Homework 9

Assigned Wednesday, November 2
Due Wednesday, November 9

Model Selection, Logistic Regression

Turn-in homework with hand-written or typed responses and include any relevant plots which you are describing.

1. The Space Shuttle Challenger

Recall at the beginning of the semester we discussed the space shuttle Challenger disaster. After looking at the data, it looked like there was information that they could have used to tell them not to take-off since the predicted temperature was 31 degrees F.

Find the shuttle data at:  http://www.stat.uiowa.edu/ rdecook/s152/datasets.html

Variables:

flight -- the number of the flight
temp -- the temperature at launch time
distress -- coded 0 if the O-rings experienced no distress
         1 if the O-rings showed signs of distress

Carry out an analysis to assess the relationship between temperature and the probability of O-ring distress. Use glm.out to hold your generalized linear model output.

(a) Provide the summary output and comment on significance.

(b) Use the following code including the ‘predict’ function to plot your fitted sigmoidal curve overtop of your plotted points in the range from 50 to 80 degrees.

> plot(temp, jitter(distress,factor=0.1))
> xvalues=seq(50,80)
> yvalues=predict(glm.out,list(temp=xvalues),type="response")
> lines(xvalues,yvalues,col="blue")

(c) Interpret the estimated regression coefficient \( \hat{\beta}_{temp} \)
   (there are a couple ways to do this, any of those we discussed is fine).
(d) Below what temperature is the probability of an O-ring being distressed greater than 0.5? (Use your $\beta$ estimates and a little algebra.)

(e) Use the following code to get the estimate of $E(Y|X = 70)$ and the SE for the estimate. Use these values to create a 95% confidence interval for the conditional mean.

```
> predict(glm.out,newdata=data.frame(temp=70),
         type="response",se.fit=TRUE)
```

(f) If you were in the meeting making a decision on the launch, what would have been your input? On what did you base your decision?

2. For this problem, you will use the *step()* function for model selection using the data set “cars1993_edit.csv” found at the class datasets website. There is an associated .txt file describing the 9 variables, as well. The response variable is:

```
Domest ---- 1 if the vehicle is a domestically made car (US)
           0 if it is foreign made car
```

So, you will be doing model selection for a logistic regression model with a candidate pool containing 8 regressors.

(a) Use the *step* function and the AIC criterion to choose a ‘best’ model. Provide a list of the terms in the ‘best’ model. ALSO, give a list of the terms that were removed in the order that they were removed from first to last. Here is a shortcut for fitting the full model with all covariates:

```
> glm.full.out=glm(Domest ~ .,family=binomial(logit),data=cars.data)
```

For (b) and (c) below, use your chosen ‘best’ model]

(b) Comment on the sign (+/−) of the estimated coefficients on the EngSize term and the Width term with respect to the response and your perception of US cars vs. foreign cars (i.e. in the context of the data). Do the signs seem expected?

(c) At first glance, the negative sign on HPW may seem odd because HPW and EngSize are positively correlated. But use the plot below, and your knowledge of how we interpret MULTIPLE REGRESSION COEFFICIENTS to explain in the context of the data why the negative coefficient is actually appropriate for the observed data.

```
> plot(HPW,EngSize,pch=(Domest+1),col=(Domest+1))
> legend(50,5,c("foreign","domestic"),col=c(1,2),pch=c(1,2))
```
3. The file called *cholesterol* at the class data website contains data from an observational study from 315 observations.

Variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>Sex (1=Male, 0=Female).</td>
</tr>
<tr>
<td>fat</td>
<td>Grams of fat consumed per day.</td>
</tr>
<tr>
<td>cholCat</td>
<td>coded 1 if the individual has high cholesterol</td>
</tr>
<tr>
<td></td>
<td>0 if the individual does not have high cholesterol</td>
</tr>
</tbody>
</table>

We are interesting in predicting the cholesterol category based on sex and fat intake.

(a) Fit the additive model. Provide the summary output, and the global null hypothesis p-value. Provide the conclusion to the global null hypothesis, which is that neither of the predictors help in predicting cholesterol category.

(b) How many iterations did it take to find the maximum likelihood estimates?

Both explanatory variables are significant in this model.

(c) For women, how does increasing their fat intake by 1 gram per day change their odds on being in the high cholesterol group? (This is an interpretation of a coefficient, please be specific).

(d) For men, how does increasing their fat intake by 1 gram per day change their odds on being in the high cholesterol group? (This is an interpretation of a coefficient, please be specific).

(e) Would the answers to (c) and (d) potentially change if we included the interaction term in this model?

(f) If you consider all people who have a given fat intake, how does changing from being a female to a male change the odds on being in the high cholesterol group? (This is an interpretation of a coefficient, please be specific).

(g) What is the estimated probability of a woman with a fat intake of 100 grams per day being in the high cholesterol group?