36-401 Class Notes and Exercise for 11/28/00
Logistic Regression (RwG ch. 7)
Howard Seltman

1. Main Idea: When the response variable is binary, regression assumptions are necessarily violated. Logistic regression changes both the mean function and the error model to produce a reasonable approach to these problems.

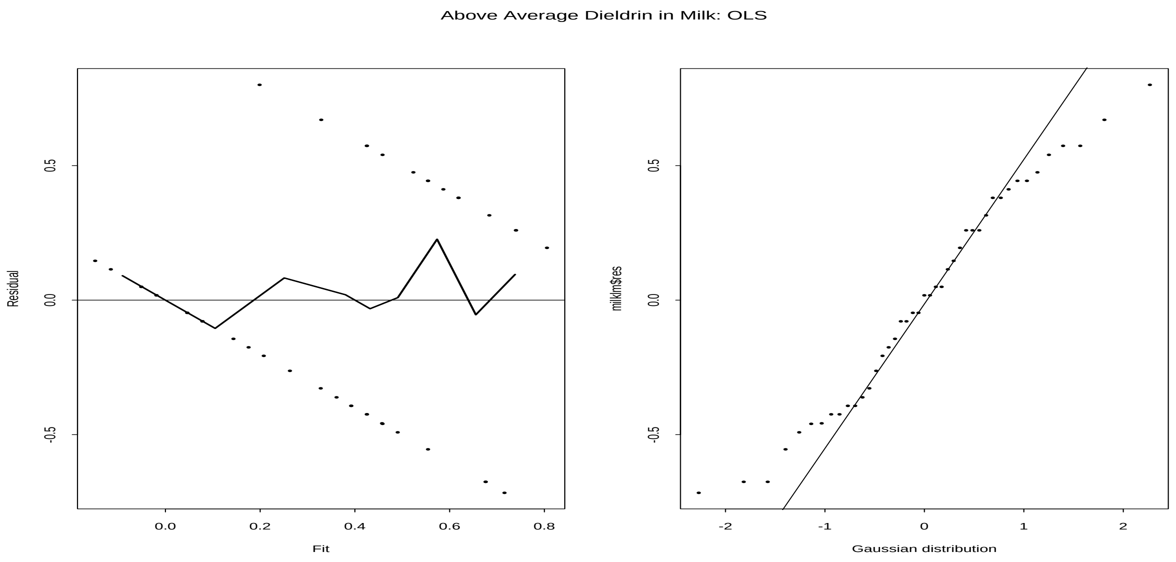
2. Example: Why OLS doesn't work

```
milk_read.table("milk.dat",T)
milk[1:5,]
# Donor Age NewSub TermTx Dieldrin
#     1  33      1      0        1
#     2  34      0      1        1
#     3  29      0      0        0
#     4  28      1     NA        0
#     5  29      1      1        1

milklm_lm(Dieldrin~Age+NewSub+TermTx,milk,na.action=na.omit)
summary(milklm)
# Coefficients: Value Std. Error t value Pr(>|t|)
# (Intercept) -0.8917  0.5389    -1.6546  0.1060
#         Age  0.0323  0.0180     1.7927  0.0808
#      NewSub  0.3476  0.1449     2.3987  0.0213
#      TermRx  0.4110  0.1304     3.1511  0.0031
#
# Residual standard error: 0.4235 on 39 degrees of freedom
# Multiple R-Squared: 0.3037
# F-statistic: 5.67 on 3 and 39 degrees of freedom, the p-value is 0.002535

par(mfrow=c(1,2), oma=c(1,1,3,1))
plot(milklm$fit, milklm$res, xlab="Fit", ylab="Residual")
lines(meanplot(milklm$fit,milklm$res,add=T))
abline(h=0)
mtext("Above Average Dieldrin in Milk: OLS", outer=T, cex=1.3)
qqn(milklm$res, main="")

predict(milklm, data.frame(Age=40, NewSub=1, TermTx=1)
# 1.16
```



3. Model for Logistic Regression

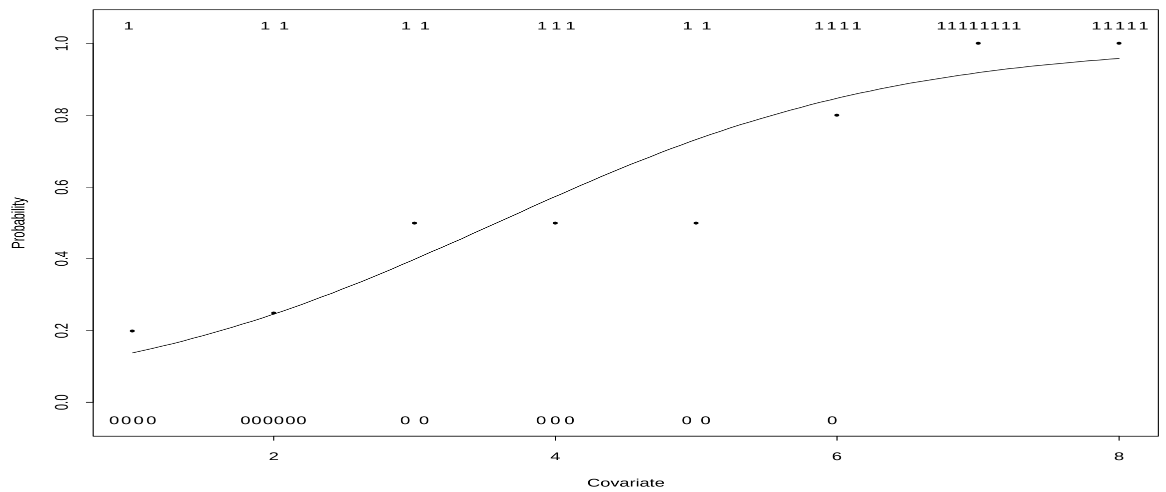
$$Y_i|X_i \sim \text{Bernoulli}(p_i)$$

$$E(Y_i|X_i) = p_i \qquad \text{Var}(Y_i|X_i) = p_i(1-p_i)$$

$$p_i = \frac{1}{1+\exp(-u_i)} = \frac{\exp(u_i)}{1+\exp(u_i)}$$

$$u_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{K-1} X_{i(K-1)}$$

$$\Rightarrow u_i = \log\frac{p_i}{1-p_i} = \log(\text{odds}_i) = \text{logit}(p_i)$$



4. Estimation is by Generalized Linear Models using IRLS

```
milkllm_glm(Dieldrin~Age+NewSub+TermTx, data=milk, na.action=na.omit,
  family=binomial)
summary(milkllm)
# Coefficients:    Value Std. Error   t value
# (Intercept) -8.9727497  3.8532397 -2.328625
#         Age  0.2083864  0.1172264  1.777640
#      NewSub  2.1406872  0.9308404  2.299736
#      TermTx  2.5969714  0.9683143  2.681951
#
# (Dispersion Parameter for Binomial family taken to be 1 )
#     Null Deviance: 56.76518 on 42 degrees of freedom
# Residual Deviance: 41.08711 on 39 degrees of freedom
# Number of Fisher Scoring Iterations: 4
```

5. Interpretation of Coefficients

- β_0 is the log(odds) of a "1" response when all covariates are at level zero
- β_k is the change in log(odds) of a "1" response when covariate k is increased by one unit and all other covariates are held constant.
- Conditional Effects Plot

$$u = \beta_0 + \beta_1 X_1 + \ldots + \beta_{K-1} X_{(K-1)}$$

$$p = \frac{1}{1 + \exp(-u)}$$

Plot a sequence of X_k's vs. corresponding p's holding the other X's constant.

Breakout Session for 11/28/00

1. Give three examples of problems with a binary response. Name the two outcome states and list at least one covariate that you expect to affect the probability of being in one state vs. the other.

2. If T subjects all have the same covariates in a logistic regression model, what is the distribution of the number of those subjects that have a "1" response?

3. Consider rolling an even number on a die, picking an ace from a deck of cards, and selecting a male from a population of computer scientists (p=0.7). Make a table showing probability, odds, and log odds. Add two columns showing the probabilities if the log odds increases or decreases by 2.

4. Sketch conditional effects plots for Age for the milk example for 1) a house with termite treatment in an old suburb and 2) a house without termite treatment in an old suburb. Plot over the age range 20 to 40.