36-401 Class Notes and Exercise for 11/20/01
Logistic Regression II

1. Coefficients for Logistic Regression

   - IRLS solutions maximize the likelihood
   $$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ P_i^{Y_i}(1-P_i)^{1-Y_i} \right\} \text{ where } P_i = \frac{1}{1+\exp(-\mathbf{x_i}\boldsymbol{\beta})}$$
   with respect to all of the β's for the observed Y's.
   - *Asymptotically* the coefficients are unbiased, efficient, and normally distributed *if*
     (a)
     (b)
     (c)
   - There must also be sufficient success and failure cases over all regions of the covariate space; otherwise, little information is provided regarding partial effects of several X variables.

2. Hypothesis Tests and Confidence Intervals

   - Wilks' Lambda: General Testing of Nested Hypotheses
   For $H_0 : \beta_1 = \beta_2 = \ldots = \beta_H = 0$, we use
   $$\chi^2_H = -2[\log_e(\mathcal{L}_{K-H}) - \log(\mathcal{L}_K)]$$
   which follows the χ^2 distribution with H degrees of freedom. Tends to be more accurate than t-tests at small n's.

     Question: what does it mean here to reject H_0?

   - Nested Test of "Null Deviance" vs. "Deviance" is Equivalent to Overall F-test
     - Deviance is -2 log(likelihood)
     - Null Deviance is for "intercept-only" model (RwG calls this "iteration-0" likelihood)
     - The test is based on the asymptotic chi-squared distribution of (null.deviance-deviance). Degrees of freedom are equal to K-1, the number of covariates.

       Explain why the degrees of freedom are $K-1$ here.

- t-statistics and confidence intervals based on the usual method of calculation are *asymptotically* correct.

3. Statistical Problems

   (a) Multicollinearity

   - Multicollinearity leads to unreliable coefficient estimates and large standard errors
   - Methods for detecting multicollinearity include:
     -
     -
     -
   - Ways to handle multicollinearity
     - Consider collecting more data
     - Drop or combine variables
     - χ^2 test can test the coefficients on a set of highly correlated variables together, when individual effects cannot be separated

   (b) Potential, Residuals, Fit and Influence

   - Logistic regression problems often have several (m_J) cases with the same X *pattern* (J denotes the number of unique X patterns in the sample) **Letting m_j denote the number of cases with the jth X pattern, what do J and m_j equal if no two cases have the same X pattern?**
   - Leverage (h_j) is like in linear regression, but X is weighted and same pattern cases are summed. h_i are the diagonal elements of

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}$$

for OLS;

$$\mathbf{H} = \mathbf{X}^*(\mathbf{X}^{*\mathbf{T}}\mathbf{X}^*)\mathbf{X}^{*\mathbf{T}}, \text{ where } \mathbf{X}^* = \mathbf{W}^{\mathbf{1/2}}\mathbf{X} \text{ and } \mathbf{W} = \text{diag}[\hat{P}_i(1-\hat{P}_i)]$$

for logit.

   - Two residuals are Pearson (r_j) and Deviance (d_j)

$$r_j = \frac{Y_j - m_j\hat{P}_j}{\sqrt{m_j\hat{P}_j(1-\hat{P}_j)}}$$

$$d_j = \pm\left\{2\left[Y_j \log\left(\frac{Y_j}{m_j\hat{P}_j}\right) + (m_j - Y_j)\log\left(\frac{m_j - Y_j}{m_j(1-\hat{P}_j)}\right)\right]\right\}^{\frac{1}{2}}$$

where $\hat{P}_j$ is the fitted value for the j^{th} pattern, Y_j is the number of 1's (successes) in the j^{th} pattern and $0 \log(0) = 0$.
The sums of squared residuals form χ^2_P and χ^2_D which are measures of *overall* goodness-of-fit. Sums of squared Pearson or deviance residuals test whether a given logit model is significantly worse than a perfect-fit (saturized) model.

$$\chi^2_P = \sum_{j=1}^{J} r_j^2$$

$$\chi^2_D = \sum_{j=1}^{J} d_j^2$$

- Goodness of fit / Outliers For each *pattern*:

  $$\Delta\chi^2_{P(j)} = \frac{r_j^2}{(1-h_j)}$$

  measures decrease in Pearson χ^2;

  $$\Delta\chi^2_{D(j)} = \frac{d_j^2}{(1-h_j)}$$

  measures change in deviance. These are analogous to standardized residuals; measure how poorly the model fits the jth pattern.

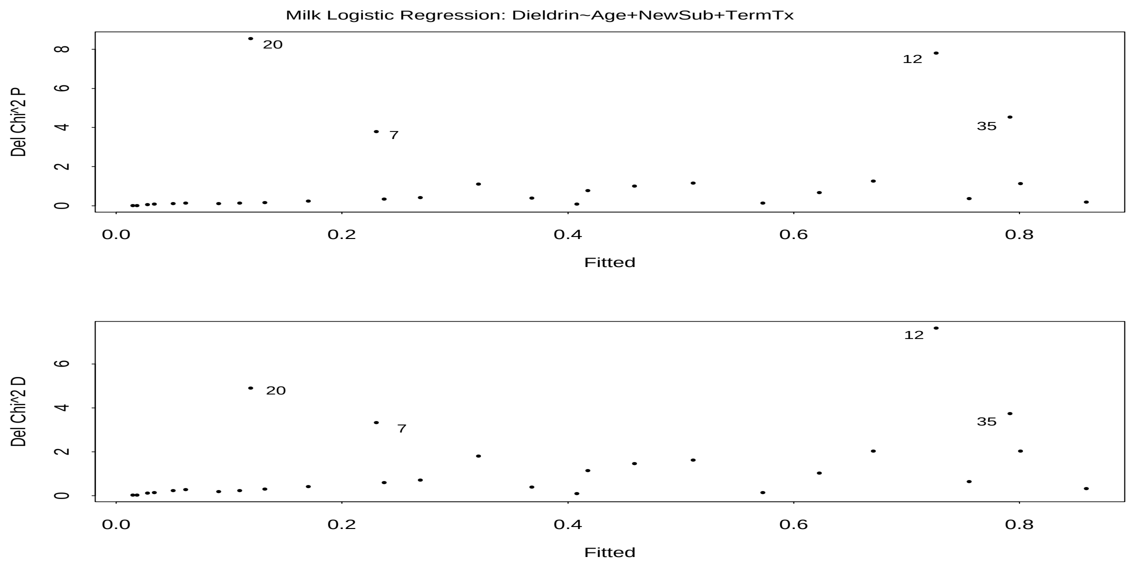
  Large values (i.e. > 4) indicate patterns that model would fit the data better if the pattern were deleted. Outliers show up nicely on plots of $\Delta\chi^2_{P(j)}$ vs. fitted probabilities.
- Influence, as indicated by $\Delta B_j \geq 1$,

  $$\Delta B_j = \frac{r_j^2 h_j}{(1-h_j)^2}$$

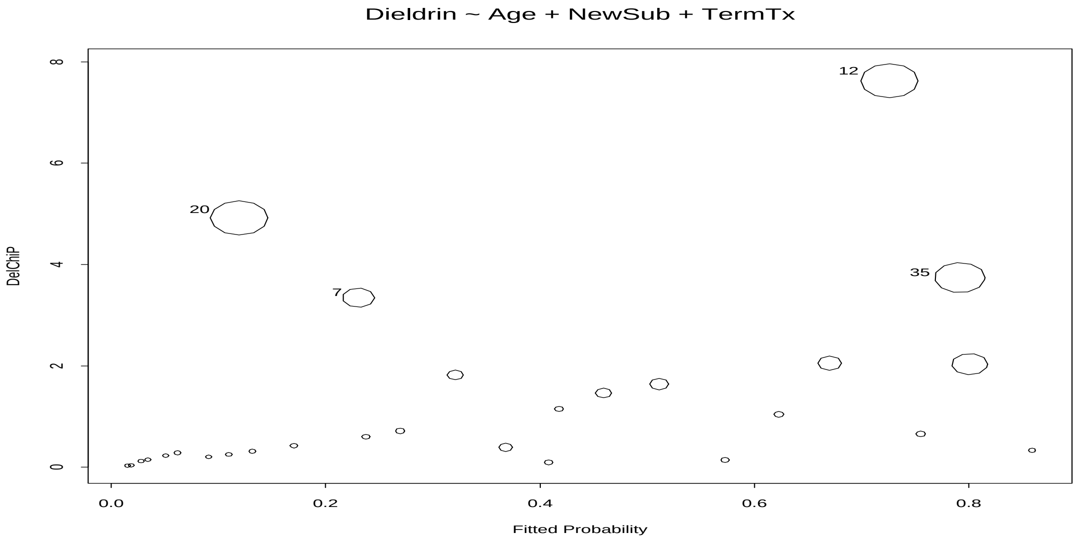
  indicates that case j has substantial influence (analogous to Cook's D). ΔB versus fitted probability plots detect influential data.

Breakout Session for 11/20/01

```
summary(milk[,c("Age","NewSub","TermTx")])     summary(milklm)$coef
 Age           NewSub        TermTx                   Value St.Err  t val     Pr
 Min.   :21.0  Min.   :0.00  Min.   :0.00       (Int) -0.892  0.539 -1.655  0.106
 1st Qu.:27.0  1st Qu.:0.00  1st Qu.:0.00         Age  0.032  0.018  1.793  0.081
 Median :28.0  Median :0.00  Median :1.00      NewSub  0.348  0.145  2.399  0.021
 Mean   :29.1  Mean   :0.29  Mean   :0.56      TermTx  0.411  0.130  3.151  0.003
 3rd Qu.:32.0  3rd Qu.:1.00  3rd Qu.:1.00
 Max.   :37.0  Max.   :1.00  Max.   :1.00
                             NA's   :2.00
OK_is.na(milk$TermTx)==F
source("logitinf.q")
inf_logitinf(milkllm, milk[OK,2:4], milk$Dieldrin[OK])  # $coll is collapsed to 28 patterns
par(mfrow=c(2,1), mai=c(1,1,0.1,0.1), oma=c(1,1,3,1))
plot(inf$coll$fit, inf$DelChiP, xlab="Fitted", ylab="Del Chi^2 P", type="n", cex=1.15)
points(inf$coll$fit, inf$DelChiP)
mtext("Milk Logistic Regression: Dieldrin~Age+NewSub+TermTx", outer=T, cex=1.1)
Bad_identify(inf$coll$fit, inf$DelChiP, inf$coll$pat, cex=1)

plot(inf$coll$fit, inf$DelChiD, xlab="Fitted", ylab="Del Chi^2 D", type="n", cex=1.15)
points(inf$coll$fit, inf$DelChiD)
Bad_c(Bad, identify(inf$coll$fit, inf$DelChiD, inf$coll$pat, cex=1))

Bad_unique(Bad)
round(cbind(inf$coll,DCP=inf$DelChiP,DCD=inf$DelChiD,DB=inf$DelB)[Bad,],2)
#  Age NewSub TermTx Y m  fit  hat pat  DCP  DCD   DB
#   21      0      1 1 1 0.12 0.14  20 8.55 4.92 1.34
#   27      1      0 1 1 0.23 0.12   7 3.80 3.34 0.53
#   25      1      1 0 2 0.73 0.32  12 7.80 7.63 3.69
#   37      0      1 0 1 0.79 0.16  35 4.53 3.74 0.88
```



```
Bplot(milkllm, milk[OK,2:4], milk$Dieldrin[OK], id=T)

noage_glm(Dieldrin~NewSub+TermTx,data=milk,fam="binomial",na.act=na.omit)
noage$deviance/-2         # -22.41348     noage$null.deviance/-2    # -28.38259
milkllm$deviance/-2       # -20.54355     milkllm$null.deviance/-2  # -28.38259

anova(noage,milkllm)
Analysis of Deviance Table
                 Terms Resid. Df Resid. Dev Test Df Deviance
       NewSub + TermTx        40   44.82696
 Age + NewSub + TermTx        39   41.08711 +Age  1  3.73985
```



1. Which patterns have the largest residuals? What in the data relates to the high residuals? Which patterns are most influential? What should you do about these?

2. Interpret the 4 values of -deviance/2.

3. Does the nested likelihood ratio test support inclusion of Age in the model? Does BIC (loglikelihood - k/2 log(n))?

4. Make substantive interpretations of the model coefficients.