Math 455 Class 20 March 11

Last time: 4 standard diagnostic plots obtained with plot(lmod)

1. Fitted vs. Residuals
2. Q-Q
3. Scale-Location (Fitted vs. $\sqrt{\text{Std. Resid.}}$)
4. Leverage vs. Standardized Residuals. (Cook's Distance)

In 1: Any visible pattern suggests an improvement to the model.

In 2: Mistake from last time: I wrote "residuals" instead of "standardized residuals". Notice: in Faraway p. 79, there is a graph of sorted residuals (not standardized) vs. theoretical quantiles $\Phi^{-1}\left(\frac{i}{n+1}\right)$, which looks similar. This is not what R produces.

R uses standardized residuals.

Q: What is a standardized residual?

A: A rescaling so that the residuals have standard deviation 1. (if model correct)

Notation: (in Faraway) $\hat{\varepsilon}_i$ (ordinary) residual [also $\hat{\varepsilon}_i$]

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

r_i: Standardized residual

where $h_i = H_{ii}$ is the ith diagonal entry of the hat matrix.

h_i is called the "leverage" of the ith data point.

Notice that the Faraway graph (p. 79) and the automatically produced graph (plot (lmod)) look very similar.

R code: standardized residuals rstandard(lmod)
ordinary residuals residuals(lmod).

What is "leverage" and why is it called that?

Idea: leverage is a measure of influence on the "outcome" of the regression.

Alternative definition of leverage: $h_i = \frac{\partial \hat{Y}_i}{\partial Y_i}$.

Why does this make sense?

Think of regression as a big function that takes as input the data and produces a vector

$$\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n)^T$$

Why is this the same as $h_i = H_{ii}$?

Remember that $\hat{Y} = HY$ where H = hat matrix.

Think of a simple case (this will generalize)

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = \begin{pmatrix} h_{11} y_1 + h_{12} y_2 \\ h_{21} y_1 + h_{22} y_2 \end{pmatrix}$$

$$\frac{\partial \hat{y}_1}{\partial y_1} = h_{11}$$

$$\frac{\partial \hat{y}_2}{\partial y_2} = h_{22}.$$

Notice: $H = X(X^TX)^{-1}X^T$.

this is the projection matrix on the column space of X.

that is, the leverage of the i^{th} data point only depends on the X-values and NOT on the Y-values.

Idea: <u>High-leverage</u> point are ones that have X-values far from the average.

<u>Notice</u>: this is NOT the same concept as "outlier". An outlier is a point that does not fit well with the model.

⚠ A data point can be high-leverage, an outlier, or both, or neither.

Why do we care about leverage?

Remember that regression = least-squares

Squaring weights large deviations more highly.

Suppose we have residuals $-1, -2, 5, \ldots, 100$.

The sum of squares is mostly about the 100.

Regression will alter the line to reduce that 100!

The point that is "producing" the 100 will have more influence on the regression line.

We want to be sure that the result of the regression does not depend on just one (or a few) data points.

(Discussion of Gino v. Harvard legal case.)

How do we know if the leverage is "too big"?

We can get the leverages with hatvalues (lmod).

Rule of thumb: leverages $> \frac{2p}{n}$

(sometimes: $\frac{3p}{n}$) should be investigated.

p = # parameters n = # data points.

Why should this be the cutoff?

A: The leverages h_i sum to p:

$$\sum_{i=1}^{n} h_i = p.$$

The rule says: investigate points whose leverage is 2x (or 3x) the average..

Why is this so? Linear algebra.

$$\sum_{i=1}^{n} h_i = \sum_{i=1}^{n} H_{ii} = \mathrm{Tr}(H).$$

Fact from Linear Algebra: $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

$$H = \underbrace{X}_{A}\underbrace{(X^TX)^{-1}X^T}_{B} \qquad \mathrm{Tr}(H) = \mathrm{Tr}(AB) = \mathrm{Tr}(BA)$$

$$= \mathrm{Tr}\left((X^TX)^{-1}(X^TX)\right)$$

$$= \mathrm{Tr}(I)$$

X^TX is a $p \times p$ matrix.

because X is $n \times p$.

Thus $\mathrm{Tr}(H) = p.$, with our standard assumption that X^TX is invertible..

Alternative argument: H is a projection matrix: $H^2 = H$. (In fact, orthogonal projection.)

The trace is the sum of the eigenvalues.

The eigenvalues of a projection matrix are 0 or 1, because if $Hv = \lambda v$, $Hv = H^2 v$ $= \lambda^2 v$, so $\lambda^2 = \lambda$, so $\lambda = 0$ or 1.

But $X^T X$ is invertible, so the col space of X has dim p, so no eigenvalue can be 0.

thus $\sum$ eigenvalues $= \sum_{i=1}^{p} 1 = p$.

Point: Average leverage is p/n.

"high-leverage" means high relative to p/n.

Another way of seeing which points are influential:

"Cook's Distance": the 4th plot obtained with plot(lmod).

This shows: leverage vs. standardized residuals.

Idea: A high-leverage point need not be a high-influence point.

An outlier need not be a high-influence point.

High-influence is a concept separate from these two.

Cook's Distance is a computable measure of influence obtained from leverage and standardized residuals.

$$(i^{\text{th}} \text{ Cook's Distance}) = D_i = \frac{1}{p} r_i^2 \frac{h_i}{1-h_i}$$

Notice: Cook's Distance is NOT one of the axes on the graph

Rule of thumb: investigate points with Cook's Distance > 0.5.