POLI SCI 231b: Problem Set 3 Solution Set

Spring 2017

University of California, Berkeley

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

1. **Show that if $z_i = a + bx_i$ for all i, then $\bar{z} = a + b\bar{x}$. How does this relate to the fact that the regression line passes through the point of averages?**

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} z_i &= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \\
\frac{z_1 + z_2 + ... + z_n}{n} &= \frac{1}{n}\sum_{i=1}^{n} a + \frac{1}{n}\sum_{i=1}^{n}(bx_i) \\
\bar{z} &= \frac{1}{n}(na) + b(\frac{1}{n}\sum_{i=1}^{n} x_i) \\
\bar{z} &= a + b\bar{x}
\end{aligned}$$

   On the first line, we divide through by n, and on the second, we distribute the sum on the right-hand side. Note that the equation for the regression line is $\hat{y}_i = a + bx_i$ for all i, which is the same equation in disguise. By the same argument, the regression line passes through the point of averages.

2. **Show that adding a constant does not change the variance: if $z_i = x_i + d$ for all i, then $\text{var}(z) = \text{var}(x)$. (Hint: what is $z_i - \bar{z}$, expressed in terms of x_i, $\bar{x}$, and d?).**

   Using the definition of the variance, we can write $\text{var}(z)$ as:

$$\text{var}(z) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \tag{1}$$

   From problem 1 (with d in place of a and $b = 1$), we know that $\bar{z} = d + \bar{x}$. Substituting into equation (1) and using the definition of z_i, we have

$$\begin{aligned}
\text{var}(z) &= \frac{1}{n}\sum_{i=1}^{n}[(x_i + d) - (\bar{x} + d)]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \text{var}(x).
\end{aligned}$$

3. (a) **Let $z_i = cx_i + d$ for all i. Show that $\text{Var}(z) = c^2\text{Var}(x)$. (Hints: use 1 and 2 to rewrite $z_i - \bar{z}$. Then, use the definition of variance, substitute for $z_i - \bar{z}$, and multiply out terms. You will also need to use the alternate definition of variance presented in class).** We have

$$\begin{aligned}
\text{var}(z) &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[cx_i + d - (c\bar{x} + d)]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[c(x_i - \bar{x})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[c^2(x_i - \bar{x})^2] \\
&= c^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= c^2\text{var}(x)
\end{aligned}$$

On the first line, we plug in for $z_i - \bar{z}$ using the result in question 1.

(b) **Now, recall that the equation for the regression line is $\hat{y}_i = a + bx_i$. So what is Var$(\hat{y}_i)$?**

By the logic above, $var(\hat{y}) = b^2 var(x)$

4. **Let the variable z_i be x_i in standard units:**

$$z_i = \frac{(x_i - \bar{x})}{\text{SD}_x}, \tag{2}$$

where SD_x is the standard deviation of x. Show that

(a) **its average is 0;**

$$
\begin{aligned}
\bar{z} &\equiv \frac{1}{n}\sum_{i=1}^{n} z_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{SDx} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{SD_x} - \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{x}}{SD_x} \\
&= \frac{\bar{x}}{SD_x} - \frac{\bar{x}}{SD_x} \\
&= 0
\end{aligned}
$$

(b) **its variance and standard deviation are equal to 1.**

$$
\begin{aligned}
\text{var}(z) &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[\frac{(x_i - \bar{x})}{SD_x} - 0]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{SD_x^2} \\
&= \frac{1}{SD_x^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{SD_x^2}\text{var}(x) \\
&= \frac{\text{var}(x)}{\text{var}(x)} \\
&= 1.
\end{aligned}
$$

Remember the SD is the square root of the variance, thus also 1.

5. **Let Y_i be an outcome variable in an experiment, and let $D_i = 1$ if unit i is assigned to treatment and 0 otherwise. Suppose m out of N units are assigned to treatment. Thus, a fraction $\overline{D} = \frac{\sum_{i=1}^{N} D_i}{N} = \frac{m}{N}$ are assigned to treatment. The equation for the regression "line" is**

$$\hat{Y}_i = a + bD_i. \tag{3}$$

(a) **Find the regression fit for a and b. (That is, use the algebra of variance and covariances to show what a linear regression returns as the intercept a and the slope b). Show your work.**

Let Y_i be an outcome variable, and let $D_i = 1$ if unit i is assigned to treatment and 0 otherwise. Suppose m out of N units are assigned to treatment. Thus, a fraction $\overline{D} = \frac{\sum_{i=1}^{N} D_i}{N} = \frac{m}{N}$ are assigned to treatment. The equation for the regression "line" is

$$\hat{Y}_i = a + bD_i. \tag{4}$$

But what are a and b? (That is, how does regression fit the intercept a and "slope" b?). For the latter, we have

$$\begin{aligned} b &= \frac{\text{Cov}(Y, D)}{\text{Var}(D)} \\ &= \frac{\overline{YD} - (\overline{Y})(\overline{D})}{\overline{D}(1-\overline{D})} \\ &= \frac{\overline{YD} - (\overline{YD})(\overline{D}) - (\overline{Y})(\overline{D}) + (\overline{YD})(\overline{D})}{\overline{D}(1-\overline{D})} \\ &= \frac{(1-\overline{D})\overline{YD}}{\overline{D}(1-\overline{D})} - \frac{(\overline{Y} - \overline{YD})(\overline{D})}{\overline{D}(1-\overline{D})} \\ &= \frac{\sum_{i=1}^{N} Y_i D_i}{\sum_{i=1}^{N} D_i} - \frac{\sum_{i=1}^{N}(1-D_i)Y_i)}{\sum_{i=1}^{N}(1-D_i)}. \end{aligned} \tag{5}$$

(Here, in the first line we use the definition of the bivariate slope coefficient; in the second line we use the alternate definition of the covariance in the numerator and the definition of the variance of a 0-1 variable in the denominator (see FPP Ch 17, section 4, "A Short-Cut"; here, 1 is the big number and 0 is the small number). Then, in the third and fourth lines we add and subtract terms in the numerator, break the expression into two, and cancel terms. From the final line, the regression coefficient is just the difference of means: the average outcome in the treatment group minus the average outcome in the control group.

What about the intercept a? This is just

$$\begin{aligned} a &= \overline{Y} - b\overline{D} \\ &= \overline{Y} - \frac{\overline{YD} - (\overline{Y})(\overline{D})}{\overline{D}(1-\overline{D})}\overline{D} \\ &= \frac{(1-\overline{D})\overline{Y} - \overline{YD} + (\overline{Y})(\overline{D})}{(1-\overline{D})} \\ &= \frac{(\overline{Y} - \overline{YD})}{(1-\overline{D})} \\ &= \frac{\sum_{i=1}^{N}(1-D_i)Y_i)}{\sum_{i=1}^{N}(1-D_i)}. \end{aligned}$$

(Here, we substitute the definition of b in the second line; collect and cancel terms in the third and fourth lines; and multiply and divide by n to get the final line). Thus, a is the average outcome in the control group, and b is the difference of the treatment and control group averages.

Note that we can also derive this result using matrix notation. Again, m is the number of units in the treatment group. Order the observations such that the first m observations indexed $1...m$ correspond to units in the treatment group and that the $N-m$ observations indexed $m+1...N$ are in the control group. We denote Y_i^C as the outcome for units assigned to the control group and Y_i^T for units assigned to the treatment group; $\overline{Y^C}$ and $\overline{Y^T}$ are the averages in the control and treatment samples, respectively. The regression fit gives

$$\hat{\gamma} = (X'X)^{-1}X'Y \tag{6}$$

where γ is a vector of length 2, composed of a and b; X is an $N \times 2$ matrix, with a first column of ones (for the intercept) and D as the second column; and $(X'X)$ is a 2 by 2 matrix,

$$\begin{pmatrix} N & m \\ m & m \end{pmatrix} \tag{7}$$

where the first element is the result of the dot of the vector of ones with itself, and thus is N and the other three elements will simply be m, the number of units in the treatment group. The determinant of this matrix is $(N * m) - (m * m) = m(N - m)$ and thus the inverse is:

$$\frac{1}{m(N-m)} \begin{pmatrix} m & -m \\ -m & N \end{pmatrix} = \begin{pmatrix} \frac{1}{(N-m)} & \frac{-1}{(N-m)} \\ \frac{-1}{(N-m)} & \frac{N}{m(N-m)} \end{pmatrix}$$

Since X' is a $2 \times N$ matrix and Y is an $N \times 1$ vector, $X'Y$ will be a 2×1 vector where the first element is equal to $\sum_{i=1}^{N} Y_i$ and the second is equal to the dot product of D_i and Y_i and is thus simply $\sum_{i=1}^{m} Y_i$. We now have,

$$\hat{\gamma} = \begin{pmatrix} \frac{1}{(N-m)} & \frac{-1}{(N-m)} \\ \frac{-1}{(N-m)} & \frac{N}{m(N-m)} \end{pmatrix} * \begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{m} Y_i \end{pmatrix} \tag{8}$$

The result of the product is a 2×1 vector where the first element is the intercept a and the second is b. Then,

$$
\begin{aligned}
a &= \frac{1}{(N-m)} * \Big[\sum_{i=1}^{N} Y_i - \sum_{i=1}^{m} Y_i \Big] \\
&= \frac{1}{(N-m)} * \sum_{i=m+1}^{N} Y_i \\
&= \overline{Y^C}
\end{aligned} \tag{9}
$$

And

$$
\begin{aligned}
b &= \frac{-1}{(N-m)} * \sum_{i=1}^{N} Y_i + \frac{N}{m(N-m)} * \sum_{i=1}^{m} Y_i \\
&= \frac{1}{(N-m)} (\frac{N}{m} * \sum_{i=1}^{m} Y_i - \sum_{i=1}^{N} Y_i)
\end{aligned} \tag{10}
$$

Note that $\sum_{i=1}^{N} Y_i = \sum_{i=1}^{m} Y_i + \sum_{i=m+1}^{N} Y_i$. And so

$$
b = \frac{1}{(N-m)} \Big[\frac{m + (N-m)}{m} * \sum_{i=1}^{m} Y_i - \sum_{i=1}^{m} Y_i - \sum_{i=m+1}^{N} Y_i \Big] \tag{11}
$$

Again, rearranging terms we have

$$
\begin{aligned}
b &= \frac{1}{(N-m)} \Big[(\frac{m + (N-m)}{m} - 1) * \sum_{i=1}^{m} Y_i - \sum_{i=m+1}^{N} Y_i \Big] \\
&= \frac{1}{(N-m)} \Big[\frac{(N-m)}{m} * \sum_{i=1}^{m} Y_i - \sum_{i=m+1}^{N} Y_i \Big] \\
&= \frac{1}{(N-m)} \frac{(N-m)}{m} * \sum_{i=1}^{m} Y_i - \frac{1}{(N-m)} * \sum_{i=m+1}^{N} Y_i \\
&= \frac{1}{m} * \sum_{i=1}^{m} Y_i - \frac{1}{(N-m)} * \sum_{i=m+1}^{N} Y_i \\
&= \frac{1}{m} * \sum_{i=1}^{m} Y_i^T - \frac{1}{(N-m)} * \sum_{i=m+1}^{N} Y_i^C \\
&= \overline{Y^T} - \overline{Y^C},
\end{aligned} \tag{12}
$$

where in the last steps we use the order of observations.

(b) **Interpret b. What is another term for this quantity?**

As noted above, a is the average outcome in the control group, and b is the difference of the treatment and control group averages. Lesson: a bivariate regression with a 0-1 variable for treatment, the "slope" coefficient is the difference of means between

treatment and control groups. This is an estimator of the average treatment effect in the experiment.

However, note that we are nowhere assuming a regression model here (nor are we contrasting such a model with the Neyman model, under which we also use a difference of means to estimate the average causal effect). The equivalence in this simple case does not imply that regression modeling is equivalent to analysis under the potential outcomes model. We will discuss this point further soon ...

6. **Consider the 2x2 matrix,**

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

   (a) **What is the determinant of this matrix?**

   The determinant is $(1 * 4) - (2 * 2) = 0$

   (b) **Find the inverse of the matrix; or, if you cannot find the inverse, say why not.**

   Let's call the 2x2 matrix X. One way to calculate the inverse is by $\frac{\text{adj(X)}}{\text{det(X)}}$. Since the determinant of X is zero, the inverse cannot be calculated (a fraction with zero in the denominator is undefined).

   (c) **How many linearly independent rows does this matrix have? How many linearly independent columns?**

   First, notice that the second row is simply the product of 2 and the first row; the second column is the product of 2 and the first column. Thus, each row and column is a linear combination of the other row or column.

   Analogously, we can use row and column operations to reduce the matrix in order to find the number of independent rows and columns. If we subtract two times the first row from the second row, we are left with:

$$\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}.$$

   Then, if we subtract two times the first column from the second column, we are left with:

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

   There are no other non-zero vectors by which we could add, subtract, or multiply the rows or columns in the matrix to reduce this matrix any further. It is fully reduced. Thus, there is one linearly independent row and one linearly independent column.

   (d) **What is the rank of the matrix?**

   The rank is 1.

7. **Let X be an $n \times 2$ matrix, where the first column is all 1's and the second column is $(x_1, x_2, ..., x_{n-1}, x_n)'$. Let Y be an $n \times 1$ column vector consisting of $(y_1, y_2, ..., y_{n-1}, y_n)'$.**

(a) **What is the size of $X'X$? Of $X'Y$? What about $(X'X)^{-1}$ and $(X'X)^{-1}(X'Y)$? Can you multiply X and Y? Why or why not?**

i. $X'X$: X' is $2 \times n$ since X is $n \times 2$. Thus, $X'X$ is 2×2.

ii. $(X'X)^{-1}$ is also 2×2.

iii. $(X'X)^{-1}(X'Y)$: X' is $2 \times n$ and Y is $n \times 1$ so $X'Y$ is 2×1. Thus, $(X'X)^{-1}(X'Y)$ is 2×1.

iv. Here, X is $n \times 2$, and Y is $n \times 1$. You cannot multiply X and Y because the inner dimensions do not conform. (Thus, the matrices are said to be not conformable).

(b) **Find $X'X$. (That is, write out $X'X$, with typical elements given by n, $\sum_{i=1}^{n}(x_i)^2$, and so on).**

$$\begin{aligned} X'X &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_{i=1}^{n}(x_i) \\ \sum_{i=1}^{n}(x_i) & \sum_{i=1}^{n}(x_i)^2 \end{bmatrix} \\ &= n \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}. \end{aligned}$$

(c) **Find $(X'X)^{-1}$.**

The inverse can be found by $\frac{\text{adj(X)}}{\det(\text{X})}$. For a 2x2 matrix the adjoint is found by flipping the items on the downward-sloping diagonal (retaining signs), and then, for the off-diagonal elements, taking the opposite sign. Thus, if the matrix is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

the adjoint will be

$$\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

For the matrix $X'X$, then, the adjoint is:

$$\text{adj(X'X)} = n \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

To find the determinant of $(X'X)$, we multiply the diagonal elements of $X'X$ and subtract from them the product of the off-diagonal elements:

$$\begin{aligned}\det(\mathrm{X'X}) &= n^2(\overline{x^2}) - n^2(\bar{x}\bar{x}) \\ &= n^2[\overline{x^2} - \bar{x}^2] \\ &= n^2\mathrm{var}(x).\end{aligned}$$

(Note $X'X = n\begin{bmatrix}1 & \bar{x} \\ \bar{x} & \overline{x^2}\end{bmatrix}$, so you need to keep the n in finding the determinant. Thus, we have $n * n\overline{x^2} - n\bar{x} * n\bar{x} = n^2(\overline{x^2}) - n^2(\bar{x}\bar{x})$, which gives the first line above).
Plugging these two elements into the equation for the inverse we get:

$$\begin{aligned}(X'X)^{-1} = \frac{\mathrm{adj(X'X)}}{\det(\mathrm{X'X})} &= \frac{n}{n^2\mathrm{var(x)}}\begin{bmatrix}\overline{x^2} & -\bar{x} \\ -\bar{x} & 1\end{bmatrix} \\ &= \frac{1}{n\mathrm{var(x)}}\begin{bmatrix}\overline{x^2} & -\bar{x} \\ -\bar{x} & 1\end{bmatrix}\end{aligned}$$

(d) **Find** $(X'Y)$**.**

$$\begin{aligned}X'Y &= \begin{bmatrix}1 & 1 & \dots & 1 \\ x_1 & x_2 & \cdots & x_n\end{bmatrix}\begin{bmatrix}y_1 \\ y_2 \\ \vdots \\ y_3\end{bmatrix} \\ &= \begin{bmatrix}\sum_{i=1}^{n}(y_i) \\ \sum_{i=1}^{n}(y_i)(x_i)\end{bmatrix} \\ &= n\begin{bmatrix}\bar{y} \\ \overline{xy}\end{bmatrix}\end{aligned}$$

(e) **Find** $(X'X)^{-1}X'Y$**.**

$$
\begin{aligned}
(X'X)^{-1}X'Y &= \frac{1}{n\text{var(x)}}\begin{bmatrix}\overline{x^2} & -\bar{x}\\ -\bar{x} & 1\end{bmatrix} n\begin{bmatrix}\bar{y}\\ \overline{xy}\end{bmatrix}\\
&= \frac{1}{\text{var(x)}}\begin{bmatrix}(\bar{y})(\overline{x^2}) - (\bar{x})(\overline{xy})\\ -(\bar{y})(\bar{x}) + \overline{xy}\end{bmatrix}\\
&= \frac{1}{\text{var(x)}}\begin{bmatrix}(\bar{y})(\overline{x^2}) - (\bar{x})(\overline{xy}))\\ \text{Cov}(x,y)\end{bmatrix}\\
&= \begin{bmatrix}(\frac{(\bar{y})(\overline{x^2}) - (\bar{x})(\overline{xy})}{\text{Var}(x)})\\ \frac{\text{Cov}(x,y)}{\text{Var}(x)}\end{bmatrix}\\
&= \begin{bmatrix}\bar{y} - b\bar{x}\\ b\end{bmatrix},
\end{aligned}
$$

where $b = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$.

To see the final step of this derivation, note that the $(1,1)$ element of $(X'X)^{-1}X'Y$ on the penultimate line is

$$
\begin{aligned}
\frac{(\bar{y})(\overline{x^2}) - (\bar{x})(\overline{xy})}{\text{Var}(x)} &= \frac{(\bar{y})(\overline{x^2}) - \bar{y}\bar{x}^2 - (\bar{x})(\overline{xy}) + \bar{y}\bar{x}^2}{\text{Var}(x)}\\
&= \frac{\bar{y}[\overline{x^2} - \bar{x}^2] - \bar{x}[\overline{xy} - \bar{y}\bar{x}]}{\text{Var}(x)}\\
&= \frac{\bar{y}\text{Var}(x) - \text{Cov}(x,y)\bar{x}}{\text{Var}(x)}\\
&= \bar{y} - \frac{\text{Cov}(x,y)}{\text{Var}(x)}\bar{x}. \qquad (13)
\end{aligned}
$$

To get to the right-hand side of (13), we add and subtract $\bar{y}\bar{x}^2$ to the numerator (first line) then factor terms (second line). Finally, we use the alternate definitions of covariance and variance (third line) and cancel $\text{Var}(x)$ in the first term, which completes the proof.

(f) **Show that the** $(2,1)$ **element of** $(X'X)^{-1}X'Y = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$**. That is, when there is a constant and one variable in an** $n \times 2$ **design matrix, the matrix representation reduces to the usual formula for the slope coefficient of the bivariate regression line.**

We saw that the $(2,1)$ element of $(X'X)^{-1}X'Y = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$ in the fifth step of part (e).

(g) **Show that the** $(1,1)$ **element of** $(X'X)^{-1}X'Y = \bar{y} - b\bar{x}$**, where** $b = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$**.**

We saw that the $(1,1)$ element of $(X'X)^{-1}X'Y = \bar{y} - b\bar{x}$ in the fourth step of part (e).

8. **Hooke's law states that when a load (weight) is placed on a spring, the length is pro-**

portional to the weight. That is,

$$\text{length under load} = \text{length under no load} + \text{constant} \cdot \text{load}$$

Physicists test this prediction in the lab and obtain the results depicted in Table 1. Do not use *R*, other than as a calculator, to answer the following questions:

(a) **Find the regression equation for predicting length from load. (Show your work, including calculations of the relevant variances and covariances!).**

We write the regression equation for predicting length from load as $y = a + bx$ where y is the length, x is the load, b is the slope—the proportionate increase in length when weight is added—and a is the intercept, i.e., the length of the string under no load. Recall that the slope of the regression line is given by

$$b = \frac{\text{cov}(x, y)}{\text{var}(x)} = r\frac{SD_y}{SD_x},$$

where r is the correlation coefficient. Substituting for the values of $\text{cov}(x, y)$ and $\text{var}(x)$ found in Table 1, we have:

$$\begin{aligned} b &= \frac{0.67}{12.86} \\ &= 0.052 \end{aligned}$$

Now, we can calculate a by remembering that the regression line passes through the point of averages, so:

$$\bar{y} = a + b\bar{x}$$

and thus, rearranging terms,

$$\begin{aligned} a &= 287.3175 - (0.052) * 3.875 \\ &= 287.13. \end{aligned}$$

Thus, the equation for the regression line is:

$$\hat{y_i} = 287.13 + 0.052 * x_i \tag{14}$$

(b) **Use your result in (a) to write out a set of eight equations, where each equation is**

$$Y_i = a + bX_i + e_i$$

for $i = 1, ..., 8$. Here, Y_i is the length of the spring (in cm); X_i is the load or weight (in kg), and e_i is the difference between the actual value of Y_i and the value on the

Table 1: Question 8

x_i	y_i	$\bar{x}$	$\bar{y}$	$x_i - \bar{x}$	$y_i - \bar{y}$	var(x)	$(x_i - \bar{x})(y_i - \bar{y})$	cov(x, y)	$\hat{b}$	$\hat{a}$
0	287.12	3.88	287.32	-3.88	-.2	12.86	.78	.67	.05	287.13
1	287.18	3.88	287.32	-2.88	-.14	12.86	.4	.67	.05	287.13
1	287.16	3.88	287.32	-2.88	-.16	12.86	.46	.67	.05	287.13
3	287.25	3.88	287.32	-.88	-.07	12.86	.06	.67	.05	287.13
4	287.33	3.88	287.32	.12	.01	12.86	0	.67	.05	287.13
4	287.35	3.88	287.32	.12	.03	12.86	0	.67	.05	287.13
6	287.4	3.88	287.32	2.12	.08	12.86	.17	.67	.05	287.13
12	287.75	3.88	287.32	8.12	.43	12.86	3.49	.67	.05	287.13

regression line, given X_i. (In each of your equations, use actual numbers in place of Y_i, a, b, X_i and e_i). What units are the e_i measured in?

There are eight equations because there are eight observations. (Here, an observation is a trial in the physics lab, in which a weight is hung on the spring and the length of the spring is measured). The equations look like this:

$$\begin{aligned}
287.11 &= 287.13 + .052 * (0) + 0046765 \\
287.17 &= 287.13 + .052 * (1) + .0124989 \\
287.15 &= 287.13 + .052 * (1) - .0074901 \\
287.24 &= 287.13 + .052 * (3) - .021844 \\
287.32 &= 287.13 + .052 * (4) + .0059674 \\
287.34 &= 287.13 + .052 * (4) + .025987 \\
287.39 &= 287.13 + .052 * (6) - .0283755 \\
287.74 &= 287.13 + .052 * (12) + .0085799.
\end{aligned}$$

e_i is the residual for observation i. It can be found by subtracting the predicted value of length for observation i from the actual length in that trial. Thus, the unit of measure for the spring (the dependent variable) determines the unit of measurement for the residual. Since length is measured in centimeters, the residual will also be in centimeters.

(c) **Show that $\bar{e} = 0$; or, if it is not, say why not.**

$\bar{e}$ is the average of the residuals. Here, $e'\mathbf{1}$ is the sum of the residuals (each element of e is multiplied by 1 and all the products are added together). Dividing by 8 gives the average. Note that since e' is 1x8 and $\mathbf{1}$ is 8x1, the product of these will be 1x1 – a scalar.

Here, the average of the residuals is

$$\begin{aligned}\bar{e} &\doteq \frac{1}{8}[0046765 + .0124989 - .0074901 - .021844 + .0059674 + .025987 - .0283755 + .0085799] \\ &= 0.0000000125.\end{aligned}$$

The average is not exactly zero due to rounding error.

(d) **Let X be the 8×2 matrix with typical element $[1 \;\; X_i]$, and e be the 8×1 vector of residuals. Show that $e'X = 0$, a $[1 \times 2]$ row vector; or if it is not, say why not.**

e' is 1x8 and X is 8x2, thus, since the inner dimensions are the same, we can carry out matrix multiplication. The result is a 1x2 matrix.

Using the values in 15 we can find $e'X$:

$$e'X = \begin{bmatrix}.0046765 & .0124989 & -.0074901 & -.021844 & .0059674 & .025987 & -.0283755 & .0085799\end{bmatrix} * \begin{bmatrix}1 & 0\\ 1 & 1\\ 1 & 1\\ 1 & 3\\ 1 & 4\\ 1 & 4\\ 1 & 6\\ 1 & 12\end{bmatrix}$$

$$= \begin{bmatrix}.0046765(1) + .0124989(1) - .0074901(1) - .021844(1) + .0059674(1) + .025987(1) - .0283755(1) + .0085799(1)\\ .0046765(0) + .0124989(1) - .0074901(1) - .021844(3) + .0059674(4) + .025987(4) - .0283755(6) + .0085799(12)\end{bmatrix}$$

$$\doteq \begin{bmatrix}0.0000001\\ 0.0000002\end{bmatrix}.$$

Here, the average of the residuals is not zero, though it is close; so, the (1,1) element of $e'X$ is not quite zero. The same goes for the sum of the products of the residuals and the values of the weights in the (2,1) element of $e'X$. Thus $e'X \doteq 0$. This discrepancy arises because of rounding error.

(e) **The two measured lengths for a load of 1 kg differ. Why might that be?**

Variation in lengths due to a not-perfectly-controlled environment is one possibility; another is measurement error. See Chapter 12, Section 2 in FPP.

(f) **Use the equation to predict length at the following loads: 2 kg, 3 kg, 5 kg, 105 kg.**

$$\begin{aligned}\hat{y}_i|x_i = 2 &= 287.13 + 0.052 * 2\\ &= 287.23\end{aligned}$$

$$\begin{aligned}\hat{y}_i|x_i = 3 &= 287.13 + 0.052 * 3\\ &= 287.29\end{aligned}$$

$$\begin{aligned} \hat{y}_i | x_i = 5 &= 287.13 + 0.052 * 5 \\ &= 287.39 \end{aligned}$$

For 105 kg, the regression equation says the length of the spring should be 287.13 + $0.052 * 105 = 292.59$ centimeters. However, the regression equation should not be used to predict the result of hanging a weight of 105 kg on the spring; this is a serious extrapolation from the data, and the results for such a heavy weight might differ substantially from the predicted length (for instance, the spring may break!).

(g) **For a load of 3 kg, the answer to (f) is different from the number in the table. Under the load of 3 kg, would you use the number in the table, or the regression equation? Explain carefully. (You may want to refer to Chapter 12 of FPP).**

This is a decent setting in which to use a regression equation to predict the result of interventions, for the range of weights we are considering. Why? There's good reason to believe Hooke's Law is linear, i.e., there is a constant of proportionality by which load affects the length of the spring. Moreover, here the interventions have actually been conducted: researchers hang different weights on the spring and record what happens. The line provides a good fit to the data as well (the r^2 is close to 1). In such a setting, you might want to use the regression equation rather than the number in the table to predict what would happen if you were to hang a load of 3 kg on the spring. The regression equation combines information from eight data points, reducing the impact of measurement error.

For more discussion of these points, see FPP Ch. 12 and Freedman and Lane (1981), section 11; section 30 presents the regression model in connection with Hooke's Law.

(h) **Do you think the regression equation provides a good basis for predicting the result of a hypothetical intervention, in which you hang a load on the spring? Why or why not?**

Not all springs are like others (tensile strength? pneumatic activity?), so we may not want to use the regression line in this experiment to predict the length from load on any spring: the slope and intercept may vary across springs. And we might not want to use the equation to make predictions for spring length under weights larger or smaller than those with which we have experimented. Yet, the relevant variables and their functional forms are known in Hooke's law, and repeated testing has suggested the relationship between load and length is quite linear. Thus, the regression equation probably provides a good basis for predicting the result of hanging a new load on this spring, within the range of weights that were tested.

9. **Hibbs (1978) is interested in changes in industrial strike activity in advanced capitalist democracies in the twentieth century. His hypothesis is that changes in political economy after the Second World War—in particular, the rise of Left-Labor governments and the welfare state—shifts the locus of conflict away from private firms. Table 2 gives the approximate numerical values for the change in strike activity and the change**

Load (kg)	Length (cm)
0	287.12
1	287.18
1	287.16
3	287.25
4	287.33
4	287.35
6	287.40
12	287.75

Table 2: A test of Hooke's Law.

in Left-Labor cabinet representation, post war vs. inter-war means. (The values are approximate because they are eyeballed from a scatterplot in Hibbs 1978).

Country	Change in Left-Labor cabinet representation (%)	Change in average strike volume
Norway	76	-1,980
Sweden	69	-1,700
U.K.	30	-1,000
Denmark	31	-650
Netherlands	19	-650
Belgium	10	-180
Italy	11	10
France	9	180
Finland	11	210
Canada	0	175
U.S.	1	190

Table 3: Change in average strike volume and average Socialist-Labour and Communist percentage of cabinet representation, inter-war to post-war period (Hibbs 1978)

(a) **Find the correlation between change in average strike volume and percentage change in Left-Labor cabinet representation, using the data in Table 3. What is the average and standard deviation of each variable? (Do this "by hand," not using R other than as a calculator).**

The mean change in Left-Labor cabinet representation is 24.3, with an SD of 24.7; the mean change in average strike volume is -490.5, with an SD of 749.9. The correlation between these variables is about -0.97. (Hibbs reports the correlation coefficient at -0.96, so the eyeballed data in the problem set table are slightly off).

(b) **Find the regression equation for predicting change in average strike volume from**

percentage change in Left-Labor cabinet representation. (Do not use R; show your work, including calculations of the relevant variances and covariances!).

Again,

$$b = \frac{\text{cov}(x,y)}{\text{var}(x)} = r\frac{SD_y}{SD_x}.$$

Thus,

$$\begin{aligned} b &\doteq -.97\frac{749.9}{24.7} \\ &\doteq -29.4. \end{aligned}$$

The intercept is

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ &\doteq -490.5 - (-29.5) * 24.3 \\ &\doteq -226.4. \end{aligned}$$

(The answers for the slope and intercept were given in the slides for lecture 4, so be sure to show your work).

(c) **Find the R^2 of the regression. Note that these data are based on differences of averages (the mean change in strike volume and the mean change in Left-Labor cabinet representation). What does this do to the R^2, for example relative to a situation in which we measure strikes and cabinet representation at yearly intervals?**

In a bivariate regression, the R^2 is just the square of the correlation coefficient: $-0.97^2 = 0.94$.

Measuring averages inflates the R^2, relative to calculating the squared correlation coefficient between strikes and cabinet representation in each year; so does taking the differences. First, within each country, there will be year-to-year variation in the volume of strikes, as well as in the cabinet representation of the Left (the latter may be stickier, though, as cabinets often last longer than a year . . .). When we take averages, we eliminate a lot of the spread, and thus generally produce tighter clustering and a stronger correlation. When r goes up, so does R^2. If this is unfamiliar, read section 4 in Chapter 9 of Freedman, Pisani and Purves (2007), "Ecological Correlations."

For similar reasons, taking the difference in the post-war and pre-war means produces a tighter correlation: it eliminates a lot of the spread. Suppose that in each country, there is a "baseline" level of strikes and Left-Labor cabinet representation, which is higher in some countries and lower in others. When we look at the change in each country (post-war minus pre-war), we remove the variation due to the differences in baselines across countries. This reduces spread and will thus drive up the correlation coefficient.

Here's another way to think about why the correlation r and thus the R^2 increases when

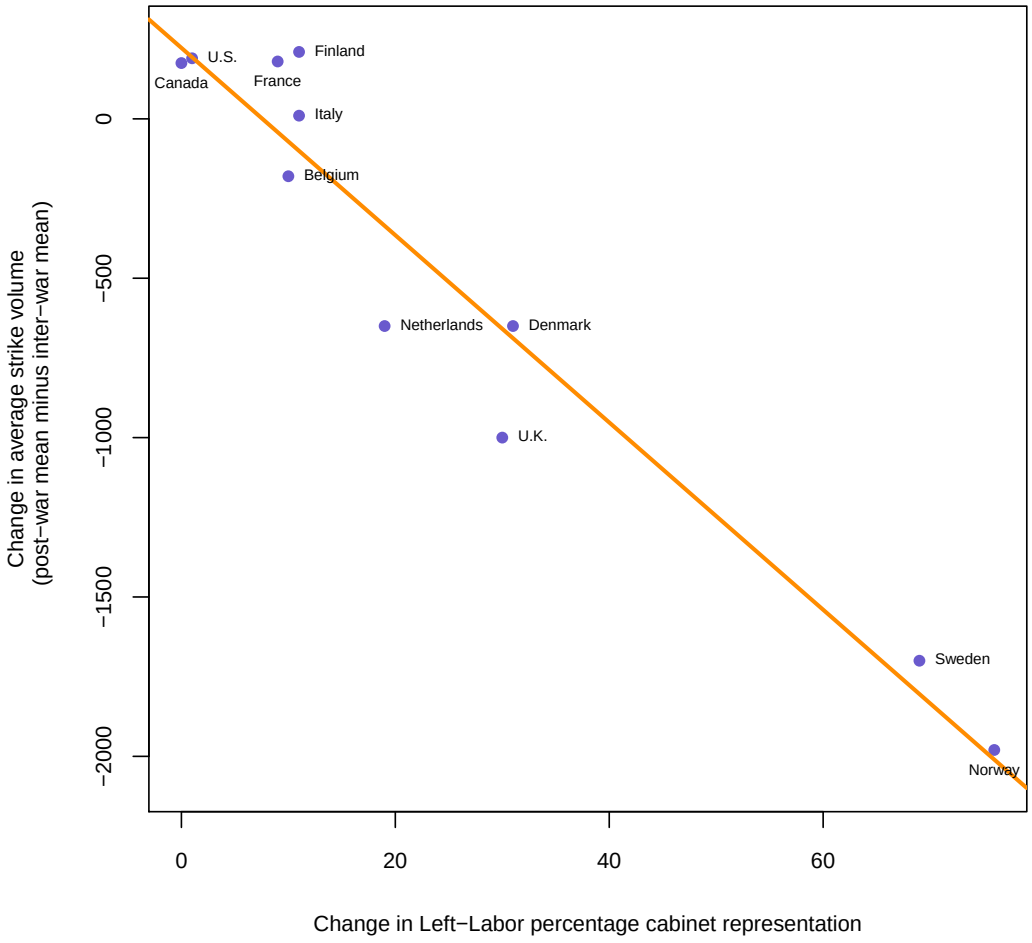
we take averages or differences. By definition, the correlation between x and y is

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{\mathrm{SD}_x \mathrm{SD}_y} \tag{15}$$

In general, taking averages is going to reduce the SDs in the denominator and thus drive up the correlation. (The sum in the numerator will have fewer terms in it after averaging observations—but the average product may or may not shift).

(d) **Now use R to create a plot showing Change in Left-Labor percentage cabinet representation on the horizontal axis and change in average strike volume, post-war mean minus inter-war mean, on the vertical axis. Label the axes, and label each point in the scatter plot with the name of the country. Superimpose the regression line you found in (b) on the plot. Create a caption describing the figure. Turn in your figure and your code.**

```
par(mar=c(5.2,4.9,4.2,2.2))
plot(hibbs$left_labor_change, hibbs$strike_vol_change, pch=16,
     col="slateblue",
     xlab="Change in Left-Labor percentage cabinet representation",
     ylab="Change in average strike volume
     (post-war mean minus inter-war mean)",
     cex.lab=.8, cex.axis=.8,
     ylim=c(min(hibbs$strike_vol_change) - 100,
            max(hibbs$strike_vol_change) + 50))
abline(lm(hibbs$strike_vol_change ~ hibbs$left_labor_change),
       col="darkorange", lwd=2.5)
text(hibbs$left_labor_change[-c(1,8,10)],
     hibbs$strike_vol_change[-c(1,8,10)],
     labels=hibbs$country[-c(1,8,10)], cex=.6, pos=4)
text(hibbs$left_labor_change[c(1,8,10)],
     hibbs$strike_vol_change[c(1,8,10)],
     labels=hibbs$country[c(1,8,10)], cex=.6, pos=1)
```



(e) **In general, do you think the regression equation provides a good basis for predicting the result of a hypothetical intervention to change Left-Labor cabinet representation? Why or why not? What are some similarities and differences between this setting and your discussion of Hooke's Law in the previous question?**

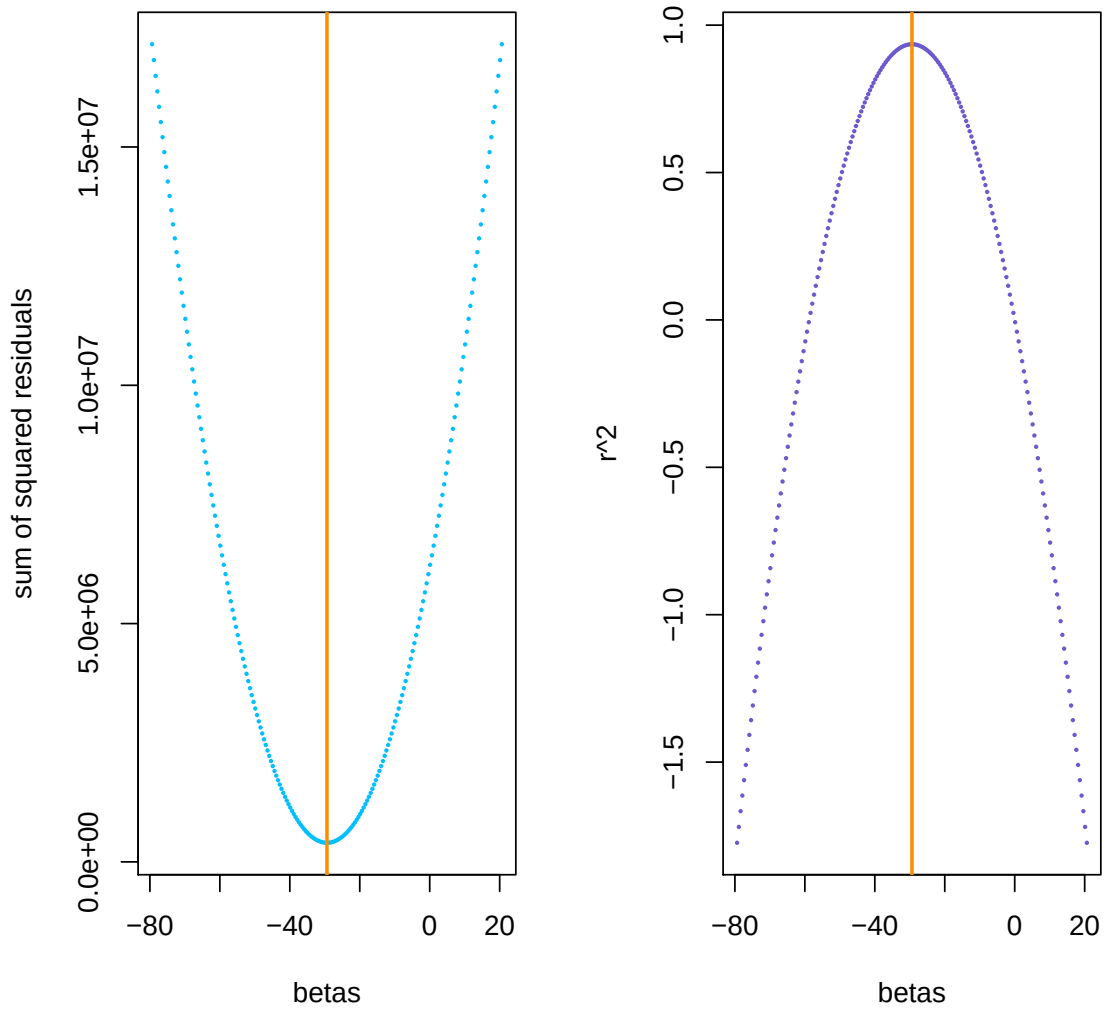
The equation is probably not very good at predicting the result of this type of intervention, because no intervention took place: these are observational data. To use the regression equation, we'd have to imagine that we could manipulate Left-Labor cabinet representation, without changing other factors that might influence strike volumes—and that the coefficient b represents the invariant effect of such a manipulation, just as in Hooke's Law. Here, this is a big thought experiment. Without the data from actual interventions, it would be hard to validate the claim that b represents the effect of an intervention.

10. **Write an R function and code to perform the following simulation: for $d = b - 50$, where**

b is the value of the slope you found in part (b), calculate for each country the residual $f_i = Y_i - a - dX_i$. (For a, use the value that ensures the line $a + dX_i$ goes through the point of averages). Calculate the sum of squared residuals across all countries i and the R^2. Now increment d by 0.5 and repeat, again calculating the sum of squared residuals as well as the R^2. Continue incrementing d and repeating until you reach $d = b + 50$. Plot the sum of squared residuals and the R^2, both as a function of d (include two separate plots). What do these show?

```
hatbeta <- with(hibbs, lm(strike_vol_change~left_labor_change))$coefficients[2]
hatbeta

## left_labor_change
##         -29.37829

# we will get a vector with the sequence from hatbeta-50 to hatbeta+50 in 0.5
# intervals
betas <- seq(hatbeta-50, hatbeta+50, by=0.5)

sum2res_rsq <- function(hatb, x, y){
  # This function takes a value for beta, calculates the intercept that ensures
  # the regression line goes through the point of averages and stores the
  # corresponding r^2 and sum of squared residuals.

  # 1. get the relevant intercept
  hat_a <- mean(y) - hatb * mean(x)

  # 2. calculate hat_y
  hat_y <- hat_a + hatb * x

  # 3. get sum of squared residuals
  sum2res <- sum((y - hat_y)^2)

  # 4. get r^2
  rsq <- 1 - (sum2res/sum((y-mean(y))^2))

  # output
  res <- c(sum2res, rsq)
  return(res)
}

#here we use sapply (instead of apply) because betas is a vector (not a matrix)
res <- sapply(betas, FUN=sum2res_rsq,
              x=hibbs$left_labor_change,
```

```
              y=hibbs$strike_vol_change)


par(mfrow=c(1,2)) # to put one plot next to the other
plot(betas, res[1,], pch=16, cex=.35, col="deepskyblue",
     ylab="sum of squared residuals")
abline(v=hatbeta, lwd=2, col="darkorange")
plot(betas, res[2,], pch=16, cex=.35, col="slateblue",
     ylab="r^2")
abline(v=hatbeta, lwd=2, col="darkorange")
```



See the orange lines for the OLS coefficient. $\hat{\beta}_{OLS}$ minimizes the sum of squared residuals

and maximizes the r^2.

11. **Write a function that takes a vector for Y and a matrix with a number of independent variables and calculates multiple regression coefficients. Use the *family.rda* data in the *problem_3_data.Rdata* file to show your function works (regress weight on height and bmi).**

```
ols <- function(Y, X){

  X_int <- cbind(1, X)

  betahat <- solve(t(X_int)%*%X_int) %*% (t(X_int)%*%Y)

  rownames(betahat) <- c("intercept", paste0("beta_", 1:ncol(X)))

  return(round(betahat, 4))
}

ols(family$weight, cbind(family$height, family$bmi))

##                [,1]
## intercept -310.3469
## beta_1       4.6755
## beta_2       6.3086

# check if the function works (comparing with coefficients using lm())
stopifnot(ols(family$weight, cbind(family$height, family$bmi))==
        round(lm(family$weight~family$height + family$bmi)$coefficients,4))
```