Some Datasets (Stat2Data package)

<table>
<thead>
<tr>
<th>Dataframe</th>
<th>Response</th>
<th>Explanatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backpack</td>
<td>BackpackProblems</td>
<td>BackpackWeight</td>
</tr>
<tr>
<td>MedGPA</td>
<td>Acceptance</td>
<td>GPA, MCAT, Sex</td>
</tr>
<tr>
<td>TMS</td>
<td>pain-free (Yes or No)</td>
<td>Group</td>
</tr>
<tr>
<td>Titanic</td>
<td>Survived</td>
<td>Age, Sex, PClass</td>
</tr>
<tr>
<td>CAFE</td>
<td>vote</td>
<td>LogContr, Party</td>
</tr>
</tbody>
</table>

1. A categorical response variable with two levels.
2. What would a model do? Two choices: classify, give a likelihood

3. Translating a quantitative response to a probability – the logistic function, computed by \( \text{ilogit()} \)

\[
p = \frac{e^y}{1 + e^y} \quad y = \log \left( \frac{p}{1 - p} \right)
\]

4. A general linear model:
   (a) fit a model \( y = f(x) \) to get a “link” value \( y \)
   (b) use a link function (e.g., \( l(y) \) to transform the response)

5. Example:

\[
\begin{align*}
> & \text{data(MedGPA)} \\
> & \text{linear = lm(Acceptance ~ GPA, MedGPA)} \\
> & \text{logistic = glm(Acceptance ~ GPA, MedGPA, family = binomial)}
\end{align*}
\]

6. How to fit model? Not least squares but maximum likelihood.

\[
\begin{align*}
> & \text{dist = c(1, 3, 5, 8, 10, 12)} \\
> & \text{made = c(1, 1, 0, 1, 0, 0)}
\end{align*}
\]
Test, Friday, April 26

In format, the test will be much like the first test. Namely, there will be a closed-book portion with short answer, multiple choice, fill-in-the-blank kinds of problems. These will focus on the conceptual aspects of the chapters covered and many problems will involve reading output from R and interpreting it correctly. There will be an open-book, open-computer portion that will require you to fit models with R and answer questions about them. All the R you need to know is in the last section of each of the four chapters together with a few extra things on the daily handouts. We've been using \texttt{lm()} a lot so you should be comfortable with it and the related functions (such as \texttt{summary()}, \texttt{fitted()}, \texttt{anova()}, \texttt{confint()}, \texttt{makeFun()}) The material covered includes:

1. In the book: Chapters 12–15. Chapter 17 will not be covered.
2. In the daily activities: Day 29 – Day 41.

Here is a brief summary of the key ideas of the four chapters covered (of course you are responsible for everything that we covered in these chapters.)

- **Chapter 12.**
The key idea here was what we learn about the precision of our estimates of the coefficients in a linear model using the sampling distribution. The key ideas are standard error and confidence interval. Be aware of the fact that we compute these using models for the sampling distribution. One way we compute these is to use the regression report or \texttt{confint()} which makes some strong assumptions about the errors in our model. Another way we compute these is to use bootstrapping. We also considered two other kinds of intervals: a confidence interval for the mean response given fixed values of the explanatory variables and a prediction interval for a new value of the response. Know the difference!

It's usually tricky for students to be able to say exactly what our confidence is in. Be able to write a sentence that says what 95% confidence means and be able to do that in the context of the particular data that you are studying.

- **Chapter 13**
Obviously the key idea here was hypothesis testing. You’ll be in good shape if you really understand the glossary of Section 13.7. Often people using statistical tests won’t explicitly say what the null hypothesis is or even what their explicit conclusion is but will just present the \textit{p}-value. Know how to go from a \textit{p}-value to a clear statement about the real world phenomena. Really try to understand the logic so that you can say confidently what the null hypothesis should be in a given situation.

- **Chapter 14**
Three huge ideas – the permutation test, the use of $R^2$ to evaluate the model, and the $F$-statistic. Really understand Figures 14.2 and 14.3 which explain why the $F$-statistic should be approximately equal to 1 if the explanatory terms in the model are no better at explanation than random terms. Don’t worry about the exact formula for computing $F^2$ but do understand it as a ratio of two slopes. Be able to explain why, if you have 20 data points, you are excited about $R^2 = 50\%$ if you have just two explanatory variables in the model but not excited at all if you have 10 explanatory variables in the model. Be able to read an ANOVA report like the one on page 263 and conclude the appropriate thing about the model from it.

- **Chapter 15 (we did not do 15.6 and 15.7 and you are not responsible for these)**
Understand what we were looking for when we were choosing between two different models. Be able to explain what the \textit{p}-values in a regression report such as the one on page 285 or in ANOVA reports that compare two models (this is something we did but the book didn’t – see the outline on April 18)

I will try to make this test so that it can be comfortably completed in 50 minutes (that was true of the first test for many of you), but you are welcome to work through the chapel break if you find that you need or want the time. Sometimes getting the right computation takes longer than we think!