Please see the HW policy on the course syllabus. Every student must write up his or her own solutions using his or her own words, symbols, calculations, etc. Copying of the work of others is a violation of academic ethics and will be handled as such.

**Homework #1 (Due September 4)**

1. Section 1 Exercises (all) Appendix A.1 of Vardeman and Jobe (pages 743-745).
2. Section 1 Exercises 1, 2, 4, 5, 6, 7, 8, 9 Section 5.1 of Vardeman and Jobe (pages 243-244).

**Homework #2 (Due September 18)**

1. Section 2 Exercises (all) Section 5.2 of Vardeman and Jobe (page 263).
2. Section 4 Exercises 2, 3, 4, 5 Section 5.4 of Vardeman and Jobe (pages 300-301).
3. Section 5 Exercises 1, 3, and 4 Section 5.5 of Vardeman and Jobe (pages 321 and 322).

**Homework #3 (Not to be Collected but Covered on Exam 1 September 24)**

1. Section 1 Exercises (all) Section 6.1 of Vardeman and Jobe (page 344).
2. Section 2 exercises (all) Section 6.2 of Vardeman and Jobe (page 361).

**Homework #4 (Due October 9)**

1. Section 3 Exercises (all) Section 6.3 of Vardeman and Jobe (page 385).

For parts a) and d) of Exercise 2, "normal probability plot" of Sections 3.2 and 5.3 are intended. (Read those sections.) There is code on the course Web page for using R to make 2 normal plots on the same set of axes (and add appropriate lines to the plots). The data vectors in the code are named x1 and x2 and you will have to replace the example data vectors with ones of interest. (You can make also the "obvious" edits to the code to make a single plot or more than 2 plots when such are needed.)

For Exercise 4 part b) do the testing first assuming that the two types of tires have the same variability of skid length (as in Example 8 page 382 of V&J). Then redo it dropping that assumption and operating as indicated on the course summary of one and two sample inference formulas and using the simplified degrees of freedom formula. For Exercise 4 part d) first use the unpleasant "Satterthwaite approximate degrees of freedom formula" and then redo the problem using the simplified degrees of freedom.

2. Section 4 Exercises (all) Section 6.4 of Vardeman and Jobe (page 399).
3. Section 5, Exercises (all) Section 6.5 of Vardeman and Jobe (page 413). For the all confidence limits problems, use the modification of the book's formulas provided on the course formula summary sheets. This means that in place of products $\hat{p}(1 - \hat{p})$ in the "plus and minus" parts of the formulas (NOT in place of center values $\hat{p}$) use products $\hat{p}(1 - \tilde{p})$. (This will have the effect of making the confidence intervals a bit larger/more conservative so that they will have actual confidence at least the nominal ones. Without this "fix" the formulas in the text will tend to be a bit too short and have actual confidence levels smaller than the nominal ones for extreme values of the unknown $p$'s.)

4. Section 6 Exercises 2 and 3 Section 6.6 of Vardeman and Jobe (pages 426 and 427).

**Homework #5 (Due October 16)**

1. Section 1 Exercises (all) Section 7.1 of Vardeman and Jobe (page 460).

2. Section 2 Exercises 1 and 2, Section 7.2 of Vardeman and Jobe (pages 471 and 472).

3. Section 4 Exercises 1b) and 2b), Section 7.4 of Vardeman and Jobe (page 495).

4. Section 1 Exercises 1,2, and 3, Section 4.1 of Vardeman and Jobe (pages 139-140).

5. Chapter 4 Exercise 1 of Vardeman and Jobe (page 203).

6. Section 1 Exercises 1a), 1b), 1c), 1d), 1f), 1h), and 2a) through 2f), Section 9.1 of Vardeman and Jobe (page 674).

**Homework #6 (Due November 13)**


Download that data set and use the first 146 cases (those corresponding to Glass #1 and Glass #2) in the following. Suppose that samples of glass are presented to an analyst in about the same relative frequencies as in the UCI data set (at fractions 70/145 of type 1 and 75/145 of type 2). Based on $x_1, x_2, \ldots, x_9$ we want to say with what probability a specimen in hand is of glass type 2. (I'll put a version of the data set on the 401B web page.)

b) Fit a logistic regression model based on all 9 predictors to the data. Identify any of the 
predictors that (in the presence of all others) might be dropped from the model. What is your 
rationale for choosing/identifying these?

c) Make side-by-side box plots for the fitted probabilities of being a Glass #2 specimen for the 2 
groups of (actual Glass #1 and actual Glass #2) specimens based on the full model of b). 
(Remember that graphics like this were produced on Lab #5.) Does it seem like logistic 
regression provides a sensible way to tell whether a specimen is of type #2? Explain.

d) Now fit a logistic regression model that includes only the predictors Na, Mg, Si, and Ca to the 
data. Make a plot of the type requested in part c). Does the plot here look much worse than the 
plot in c)? Explain. The "residual deviance" for a fitted logistic regression model is more or less 
an analogue of the "error sum of squares" for ordinary MLR. Locate and report values for this 
for both the full model of b) and for this reduced logistic regression model. Is the increase in 
"residual deviance" observed here consistent with the plots here and in c)? Explain. Do you 
consider the increase to be "severe"? (The "null deviance" is more or less the analogue of the 
"total sum of squares" for ordinary MLR.)

e) Interpret the signs of the coefficients in the fit of part d). (Relative to Glass#1, do Glass#2 
specimens tend to have higher or to have lower values of the predictors?)

f) Purely for purposes of exercise, not because it's a good model here, fit a logistic regression 
model that includes only the predictors Na and Ca to the data. Then make a plot on the Na-Ca 
plane locating the lines (the sets of (Na, Ca) pairs) for which the model says that the 
probabilities of a specimen being a Glass #2 specimen are .1,.3,.5,.7, and .9.

2. The book *Nonlinear Regression Analysis and its Applications* by Bates and Watts contains a 
small data set taken from an MS thesis of M.A. Treloar "Effects of Puromycin on 
Galactosyltransferase of Golgi Membranes." It is reproduced below. $y$ is reaction velocity (in 
counts/min$^2$) for an enzymatic reaction and $x$ is substrate concentration (in ppm) for untreated 
enzyme and enzyme treated with Puromycin.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0.02</th>
<th>0.06</th>
<th>0.11</th>
<th>0.22</th>
<th>0.56</th>
<th>1.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untreated</td>
<td>67, 51</td>
<td>84, 86</td>
<td>98, 115</td>
<td>131, 124</td>
<td>144, 158</td>
<td>160</td>
</tr>
<tr>
<td>Treated</td>
<td>76, 47</td>
<td>97, 107</td>
<td>123, 139</td>
<td>159, 152</td>
<td>191, 201</td>
<td>207, 200</td>
</tr>
</tbody>
</table>

Apparently a standard model here (for either the untreated enzyme or for the treated enzyme) is 
the "Michaelis-Menten model"
\[
y_i = \frac{\theta_1 x_i}{\theta_2 + x_i} + \epsilon_i \quad \text{(*)}
\]

Note that in this model, 1) the mean of \( y \) is 0 when \( x = 0 \), 2) the limiting (large \( x \)) mean of \( y \) is \( \theta_1 \), and 3) the mean of \( y \) reaches half of its limiting value when \( x = \theta_2 \).

Begin by considering only the "Treated" part of the data set (and an iid N\( (0, \sigma^2) \) \( \epsilon_i \)'s version of the model). Of course, use R to help you do all that follows. Begin by reading in 12\( \times \)1 vectors \( y \) and \( x \).

\textbf{a}) Plot \( y \) vs \( x \) and make "eye-estimates" of the parameters based on your plot and the interpretations of the parameters offered above. (Your eye-estimate of \( \theta_1 \) is what looks like a plausible limiting value for \( y \), and your eye-estimate of \( \theta_2 \) is a value of \( x \) at which \( y \) has achieved half its maximum value.)

\textbf{b}) Add the \texttt{nls} package to your R environment. Then issue the command

\begin{verbatim}
> REACT.fm<-nls(formula=y~theta1*x/(theta2+x),start=c(theta1=#,theta2=##),trace=T)
\end{verbatim}

where in place of \# and ## you enter your eye-estimates from \textbf{a}). This will fit the nonlinear model (\*) via least squares. What are the least squares estimate of the parameter vector and the "deviance" (error sum of squares)

\[
\hat{\theta}_{OLS} = \left( \frac{\hat{\theta}_1}{\hat{\theta}_2} \right) \quad \text{and} \quad SSE = \sum_{i=1}^{12} \left( y_i - \left( \frac{\hat{\theta}_1 x_i}{(\hat{\theta}_2 + x_i)} \right) \right)^2
\]

\textbf{c}) Re-plot the original data with a superimposed plot of the fitted equation.

\textbf{d}) Get more complete information on the fit by typing

\begin{verbatim}
> summary(REACT.fm)
\end{verbatim}

\textbf{e}) The concentration, say \( x_{100} \), at which mean reaction velocity is 100 counts/min\(^2\) is a function of \( \theta_1 \) and \( \theta_2 \). Find a sensible point estimate of \( x_{100} \).

\textbf{f}) As a means of visualizing what function the R routine \texttt{nls} minimized in order to find the least squares coefficients, do the following. First set up a grid of \((\theta_1, \theta_2)\) pairs as follows. Type

\begin{verbatim}
> theta<-coef(REACT.fm)
> se<-sqrt(diag(vcov(REACT.fm)))
> dv<-deviance(REACT.fm)
\end{verbatim}
> gsize<-101
> th1<-theta[1]+seq(-4*se[1],4*se[1],length=gsize)
> th2<-theta[2]+seq(-4*se[2],4*se[2],length=gsize)
> th<-expand.grid(th1,th2)

Then create a function to evaluate the sums of squares

> ss<-function(t){sum((y-t[1]*x/(t[2]+x))^2)}

As a check to see that you have it programmed correctly, evaluate this function at $\hat{\theta}_{\text{OLS}}$ for the data in hand, and verify that you get the deviance. Then evaluate the error sum of squares over the grid of parameter vectors $\theta$ set up earlier and produce a contour plot using

> SumofSquares<-apply(th,1,ss)
> SumofSquares<-matrix(SumofSquares,gsize,gsize)
> plot(th1,th2,type="n",main="Error Sum of Squares Contours")
> contour(th1,th2,SumofSquares,levels=c(seq(1000,4000,200)))

What contour on this plot corresponds to an approximately 90% approximate confidence region for the parameter vector $\theta$?

g) Now redo the contour plotting, placing only two contours on the plot using the following code.

> plot(th1,th2,type="n",main="Error Sum of Squares Contours")
> contour(th1,th2,SumofSquares,levels=dv*c((1+.1*qf(.95,1,10)),
            (1+.2*qf(.95,2,10))))

Identify on this plot an approximately 95% (joint) confidence region for $\theta$ and individual 95% confidence intervals for $\theta_1$ and $\theta_2$. (If you draw a rectangle with sides parallel to the coordinate axes around the region defined by the first (lower) contour, the extent of the box in each direction gives the 95% individual confidence limits. The second (higher) contour defines the joint 95% confidence region.)

(By the way, it would have been possible to simply add these contours to the first plot, by making the second call to `contour()` as above, except for setting "add=T" as a parameter of the call.)
h) Use the standard errors for the estimates of the coefficients produced by the routine \texttt{nls()} and make 95\% \textit{t} intervals for $\theta_1$ and $\theta_2$. How much different are these from your intervals in g)?

(Notice that the sample size in this problem is small and reliance on any version of large sample theory to support inferences is tenuous. I would take any of these inferences above as very approximate.

i) Make an approximate 95\% confidence intervals for $\sigma$ by carrying over the MLR linear model result that $\text{SSE} / \sigma^2 \sim \chi^2_{n-k}$.

j) Use the R function \texttt{confint()} to get 95\% intervals for $\theta_1$ and $\theta_2$. That is, add the \texttt{MASS} package in order to get access to the function. Then type

> \texttt{confint(REACT.fm, level=.95)}

How do these intervals compare to the ones you found in part g)?

k) Apparently, scientific theory suggests that treated enzyme will have the same value of $\theta_2$ as does untreated enzyme, but that $\theta_1$ may change with treatment. That is, if

$$z_i = \begin{cases} 0 & \text{if treated (Puromycin is used)} \\ 1 & \text{otherwise} \end{cases}$$

a possible model is

$$y_i = \frac{(\hat{\theta}_1 + \hat{\theta}_2 z_i) x_i}{\hat{\theta}_2 + x_i} + \epsilon_i$$

and the parameter $\theta_3$ then measures the effect of the treatment. Go back to the data table and now do a fit of the (3 parameter) nonlinear model including a possible Puromycin effect using all 23 data points. Make 2 different approximately 95\% confidence intervals for $\theta_3$. Interpret these. (Do they indicate a statistically detectable effect? If so, what does the sign say about how treatment affects the relationship between $x$ and $y$?) Plot on the same set of axes the curves

$$y = \frac{\hat{\theta}_1 x}{\hat{\theta}_2 + x} \quad \text{and} \quad y = \frac{(\hat{\theta}_1 + \hat{\theta}_3) x}{\hat{\theta}_2 + x} \quad \text{for} \quad 0 < x < 2$$

3. This question concerns the analysis of a set of home sale price data obtained from the Ames City Assessor’s Office. Data on sales May 2002 through June 2003 of $1\frac{1}{2}$ and 2 story homes built 1945 and before, with (above grade) size of 2500 sq ft or less and lot size 20,000 sq ft or less, located in Low- and Medium-Density Residential zoning areas. (The data are in an
Excel® spreadsheet on the Stat 401B Web page. These need to be loaded into R for analysis.)

\( n = 88 \) different homes fitting this description were sold in Ames during this period. (2 were actually sold twice, but only the second sales prices of these were included in our data set.) For each home, the value of the response variable

\[
\text{Price} = \text{recorded sales price of the home}
\]

and the values of 14 potential explanatory variables were obtained. These variables are

- **Size**: the floor area of the home above grade in sq ft,
- **Land**: the area of the lot the home occupies in sq ft,
- **Bedrooms**: a count of the number in the home
- **Central Air**: a dummy variable that is 1 if the home has central air conditioning and 0 if it does not,
- **Fireplace**: a count of the number in the home,
- **Full Bath**: a count of the number of full bathrooms above grade,
- **Half Bath**: a count of the number of half bathrooms above grade,
- **Basement**: the floor area of the home's basement (including both finished and unfinished parts) in sq ft,
- **Finished Bsmnt**: the area of any finished part of the home's basement in sq ft,
- **Bsmnt Bath**: a dummy variable that is 1 if there is a bathroom of any sort (full or half) in the home's basement and is 0 otherwise,
- **Garage**: a dummy variable that is 1 if the home has a garage of any sort and is 0 otherwise,
- **Multiple Car**: a dummy variable that is 1 if the home has a garage that holds more than one vehicle and is 0 otherwise,
- **Style (2 Story)**: a dummy variable that is 1 if the home is a 2 story (or a \( \frac{2}{2} \) story) home and is 0 otherwise, and
- **Zone (Town Center)**: a dummy variable that is 1 if the home is in an area zoned as "Urban Core Medium Density" and 0 otherwise.

(The effect of using a dummy variable in a MLR context is to add the corresponding beta to the mean function for those cases where the dummy variable is 1. So, for example, you are allowed to just add a fixed amount for having a garage.)

Use **leaps** to find the single models with highest \( R^2 \) values for all numbers of predictors 1 through 14. Then compare these 14 models on the basis of 11-fold cross validation. Ultimately, what model do you like best? Fit it to the entire data set. Interpret its fitted coefficients in the context of the problem and estimate the standard deviation of selling price for fixed values of the predictors (using 95% confidence limits).

**Homework #7 (Due November 19)**

1. Section 3 Exercises 1 and 2, Section 4.3 of Vardeman and Jobe (page 190-191).
2. Section 1 Exercise 2 parts a) through f) Section 8.1 of Vardeman and Jobe (pages 567-569).

3. Section 2 Exercises 1a, 2 (don't answer the second part of c)), and 3, Section 8.2 of Vardeman and Jobe (pages 590-591).