EEP/IAS 118
Andrew Dustan
Section Handout 8

1. Revisiting Standard Errors (for your reference)

There seems to be some confusion over what the "standard error" is in the context of various hypothesis tests (and confidence intervals).

Remember why we do statistics:

1. There is a **population parameter** (μ, $D = \mu_1 - \mu_2$, or β_j) that we want to know but can't see.
2. So we estimate that population parameter with an **estimator** ($\bar{x}$, $\ddot{D} = \bar{x}_1 - \bar{x}_2$, or $\hat{\beta}_j$).
3. We want to know how precise this estimator is. In other words, we want to know the **standard deviation of the estimator**. This is the number we would get if we took an infinite number of samples from the population, calculated the estimator for each sample, and then took the standard deviation of all of those calculated estimators.
4. Problem: we don't have an infinite number of samples, so we must find a way to estimate the standard deviation of the estimator using our one sample. This estimate is called the **standard error**. We use the standard error as our best guess of the estimator's standard deviation.

To reiterate: **The standard error (SE) is our estimate of the estimator's standard deviation.**

Example:

1. Want to know a population mean, μ.
2. Estimate μ by using $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
3. Want to know how precise of a guess $\bar{x}$ is. For this we want to know the standard deviation of the estimator: $\sigma(\bar{x})$. We know from a derivation in lecture that $\sigma(\bar{x}) = \sqrt{\frac{var(x)}{n}} = \frac{sd(x)}{\sqrt{n}}$.
4. We don't have $\sigma(\bar{x})$ so we must estimate it. Specifically, we don't know $var(x)$. Statisticians found out that a good estimator of $var(x)$ is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Then our estimator for $\sigma(\bar{x})$ is

$$\widehat{\sigma(\bar{x})} = \boldsymbol{SE} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

Example:

1. Want to know a population proportion, p.
2. Estimate p by using $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Note the special case here: $x_i = 0$ or $x_i = 1$.
3. Want to know how precise of a guess $\hat{p}$ is. For this we want to know the standard deviation of the estimator: $\sigma(\hat{p})$. We know from a derivation in lecture that $\sigma(\hat{p}) = \sqrt{\frac{var(x)}{n}} = \sqrt{\frac{p(1-p)}{n}}$.
4. We don't have $\sigma(\hat{p})$ so we must estimate it. Specifically, we don't know $p(1-p)$. Statisticians found out that a good estimator of $p(1-p)$ is $\hat{p}(1-\hat{p})$. Then our estimator for $\sigma(\hat{p})$ is

$$\widehat{\sigma(\hat{p})} = \boldsymbol{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

2. Practice: Interpreting Regression Results

Economists are social scientists. We are not robots or pure mathematicians. When we interpret the results of a regression, we need to **think** about what they mean in terms of real-world implications, not just say "ceteris paribus, a ___ change in *x* causes a ____ change in *y*." That's not good enough for an economist; we need to say whether this effect is statistically significant and whether the size of the effect is economically important.

Problem Set 3 presented a great chance to use your economist skills to interpret an interesting regression. We compared census tracts that had toxic waste sites, 200 of which were bad enough to be put on the NPL list for cleanup and 300 of which were not quite bad enough for the NPL:

```
regress lmdval2000 npl2000 lmdval80 povrat80 pop80;
```

```
      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  4,   495) =  203.19
       Model |  105.267994     4  26.3169984           Prob > F      =  0.0000
    Residual |  64.1109437   495  .129517058           R-squared     =  0.6215
-------------+------------------------------           Adj R-squared =  0.6184
       Total |  169.378937   499  .339436748           Root MSE      =  .35988

------------------------------------------------------------------------------
  lmdval2000 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     npl2000 |    .092683   .0342967     2.70   0.007     .0252979    .1600681
    lmdval80 |   .7670555   .0355218    21.59   0.000     .6972634    .8368475
    povrat80 |   -1.07724   .1908524    -5.64   0.000    -1.452221   -.7022597
       pop80 |   .0000171   2.11e-06     8.10   0.000     .0000129    .0000212
       _cons |   3.377376   .3953333     8.54   0.000     2.600638    4.154114
------------------------------------------------------------------------------
```

mdval80:	Housing value in 1980 (measured by the median house value, in current $)
mdval2000:	Housing value in 2000 (measured by the median house value, in current $)
npl2000:	Indicator for whether the census tract contains a toxic waste site that is listed on the NPL in 2000
pop80:	population density in 1980 (people/square mile)
povrat80:	poverty rate in 1980 (between 0 and 1. povrat80=1 means everyone is in poverty)

lmdval80 is log(mdval80) and lmdval2000 is log(mdval2000).

To help guide our discussion, here are summary statistics (always look at these when you interpret regressions):

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     npl2000 |       500          .4    .4903886          0          1
       pop80 |       500    4250.827    8450.268   1.872531   95033.34
    povrat80 |       500    .1141938    .1005672    .002267    .695319
   mdval2000 |       500      139937    100856.6      14600     964000
     mdval80 |       500    60592.76    94181.21   11284.48    2044114
```

The best interpretation for each these coefficients isn't obvious and will take some thinking.

Let's talk about the results for each variable, one at a time:

lmdval80 (log of median house value in 1980)	
npl2000 (=1 if tract on NPL by 2000, =0 otherwise)	
povrat80 (poverty rate in 1980)	
pop80 (population density in 1980, people/mi^2)	