Math 455 Class 21 March 13.

Last time we discussed 3 concepts:

leverage, outliers, influence.

These concepts are relevant to regression because we do not want a result that depends only on a few data points: such results are questionable.

A high-leverage point is one whose X-coordinates are far from the mean of the data.

An outlier is a point which does not fit well with the model.

An influential point is one whose removal from the data set would cause a large change in the fit.

Our measure of influence, at least the one in the 4th standard diagnostic plot, is Cook's Distance

This 4th plot is leverage vs. standardized residuals. Note that Cook's Distance is NOT one of the axes of the plot.

Recall that the standardized residuals

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}$$

where $\hat{\varepsilon}_i$ [e_i in many other sources] are the (ordinary) residuals.

h_i are the leverages: $h_i = H_{ii}$ - the i^{th} diagonal entry of the hat matrix. $\left(\frac{\partial \hat{Y}_i}{\partial Y_i}\right)$

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\text{Sum of squared residuals}}{n = \#\text{ parameters}}$$

If the model assumptions of linear regression are true, $E[r_i] = 0$ and $V[r_i] = 1$, whereas $V[\hat{\varepsilon}_i] = \sigma^2(1-h_i)$.

It is also often true that $Cor(r_i, r_j)$ tends to be small.

This is the basis for the standard Q-Q plot: the r_i are a lot like independent standard normal RVs, so to check this we compare sorted standardized residuals with what we would expect to get from the order statistics of a standard normal distribution.

Many Q-Q

Note: ^ plots in Faraway Ch. 6 use "raw" residuals rather than standardized residuals.

See p. 85.

An issue with $\hat{\varepsilon}_i$: if the i^{th} data point is "influential" it "pulls" the line towards it, decreasing $\hat{\varepsilon}_i$.

We can adjust for this with "studentized" or "jackknife" residuals.

Remember that the "jackknife means "leave one out."

We could exclude the ith data point and compute coeff. estimates $\hat{\beta}_{(i)}$, and compute $\hat{\sigma}^2_{(i)}$

The (i) is shorthand for "exclude the i^{th} data point".

Then we could treat the ith data point as new data and compute

$$\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$$

It can be shown that

$$\text{Var}\left[y_i - \hat{y}_{(i)} \right] = \hat{\sigma}^2_{(i)} \left(1 + x_i^T \left(X_{(i)}^T X_{(i)} \right)^{-1} x_i \right)$$

Motivated by this we define the jackknife residuals as

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\text{Var}\left[y_i - \hat{y}_{(i)} \right]^{1/2}}$$

If the model is true (in particular $\varepsilon \sim N(0, \sigma^2 I)$)

t_i has the student t-dist with $n-p-1$ d.o.f.

This justifies the name "studentized".

Mercifully, there is a way to compute these without doing n separate regressions.

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}} = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2}.$$

In R, we have rstudent(lmod)

analogous to rstandard(lmod) and residuals(lmod).

We can now define Cook's dist. of the i^{th} point:

$$D_i = \frac{\| \hat{y} - \hat{y}_{(i)} \|^2}{p\hat{\sigma}^2} = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{p\hat{\sigma}^2}$$

$$= \frac{1}{p} r_i^2 \frac{h_i}{1-h_i}$$

r_i^2: standardized residual

$\frac{h_i}{1-h_i}$: leverage.

No need to compute; in R: cooks.distance(lmod)

What should be done about outliers / high-leverage / high-influence points?

(1) Check for data entry errors.

(2) Examine the context: Why is this measurement what it is?

(3) ⚠ This can be wrong sometimes:

Exclude the point.

(4) Robust regression: the assumption that errors are normal, is just that: an assumption.

We can assume that the errors have some other distribution.

We can do least - < something else >

instead of least-squares, etc. See Faraway Ch. 8.

(5) ⚠ Dangerous to exclude outliers automatically.

Examples from 3(c):

(A) "Even if a predictor is not statistically significant, it may become statistically significant when combined with other predictors".

Why this is not right:

It is true that it could be that the p-value associated to the statistical test of

$H_0: \beta_{X2} = 0$ vs. $H_a: \beta_{X2} \neq 0$

could be less than 0.05 (say) (standard threshold for significance)

While if we add another predictor X3, that the p-value associated to this new test of

$H_0: \beta_{X2} = 0$ vs. $H_a: \beta_{X2} \neq 0$

is less than 0.05.

But this is NOT what is going on in 3(c).

Notice that we have interpreted

"a predictor is not statistically significant"

to mean

"the p-value associated to the statistical test of $H_0: \beta_{X2} = 0$ vs. $H_a: \beta_{X2} \neq 0$ is less than 0.05 (in the context of the current model)".

(B) "The p-value 0.216 is associated to a different hypothesis test from the hypothesis test corresponding to the p-value 0.000203 for the F-test".

Why this is not right:

Again, the statement is true, but the problem with this answer is that it doesn't say what the hypothesis tests are. In principle, the rejection of the null for the F-test

Could imply the failure of the null for the test of $H_0: \beta_{X2} = 0$ vs. $H_a: \beta_{X2} \neq 0$.

Examples from 2 (b):

"Holding size of the house constant, an increase of \$1000 in the sale price of the house corresponds to a decrease in the age of the house by 2 years."

Why this is not right:

The response and the predictor are reversed.

"An increase of 1 year in the age of the house makes the price fall by \$2,000."

Why this is not right:

Causal relationship is implied. No statement about the variable size in the regression.