Biostatistics 201a - Lecture 22 - 11/16/11

- Model Fit Example
- Regression Assumptions
- Histograms, QQ plots, Residual Plots

pages = 8

Note: Project Materials Posted earlier this week - due on last day of class

Evaluating Model Fit II: Regression Assumptions

11/16/11

1. Model Assumptions
2. Outliers, Leverage Points and Influential Points

Regression Assumptions: 4 main ones, all about distribution of the error terms, ε_i's

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

this is where all the randomness/uncertainty is

We assume the εs are

(a) <u>Mean 0</u> for all possible combinations of the X values: this tells us if we have the right shaped model or are systematically off somewhere; leads to unbiased estimates of the βs

(b) Normally distributed at each value of X; lets us use t and F distributions for our tests and also makes sure that the least squares estimates of βs are the maximum likelihood estimates

(c) Independent: no relationship between error at one point and error at another point; can think of this as like saying (conditional on right model) the points represent a random sample. If you fit the wrong model this shows up as a systematic pattern in your errors

(d) Homoscedastic: variation of the points about the regression line (or plane) is the same (constant) for all combinations of X values. Nonconstant variance is called heteroscedasticity. Homosc. makes computing s.e.s easier

This also makes RMSE a sensible number - it's our estimate of that common standard deviation about the line.

How do we check these assumptions? We use "estimated" errors, called residuals - i.e. what's left over after we fit the regression equation. For subject i, the residual is

$$e_i = Y_i - \hat{Y}_i = \text{how far is point from line}$$

Y_i: observed value for subject i

$\hat{Y}_i$: value predicted by the model for subject i

We draw pictures of the e_i's and also compute various statistics to see how big they are (big is bad.)

(b) Normality: 2 plots for this - a histogram and a qq- or normal quantile plot.

Histogram: counts the # of errors in each of a set of ranges or bins

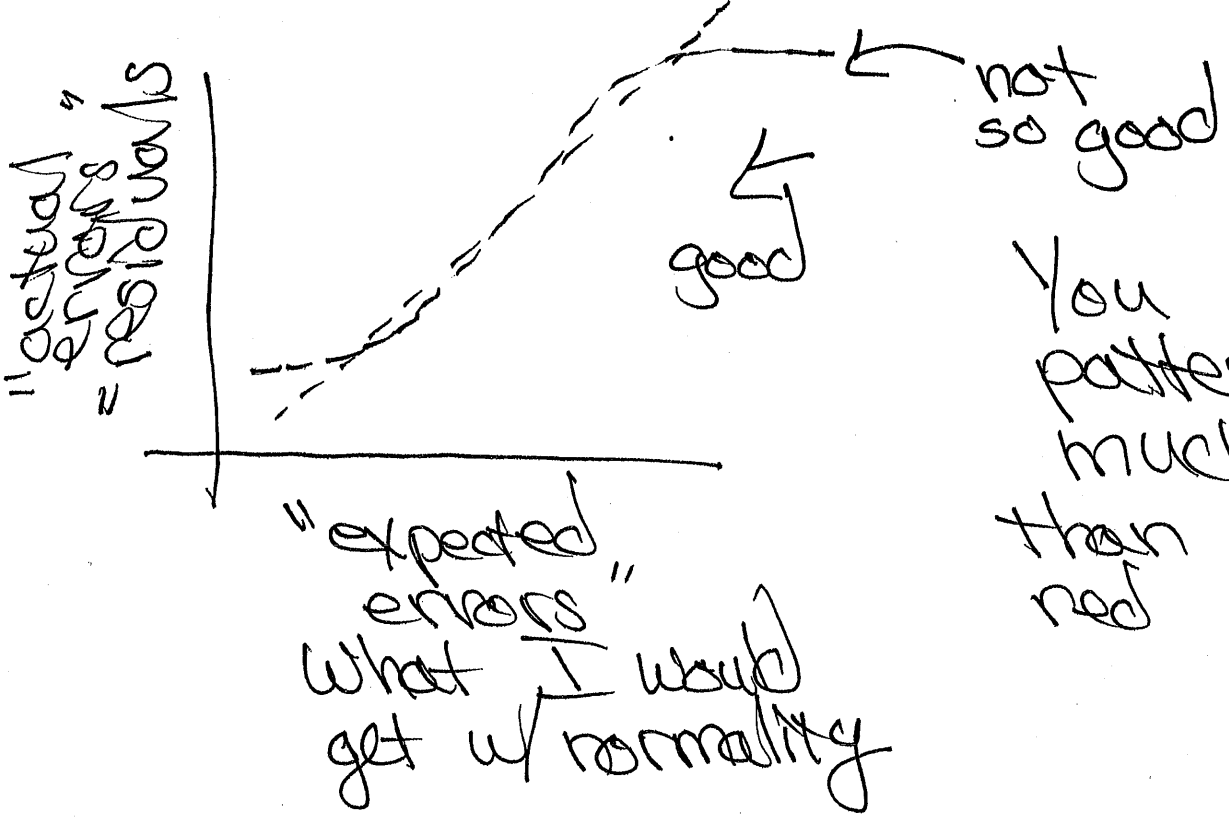
Want this to look bell-shaped!

Hard to judge if shape is really normal, especially for small samples.

Alternative is a quantile-quantile or a normal quantile plot.

Idea: plot the errors you expected or should have gotten if they were normal versus the ones you actually got.

If normality is correct this should produce a straight line



You can get patterns much worse than the red line

Usually regression inference is pretty robust to moderate violations of normality so you only get upset if line is really not straight.

Other three assumptions (a), (c) & d)) are checked either with a scatterplot (in SLR, one X) or a "residual plot"

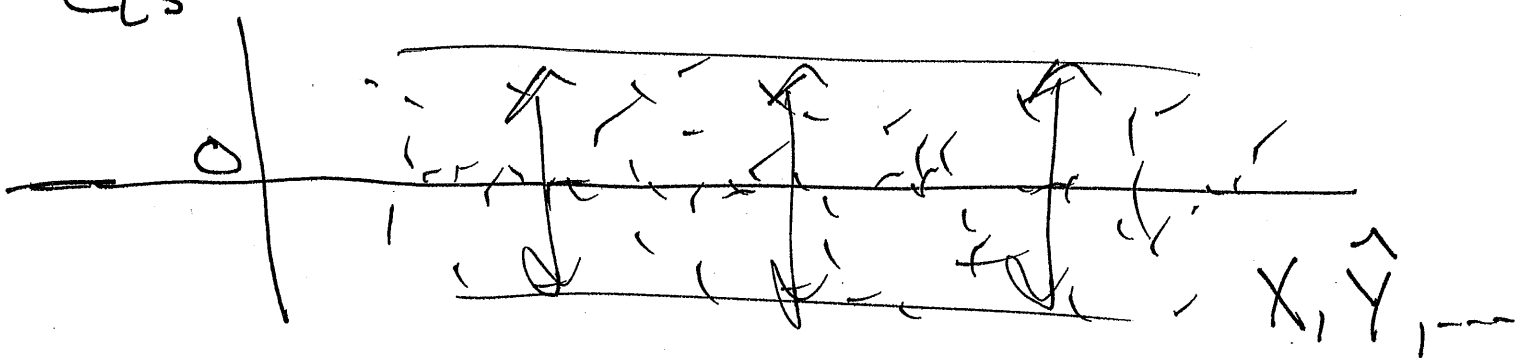
Residual plots show residual values on the Y axis versus the following sorts of things on the X axis

- each predictor variable
- fitted values, $\hat{Y}$'s
- Index variables - e.g. order in which data were obtained.

We want to see "random scatter" - a clear pattern means we have a problem

"Random scatter" means centered about 0 with the same spread for each X-axis value on the plot, more points near 0, no clear patterns

e_i's



Not good enough to judge model just on F, RMSE, R^2_{adj}

$\hat{Y} = 3 + .5X$ in all models (pg. 117 text)

Same F, R^2, RMSE, fitted line.