Last Time:

- Data cleaning considerations (see online handout)
  - Always identify and consider explanations for outliers, may not be able to remove
  - Document explanations for any observations removed (entire group)
- Prediction Intervals and Confidence Intervals
  - Confidence interval for $\mu_y|x^*$: $\hat{y} \pm t_{n-2} se_{fit}$
  - Prediction interval for $y$ at $x^*$: $\hat{y} \pm t_{n-2} \sqrt{se_{fit}^2 + s^2}$

Factors that affect width of intervals: $s$, $n$, $sx^2$ (which affect variability in lines) and $(x^* - \bar{x})$

The basic regression model has the following conditions:

L: There is a linear relationship between the means at each $x$ and $x$: $E(Y|X=x) = \beta_0 + \beta_1 x$

I: The observations (errors) are independent (uncorrelated)

N: The responses follow a normal distribution for each value of $x$: $Y|X \sim \text{Normal}$

E: The variance of the responses is the same for each value of $x$: $V(Y|X=x) = \sigma_e^2$

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim \text{N}(0, \sigma_e)$

How do we verify the distribution of the $\epsilon_i$?

- Can tell a lot from the scatterplot of $y$ vs. $x$
- Can tell even more from a scatterplot of the residuals ($e_i = y_i - \hat{y}_i$) vs. $x$.

Though not uncorrelated and have to worry about estimating variability…

Demo: Scatterplots vs. Residual plots (stride.mtw)
Residual Plots: *(see online instructions)*
Typically you can use either residuals or standardized or studentized residuals

- **L:** If the relationship between $E(Y)$ and $X$ is linear, there should not be any pattern in the residuals (what’s leftover after you take out the linear component) vs. $X$ graph.
- **E:** If the variability is constant, the graph of residuals vs. $X$ should show a band of equal width. That is, the distances of the residuals away from the 0 line should be the same across the length of the graph (as you change $x$ values). Don’t worry about a few values outside the box. Slightly easier to read: $|e|$ or $e^2$ vs. $\hat{y}$ or $x$.

- **N:** If the distribution of the response variable at each $x$ value is normal, then the residuals themselves should follow a roughly normal distribution. “It is a matter of determining what constitutes an extreme configuration of residuals relative to the boundaries, and that determination has to be somewhat of an art.”

- **I:** Independence is judged less by looking at graphs and more by looking at data collection methods. But if there is an ordering to your data you can look at residuals vs. observation number. You want this to look random.

**Example 1: TVs and Life Expectancy**
Data from the *2006 World Almanac and Book of Facts* includes the life expectancies of people in different countries and the number of televisions per thousand people in the country (TVLife.txt).
(a) Create a scatterplot predicting life expectancy from number of televisions per thousand people. Identify the country with the most TVs per 1000 people.

(b) **Add a smoother to the plot.** Describe the nature of the relationship (direction, form, strength, unusual observations).

(c) Fit a linear model, storing the residuals and the predicted (fitted) values *(see online instructions)*. Is this relationship statistically significant? What does that tell you?

(d) Create a scatterplot of the residuals vs. the explanatory variance (*TVs per K*). Comment on the behavior observed in this plot.
(e) Now examine a histogram of the residuals or, better yet, look at a normal probability plot of the residuals. Comment on the behavior observed in these graphs. Also examine and evaluate a statistical test of normality. (Keep in mind that the graphs are often more useful than the test for pinpointing the issue and suggestion solutions.)

(f) Now examine a scatterplot of the residuals vs. the fitted values for the TV life data. How does the behavior of this plot differ from the plot in (d), residuals vs. explanatory variable?

(g) Now examine the “four in one” (Minitab) or “plot residuals” (JMP Fit Y by X) or “plot(model)” (R) options. Notice the “advanced moves” in the online instructions for Minitab and JMP.

Example 2: Return to the cleaned NBA point guard efficiency data.
(a) Examine the residual plots and write a summary of your analysis.

(b) Not fit a quadratic model (see online instructions). Summarize the behavior of the model. Does the model fit better? How are you deciding? Does the model appear to be more valid? How are you deciding? Are Chris Paul and Russell Westbrook still outside the prediction bands?

Practice Problem (due by 9am Tuesday)
(c) Compare the linear and the quadratic models you find if you first remove Chris Paul and Russell Westbrook from the data set. Which model (linear vs. quadratic, with or without these observations) would you recommend for predicting future performances? Explain.

Notes:
- The normal correlation model assumes bivariate normality for X and Y. This assumption implies conditional normality at each x.
- Don’t worry too much about small departures from the normality assumption, especially with the extreme values (is it a systematic problem or an outlier or two?)
- “Some experience is required to interpret normal probability plots … Usually 20 points are required to produce normal probability plots that are stable enough to be easily interpreted.”
- We will come back to standardized residuals.
- Worry about the linearity and constant variance conditions first. Can often be less concerned with the normality condition (unless making prediction intervals). Though might impact your willingness to focus on the means of the conditional distributions.
Which of these are considered “linear”?

- \( E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 \)
- \( E(Y) = \beta_0 + \beta_1 10^x \)
- \( E(Y) = (\beta_0 + \beta_1 x)/(\beta_0 + \beta_2 x) \)
- \( E(Y) = \beta_0 e^{\beta_1 x} \)

Display 8.6 from *Statistical Sleuth* (Ramsey & Schafer, 2002)

Some hypothetical scatterplots of response versus explanatory variable with suggested courses of action; (A) is ideal.