Math 455 Class 23 March 18

Added-variable vs. Partial Residual.

The big picture: we would like to draw our regression line on a scatter plot.

If we have 2 variables `lmod <- lm(Y ~ X)` we can do this.

If we have more variables, it's not possible.

What can we do instead?

Added-Variable and Partial-Residual plots.

What are these?

Suppose our original model is:

$$Y \sim \cancel{\beta_0} \; X_1 + X_2 + X_3 .$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Added-Variable: (for each of X_1, X_2, X_3)

Regress Y on all Xs except X_i.

get residuals $\hat{\delta}$.

Regress X_i on all <u>other</u> X_j. $j \neq i$

get residuals $\hat{\gamma}$.

Now plot $\hat{\gamma}, \hat{\delta}$ as a scatter plot

with regression line from $\hat{\delta} \sim \hat{\gamma}$.

Note: with 3 explanatory variables, we will have 3

added-variable plots.

Partial-Residual: (Again 3 explanatory vars = 3 plots)

Construct (for each explanatory X_i)

the response Y with the predicted effect of

the other X_j s <u>removed</u>: $(j \neq i)$

$$Y - \sum_{j \neq i} x_j \hat{\beta}_j = \hat{Y} + \hat{\varepsilon} - \sum_{j \neq i} x_j \hat{\beta}_j = X_i \hat{\beta}_i + \hat{\varepsilon}.$$

Plot this against X_i.

With 3 variables we will have 3 partial-residual plots.

Why? What are these for?

Added-Variable: What does X_i add to the model given that the other X_js are already included? (Better for detecting influential points.)

Partial-Residual: Is the relation between Y and X_i linear, after accounting for other predictors?

In the big picture, this is another test to make sure we are not getting our regression result for stupid reasons.

There is no end to tests of this type.

Code for these plots is in 6.3 done "by hand".
There is a package "car" with functions avPlots and crPlots that does it for you.

More about checking for problems in a regression.

What if Y or X is measured with an error? What happens, and what should we do about that?

Suppose the model $Y = \beta_0 + \beta_1 X + \varepsilon$ is actually true, but we can only measure

$Y^* = Y + \text{error}$ and/or $X^* = X + \text{error}$.
"δ".

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

that is: lm$(Y \sim X^*)$

How do we expect the estimate $\hat{\alpha}_1$ to compare to true β_1 (which is what we hope to find)?

Similarly, we could ask, what if we do the regression: $Y^* = \gamma_0 + \gamma_1 X + \varepsilon$

that is: lm$(Y^* \sim X)$

How do we expect the estimate $\hat{\gamma}_1$ to compare to the true β_1 (which is what we want)?

This is the topic of Ch. 7 "Problems with the predictors".

We could even ask: what if we run the regression $Y^* = \eta_0 + \eta_1 X^* + \varepsilon$

that is lm($Y^* \sim X^*$): how does our estimate $\hat{\eta}_1$ compare to the "true" β_1?

We can address this either with calculations in the manner of Math 447 (midterm problem alert!)

We can also address this via simulation: make up some data and errors and let R show us the results.

The answer goes by a name: "Attenuation Bias".

In the first model: $lm(Y \sim X^*)$

$\hat{\alpha}_1$ is NOT an unbiased estimator of β_1.

We expect $|\hat{\alpha}_1| < |\beta_1|$.

$\hat{\alpha}_1$ "shrinks β_1 toward 0".

We can see with simulated data and graphs why this is.

What if we look at $lm(Y^* \sim X)$?

Is $\hat{\gamma}_1$ an unbiased est of β_1?

YES. Note that regression is NOT symmetric with respect to exchanging X for Y.

Examples in R.