PLSC 503: "Multivariate Analysis for Political Research"

Bivariate Regression, I
January 19, 2017

Motivation

We'll spend a few days talking about bivariate regression – that is, regressing a single variable Y on a single variable X. Moreover, we'll use the 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 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:

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

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. In addition, we'll also want/need to know a few things about model "fit" and other diagnostics. We'll focus on the point estimates today, and on standard errors and inference (and other topics) next week.

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:

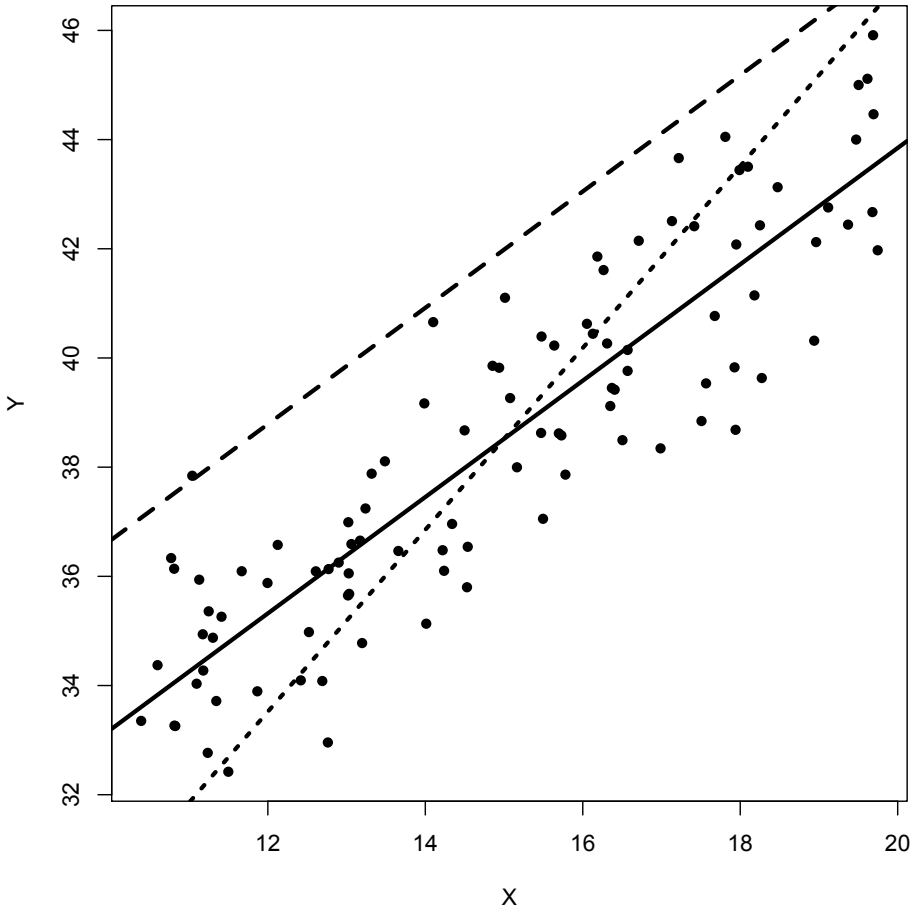
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3}$$

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 "gets it right" vis-a-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.

Least Squares Regression

Ordinary least squares ("OLS") regression starts with a simple premise for estimation: to minimize $\hat{S} = \sum_{i=1}^{N} \hat{u}_i^2$. To do this, we can start out by noting that

$$\begin{aligned}
\hat{S} &= \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
&= \sum_{i=1}^{N} (Y_i^2 - 2Y_i\hat{\beta}_0 - 2Y_i\hat{\beta}_1 X_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 X_i + \hat{\beta}_1^2 X_i^2) \qquad (5)
\end{aligned}$$

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

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

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

$$\sum_{i=1}^{N} Y_i = N\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{N} X_i \tag{6}$$

and

$$\sum_{i=1}^{N} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{N} X_i + \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 \tag{7}$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{8}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{9}$$

A Bit of Intiuition

Equation (8) 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

$$\hat{\beta}_1 = \frac{\text{Covariance of } X \text{ and } Y}{\text{Variance of } X}$$

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: Immunizations and Infant Mortality

Continuing with the data we used from last time ($\approx$ 190 countries for the year 2000), I thought we'd examine the relationship between infant mortality (Y) and the percentage of the population which has received DPT immunizations (X).:

```
> library(psych)
> describe(infantmortalityperK)
  var   n  mean    sd median   mad min max range skew kurtosis   se
1   1 179 43.83 40.39     29 34.26 2.9 167 164.1    1     0.06 3.02
> describe(DPTpct)
  var   n  mean    sd median   mad min max range  skew kurtosis   se
1   1 181 81.71 19.77     90 11.86  24  99    75 -1.31     0.57 1.47
```

Estimating an OLS regression model is straightforward; we use the `lm` command (for 'linear model") in R (or the `regress` command in Stata):

```
> IMDPT<-lm(infantmortalityperK~DPTpct)
> summary.lm(IMDPT)

Call:
lm(formula = infantmortalityperK ~ DPTpct)

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
  (14 observations deleted due to missingness)
Multiple R-Squared: 0.582,Adjusted R-squared: 0.58
F-statistic:  244 on 1 and 175 DF,  p-value: <2e-16
```

It is also useful to look at the "analysis of variance" (ANOVA) of the regression. We can get it in R using the `anova()` command:

```
> 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
```

Together, these results tell us several things about the (linear) relationship between these two variables:

- We would expect a country with a zero value on the `DPTpct` variable (that is, a country with no immunization whatsoever) to have an infant mortality rate of about 173 per 1000 births.
- Each one-unit increase in `DPTpct` is associated with a corresponding decrease of about 1.6 in the expectation of the `infantmortalityperK` variable.
    - Note that because `DPTpct` ranges from 24 to 99, a one-unit increase is a pretty small increase. However,
    - Because the relationship is linear, it rescales perfectly.
    - That means that (e.g.) an 20 percent increase in `DPTpct` is associated with a (1.58 $\times$ 20) = 31.2-unit change in the expectation of `infantmortalityperK`.
- The "total sum of squares" (TSS) is equal to (167423 + 120033) = 287456; 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 120033; that is the amount of "residual" / "error" / stochastic variability left in `infantmortalityperK` after the effect of `DPTpct` is accounted for.
- The "model sum of squares" is 167423; this reflects the amount of (mean-centered) variation "explained" by the `DPTpct` 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:

$$\hat{s}^2 = \frac{\text{RSS}}{N-k} = \frac{\sum_{i=1}^{N} \hat{u}^2}{N-2} \tag{10}$$

  We can think of this as an "average" value of the squared residuals. As reported in the table, in the model above this is equal to 686. 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

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

  Think of this as the "average" residual; in our infant mortality model, this is equal to $\sqrt{686} = 26.2$. This tells us that, on average, our model mis-predicts the `infantmortalityperK` variable by about 26 infant deaths per 1000 births.

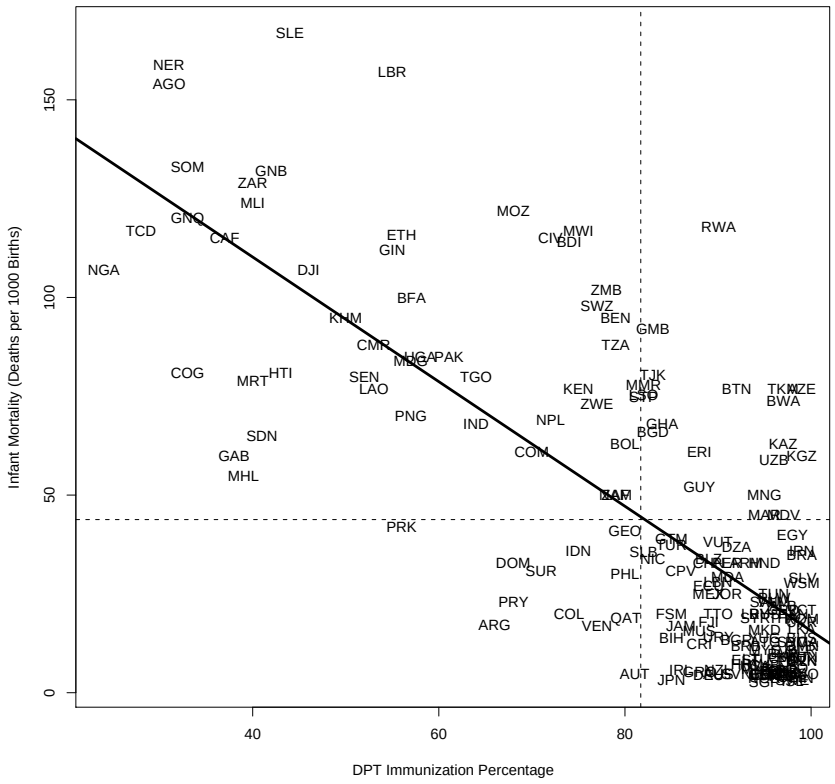
- There are also a number of other things in there that we'll be returning to over the next few of classes.

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.

The OLS regression line estimated is plotted above. The horizontal and vertical dotted lines represent the values of $\overline{\texttt{infantmortalityperK}}$ and $\overline{\texttt{DPTpct}}$, respectively; note that the OLS fit line passes through the point $(\bar{X}, \bar{Y})$.

Figure 2: Regression of Infant Mortality on DPT Immunization Rates



Predicted Values, Residuals, etc.

We can generate predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ straightforwardly. The predictions are just the values of Y that we would expect, conditional on the (observed) value of X; so:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i. \tag{12}$$

Similarly, the estimated residual for a given observation i is just the difference between the actual value Y_i and the predicted value:

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i. \end{aligned} \tag{13}$$

Note as well that it is sometimes useful to express the fitted value as the difference between the observed value of Y and the residual:

$$\hat{Y}_i = Y_i - \hat{u}_i.$$

Take a look at the summary statistics for these:

```
> describe(IR2000$IMDPTres)

  var   n mean    sd median   mad   min   max range skew kurtosis   se
1   1 177    0 26.12   -5.1 19.42 -56.8 86.59 143.4 0.75     0.44 1.96

> describe(IR2000$IMDPThat)

  var   n  mean    sd median  mad   min   max range skew kurtosis   se
1   1 177 44.26 30.84  31.41 18.7 17.22 135.4 118.2  1.3     0.59 2.32
```

Notice that:

- The mean of the predicted values $\hat{Y}_i$ is the same as the mean of Y (that is, about 44). (Why?)
- The mean of the residuals is zero. (Again, why?)

Moreover, if we correlate them:

```
> cor(IR2000$infantmortalityperK,IR2000$DPTpct,use="complete.obs")
[1] -0.7632

> cor(IR2000$IMDPTres,IR2000$infantmortalityperK,use="complete.obs")
[1] 0.6462

> cor(IR2000$IMDPTres,IR2000$DPTpct,use="complete.obs")
[1] 9.573e-17

> cor(IR2000$IMDPThat,IR2000$infantmortalityperK,use="complete.obs")
[1] 0.7632

> cor(IR2000$IMDPThat,IR2000$DPTpct,use="complete.obs")
[1] -1

> cor(IR2000$IMDPTres,IR2000$IMDPThat,use="complete.obs")
[1] 5.302e-17
```

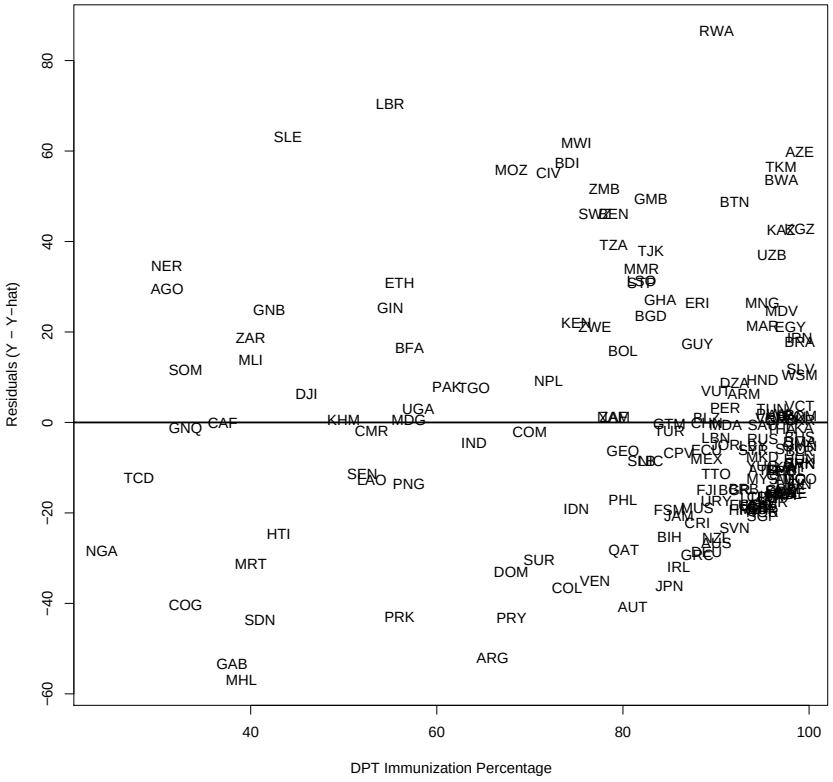
we can see that:

- The correlation between `infantmortalityperK` and `DPTpct` is the same as the (negative) square root of the reported R^2 of the regression (more on that next week...).

- The correlation between `DPTpct` and the fitted values (that is, between $\hat{Y}$ and X) is 1.0. (Why?)
- The correlation between `infantmortalityperK` and the fitted values is the same as that between `infantmortalityperK` and `DPTpct`. That is, $\text{corr}(Y, \hat{Y}) = \text{corr}(Y, X)$. (Again, why?)
- Finally, the correlations between both X and $\hat{Y}$ and the residuals $\hat{u}$ are zero. (Why?)

Over the course of the semester, we'll be doing a lot more with residuals. But, for now, there are density plots of both the fitted values and the residuals in the handout. It can also be useful to look at a plot of the residuals against the covariate `DPTpct`, just to get a sense of "how the model is doing":

Figure 3: Regression Residuals ($\hat{u}$) vs. DPT Immunization Rates (X)



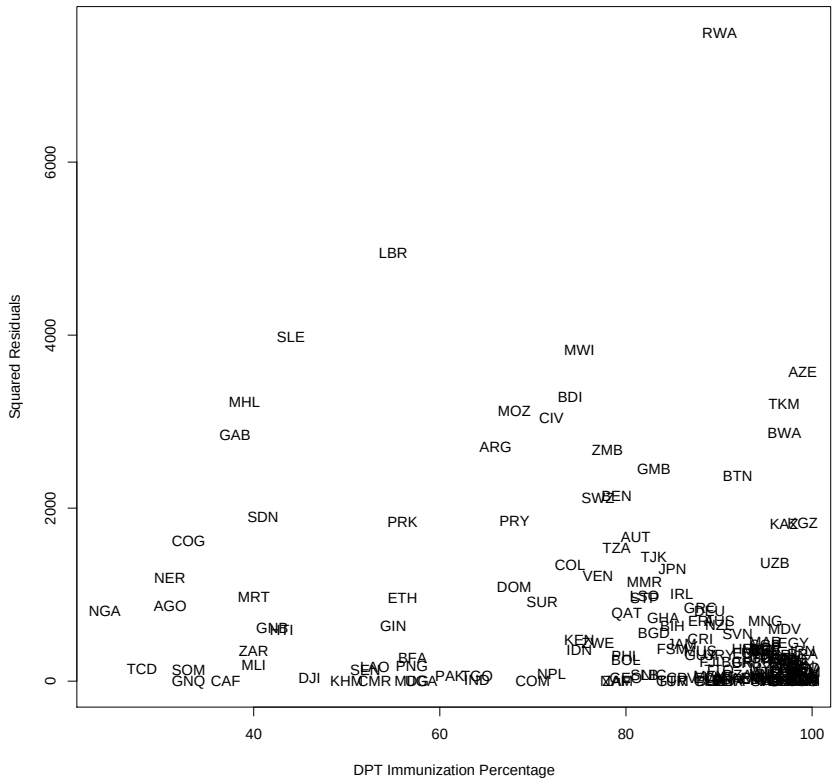
Remember that:

- Residuals greater than zero reflect countries with higher levels of infant mortality than their values of `DPTpct` values would predict;

- Residuals less than zero are the opposite (countries whose infant mortality was lower than one would predict on the basis of their DPT immunization percentage).

We can also look at the squared residuals – that is, the things that OLS minimizes in estimating $\hat{\beta}_0$ and $\hat{\beta}_1$. Here's an example:

Figure 4: Squared Residuals vs. Fitted Values



There's nothing in particular you need to look at here, at least not now (we'll discuss residuals plots as diagnostic tools a bit later...).

Next Tuesday, we'll discuss standard error estimates, confidence intervals, and *inference*...