Reading: Section 12.2 and 12.3 of Devore.

Summary: In this class we study how to evaluate, more mathematically, whether it is worth it to fit the simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.

To mathematically determine whether the simple linear regression model accounts for a significant amount of variability in the response, $y$, we test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. This makes sense because if we fail to reject $H_0$ that means there is no evidence that $\beta_1 \neq 0$, i.e., there is no evidence that $y$ varies with $x$. The hypothesis $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ can be tested in two ways. Today we discuss how to test this hypothesis using analysis of variance (ANOVA) methods.

To test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ using ANOVA methods, we try to compare how much $y$ varies as a result of just random error to how much $y$ varies as a result of the variation in $x$. The variation in $y$ as a result of random error is estimated as $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ and is referred to as the error sums of squares (SSE). The variation in $y$ as a result of the variation in $x$ is calculated as $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ and is referred to as the regression sums of squares (SSR). The average variation as a result of random error is referred to the mean squared error (MSE) and is calculated as $\text{SSE}/(n-2)$, and the average variation as a result of the variation in $x$ is referred to as the mean squared regression (MSR) and is calculated as $\text{MSR} = \text{SSR}/1 = \text{SSR}$. If $H_0$ is true and $\beta_1 = 0$, then $\text{MSR}/\text{MSE} \sim F_{1, n-2}$. The p-value is thus calculated as $P(F > \text{MSR}/\text{MSE})$ where $F \sim F_{1, n-2}$. All of these methods are organized in the analysis of variance (ANOVA) table given below:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR</td>
<td>MSR</td>
<td>$\text{MSR}/\text{MSE}$</td>
</tr>
<tr>
<td>Error</td>
<td>$n-2$</td>
<td>SSE</td>
<td>MSE</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td>SST</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This ANOVA table is simple to produce in R. If you are regressing $y$ on $x$ you can simply enter the following commands:

```r
> reg = lm(y ~ x)
> anova(reg)
```

**Homework:** This assignment is due at the beginning of class on May 12, 2010

1. Look at the data set on problem 5 on page 453.
   (a) Plot $x$ vs. $y$ in R.
   (b) Fit the simple linear regression $Y = \beta_0 + \beta_1 x + \epsilon$ to the data. What are your estimated values of $\beta_0$ and $\beta_1$?
   (c) Interpret the values of $\beta_0$ and $\beta_1$ in this model.
   (d) Based on the graph, do you believe the linear model you fit to the data is appropriate?

2. Devore, page 454, #10. Recall that $Y|x \sim (\beta_0 + \beta_1 x, \sigma^2)$

3. Devore, page 467, #26. All you have to do in this problem is show that the point $(\bar{x}, \bar{y})$ goes through the line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

4. The data for this problem will either be sent to you, or will be on the course website. The data set contains the stopping distances of cars traveling at certain speeds. With this data set, use simple linear regression to predict the stopping distance for a car traveling 40mph. Note: to successfully complete this problem, you might have to re-express the data until you get a pattern that looks straight!

5. A student experimenting with a pendulum counted the number of full swings the pendulum made in 20 seconds for various lengths of string. I will either e-mail you this data set or put it on the web. With this data set, use simple linear regression to predict the number of swings for a pendulum with a 20 inch string.