# PLSC 503: "Multivariate Analysis for Political Research"

**Bivariate Regression, II: Inference**
January 24, 2017

## Inference: Introduction

Point estimates are nice, but they only get us so far...

- If we want to make inferences about the population from which our sample of data is drawn, we need to know the *variability* (or *precision*) of our estimates.
- This will also allow us to say some things about the *sampling variability properties* of those estimates as well...

The key thing to remember – from this entire course, frankly – is that **parameter estimates like $\hat{\beta}_0$ and $\hat{\beta}_1$ are themselves random variables**. This means that they have their own variability, as well as also having co-variability with each other. Moreover, remember the two things I mentioned that we were looking for in an estimator last time:

1. *Unbiasedness* – that is, $\mathrm{E}(\hat{\beta}) = \beta$, and
2. *Efficiency* – that is, "small" variance in repeated samples.

Ideally, we'd like to be able to show that the estimator we use and favor scores well on both of these two fronts.

## Some Maths

For the simple linear model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

where $u_i \sim N(0, \sigma^2)$, recall that our estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{2}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}, \tag{3}$$

where (for easier notation) I'll assume from now on that summation is over $\{1, 2, ...N\}$ unless otherwise noted. To conduct inference, we need to figure out the sampling distributions of these two random variables. That means we need to know the sampling variance of each, as well as the covariance between them. Recall that we can rewrite (3) as:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \\
&= \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} - \frac{\bar{Y}\sum(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} \\
&= \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} \qquad (4)
\end{aligned}$$

where the latter relationship holds because (by definition) $\sum(X_i - \bar{X}) = 0$. Now, think about the variability in $\hat{\beta}_1$:

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left[\frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right]. \qquad (5)$$

Because the $X$s are "fixed" (more on this later), the only source of stochastic variability is $Y$. And because we've parsed the total variability of $Y$ into the systematic (i.e., related to $X$) part and the random part $u$, all we need to know now is how $u$ varies. In practice, we assume that

$$u_i \sim \text{i.i.d. } N(0, \sigma^2); \qquad (6)$$

that is, the *stochastic* variability in $Y$ is

$$\mathrm{Var}(Y|X, \beta) = \sigma^2. \qquad (7)$$

That in turn means that:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left[\frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right] \\
&= \left[\frac{1}{\sum(X_i - \bar{X})^2}\right]^2 \sum(X_i - \bar{X})^2 \,\mathrm{Var}(Y) \\
&= \left[\frac{1}{\sum(X_i - \bar{X})^2}\right]^2 \sum(X_i - \bar{X})^2 \,\sigma^2 \\
&= \frac{\sigma^2}{\sum(X_i - \bar{X})^2}. \qquad (8)
\end{aligned}$$

In a somewhat analogous fashion, starting from (2), we can show that the variance of $\hat{\beta}_0$ is:

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sum X_i^2}{N\sum(X_i - \bar{X})^2}\sigma^2, \qquad (9)$$

and further that:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{X}}{\sum(X_i - \bar{X})^2}\sigma^2. \tag{10}$$

(These derivations are not hard, and a bit tedious, so we'll skip them). The estimates for the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are simply the square roots of (9) and (8).

Note several things about these formulas:

- **The variance of both estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ is directly proportional to $\sigma^2$.**
    - All else equal, the more variability there is in the errors, the greater the variability in our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.
    - This makes sense: Greater variability in the $u_i$s means that our regression isn't "predicting" very well; thus, we can't be as sure that our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are accurate...
- **The variance of both estimates is inversely proportional to $\sum(X_i - \bar{X})$.**
    - That is, all else (including $\sigma^2$) equal, the more variation there is in $X$, the more precise our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.
    - Again, this is not surprising: The less variability we have in $X$, the worse job $X$ can do of "explaining" variation in $Y$. (Think of it this way: In the limit, if $X$ is constant, it can tell us nothing about variability in $Y$).
- **As $N$ increases, the variability of our estimates will go down.**
    - This is directly apparent for $\text{Var}(\hat{\beta}_0)$.
    - It is also true for $\text{Var}(\hat{\beta}_1)$, since $\sigma^2$ is also a decreasing function of $N$.
- **The covariance of the two estimates depends on the sign of $\bar{X}$.**
    - The importance of the covariance of the estimates will be made a bit more obvious later in the term.

Now that we know the variability of our two parameter estimates, we can begin to say some useful things about them.

## The Gauss-Markov Theorem

We can show that – under some reasonable conditions – the least-squares estimator is a really good choice for estimating $\hat{\beta}_0$ and $\hat{\beta}_1$. In fact, it's the *best* choice possible, in a specific statistical sense:

*"Given the assumptions of the classical linear regression model, the least squares estimators are the minimum variance estimators among the class of unbiased linear estimators. (They are BLUE)."*

How do we know this?

To answer this, first think again about the estimator for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Recall that the presence of $\bar{Y}$ in the numerator means that we can rewrite (3) as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}. \tag{11}$$

Equation (11) suggests that we can think of $\hat{\beta}_1$ as a weighted combination of the $Y_i$s, with "weights" corresponding to $\frac{\sum(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$. Call this weight $k$:

$$\hat{\beta}_1 = \sum k_i Y_i \tag{12}$$

Now, suppose we were to derive some other estimator, using some "weight" other than $k_i$ instead; call that weight $w_i$:

$$\tilde{\beta}_1 = \sum w_i Y_i \tag{13}$$

For this new weight, we can see what needs to hold for $\tilde{\beta}_1$ to be unbiased:

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_1) &= \sum w_i \mathrm{E}(Y_i) \\ &= \sum w_i(\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i X_i \end{aligned} \tag{14}$$

That is, in order for this new estimator to be unbiased, it has to be the case that $\sum w_i = 0$ and $\sum(w_i X_i) = 1$. Otherwise, $\mathrm{E}(\tilde{\beta}_1)$ will equal something other than $\beta_1$.

That's fine for bias, but what about efficiency? That is, what can we say about the variance of $\tilde{\beta}_1$? Turns out that the variance of this new estimator is:

$$
\begin{aligned}
\mathrm{Var}(\tilde{\beta}_1) &= \mathrm{Var}\left(\sum w_i Y_i\right) \\
&= \sigma^2 \sum w_i^2 \\
&= \sigma^2 \sum \left[ w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} + \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 \\
&= \sigma^2 \sum \left[ w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 + \sigma^2 \left[ \frac{1}{\sum (X_i - \bar{X})^2} \right] \qquad (15)
\end{aligned}
$$

Note that the last term of this last equation is a constant, as is $\sigma^2$ itself. This means that the estimator with the smallest variability has weights which minimize

$$\sum \left[ w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 .$$

Obviously, this term is minimized when it equals zero (since, as a sum of squares, it can never be negative) – that is, when:

$$w_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}. \qquad (16)$$

When this happens, the first term in (15) drops out, and the variance of our estimator becomes

$$\mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}. \qquad (17)$$

that is, the variance of the least-squares estimator.[1] Thus, the least-squares estimator is the minimum-variance estimator in the class of linear estimators. □

[1] We can show a similar property for any given estimator for $\hat{\beta}_0$.

## Inference

To begin talking about inference, we can make use of the properties of the least-squares linear-regression estimator that we just derived.

If, as we did above, we assume that our errors $u_i$ are distributed normally (that is, $u_i \sim N(0, \sigma^2)$), then our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ (which are also random variables, and which are functions of the $u_i$s) will also be normally distributed; i.e.,

$$\hat{\beta}_0 \sim N[\beta_0, \text{Var}(\hat{\beta}_0)] \tag{18}$$

and

$$\hat{\beta}_1 \sim N[\beta_1, \text{Var}(\hat{\beta}_1)] \tag{19}$$

This is because the only thing stochastic about the $Y$s (and therefore the $\beta$s) are the disturbance terms $u$. Moreover, it means that inference on the $\beta$s is really, really easy. For example, since $\hat{\beta}_1$ is normally distributed, then we ought to be able to convert it into a $z$-score:

$$\begin{aligned} z_{\hat{\beta}_1} &= \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\text{Var}(\hat{\beta}_1)}} \\ &= \frac{(\hat{\beta}_1 - \beta_1)}{\text{s.e.}(\hat{\beta}_1)} \end{aligned}$$

This $z$-variable is distributed as $N(0, 1)$, because it is simply the "standardized" version of $\hat{\beta}_1$. We can then do hypothesis testing, inference, and so forth in a straightforward way – that is, by (e.g.) substituting hypothesized values of $\beta_1$ into (20) and comparing the result to a standard normal distribution.

### A (Small) Problem

All would seem pretty easy at this point – but, alas, it is not to be. Note that to calculate s.e.$(\hat{\beta}_1)$, we need to know $\sigma^2$ – that is, the variance of the errors in the population (or the "true" error variance).

As a practical matter, we really never know this, except in some very unusual cases. So, instead, we have to use some other value in place of $\sigma^2$. In particular, we need some *estimate* of that quantity, preferably one that is itself unbiased (or at least consistent).

How do we go about estimating $\sigma^2$? Happily, there is a ready-to-use substitute: $\hat{\sigma}^2$, the *estimated* variance of the errors $u_i$. Fox (section 10.3) shows that an unbiased estimator of $\sigma^2$ is:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{N-k} \tag{20}$$

where $\hat{u}_i^2$ is simply the estimated $u_i^2$ from the regression (i.e., the square of the observed minus the expected values) and $k$ is the number of regressors ("independent variables"), including the constant term. Plugging $\hat{\sigma}^2$ in for $\sigma^2$ in (8) yields:

$$\widehat{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}, \tag{21}$$

and similarly, for $\hat{\beta}_0$:

$$\widehat{\text{Var}(\hat{\beta}_0)} = \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \hat{\sigma}^2 \tag{22}$$

We can then just take its square root, to get the estimated standard error of (say) $\hat{\beta}_1$:

$$\begin{aligned} \widehat{\text{s.e.}(\hat{\beta}_1)} &= \sqrt{\widehat{\text{Var}(\hat{\beta}_1)}} \\ &= \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}} \\ &= \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}} \end{aligned} \tag{23}$$

If we now calculate our previously-discussed "$z$-score" $\left(\frac{(\hat{\beta}_1 - \beta_1)}{\text{s.e.}(\hat{\beta}_1)}\right)$ using this formula, we get a statistic:

$$\begin{aligned} t_{\hat{\beta}_1} \equiv \frac{(\hat{\beta}_1 - \beta_1)}{\widehat{\text{s.e.}(\hat{\beta}_1)}} &= \frac{(\hat{\beta}_1 - \beta_1)}{\frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}} \\ &= \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum (X_i - \bar{X})^2}}{\hat{\sigma}} \end{aligned} \tag{24}$$

Unlike the case of our $z$-score previously, which was normally distributed, the distribution for (24) is a bit more complicated. In particular, in the bivariate case, (24) follows a $t$ distribution, with $N-2$ degrees of freedom. This is because:

- as we noted before, the numerator is a standard normal variable, and
- the errors $u_i$ are also distributed as independent standard normal variables. Thus,
- the squared errors $u_i^2$ are all independent chi-squared variables with one degree of freedom each; this in turn means that

- the denominator $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{N-k}$ is a sum of $N-2$ independent chi-squared variables, and so
- $\hat{\sigma}^2$ is itself distributed as chi-squared with $N-2$ degrees of freedom. So,
- we have a standard normal variate divided by the square root of a chi-square variable with $N-2$ d.f., yielding our old friend, the $t$ distribution...

In case you were ever wondering, this is where the regression "$t$-test" for the coefficient estimates comes from. One can derive a similar distribution for the estimate of $\hat{\beta}_0$, using $\hat{\sigma}^2$ there as well.

## Practical Inference

Given that we know this, how do we test regression hypotheses? It's commonplace to think of hypothesis testing in two ways: in terms of *confidence intervals* (which Bayesians call *credible intervals*), and in terms of *statistical significance* (which Bayesians don't think about at all...). I'll discuss these briefly in terms of a generic parameter estimate $\hat{\beta}$, since everything here can apply to either $\hat{\beta}_0$ or $\hat{\beta}_1$.

### Significance Testing

In the vast majority of cases, we (as frequentists) can think of estimation as using the data to make a "guess" about the "true" (or "population") parameter $\beta$. The idea behind significance testing is to test a specific hypothesis about the true value of $\hat{\beta}$ against our findings from the data. In essence, we

1. "plug in" a value for $\beta$ into the $t$-test formula in (24), and
2. choose some significance level (typically called $\alpha$) at which we wish to "reject" that hypothesis. We can then
3. evaluate how likely it is that we'd have drawn that particular sample of data from a population with "true" value $\beta$ given the estimated value $\hat{\beta}$.

The most common approach is to consider the *null hypothesis*: i.e., the hypothesis that $\beta = 0$. This yields a test for whether or not the estimate is "statistically significant" at some level. Put simply, it is a bad approach (see, e.g., Gill's 1999 article in *Political Research Quarterly*, for starters), and – while it is used a lot – I don't recommend it.

### Confidence Intervals

Inference via confidence intervals moves the emphasis of estimation away from the point estimate $\hat{\beta}$ and toward a range of potential values. It involves drawing/estimating a "confidence interval" around our point estimate $\hat{\beta}$, the width of which is defined by two quantities:

1. The estimate of the variability of the parameter estimate (that is, $\widehat{\text{s.e.}(\hat{\beta})}$), and

2. the desired level of confidence.

Since we know the distribution of the estimate, including both its center point and its variability, we (as frequentists) can talk about how often – in repeated sampling – the confidence interval constructed in this way will contain the "true" (population) value. For example, we know that roughly 95 percent of the "mass" of a $t$-distribution occurs within 1.96 standard deviations of its mean. Accordingly, given a confidence interval of 95 percent, we can be sure that, 95 times out of 100, an interval about two standard deviations wide around a given point estimate will contain the "true" value $\beta$.[2]

## Predictions and Inference

Recall that the predicted value for $Y$ given a particular value of $X$ (say, $X_k$) is just:

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k \tag{25}$$

This is a *point prediction* – a single value equal to the expected value of $Y$ associated with a particular set of values for $X_k$, $\hat{\beta}_0$ and $\hat{\beta}_1$. Note in particular that, since this prediction depends on the values of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ (which are themselves random variables), **it is also a random variable.**

OK, so what are the properties of this random variable $\hat{Y}_k$?

Well, its easy to show that it is an *unbiased estimator of* $Y_k$:

$$\begin{aligned} \text{E}(\hat{Y}_k) &= \text{E}(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\ &= \text{E}(\hat{\beta}_0) + X_k \text{E}(\hat{\beta}_1) \\ &= \beta_0 + \beta_1 X_k \\ &= \text{E}(Y_k) \end{aligned} \tag{26}$$

That is, the expected value of the prediction is equal to the expected value of $Y$, conditional on the values of $X$. This shouldn't be too surprising, since we already know that the estimators for the parameters are unbiased.

The next thing we might want to know is, How "good" a prediction is this? That is, how much variability is there in this prediction? Recall that, for two random variables $A$ and $B$, the variance of their sum is equal to:

[2] Put a bit differently: If we took an infinite number of samples of size $N$, and calculated our $\alpha$% confidence intervals for each one, we would know that $\alpha$ percent of those intervals contained $\beta$.

$$\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B) + 2\,\text{Cov}(A, B)$$

This is useful, in that we want to know $\text{Var}(\hat{Y}_k)$, which is the same as:

$$\begin{aligned}
\text{Var}(\hat{Y}_k) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\
&= \frac{\sum X_i^2}{N\sum(X_i - \bar{X})^2}\sigma^2 + \left[\frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right] X_k^2 + 2\left[\frac{-\bar{X}}{\sum(X_i - \bar{X})^2}\sigma^2\right] X_k
\end{aligned}$$

A little bit of algebra yields a more visually-satisfying representation:

$$\text{Var}(\hat{Y}_k) = \sigma^2\left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{27}$$

*What does equation (27) tell us?*

- The variability of a prediction decreases as $N$ increases.
- The variability of the predicted $Y$ decreases as the variability of $X$ increases.
- The variability of the prediction increases as the value of $X$ at which we are predicting (that is, $X_k$) gets farther away from $\bar{X}$.

All of these ought to strike you as pretty reasonable characteristics.

**Inference and Predictions**

The square root of $\text{Var}(\hat{Y}_k)$ is what is known as the *standard error of the prediction*:

$$\widehat{\text{s.e.}(\hat{Y}_k)} = \sqrt{\sigma^2\left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]} \tag{28}$$

As with estimation of the $\hat{\beta}$s, since we typically don't know $\sigma^2$, we replace it with its unbiased estimator $\hat{\sigma}^2$. This in turn means that, as with our coefficient estimates, we can use the $t$ distribution to make inferences about the predictions; this is because the predictions – like the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ themselves – follow a $t$ distribution. So, for example, we can calculate (say) the 95-percent confidence interval around our prediction $\hat{Y}_k$ as:

$$95\%\,\text{c.i.}(\hat{Y}_k) = \hat{Y}_k \pm [1.96 \times \widehat{\text{s.e.}(\hat{Y}_k)}]$$

Also: Most statistical packages (including R and Stata) will generate the value(s) of $\widehat{\text{s.e.}(\hat{Y}_k)}$ for you, along with the predicted values $\hat{Y}$ themselves. We'll demonstrate this in a bit.

### The Use and Misuse of Predicted $\hat{Y}$s

Predicted values of $Y$ are often a good way to discuss the results of a regression model. For example, it can be useful to make statements like "The model indicates that when $X$ equals 10, the predicted value of $Y$ is 235," or whatever. Similarly, we often want to make statements along the lines of "When $X$ increases from zero to one, the predicted value if $Y$ decreases from 76 to 54."

There's nothing wrong with such statements in principle. However, it is crucially important when using predicted values to interpret a regression model also to include a discussion of the variability around those predictions. Doing so tells us how "accurate" (or "precise") the predictions of our model are. So, instead of saying:

*"The results indicates that when $X$ equals 10, the predicted value of $Y$ is 235."*

it is *always* better to say:

*"The results indicates that when $X$ equals 10, the 95 percent confidence interval for the predicted value of $Y$ ranges from 210 to 260."*

Similarly, one should always include some mention of the variability of the predicted values in "change" statements as well. So, instead of saying:

*"When $X$ increases from zero to one, the predicted value if $Y$ decreases from 76 to 54."*

it is preferable to say that:

*"An increase in $X$ from zero to one decreases the range of 95-percent credible values of $\hat{Y}$ from $[72, 80]$ to $[49, 59]$."*

Note also that implicit in the latter of these discussions is a statement about the "statistical significance" of the effect of $X$ on $Y$. Because the ranges $[72, 80]$ and $[49, 59]$ do not overlap (or even come close), it indicates that the estimated effect of $X$ on $Y$ is highly unlikely to be due to sampling error / chance.

Finally: A truly excellent way to present such results – provided $X$ is continuous – is to **plot** the predicted values and confidence intervals over some range of $X$; we'll do this in the example below.

## Back to Our Example: More Infant Mortality

Remember this?

```
> IMdata<-na.omit(IR2000[c("infantmortalityperK","DPTpct")])
> IMDPT<-lm(infantmortalityperK~DPTpct,na.action=na.exclude)
> summary(IMDPT)

Call:
lm(formula = infantmortalityperK ~ DPTpct, data = IMdata)

Residuals:
   Min     1Q Median     3Q    Max
 -56.8  -16.3   -5.1   11.8   86.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  173.277      8.489    20.4   <2e-16 ***
DPTpct        -1.576      0.101   -15.6   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 26.2 on 175 degrees of freedom
Multiple R-Squared: 0.582,Adjusted R-squared: 0.58
F-statistic:  244 on 1 and 175 DF,  p-value: <2e-16



> anova(IMDPT)
Analysis of Variance Table

Response: infantmortalityperK
           Df Sum Sq Mean Sq F value Pr(>F)
DPTpct      1 167423  167423     244 <2e-16 ***
Residuals 175 120033     686
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We can now ask again: What do these results tell us?

First, remember that

- What we called *TSS* is the total variability in $Y$ around its mean – that is, $\sum(Y_i - \bar{Y})^2 = 167423 + 120033 = \mathbf{287456}$.
- *MSS* (which R calls `DPTpct` and Stata calls `Model`) is the model ("explained" or "regression") sum of squares – that is, $\sum(\hat{Y}_i - \bar{Y})^2 = \mathbf{167423}$.
- *RSS* (which R calls `Residuals` and Stata calls `Residual`) is the residual ("unexplained" of "error") sum of squares – that is, $\sum \hat{u}_i^2 = \mathbf{120033}$.
- This latter value, divided by $N$ minus the number of independent variables $k$ (including the constant term), gives us our estimate $\hat{\sigma}^2$. So, here, $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{N-2} = \frac{120033}{175} = \mathbf{686}$, and
- the standard error of our estimate ("SEE," which R calls the `Resisual standard error` and Stata calls the "`Root MSE`," for "root mean squared error") is $\hat{\sigma} \equiv \sqrt{\hat{\sigma}^2} = \sqrt{686} = \mathbf{26.2}$.

Now, we can calculate the variability in $X$ as $\sum X_i^2 = \mathbf{1253105}$, and its variability around its mean as $\sum(X_i - \bar{X})^2 = \mathbf{67381}$. This means that:

- The variance estimate for the constant term is

$$\begin{aligned}
\widehat{\mathrm{Var}(\hat{\beta}_0)} &= \frac{\sum X_i^2}{N\sum(X_i - \bar{X})^2}\hat{\sigma}^2 \\
&= \frac{1253105}{177(67381)} \times 686 \\
&= \mathbf{72.1}
\end{aligned}$$

  and the square root of this is the reported standard error estimate of the constant term (that is, about **8.49**).

- Similarly, variance estimate for the slope term is

$$\begin{aligned}
\widehat{\mathrm{Var}(\hat{\beta}_1)} &= \frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2} \\
&= \frac{686}{67381} \\
&= \mathbf{0.010}
\end{aligned}$$

  and the square root of this is the reported estimated standard error for the slope coefficient (that is, about **0.10**).

- Finally, the covariance of the estimated slope and intercept is

$$\begin{aligned}\widehat{\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)} &= \frac{-\bar{X}}{\sum(X_i - \bar{X})^2}\hat{\sigma}^2 \\ &= \left(\frac{-81.8}{67381}\right) \times 686 \\ &\approx \mathbf{-0.83}\end{aligned}$$

Note that R (and Stata for that matter) will tell us all these things after estimation, if we ask nicely:

```
> vcov(IMDPT)

            (Intercept)   DPTpct
(Intercept)     72.0677 -0.83317
DPTpct          -0.8332  0.01018
```

### Confidence Intervals

A simple `summary()` of an `lm` object gives us $p$-values for each of the $\hat{\beta}$s, but not confidence intervals around the $\hat{\beta}$s. We can obtain the latter by using `confint()`:

```
> confint(IMDPT)

               2.5 %  97.5 %
(Intercept) 156.523 190.032
DPTpct       -1.775  -1.377
```

If we want confidence intervals different from the 95% two-tailed defaults, we can specify the `level` option:

```
> confint(IMDPT,level=0.99)

               0.5 %  99.5 %
(Intercept) 151.169 195.385
DPTpct       -1.839  -1.314
```

Note that Stata provides 95% confidence intervals in the output by default; that can be changed as well.

### Practical Prediction

It's also pretty easy to use R to generate standard errors and confidence intervals for predicted values of $Y$. The `predict` function creates either a list or a matrix of predicted values, standard errors, and confidence / prediction intervals for those predictions; see `?predict` for details (particularly `?predict.lm`). E.g.:

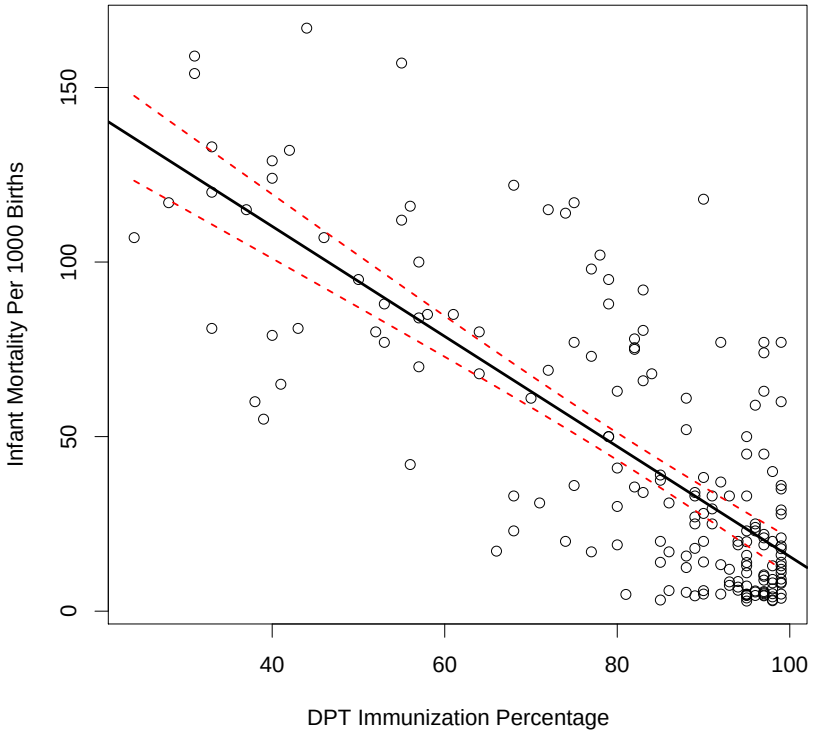
```
> SEs<-predict(IMDPT,interval="confidence")
> SEs
       fit    lwr    upr
1    25.10  20.53  29.68
3    17.22  12.05  22.40
4    23.53  18.84  28.21
.
.
<rows omitted>
.
.
189  21.95  17.15  26.75
190  39.29  35.36  43.23
191  17.22  12.05  22.40
```

Here, the object `SEs` is a matrix containing three columns; `fit` is just the fitted values (i.e., the $\hat{Y}$s), while `lwr` and `upr` are the lower and upper 95% confidence intervals, respectively. We can use the latter to create a useful little plot:

```
> # Sort the data for a prettier plot...
> Sort<-order(IMdata$DPTpct)
> plot(IMdata$DPTpct,IMdata$infantmortalityperK,xlab="DPT Immunization Percentage",
  ylab="Infant Mortality Per 1000 Births")
> abline(IMDPT,lwd=3)
> lines(sort(IMdata$DPTpct),SEs[Sort,2],col="red",lwd=2,lty=2)
> lines(sort(IMdata$DPTpct),SEs[Sort,3],col="red",lwd=2,lty=2)
```

which yields the rather nice-looking:

Figure 1: Scatterplot of Infant Mortality and DPT Immunizations, along with Least-Squares Line and 95% Prediction Confidence Intervals




Note a few things about Figure 1:

- We can use the created confidence intervals to make the sorts of verbal statements we discussed above. So, for example, we might way something like

  > "The results indicate that an increase in the *DPT Percentage* from 40 to 80 corresponds to a decrease in the predicted (95-percent confidence) range of infant mortality from [100,125] to [45,55]."

- Similarly, we can use the confidence intervals to observe ranges of $X$ over which particular values of $\hat{Y}$ are plausible. Here, for example:

  - a predicted infant mortality rate of 100 per 1000 births (that is, ten percent) is within the 95 percent confidence interval in the range `DPTpct` $\approx$ [40,55], while
  - that for an infant mortality rate of 50 per 1000 (five percent) is within the 95 percent confidence interval in the range DPTpct $\approx$ [78,82].

- Finally, we can do "out of sample" as well as "in-sample" predictions; we'll talk more about that a bit later.