In-class Data Analysis Exercise

We will use this class to explore the material on residuals and diagnostics for logistic regression in Section 14.8 (pp 591-601). The material and code is largely available on the slides that Prof. Hanson prepared, but is best understood as a class exercise. Start by inputting the tab-delimited text file for the horseshoe crab data (this code is also available on the webpage) and transforming a couple variables. E.g.:

```
proc import out=horseshoecrab
            datafile="C:\Grego\My Documents\STAT705\HorseshoeCrab.txt"
replace;
run;

data crabs; set horseshoecrab;
weight=weight/1000;
color=color-1;
y=0;
if satellite>0 then y=1;
id+_n_;
run;
```

The session will start with a look at plots of the residuals and link function for a logistic regression of presence/absence of satellites on Color and Width. We could obtain most of the following plots simply by specifying `plots=all`, but let's first use a more targeted approach and then return to the (copious) default output display. These commands will save the standardized Pearson chi-squared residuals (defined on page 49 of your slides), the linear predictor `eta` ($\hat{\eta}$), predicted probabilities, and C_j (the book uses D_i), an analog to Cook's distance.

```
proc logistic data=crabs;
class color/param=ref;
model y(event=’1’)=color width;
output out=diag1 stdreschi=r xbeta=eta p=p c=c;
run;
```

The SAS code available on the course webpage following the commands above steps through four different graphical and print procedures. Let's consider the series of plots from the first procedure. Does the loess line for the first two plots provide any indication of lack of fit? For the residual plot against $\hat{\eta}$ (the second plot), explain why the residuals appear as a pair of ordered bands. Why does the residual plot against Width generate multiple pairs of ordered bands? Discuss the residual plot against Color. Why is there a gap between residuals for each level of Color? How do you interpret the large positive outlier for Color=4?

In the second set of plots, why would we plot diagnostics against the ID variable? What do we detect with Cook's D that we do not detect from a standardized residual plot?

The third plot plots the fitted values for fixed levels of color (indexed here by j) $\{1+\exp(-\hat{\eta}_j)\}^{-1}$ against Width. Relate the ordering of the curves in your graph to estimated parameter values for Color.

The fourth procedure prints records with large residuals, high influence or both. Explain why each of these records was flagged–what makes these female horseshoe crabs so very special?

Now simply add `plots=all` to the initial `proc logistic` statement. Since we are focusing on diagnostics, skip the first two displays (odds ratios, ROC curve); the next five displays are all labelled Influence Diagnostics; answer these questions corresponding to each display.

1. We generated one of these plots already. Do you see any differences between residuals/standardized residuals? Between Pearson residuals/Deviance residuals?
2. This display includes residuals, leverage and influence diagnostics, including the Cook's D display we generated earlier. Remember our STAT 704 rule-of-thumb for high leverage? I don't either, but I looked it up ($h_j > 2p/n$). How many leverage values appear significant here?
3. The next two plots measure difference in the X^2 and deviance goodness of fit statistics when each observation is deleted in turn (see DIFCHISQ_j in your notes). The book suggests there are no easy rules-of-thumb for these diagnostics. Do they seem to be flagging different cases from the residuals, influence and leverage plots?
4. We will skip the two DFBETA's plots since we did not study this diagnostic in STAT 704.

The next set of plots is entitled Predicted Probability Diagnostics. Four of our earlier influence diagnostics are plotted against $\hat{\pi}_{\mathbf{x}}$. The book actually discusses the first two plots; extreme values will appear as outliers in the upper corners of the plots–how many extreme values appear to be present? For the fourth plot (leverage vs. predicted probability), why do you suppose there are four distinct curves, and why are 1's and 0's intermingled in these curves, unlike the pattern we saw for residuals? Equation (14.80) and the definition of $\mathbf{W}$ below (14.80) are helpful in answering these questions.

The next set of plots is entitled Leverage Diagnostics; these are quite unusual, aren't they? Recall that diagnostics tend to be functions of residuals, leverage (and predicted probabilities, in this case), so it is not surprising to see underlying functional relationships between them. It's interesting that the first three have roughly the same pattern, albeit with different y scales. And neither leverage nor C_j nor DIFCHISQ_j are signed, so it's interesting that the positive and negative values generate different curves, isn't it?

The final plots are the same as the first two plots in Predicted Probability Diagnostics, though a heat bar has been added for the Cook's distance diagnostic. Cook's distance is related to Pearson Chi-squared Difference by $C_j = \text{DIFCHISQ}_j \times \frac{h_j}{1-h_j}$, so it's not clear why it is presented as a *linear* heat scale opposite the Deviance Difference rather than the Pearson Chi-squared Difference.

What do you feel you learned from using the `plots=all` option as opposed to Prof. Hanson's more narrowly targeted commands?