Math 455 Class 24 March 20.

Last time we talked about the case when variables are measured with errors:

Instead of the "true" X we can only see

$X^* = X + \delta$ where δ is a RV,

say $E[\delta] = 0$ and $V[\delta] = \sigma_\delta^2$.

We WANT to estimate β_1 in:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

But the regression we run is:

$$Y = \alpha_0 + \alpha_1 X^* + \varepsilon$$

We saw that there was something we called "Attenuation Bias": our estimate $\hat{\alpha}_1$ will have the property that $|\hat{\alpha}_1| < |\beta_1|$.

We saw this in simulation: we produced fake data in R and observed the results.

This phenomenon did NOT occur for the case where Y was replaced by $Y^* = Y + \delta$. Errors in the measurement of Y do not cause attenuation bias.

This was a special case of the treatment in Faraway. See p. 100.

His formula on p. 100 can be obtained by techniques of Math 447.

In our regression: $Y = \alpha_0 + \alpha_1 X^* + \varepsilon$

We will assume that the error δ is independent of everything else, and that

$E[\delta] = 0 \quad V[\delta] = \sigma_\delta^2$.

We propose to compute $E[\hat{\alpha}_1]$ and compare this to β_1. The true relationship (assumed in this calculation) is $Y = \beta_0 + \beta_1 X + \varepsilon$.

In the regression $Y = \alpha_0 + \alpha_1 X^* + \varepsilon$

$$E[\hat{\alpha}_1] = \frac{Cov(X^*, Y)}{Var(X^*)} = \frac{Cov(X+\delta, Y)}{Var(X+\delta)}$$

With the assumption that δ is indep of everything,

$$= \frac{Cov(X,Y) + \cancel{Cov(\delta, Y)}^{\,0 \text{ (indep.)}}}{Var(X) + Var(\delta)}$$

$$= \frac{Cov(X,Y)}{Var(X) + \sigma_\delta^2}$$

Same as value of β_1 but $(+\sigma_\delta^2)$.

in denominator. Thus we have:

$$|E[\hat{\alpha}_1]| < |\beta_1|.$$

In our R code from last class, we got errors by introducing them in R.

What is an error? In practice, where do errors come from?

We'll talk about the case of the BLS NFP report.

The estimate of jobs gained/lost is just that: an estimate. It is obtained by sending surveys to a sample of employers, then making adjustments.

Not all employers return the surveys!
Not all employers get a survey.

Small, so recently created businesses are not covered. This is estimated using a "birth-death model": they make it up.

This is only reconciled with tax/insurance records once per year.

The difference between the "true" tax/ins number and the estimate from BLS is an error.

1. You fit a linear regression in R and get the following summary output:

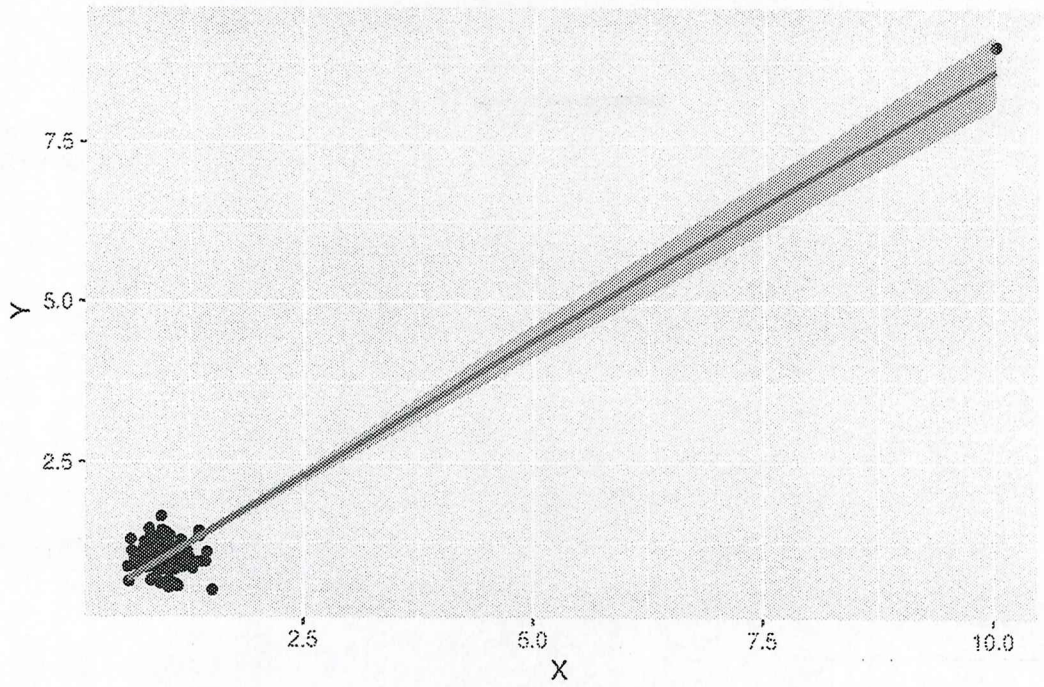
```
> summary(lmod)
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94484 -0.18748  0.00193  0.19895  0.66746

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17256    0.04469   3.861 0.000204 ***
X            0.84458    0.03118  27.086  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2853 on 97 degrees of freedom
Multiple R-squared:  0.8832,Adjusted R-squared:  0.882
F-statistic: 733.6 on 1 and 97 DF,  p-value: < 2.2e-16
```

Then you do a scatter plot of the data with the regression line and see:



Answer the following questions about this regression:

(a) (4 points) The fourth standard diagnostic plot that R produces for a regression (using `plot(lmod)`) is sometimes called the "Cook's Distance" plot, because it has this label. What are the two variables defining the "Cook's Distance" plot? (That is, what are the labels on the axes of the "Cook's Distance" plot?)

(b) (4 points) There is a point on the top right of the scatter plot on the previous page. Where, in the "Cook's Distance" plot, would you expect the point corresponding to this row of data to show up?

(c) (6 points) Suppose the row of data corresponding to the point in (b) were omitted and we refit the model. How might the intercept, the R^2, and the coefficient of `X` change? (At a minimum, your answer should clearly indicate whether each of these numbers increases, decreases, or remains the same.)

(d) (4 points) Looking at the regression summary output on the previous page, another student asserts that the regression result is very reliable, because the t- and F-statistics are so significant and the R^2 is so high. Comment on this assertion.