beta_j \)
   ii. General Definition: The product of a \( n \times k \) matrix and a \( k \) vector is
      • an \( n \) vector
      • component \( i \) is sum of elements in row \( i \) of the matrix times corresponding elements of vector.

e. \( \epsilon \) is \( n \) vector of entries of independent identically-distributed errors.

III. The model in matrix form is \( Y = X \beta + \epsilon \)

a. \( Y \) represents the ordered list of observations
   i. Called a vector.
   ii. To incorporate information that it has \( n \) entries, call it an \( n \) vector
b. \( X \) represents the explanatory variables
   i. Represented in a grid
   ii. Rows represent subject, columns represent regressors
   iii. In order to use a similar notation for the intercept parameter as we use for the slope parameter,
      • define \( X \) to have as many columns as it has slope parameters, plus 1.
      • Make the first column all 1s, because \( \beta_0 = 1 \times \beta_0 \)
      • Denote this by \( x_{i0} = 1 \) for all \( i \)

c. \( \beta \) represents the ordered list of slope parameters
   i. Put the intercept parameter as the first value
   ii. So with two explanatory variables, \( \beta = (\beta_0, \beta_1, \beta_2) \).

d. Juxtaposition of \( X \) and \( \beta \) represents matrix multiplication
   i. That is, if \( X \beta \), then \( \beta \) is the vector
      • with as many components as there are observations
      • Value in slot \( i \) is \( \beta_0 + \sum_{j=1}^{k} x_{ij} \beta_j = \sum_{j=0}^{k} x_{ij} \beta_j \)
   ii. General Definition: The product of a \( n \times k \) matrix and a \( k \) vector is
      • an \( n \) vector
      • component \( i \) is sum of elements in row \( i \) of the matrix times corresponding elements of vector.

e. \( \epsilon \) is \( n \) vector of entries of independent identically-distributed errors.

f. \(+\) is vector addition.
   i. need both sides to have same number of components
   ii. Result is component-wise sum.

MPV: 3.2

B. Estimation of Coefficients

1. Multivariate Normal Equations

a. \( S = \sum(Y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \)
   i. Recall that \( x_{i0} \) are all 1.

b. \( \frac{\partial}{\partial \beta_0} = -\sum_i 2(Y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \beta_2 x_{i2}) x_{i0} \)

i. Because we introduced \( x_{i0} \), we don’t have to treat the intercept differently

c. Advantages:
   i. Algebraically simpler
   ii. Delegates parts of the problem that statisticians aren’t likely so good at to those who are.
   iii. Allows for dynamic updating of the solution, as the field of matrix algebra evolves

d. Disadvantages:
   i. Removes insights that we might gain from getting fingers into more of the details.

2. Some matrix facts:

a. Just as with scalars, matrix multiplication is
   i. This is not really multiplication: here if \( n = 10 \) and \( k = 3 \), read this as “10 by 1” and not “30”.

c. \( \beta \) represents the ordered list of slope parameters
   i. Put the intercept parameter as the first value
   ii. So with two explanatory variables, \( \beta = (\beta_0, \beta_1, \beta_2) \).

d. Juxtaposition of \( X \) and \( \beta \) represents matrix multiplication
   i. That is, if \( X \beta \), then \( \beta \) is the vector
      • with as many components as there are observations
      • Value in slot \( i \) is \( \beta_0 + \sum_{j=1}^{k} x_{ij} \beta_j = \sum_{j=0}^{k} x_{ij} \beta_j \)
   ii. General Definition: The product of a \( n \times k \) matrix and a \( k \) vector is
      • an \( n \) vector
      • component \( i \) is sum of elements in row \( i \) of the matrix times corresponding elements of vector.

e. \( \epsilon \) is \( n \) vector of entries of independent identically-distributed errors.

f. \(+\) is vector addition.
   i. need both sides to have same number of components
   ii. Result is component-wise sum.

MPV: 3.2

The result of doing these two sequential multiplications of the vector can be expressed as a matrix.

ii. Formulate more generally: Find \( C \) so that
   \( A(B) = C \beta \)
   i. Let \( A \) have entries in row \( i \) and column \( j \) \( a_{ij} \)
   ii. Let \( B \beta \) have entries in row \( i \) and column \( j \) \( b_{ij} \)
   iii. Recall that entry \( i \) in \( B \beta \) is \( \sum_{j} b_{ij} \beta_j \)
   iv. Then entry \( l \) in \( A(B) \) is \( \sum_{j} a_{ij} (\sum_{j} b_{ij} \beta_j) \)
   v. Rearrange terms in sum to do summation over \( i \) for \( j = 1 \) first, then summation over \( i \) for \( j = 2 \), then ... \( \sum_{j} (\sum_{i} a_{ij} b_{ij} \beta_j) \)
      Commutative property of addition
Factor out $\beta_j$ from multiple terms that contain it:
\[
\sum_j (\sum_i a_{ij} b_{ij}) \beta_j \text{: Distributive Property}
\]
\[
\sum_j c_{ij} \beta_j \text{ for } c_{ij} = \sum_i a_{ij} b_{ij}
\]
iii. So define the matrix product $AB$ to be the matrix with entry $c_{ij} = \sum_i a_{ij} b_{ij}$.

- Makes sense only of number of rows of $B$ is the same as the number of columns of $A$.
- Result has number of rows of first matrix and number of columns of second matrix.

iv. The definition of multiplication of a matrix and a vector is a special case, if the vector is viewed as having one column.

v. The same argument shows that matrix multiplication is associative: $A(BC) = (AB)C$.

vi. Matrix multiplication is NOT commutative.

- The product of a $2 \times 3$ and a $3 \times 4$ matrix is a $2 \times 4$ matrix.
- But the product with the orders reversed is not defined, because $3 \times 4$ and $2 \times 3$ matrices do not have the number of columns of first matching number of rows of second.

3. Application of Matrix Operations to Model Fitting

a. Then $0 = X^T Y - X^T (X \beta) = X^T Y - (X^T X) \beta$.

b. Then $X^T Y = (X^T X) \beta$.

4. Case of regression through the origin, one explanatory variable

a. Write values of explanatory variable $x_{11}, x_{21}, \ldots, x_{n1}$

i. In this case the second subscript always takes the same value

6. Case of standard regression, one centered explanatory variable

a. $X_1 = 0$

b. Normal equation is
\[
\left( \sum_i x_{1i} \right) Y_i = \left( \begin{array}{c} n \\ \sum_i x_{1i}^2 \end{array} \right) \begin{array}{c} \beta_0 \\ \beta_1 \end{array} = \begin{array}{c} \beta_0 \sum_i x_{1i}^2 \\ \beta_1 \sum_i x_{1i} \\ \end{array}
\]

c. Solution is easy, since $X^T X$ has only one non-zero entry in each row and column.

i. Problem for each parameter looks like the 1-parameter no-intercept case.

ii. Happens because $X^T X$ has only one non-zero entry in each row and column.

7. Plan for General Case

a. Normal equations will reduce to a univariate equation in general by multiplying each side of the normal equation by a matrix to remove multiple parameters being involved in one of the normal equations.

b. Find $C$ such that $C(X^T X)$ is of form
\[
\begin{pmatrix}
1 & 0 & \ldots \\
0 & 1 & \ldots \\
\vdots & \vdots & \ddots
\end{pmatrix}
\]

i. All zeros except 1 if row number matches column number.

ii. Such positions are called the diagonal.

iii. A matrix of all zeros except all 1 on the diagonal is called an identity matrix, because

- if $B$ is a $n \times m$ matrix, and if $I$ is a $n \times n$ identity matrix, then $IB = B$.

- Retained for consistency with larger models.

b. $X$ is a vector with one column, $n$ rows.

c. $X^T Y$ is a vector with entry, entry $\sum_{i=1}^n x_{1i} y_i$.

d. $X^T X$ is a $1 \times 1$ matrix, entry $\sum_{i=1}^n x_{1i}^2$.

e. Normal equation is $\sum_{i=1}^n x_{1i} y_i = (\sum_{i=1}^n x_{1i}^2) \beta_1$.

f. Only solution is $\hat{\beta}_1 = (\sum_{i=1}^n x_{1i} y_i) / (\sum_{i=1}^n x_{1i}^2)$, after dividing by element of $(X^T X)$.

5. Case of standard regression, one explanatory variable

a. $X$ is $n \times 2$, first column all 1.

b. $X^T Y$ is vector with entry $j$ equal to $\sum_{i=1}^n x_{1i} Y_i$

c. $X^T X$ is a $2 \times 2$ matrix.

i. Upper left entry is $n$.

ii. Upper right entry is

- sum of 1 times $x_{1i}$; sum is $\sum_i x_{1i}$.

iii. Lower left entry is same.

iv. Lower right entry is $\sum_i x_{1i}^2$.

d. Normal equation is
\[
\left( \sum_i x_{1i} Y_i \right) = \left( \begin{array}{c} n \\ \sum_i x_{1i}^2 \end{array} \right) \begin{array}{c} \beta_0 \\ \beta_1 \end{array} = \begin{array}{c} n \beta_0 + \beta_1 \sum_i x_{1i} \\ \beta_0 \sum_i x_{1i} + \beta_1 \sum_i x_{1i}^2 \end{array}
\]

e. By inspection, these solve the equations:

ii. $\hat{\beta}_1 = \sum_{i=1}^n (Y_i - \bar{Y}) X_{1i} / \sum_{i=1}^n (X_{1i} - \bar{X})^2$.

- Uses $\sum_i (X_{1i} - \bar{X})^2 = \sum_i X_{1i}^2 - 2 \bar{X} \sum_i X_{1i} + n \bar{X}^2 = \sum_i X_{1i}^2 - n \bar{X}^2$

8. Interpretation of $\beta_j$

a. For $j \geq 1$ is the amount by which the fitted value increases per unit increase in explanatory variable $j$.

i. If coefficient is negative, indicates that fitted value decreases as explanatory variable increases.

ii. All other explanatory variables are assumed held fixed.

- This may not be realistic.

b. For $j = 0$, indicates fitted value when all explanatory variables are zero.

i. This may be made more useful if variables have a baseline value subtracted off first.

- Often more interested measuring the strength of an effect of an explanatory variable on a response.

i. Often useful to compare effect to the variability in the data.
ii. Make this comparison easier by subtracting off the mean and dividing by standard deviation of explanatory and response variables.
iii. Resulting regression parameters are called standardized.

9. Fitted Values and Residuals
a. Fitted values are $X\hat{\beta} = X(X^\top X)^{-1}X^\top Y$

b. Matrix $H = X(X^\top X)^{-1}X^\top$ is called the Hat matrix.

c. Note $H$ is symmetric
d. Residuals are $\hat{\epsilon} = (I - H)Y$.

10. Hat Matrix Times Itself
a. Note $HH = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = H$

c. A matrix that is its own square is called idempotent.

11. Unbiasedness of $\hat{\beta}$

a. Let $C = (X^\top X)^{-1}X^\top$

b. Because either $X$ is fixed, or we condition on $X$, for the purposes of calculating expected value, $C$ acts like a constant.
c. Because each component of $CY$ is the sum of constants times components of $Y$, then $E[\hat{\beta}] = C E[Y] = C (X\beta) = (X^\top X)^{-1}(X^\top X)\beta = \beta$.

12. A General Covariance Fact:

a. Suppose $Y$ is a random vector with $k$ components.
b. Define its covariance matrix $\text{Cov}[Y]$ as the matrix with $\text{Cov}[Y_i, Y_j]$ in row $i$, column $j$.

i. $\text{Cov}[Y]$ also contains variances, which are a special case of covariances.
ii. I will probably forget and sometimes write this as $\text{Var}[Y]$.
iii. Let $\Sigma = \text{Cov}[Y]$.
iv. This matrix is $k \times k$.
c. Let $U = AY$ for a $m \times k$ matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ a_{21} & \cdots & a_{2k} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mk} \end{pmatrix}.$$ 

d. Calculate $\text{Cov}[U]$ 

i. Pick two $i$, $j$ and calculate $\text{Cov}[U_i, U_j]$

• picking $i = j$ gives a variance.
ii. $\text{Cov}[U_i, U_j] = \text{Cov}[\sum_l a_{il}Y_l, \sum_m a_{jm}Y_m]$

iii. Sums come out front, so $\text{Cov}[U_i, U_j] = \sum_l \sum_m a_{il} \text{Cov}[Y_l, Y_m] a_{jm}$

iv. This is element in row $i$, column $j$ of $A\Sigma A^\top$.
v. Here we use the fact $(AB)^\top = B^\top A^\top$ (check)