PLSC 502: "Statistical Methods for Political Research"

Linear Regression, I
November 10, 2016

Regression

We'll spend a few days talking about *bivariate regression* – that is, regressing a single variable Y on a single variable X. We'll use our standard notation, whereby we have data on a sample of N observations, indexed by i: $i = \{1, 2, ...N\}$.

It's common to think of a random variable Y as having *systematic* and *stochastic* parts:

$$Y_i = \mu + u_i \tag{1}$$

This is a general description of a random variable; from this, we can "get to" just about any regression-type model you can name. If we think of a linear relationship between some covariate X and the response variable Y, the standard approach is to treat X as influencing the "systematic part" of Y in a linear way:

$$\mu_i = \beta_0 + \beta_1 X_i$$

so that we get

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2}$$

From this, our goal is to come up with two sets of things:

1. *Point estimates* of β_0 and β_1 (which we'll refer to as $\hat{\beta}_0$ and $\hat{\beta}_1$), and
2. Estimates of the variability of our point estimates; that is, *standard errors* for $\hat{\beta}_0$ and $\hat{\beta}_1$.

The former are our direct indicators of the relationship between X and Y; the latter tell us how precise our estimates are, and also allow us to engage in inference. We'll focus on the point estimates today, and on standard errors and inference (and other topics) next time.

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

Our goal, then is to come up with estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$. Consider what would happen if we had such estimates: we could then "plug them in" to Eq. (2) to get:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3}$$

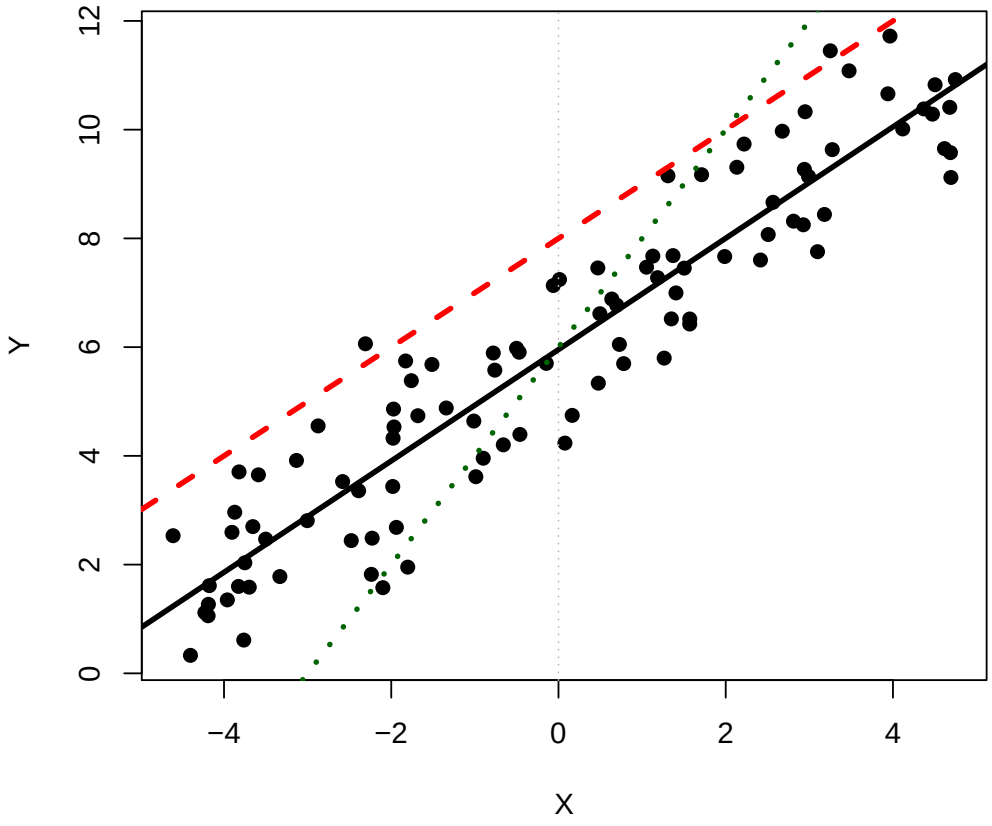
That is, we can get a *predicted* (or *expected*) value of Y for each of the N observations in the data, which is a function of that observation's value of X_i and the two estimated parameters

$\hat{\beta}_0$ and $\hat{\beta}_1$. Moreover, the difference between these two values – observed Y and predicted $\hat{Y}$ – constitutes our estimate of the stochastic component of the model:

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \end{aligned} \tag{4}$$

How do we go about estimating $\hat{\beta}_0$ and $\hat{\beta}_1$? Obviously, there are a lot of possibilities, and a lot of criteria we might consider in choosing among them. Consider the following scatterplot of some made-up data ($N = 100$), along with some lines representing the possible relationship between the two variables:

Figure 1: Scatterplot: X and Y (with regression lines)



If we look at these data, we'd probably all agree that:

- The solid line is the best "fit" to the data,

- The long-dashed line is systematically overpredicting Y (that is, the *intercept* (β_0) is incorrect), and
- The short-dashed line underpredicts Y at low values of X, and overpredicts Y at high values of X – that is, it gets the *slope* of the line (β_1) wrong.

This gives us some intuition – that, all else equal, we want an estimator that will somehow minimize the distance (in some sense) between the actual values of Y_i and the predicted/expected (based on the value of X_i) values $\hat{Y}_i$. More specifically, we'd prefer an estimator that had a couple properties:

1. *Unbiasedness* – that is, one for which $E(\hat{\beta}) = \beta$. Put differently, we want an estimator that (on average) "gets it right" vis-á-vis β.
2. *Efficiency* – one which has the smallest variance. Intuitively, we would prefer an estimator that – in addition to "getting it right" on average – was also never too far off from "right" in any given sample.

Some Estimators

Since the "distance" in question is in fact the stochastic part of Y (or the "error term" u_i), we can think of estimators in terms of what function of $\hat{u}_i$ they minimize. Three possibilities come to mind:

1. **Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ so as to minimize $\sum_{i=1}^{N} \hat{u}_i$.**
   - This seems like the most intuitive at first, but it has a serious problem: Large positive values of $\hat{u}_i$ can offset large negative values of $\hat{u}_i$.
   - This is the problem with the short-dashed line in Figure 1: the large positive residuals at low values of X are offset by large negative values at high levels of X. In fact, the sums of residuals for the solid and short-dashed lines in Figure 1 are exactly the same.
   - In fact, as Fox notes, *any* line that passes through the point $(\bar{X}, \bar{Y})$ will have the same sum of residuals $\sum_{i=1}^{N} \hat{u}_i$ (that is, zero) as any other, *no matter what its slope or intercept.*
2. **Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ so as to minimize $\sum_{i=1}^{N} |\hat{u}_i|$.**
   - This is usually known as the "minimum absolute distance" (MAD) estimator, because it minimizes the "distance" between Y_i and $\hat{Y}_i$.
   - It is a perfectly reasonable choice, but not the one that most people use (we'll see why in a minute). However,
   - It does have the nice property that it is more resistant to influence by outliers than are some other estimators, which can be valuable.
   - We'll come back to the MAD estimator a bit later in the semester.

3. **Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ so as to minimize $\sum_{i=1}^{N} \hat{u}_i^2$.**

    - This is (understandably) known as the *least-squares* estimator.
    - It is the basis for just about everything we'll do in PLSC 503.
    - As we'll see, under some relatively general conditions, it has some very useful properties.

We'll work through a couple examples of least-squares regression, beginning with one that is – literally – as simple as it can possibly be.

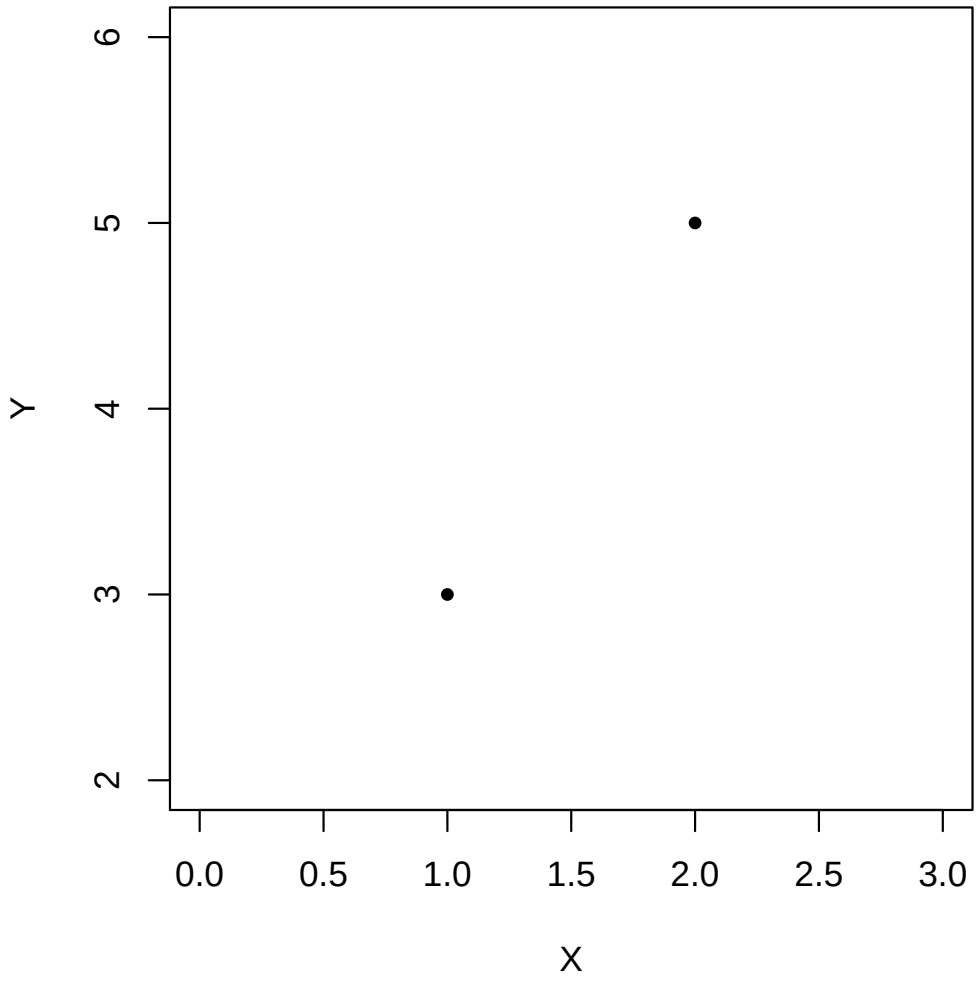
The Simplest Regression In Human History

As a very preliminary example, consider some data on two variables X and Y where we have only two data points: (1,3) and (2,5):

```
> d
  x y
1 1 3
2 2 5
```

Here's what the scatterplot looks like:

Scatterplot of the Simplest Regression in Human History



Now suppose we want to estimate the linear relationship between Y and X using a least-squares estimator. We start with our linear equation:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

and remember from (3) that:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

So, for the first observation,

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 (1),$$

and for the second

$$\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1(2)$$

Note also that

$$\begin{aligned}
\hat{u}_i &= Y_i - \hat{Y}_i \\
&= 3 - \hat{\beta}_0 + \hat{\beta}_1(1) \text{ for the first observation, and} \\
&= 5 - \hat{\beta}_0 + \hat{\beta}_1(2) \text{ for the second observation.}
\end{aligned}$$

From this, we can see that the sum of squared residuals $\hat{S} = \sum_{i=1}^{N} \hat{u}_i^2$ is

$$\begin{aligned}
\hat{S} &= u_1^2 + u_1^2 \\
&= [3 - \hat{\beta}_0 + \hat{\beta}_1(1)]^2 + [5 - \hat{\beta}_0 + \hat{\beta}_1(2)]^2 \\
&= (9 + \hat{\beta}_0^2 + \hat{\beta}_1^2 - 6\hat{\beta}_0 - 6\hat{\beta}_1 + 2\hat{\beta}_0\hat{\beta}_1) + (25 + \hat{\beta}_0^2 + 4\hat{\beta}_1^2 - 10\hat{\beta}_0 - 20\hat{\beta}_1 + 4\hat{\beta}_0\hat{\beta}_1) \\
&= 2\hat{\beta}_0^2 + 5\hat{\beta}_1^2 - 16\hat{\beta}_0 - 26\hat{\beta}_1 + 6\hat{\beta}_0\hat{\beta}_1 + 34 \qquad (5)
\end{aligned}$$

Now that we know what we're after, the goal is to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize this function in Eq. (5). To do that, we do what any right-minded person does who wants to find the extremum of some function: take the partial derivatives of (5) with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, set them equal to zero, and solve:

$$\begin{aligned}
\frac{\partial \hat{S}}{\partial \hat{\beta}_0} &= 4\hat{\beta}_0 + 6\hat{\beta}_1 - 16 \\
\frac{\partial \hat{S}}{\partial \hat{\beta}_1} &= 6\hat{\beta}_0 + 10\hat{\beta}_1 - 26
\end{aligned}$$

So,

$$\begin{aligned}
4\hat{\beta}_0 + 6\hat{\beta}_1 - 16 = 0 \quad &\Rightarrow \quad 2\hat{\beta}_0 = -3\hat{\beta}_1 + 8 \\
&\Rightarrow \quad \hat{\beta}_0 = -3/2\hat{\beta}_1 + 4
\end{aligned}$$

$$\begin{aligned}
6\hat{\beta}_0 + 10\hat{\beta}_1 - 26 = 0 \quad &\Rightarrow \quad 5\hat{\beta}_1 - 3(-3/2\hat{\beta}_1 + 4) - 13 = 0 \\
&\Rightarrow \quad 5\hat{\beta}_1 - 9/2\hat{\beta}_1 + 12 - 13 = 0 \\
&\Rightarrow \quad \frac{1}{2}\hat{\beta}_1 - 1 = 0 \\
&\Rightarrow \quad \boldsymbol{\hat{\beta}_1 = 2}
\end{aligned}$$

$$\begin{aligned} 4\hat{\beta}_0 + 6(2) - 16 = 0 &\Rightarrow 4\hat{\beta}_0 = 4 \\ &\Rightarrow \boldsymbol{\hat{\beta}_0 = 1} \end{aligned}$$

So the least-squares estimate of the parameters is $Y_i = 1 + 2X_i + u_i$. Note that if we'd just calculated the "slope" of a line that connected the two points, we'd have gotten the same thing:

$$\begin{aligned} \hat{\beta}_1 &= (5-3)/(2-1) \\ &= 2, \text{ and} \end{aligned}$$

$$\begin{aligned} \hat{\beta}_0 &= -2(2) + 5 \\ &= 1 \end{aligned}$$

Least Squares with > 2 Observations

That's a nice exercise and all, but not very useful; we generally want to calculate regression estimates for datasets with more than two observations. In that more general case with N observations, the intuition is identical: to minimize $\hat{S} = \sum_{i=1}^{N} \hat{u}_i^2$, we can start out by noting that

$$\begin{aligned} \hat{S} &= \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ &= \sum_{i=1}^{N} (Y_i^2 - 2Y_i\hat{\beta}_0 - 2Y_i\hat{\beta}_1 X_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 X_i + \hat{\beta}_1^2 X_i^2) \end{aligned} \tag{6}$$

This somewhat complicated thing is our least-squares function in the general (bivariate) case. To obtain our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$, we again partially differentiate this equation w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\begin{aligned} \frac{\partial \hat{S}}{\partial \hat{\beta}_0} &= \sum_{i=1}^{N} (-2Y_i + 2\hat{\beta}_0 + 2\hat{\beta}_1 X_i) \\ &= -2 \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ &= -2 \sum_{i=1}^{N} \hat{u}_i \end{aligned}$$

and

$$\begin{aligned}\frac{\partial \hat{S}}{\partial \hat{\beta}_1} &= \sum_{i=1}^{N}(-2Y_iX_i + 2\hat{\beta}_0X_i + 2\hat{\beta}_1X_i^2) \\ &= -2\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1X_i)X_i \\ &= -2\sum_{i=1}^{N}\hat{u}_iX_i\end{aligned}$$

Setting these two equations equal to zero and doing a little algebra yields:

$$\sum_{i=1}^{N}Y_i = N\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{N}X_i \tag{7}$$

and

$$\sum_{i=1}^{N}Y_iX_i = \hat{\beta}_0\sum_{i=1}^{N}X_i + \hat{\beta}_1\sum_{i=1}^{N}X_i^2 \tag{8}$$

These are what is known as the *OLS normal equations*; solving them simultaneously for $\hat{\beta}_0$ and $\hat{\beta}_1$ yields:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{9}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}. \tag{10}$$

Some Intuition

Equation (9) is of particular interest here, since it is the general least-squares estimator for the slope of a bivariate relationship. It can be though of as essentially

$$\frac{\text{Covariance of } X \text{ and } Y}{\text{Variance of } X}$$

The fact that we don't have Y in the denominator means that Y isn't "normed out" of the estimate of $\hat{\beta}_1$. In fact, $\hat{\beta}_1$ is expressed in terms (units) of Y – a common interpretation of $\hat{\beta}_1$ is that it is "the effect on Y of a one-unit change in X."

$\hat{\beta}_0$, meanwhile, is the "Y-intercept" of the model; that is, the place where the regression line crosses the Y-axis. This means that it can also be interpreted as the expected value of Y when $X = 0$. Note that sometimes this has some substantive meaning, while other times it does not.

An Example: The U.S. Supreme Court

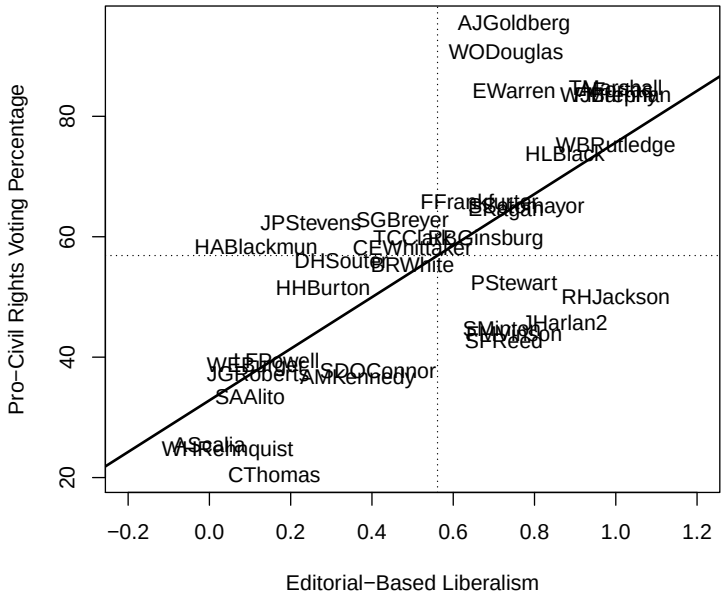
I thought we'd get away from the Africa data for a while, and focus on something that is closer to my own interests (it's my prerogative, I suppose...). Here are some data on some Supreme Court justices[1] ($N = 38$) on two variables:

- `ideology_score` is a rating of U.S. Supreme Court justices in terms of how liberal they are.
    - It ranges from 0 (very conservative) to 1 (very liberal), and, importantly
    - It was coded from editorials written in major newspapers at the time of their nomination to the Court; this means it is logically and temporally prior to (and, at least potentially, independent of) any of their votes on the Court.
- `civlibs` is a summary measure of the percentage of cases in which that justice voted to support a liberal (that is, pro-civil rights) outcome in cases decided by the Court.

FYI, here's what the data looks like when we plot them:

[1] Note that these data cover the Warren, Burger, Rehnquist, and Roberts Courts.

Figure 2: Scatterplot of Justice Liberalism and Liberal Voting in Civil Rights Cases



The horizontal and vertical lines represent the values of $\overline{\texttt{civlibs}}$ and $\overline{\texttt{score}}$, respectively. We might imagine using these data to investigate the relationship between justices' ideologies and their voting records.

We can estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ "by hand," as an illustration:

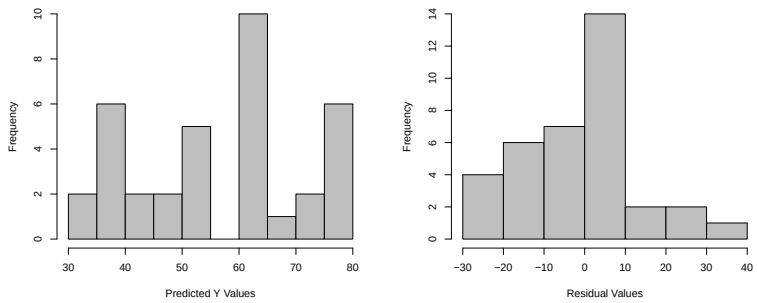
```
> Beta1 <- with(SCOTUS, (sum((ideology_score - mean(ideology_score)) *
+                            (civlibs - mean(civlibs))) /
+                          sum((ideology_score - mean(ideology_score))^2)))
> Beta1
[1] 43

> Beta0 <- with(SCOTUS, mean(civlibs) - (Beta1 * mean(ideology_score)))
> Beta0
[1] 33
```

From these, we can generate estimates of the residuals ($\hat{u}_i$s), from which we can learn other things. Here are the residuals along with the predicted and actual Ys:

```
> SCOTUS$Yhats <- with(SCOTUS, Beta0 + Beta1*ideology_score)
> SCOTUS$Uhats <- with(SCOTUS, civlibs - Yhats)
> describe(SCOTUS$civlibs)
   vars  n mean sd median trimmed mad min max range skew kurtosis  se
```

```
X1    1 36   57 20     57      57  23  21  95    75 0.14    -0.94 3.3
> describe(SCOTUS$Yhats)
   vars  n mean sd median trimmed mad min max range  skew kurtosis  se
X1    1 36   57 14     62      57  18  33  76    43 -0.21     -1.4 2.3
> describe(SCOTUS$Uhats)
   vars  n mean sd median trimmed mad min max range skew kurtosis  se
X1    1 36    0 14   0.44   -0.21  13 -26  30    56 0.02    -0.58 2.3
```



If we fit a least-squares regression to the data (using `lm`), we get the following estimates:

```
> with(SCOTUS, summary(lm(civlibs~ideology_score)))

Call:
lm(formula = civlibs ~ ideology_score)

Residuals:
    Min      1Q  Median      3Q     Max
-25.662 -10.715   0.437   8.139  30.374

Coefficients:
               Estimate Std. Error t value    Pr(>|t|)
(Intercept)       32.83       4.78    6.86 0.000000067 ***
ideology_score    42.84       7.40    5.79 0.000001628 ***
---
Signif. codes:  0 *** 0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 14 on 34 degrees of freedom
Multiple R-squared:  0.496,Adjusted R-squared:  0.481
F-statistic: 33.5 on 1 and 34 DF,  p-value: 0.00000163
```

This tells us several things about the (linear) relationship between these two variables:

- We would expect a justice who has a zero value on the `ideol_score` variable (that is, an extreme conservative) to vote in favor of a liberal position in civil rights and liberties cases about 33 percent of the time.

- Each one-unit increase in `score` is associated with a corresponding increase of 42.8 in the expectation of the `civlibs` variable.
    - Note that because `score` ranges from 0 to 1, a one-unit increase is quite a lot; however,
    - Because the relationship is linear, it rescales perfectly.
    - That means that (e.g.) an 0.2 unit increase in `score` is associated with a $(42.8 \times 0.2) = 8.6$-unit change in the expectation of `civlibs`.
- The "total sum of squares" (TSS) is equal to 13651; this is the same as $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. Of this,
    - The "residual sum of squares" (RSS; that is, $\hat{S} = \sum_{i=1}^{N} \hat{u}_i^2$) is 6877; that is the amount of "residual" / "error" / stochastic variability left in `civrts` after the effect of `score` is accounted for.
    - The "model sum of squares" is 6774; this reflects the amount of (mean-centered) variation "explained" by the `score` variable, and is equal to TSS - RSS.
- Similarly, if we divide the RSS by $N-k$ (where k is the number of regressors, including the constant – here, two) we get the variance of the residuals:

$$\hat{s}^2 = \frac{\text{RSS}}{N-k} = \frac{\sum_{i=1}^{N} \hat{u}^2}{N-2} \tag{11}$$

  We can think of this as an "average" value of the squared residuals. In the model above this is equal to 196.5. More valuable still is...

- ... the *standard error of the regression*; this can be thought of as the standard deviation of the residuals, and is equal to

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{N} \hat{u}^2}{N-2}} \tag{12}$$

  Think of this as the "average" residual; in our Supreme Court model, this is equal to $\sqrt{196.5} \approx 14$. This tells us that, on average, our model mis-predicts the `civlibs` variable by about 14-15 percentage points.

- There are also a number of other things in there that we'll be returning to over the next couple of classes.

Note as well a number of things that the regression results *don't* tell you:

- The regression results don't tell you if the model is correctly specified (that is, whether you have included all the "right" covariates and none of the "wrong" ones, have included interaction effects when and only when necessary, have gotten the functional forms right, etc.).
- They also don't tell you if the relationship is linear or not – remember: the fact that a linear model appears to "fit" data adequately (i.e., has statistically significant coefficients, etc.) does not mean that the relationship is linear.