Modeling a Binary Response

Bruce A Craig

Department of Statistics
Purdue University

Reading: Faraway Ch. 2, Agresti Ch. 5-6, KNNL Ch. 14

Background

- In many applications, the response variable Y has only two possible outcomes, labeled numerically 0 and 1
  - Not heart diseased ($Y = 0$) vs heart diseased ($Y = 1$)
  - Unemployed ($Y = 0$) vs Employed ($Y = 1$)
  - PhD student ($Y = 0$) vs MS student ($Y = 1$)
- Response is *binary* or *dichotomous*
- Can model response using Bernoulli distribution

$$\begin{array}{l} \Pr(Y_i = 1) = p_i \\ \Pr(Y_i = 0) = 1 - p_i \end{array} \quad \rightarrow \quad \mathrm{E}(Y_i) = p_i$$

- Want to describe $E(Y_i) = p_i$ in terms of covariates $\mathbf{x}_i$
- Clearly important to keep track of what is labeled $Y = 1$

Why Not Fit Using an LM?

- Consider the linear model (with one covariate x_i)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- With Y_i only taking values 0 and 1, then
  - There error terms are non-Normal (only two values)

$$\text{when } Y_i = 0 : \varepsilon_i = -\beta_0 - \beta_1 x_i$$
$$\text{when } Y_i = 1 : \varepsilon_i = 1 - \beta_0 - \beta_1 x_i$$

  - There is nonconstant variance

$$\text{Var}(Y_i) = p_i(1 - p_i)$$

  - Need parameter bounds so $0 \leq p_i \leq 1 \ \ \forall i$
- Note: When $0.2 < p < 0.8$, variances not that different so LM may be a reasonable approximate model

Fit Using a GLM

- Can show Bernoulli distribution is an EFD!!
- Key is then determining $g(p_i) = \mathbf{x}_i\boldsymbol{\beta}$
- Most common link functions are
  - logit link $\longrightarrow$ logistic regression
  - probit link $\longrightarrow$ probit regression
  - complementary log-log link
- Both logit and probit are symmetric inverse CDF links
  - logit link $\longrightarrow$ logistic distribution
  - probit link $\longrightarrow$ Normal distribution
- Log-log link an asymmetric inverse CDF link
  - Based on extreme value (Gumbel) distribution

Latent Variable Formulation

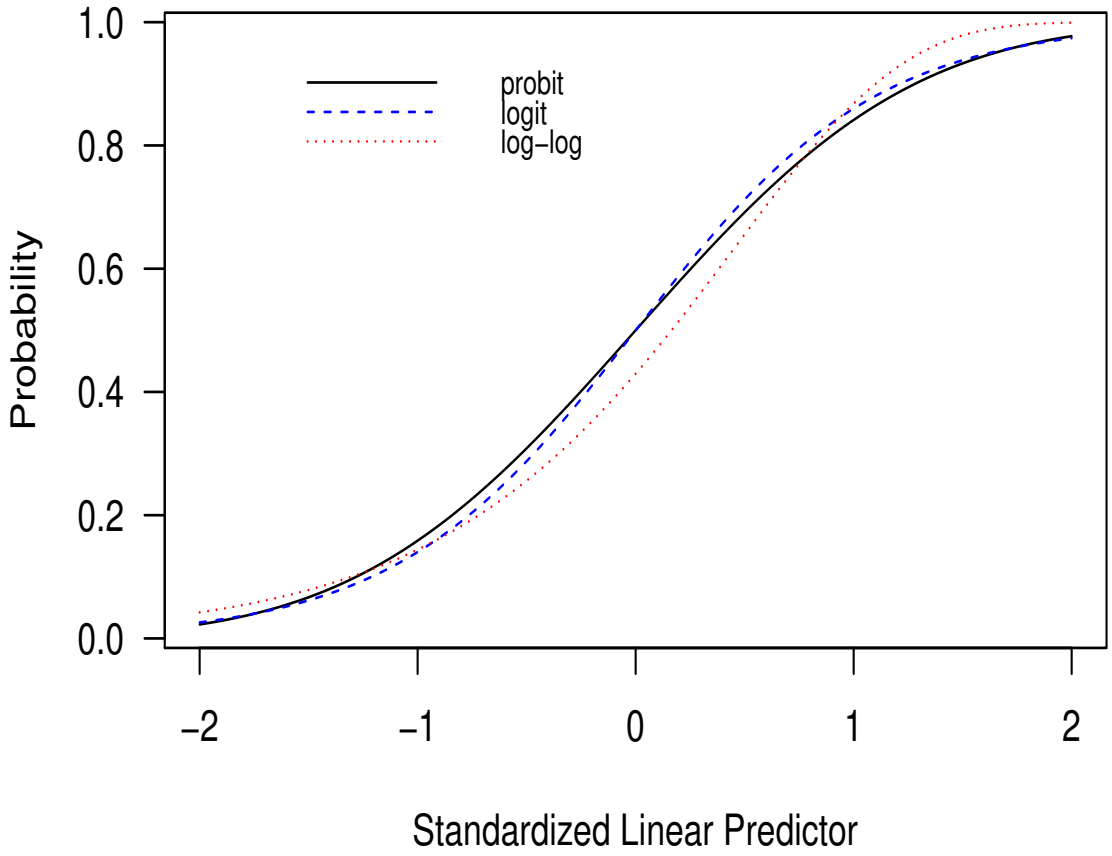
- Suppose there is a latent response Z_i with standard deviation one such that $P(Y_i = 1) = P(Z_i > 0)$
- Let Z_i depend on covariates $\mathbf{x}_i$ through the model

$$Z_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i = \eta_i + \varepsilon_i$$

- Thus, $p_i = P(Y_i = 1) = P(\varepsilon_i > -\eta_i) = 1 - F(-\eta_i)$
- Defines family of GLMs with $\eta_i = -F^{-1}(1 - p_i)$
  - When F symmetric, $\eta_i = F^{-1}(p_i)$
- Note: Choice of threshold zero and standard deviation one are made for model identity
  - Example: If $Z_i \sim N(0, \sigma)$, then $p_i = \Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma)$. Cannot separately estimate $\boldsymbol{\beta}$ and σ

Comparison of Link Functions

- The differences between links are subtle
- Likely need large data set to discriminate
- Unique features and domain area may dictate choice
  - Logistic regression: Odds ratio
  - Complementary Log-Log: Hazard ratios
  - Probit regression: Use of Normal dists
- Following plot shows p_i versus standardized η_i
  - Logit and probit quite similar
  - Complementary log-log starts slow but rapidly increases


1.0
0.8
0.6
0.4
0.2
0.0
Probability
probit
logit
log−log
−2
−1
0
1
2
Standardized Linear Predictor

Logistic Response Function

- Symmetric sigmoidal function (nonlinear model)
- Monotonically increasing/decreasing function

$$\begin{aligned} E(Y_i) &= \frac{\exp\{\mathbf{x}_i\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i\boldsymbol{\beta}\}} \\ &= (1+\exp\{-\mathbf{x}_i\boldsymbol{\beta}\})^{-1} \end{aligned}$$

- Based off of the logit link function

$$\log\left(\frac{E(Y_i)}{1-E(Y_i)}\right) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i\boldsymbol{\beta}$$

Relationship with the Odds Ratio

- In a simple logistic regression model, the log odds are linearly related to a covariate x

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

- The slope β_1 then represents the change in log odds for a unit increase in x

$$\begin{aligned}\beta_1 &= \log\left(\frac{p(x+1)}{1-p(x+1)}\right) - \log\left(\frac{p(x)}{1-p(x)}\right) \\ &= \log\left(\text{odds}(x+1)\right) - \log\left(\text{odds}(x)\right)\end{aligned}$$

Relationship with the Odds Ratio

- A difference in logs can be expressed as the log of a ratio

$$\log\left(\frac{A}{B}\right) = \log(A) - \log(B)$$

- Thus

$$\begin{aligned} \beta_1 &= \log\left(\frac{\text{odds}(x+1)}{\text{odds}(x)}\right) \\ &\downarrow \\ \exp\{\beta_1\} &= OR(x) \end{aligned}$$

- In multiple logistic regression, $\exp\{\beta\}$'s called adjusted ORs

Logistic model

- Y_i are independent but not identically distributed Bernoulli random variables with the means $P(Y_i = 1) = p_i$ linked to a linear combination of the covariates using the logit function
- Can express as

$$Y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \mathbf{x}_i\boldsymbol{\beta}$$

- Probability density of Y_i

$$\begin{aligned} f(Y_i) &= p_i^{Y_i}(1-p_i)^{1-Y_i} \\ &= \left(\frac{p_i}{1-p_i}\right)^{Y_i}(1-p_i) \end{aligned}$$

Estimation

- The log likelihood function is

$$\begin{aligned} l(\boldsymbol{\beta}) &= \log\left(\prod p_i^{Y_i}(1-p_i)^{1-Y_i}\right) \\ &= \sum Y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum \log(1-p_i) \\ &= \sum Y_i(\mathbf{x}_i\beta) - \sum \log(1+\exp\{\mathbf{x}_i\beta\}) \end{aligned}$$

- MLEs do not have closed form
- Must use IRWLS or other iterative procedure

Example

- What factors are associated with coronary heart disease?
- An early study investigating this started in 1960 and followed 3154 men in San Francisco aged from 39-59
- Initial examination provided predictors
- Men were classified with (chd=yes) or without (chd=no) heart disease 8.5 years later
- Western Collaborative Group Study (WCGS) is a cohort, or prospective, study
- Typical for studies of that time to focus on men (boo!)

Summarizing the wcgs data set

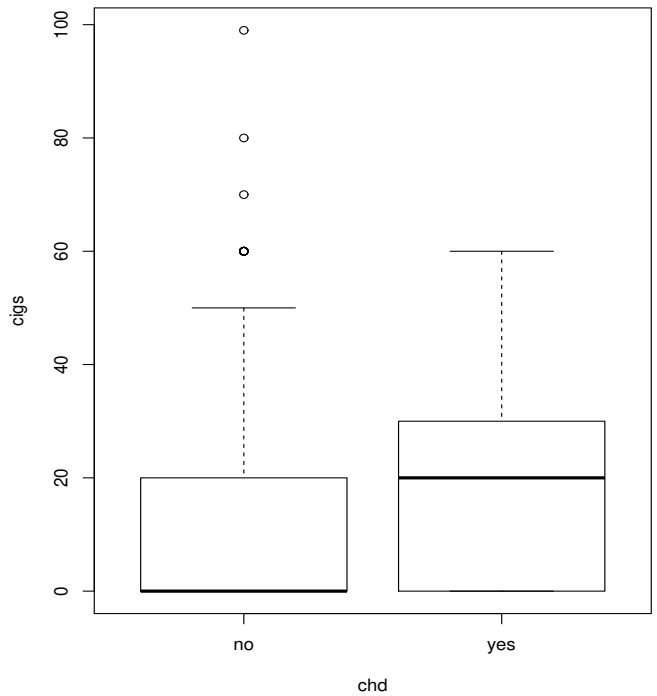
```
> library(faraway)
> attach(wcgs)
> str(wcgs)

'data.frame': 3154 obs. of  13 variables:
 $ age    : int  49 42 42 41 59 44 44 40 43 42 ...
 $ height : int  73 70 69 68 70 72 72 71 72 70 ...
 $ weight : int  150 160 160 152 150 204 164 150 190 175 ...
 $ sdp    : int  110 154 110 124 144 150 130 138 146 132 ...
 $ dbp    : int  76 84 78 78 86 90 84 60 76 90 ...
 $ chol   : int  225 177 181 132 255 182 155 140 149 325 ...
 $ behave : Factor w/ 4 levels "A1","A2","B3",..: 2 2 3 4 3 4 4 2 3 2 ...
 $ cigs   : int  25 20 0 20 20 0 0 0 25 0 ...
 $ dibep  : Factor w/ 2 levels "A","B": 2 2 1 1 1 1 1 2 1 2 ...
 $ chd    : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
 $ typechd: Factor w/ 4 levels "angina","infdeath",..: 3 3 3 3 2 3 3 3 3 3 ...
 $ timechd: int  1664 3071 3071 3064 1885 3102 3074 3071 3064 1032 ...
 $ arcus  : Factor w/ 2 levels "absent","present": 1 2 1 1 2 1 1 1 1 2...
```

WCGS Variables

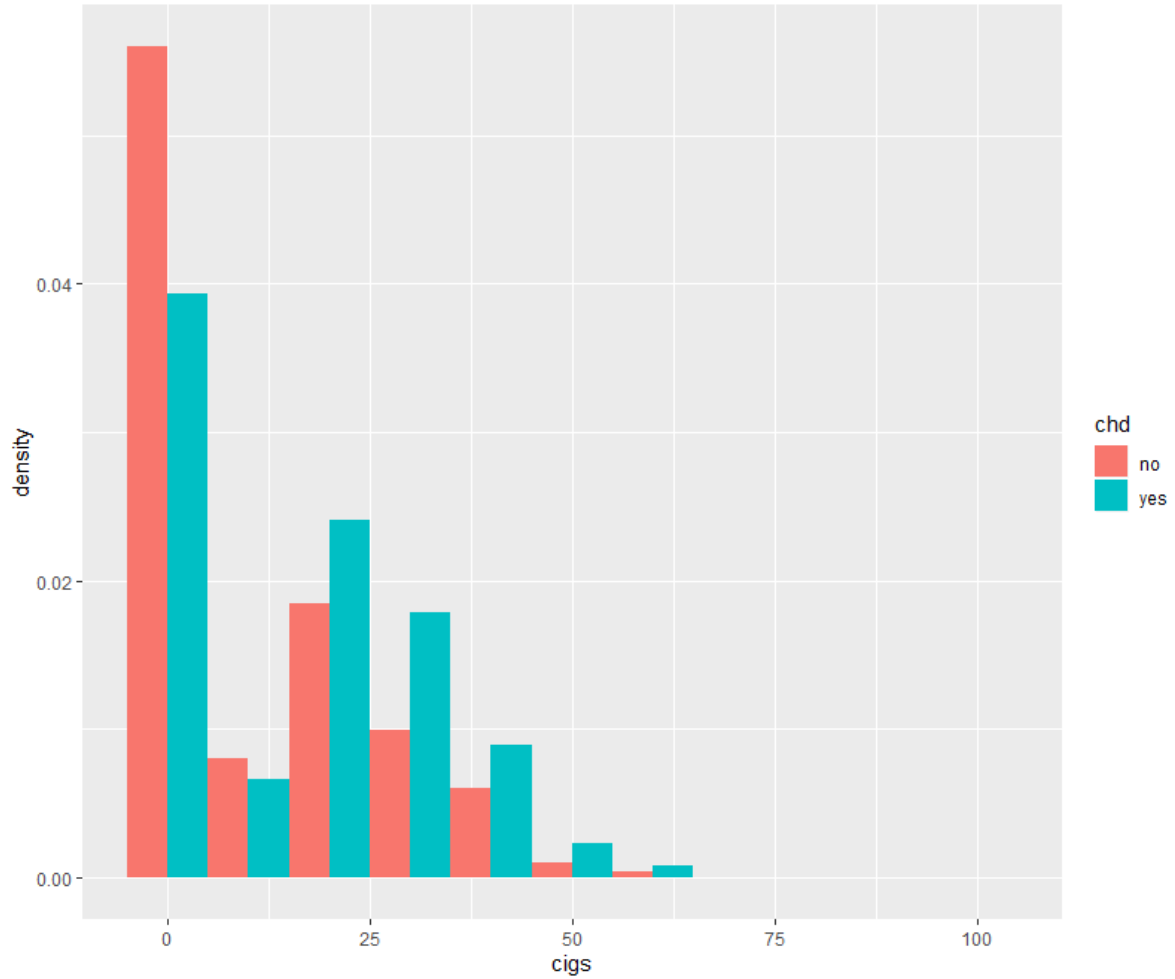
- Some are responses, the others are predictors
- Predictors
  - Age, height, and weight
  - Systolic and diastolic blood pressure
  - Cholesterol
  - Personality behavior* (2 or 4 levels)
  - Number of cigarettes per day
  - Presence of arcus senilis
- Responses
  - Developed coronary heart disease
  - Type of CHD
  - Time of CHD event
- Let's first examine CHD versus smoking

Exploratory Analysis I

```
> summary(chd)
  no  yes            ***Only 8% of men developed CHD
2897  257

> plot(cigs ~ chd)
```

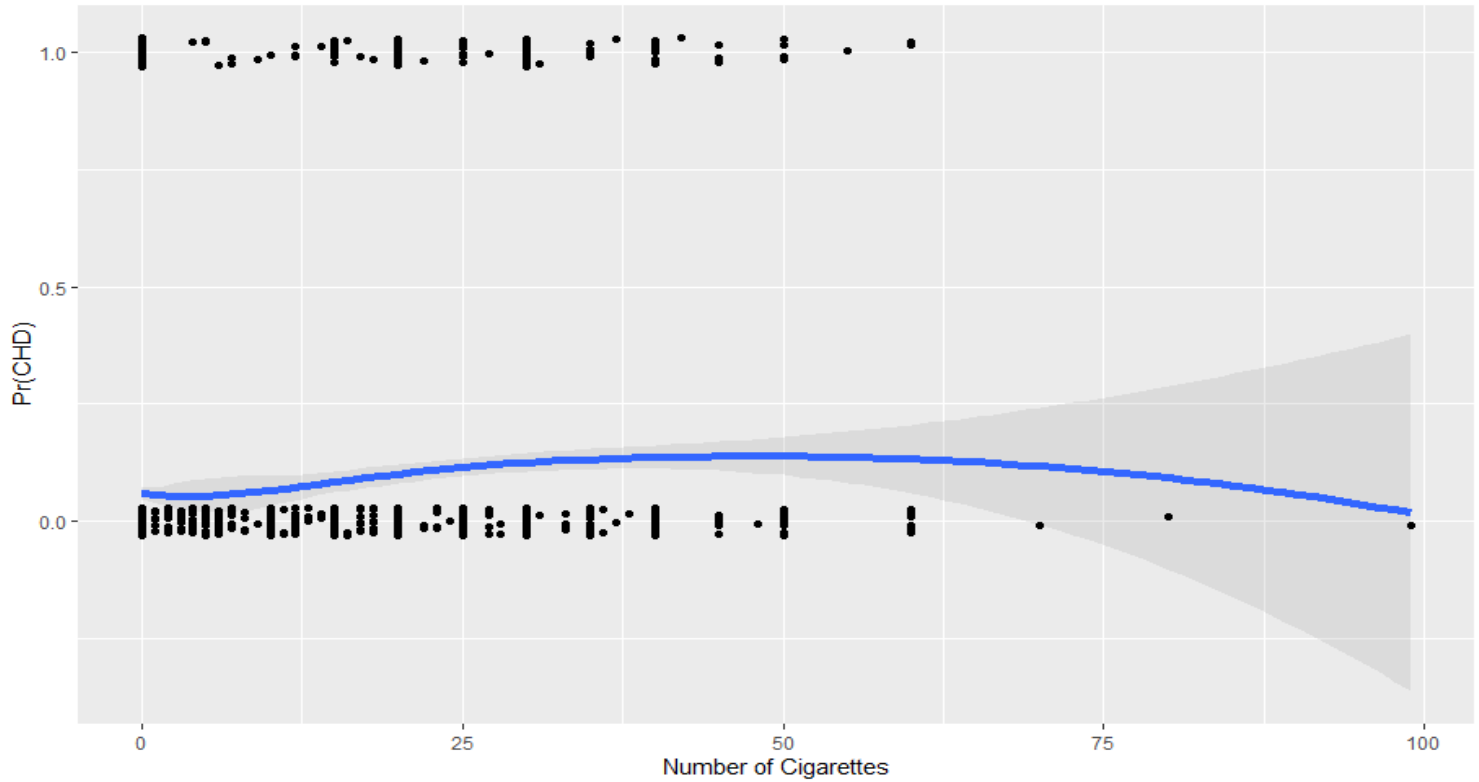


```
> ggplot(wcgs, aes(x=cigs, fill=chd)) + geom_histogram(position="dodge"
        ,binwidth=10, aes(y=..density..))
```



Exploratory Analysis II

```
> ggplot(wcgs, aes(cigs, as.numeric(chd)-1)) +
  stat_smooth(method="loess", formula=y~x, alpha=0.2, size=2) +
  geom_point(position=position_jitter(height=0.03, width=0)) +
  xlab("Number of Cigarettes") + ylab("Pr(CHD)")
```



Fit a Simple Logistic Model

```
> model1 = glm(chd ~ cigs, family=binomial, wcgs)
> summary(model1)

Call:
glm(formula = chd ~ cigs, family = binomial, data = wcgs)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.742160   0.092111 -29.770  < 2e-16 ***
cigs         0.023220   0.004042   5.744 9.22e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1781.2  on 3153  degrees of freedom
Residual deviance: 1750.0  on 3152  degrees of freedom
AIC: 1754

Number of Fisher Scoring iterations: 5
```

Interpreting Results

- Should really check diagnostics first but...
- Difference in deviances is 31.2 on 1 df
- Wald test for β_1 results in $z = 5.744$
- Both tests have an extremely small P-value
- Pack a day (cigs=20) increases odds 59% (35.6%, 86.3%)

```
> exp(coef(model1)[2]*20)
    cigs
1.591053
> exp(confint(model1)[2,]*20)
   2.5 %   97.5 %
1.356293 1.862701
```

Diagnostics

- Plots of residuals vs linear predictor not that informative because residual can only take two values
- Faraway suggests binning cases based on linear predictor
  - Number of bins will depend on the size of the data set
  - Compute the average linear predictor and average residual in each bin and plot them
- These take some effort to create
- May smooth out outliers and/or patterns
- Alternative is to consider randomized quantile residual

Randomized Quantile Residuals

- We are comfortable assessing residuals from Normal dists
- Residuals from GLMs can be highly non-Normal, making it difficult to interpret plots
- Dunn and Smyth (1996) proposed randomized quantile residuals

$$r_{i,q} = \Phi^{-1}(F(Y_i|\hat{\mu}_i, \hat{\phi}))$$

  where F is the cdf of the EFD
- Premise: If Y_i are from EFD, then $F(Y_i) \sim U(0,1)$ and $r_{i,q} \sim N(0,1)$
- If Y_i are from EFD, then variation due to estimation of μ and ϕ

Randomized Quantile Residuals

- If F not continuous, a more general definition is needed
- Let $a_i = F(\max(0, Y_i - 1)|\hat{\mu}_i, \hat{\phi})$ and $b_i = F(Y_i|\hat{\mu}_i, \hat{\phi})$
- The randomized quantile residual is

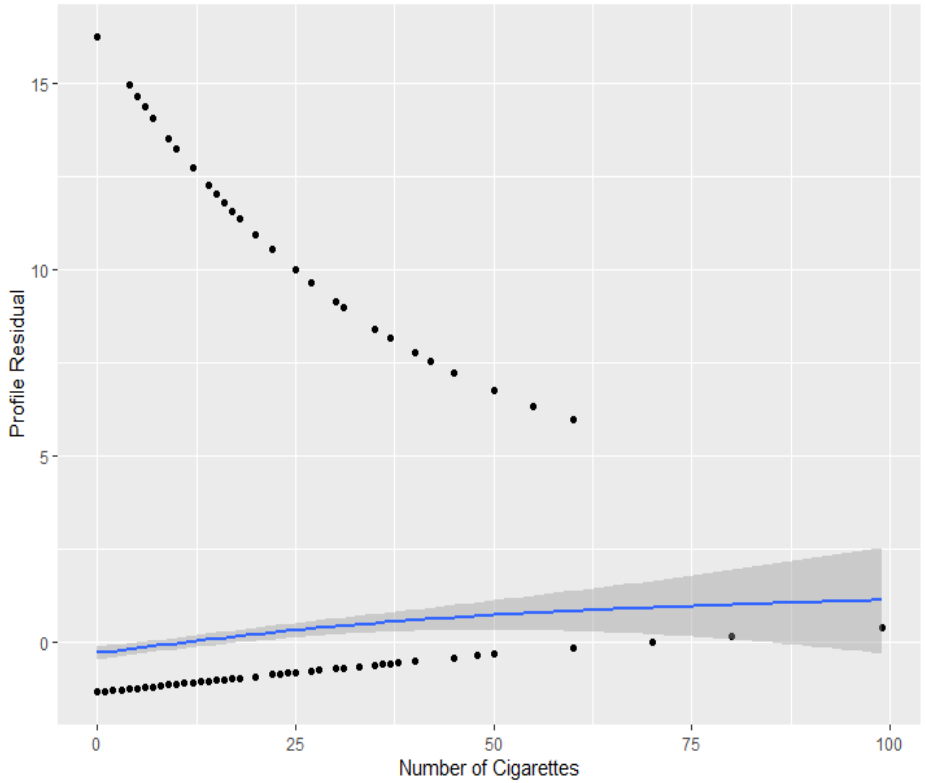
$$r_{i,q} = \Phi^{-1}(u_i)$$

  where $u_i \sim U(a_i, b_i)$
- Due to this randomness, the residuals will vary each plot to plot but the general patterns in them should remain
- Does not address influence but could "leave out" case to get $\hat{\mu}_{(i)}$ and $\hat{\phi}_{(i)}$ but far more computation

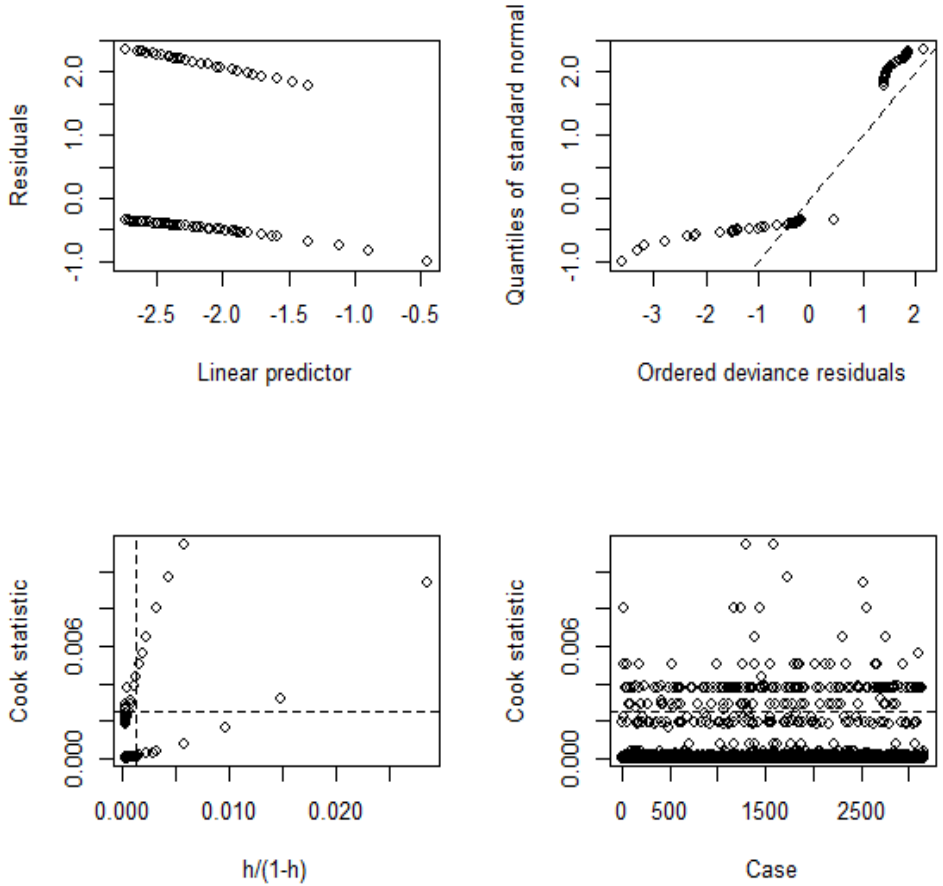
Examining Residuals

```
qplot(cigs,residuals(model1,type="partial"),geom=c('point','smooth'),
      xlab="Number of Cigarettes",ylab="Profile Residual")

glm.diag.plots(model1)

install.packages("statmod")
library(statmod)

###Residual plot
plot(cigs,qres.binom(model1),geom=c('point','smooth'),
     xlab="Number of Cigarettes",ylab="Quantile Residual")

###Assess Normality of the residuals
qqnorm(qres.binom(model1))
abline(a=0,b=1,col="red")


***** Based on the following randomized quantile regression plots *****
*****  I'd argue that the model fits the data well. Don't know    *****
*****  to conclude using standard plots with loess smoother        *****
```

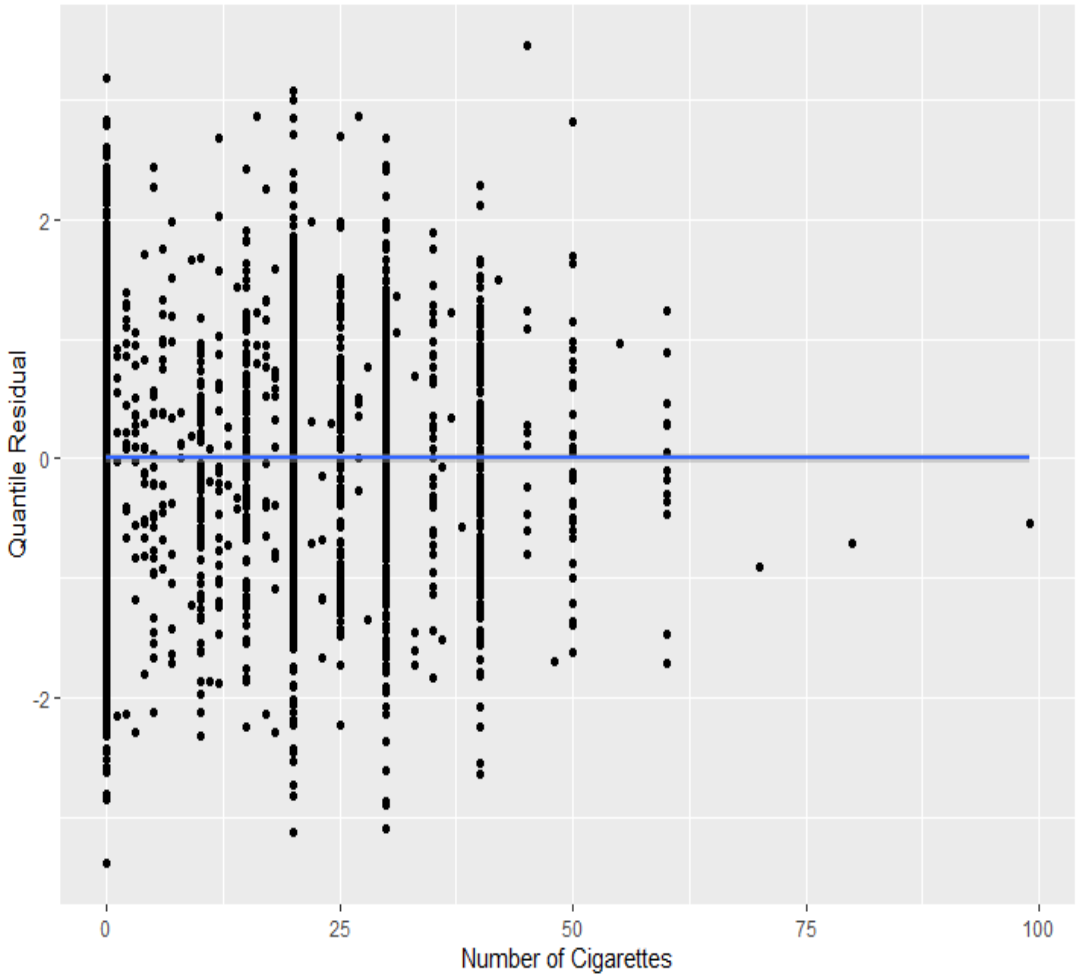
Partial Residual Plot



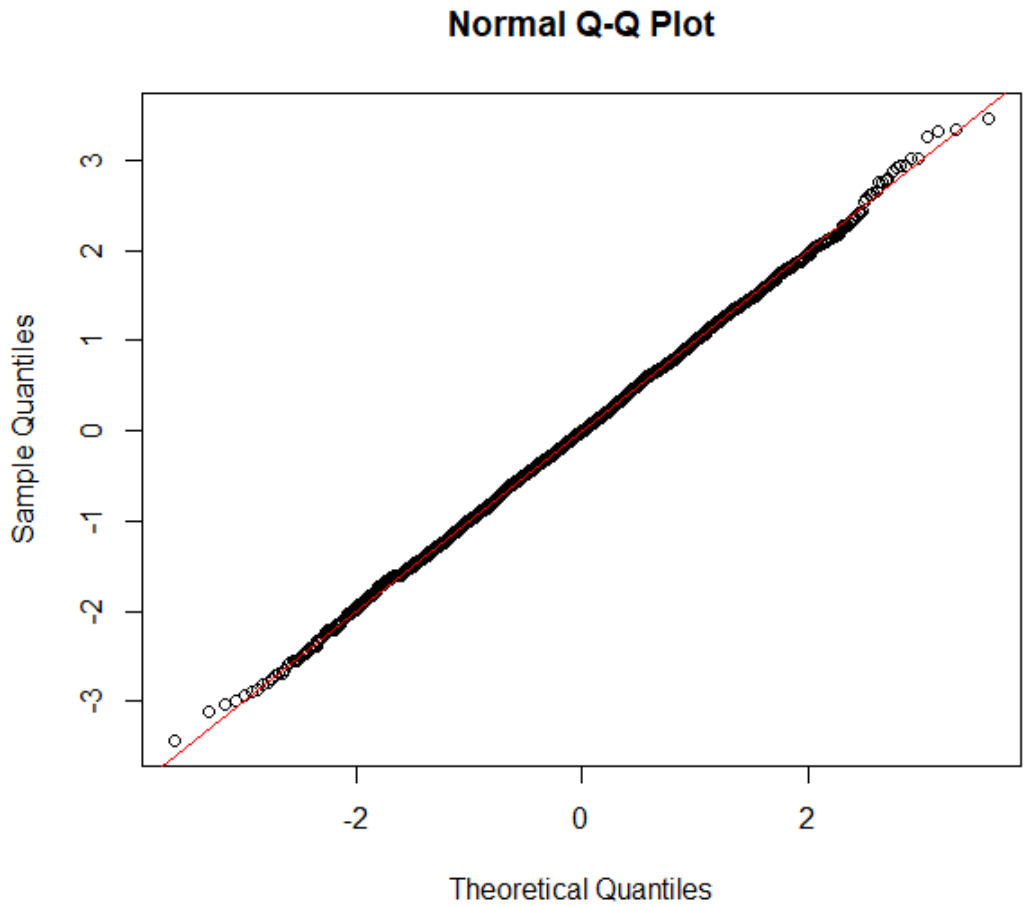
Diagnostic Set of Plots



Quantile Residual Plot



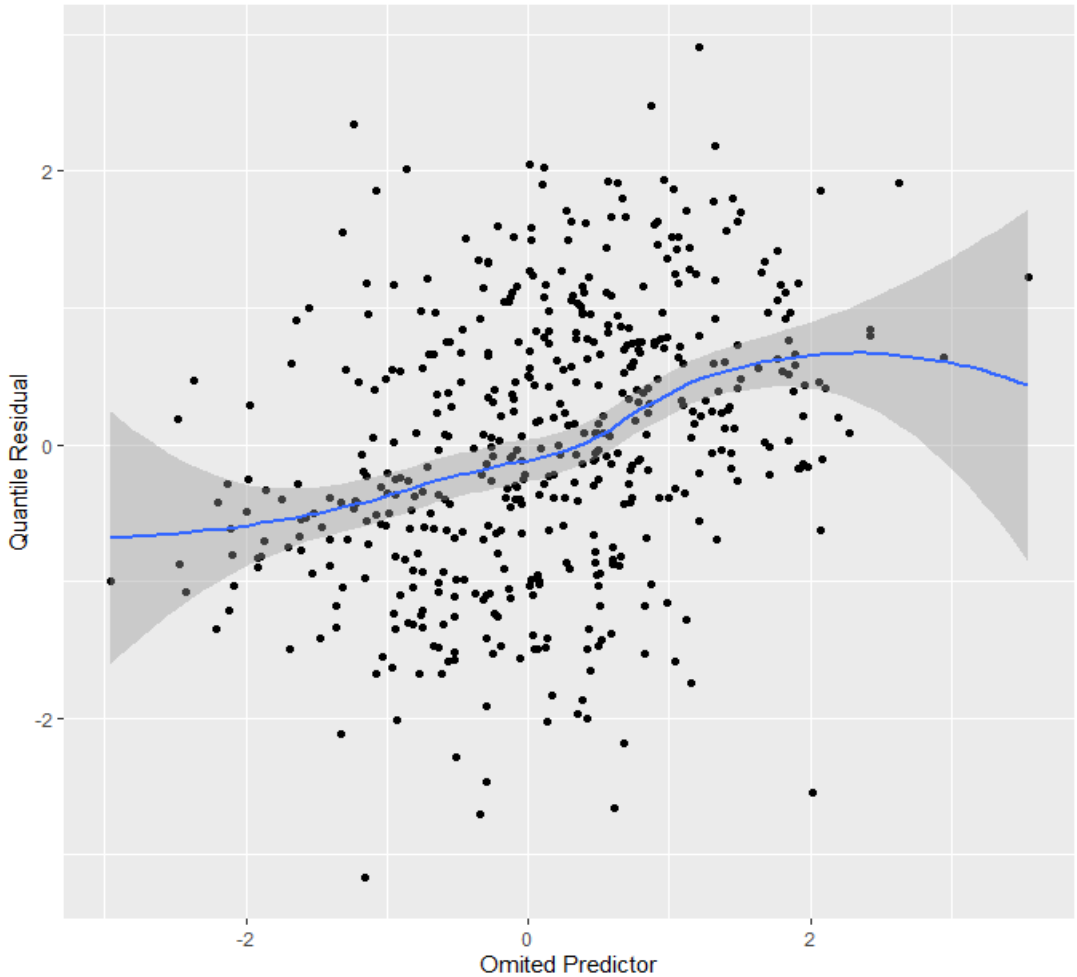
QQPlot – Is this helpful?



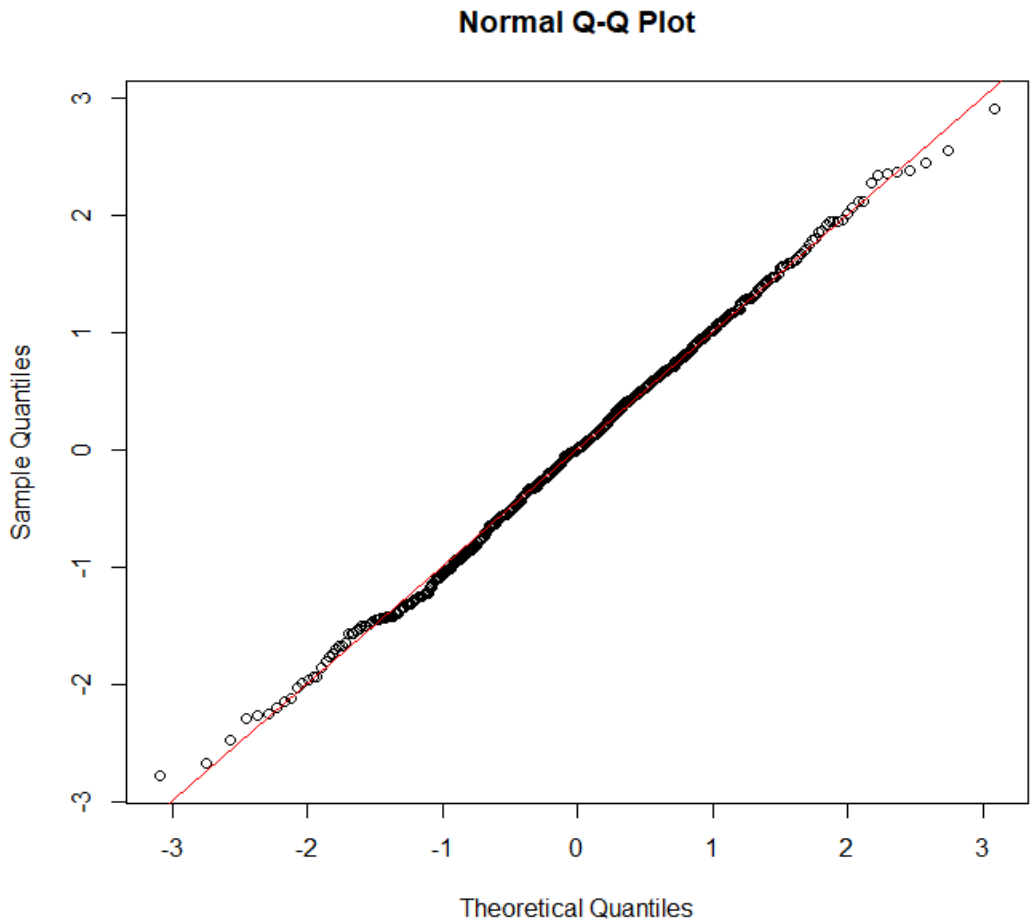
Perils of QQplot With Binary Data

```
set.seed(612)

x1 = rnorm(500,0,1)
logitp = 0.35 + 1*x1
y = rbinom(500,1,(1+exp(-logitp))^(-1))


model1 = glm(y ~ 1,family=binomial)    ### Fitting null model instead

###Residual plot
qplot(x1,qres.binom(model1),geom=c('point','smooth'),
      xlab="Omited Predictor",ylab="Quantile Residual")

###QQplot
qqnorm(qres.binom(model1))
abline(a=0,b=1,col="red")
```

Quantile Residual Plot



QQPlot


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3

Summary

- Residual plot adequately detects missing predictor
- QQplot looks as if model fits appropriately
  - Without replicates, model fit will adjust to missing covariates or incorrect variance specification by altering the mean structure to best fit the data
- Thus QQplot assessment of Normality only applicable when possible to separately estimate the mean and variance

Goodness of Fit I

- In Topic 4, we discussed assessing goodness of fit
- For the Bernoulli distribution, the log likelihood is

$$\begin{aligned} l(\beta) &= \sum y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum \log(1-p_i) \\ &= \sum y_i \mathbf{x}_i \beta - \sum \log\left(1+\exp\{\mathbf{x}_i\beta\}\right) \\ &\downarrow \\ \frac{\partial l}{\partial \beta_j} &= \sum y_i x_{ij} + \sum p_i x_{ij} \end{aligned}$$

- Thus the deviance can be written

$$D(\hat{\beta}) = -2\sum \hat{p}_i \text{logit}(\hat{p}_i) + \log(1-\hat{p}_i)$$

- This is just a function of the fitted values and thus cannot be used to assess model fit

Goodness of Fit II

- In addition, given that $m_i = 1$, Pearson residual not asymptotically Normal for large n
- This means the Pearson X^2 statistic is also not approximately χ^2
- Thus, alternative goodness of fit approaches are needed for binary logistic model
- Simulation (see HW #3) can be used to demonstrate this poor approximation to a chi-square
- These issues subside with replicates (as we will see with binomial data) but issues when m_i small and/or p_i generally close to 0 or 1 still remain

Goodness of Fit: Pearson χ^2

- When there are replicates, we can perform a χ^2 test

$$H_0: \quad \text{logit}\left(p(\mathbf{x}_i)\right) = \mathbf{x}_i\beta$$
$$H_a: \quad \text{logit}\left(p(\mathbf{x}_i)\right) \neq \mathbf{x}_i\beta$$

- Assume $j = 1, 2, \ldots, C$ covariate sets of n_j cases
- Observed counts
  - O_{j1}: # of $Y_i = 1$ and O_{j0}: # of $Y_i = 0$ in set j
- Expected counts
  - $E_{j1} = n_j\hat{p}(\mathbf{x}_j)$ and $E_{j0} = n_j - E_{j1}$
- Test statistic: reject H_0 if

$$X^2 = \sum_{j=1}^{C}\sum_{k=0}^{1}\frac{(O_{jk} - E_{jk})^2}{E_{jk}} > \chi^2(1-\alpha, C-p)$$

Lack of Fit: Deviance

- Similarly, can perform deviance test if replicates
- Again C covariate sets, each with n_j cases
  - H_0: $E(Y_{ij}) = (1 + \exp\{-\mathbf{x}_i\boldsymbol{\beta}\})^{-1}$
  - H_a: $E(Y_{ij}) = p_j$
- The MLEs for these models are
  - H_0: $\hat{E}(Y_{ij}) = \left(1 + \exp\{-\mathbf{x}_i\hat{\boldsymbol{\beta}}\}\right)^{-1}$
  - H_a: $\hat{E}(Y_{ij}) = Y_{\cdot j}/n_j$
- Test statistic: Likelihood ratio

Deviance Goodness of Fit Test

$$
\begin{aligned}
G^2 &= DEV(X_0, X_1, \ldots, X_{p-1}) \\
&= -2\left[\, \log L(\text{current model}) - \log L(\text{saturated model}) \,\right] \\
&= -2\sum_{j=1}^{c}\left[\, Y_{\cdot j}\log\left(\frac{\hat{p}_j}{\tilde{p}_j}\right) + (n_j - Y_{\cdot j})\log\left(\frac{1-\hat{p}_j}{1-\tilde{p}_j}\right)\,\right]
\end{aligned}
$$

- Reject H_0 if $G^2 > \chi^2(1-\alpha, C-p)$
- For large n_j, this statistic also follows χ^2 with $C-p$ df
- In both tests, assuming $C > p$
- Also, in assuming replicates at each covariate set, we could treat the data as binomial

General Binary Goodness of Fit Test

- In WCGS there are 23 cigs values with fewer than 10 replicates so we can't comfortably use these two approaches
- As with other χ^2 tests, typically want $E_{jk} > 5$ and none smaller than 1
- General approach is to bin data together and treat the data in each bin as replicates
- Hosmer and Lemeshow proposed this test for both unreplicated data sets or data sets that have only a limited number of sets with replicates

Hosmer-Lemeshow Goodness of Fit

- Used when few or no replicate Y_i's per covariate set
- Groups cases based on $\hat{p}_i$ (or $\hat{\eta}_i$'s)
- No definitive rules on how many groups or how to group
  - Choices really depend on distribution of $\hat{p}$ (or $\hat{\eta}$)
  - Want groups with similar $\hat{p}_i$'s
  - Want groups that are approximately the same size
  - Grouping by quantiles guarantees only the latter
  - Want C so that E_{j0} and E_{j1} are mostly > 5
- Compute Pearson χ^2 statistic based on the sets
- Reject H_0 if $X^2 > \chi^2(1-\alpha, C-2)$
  - Showed this distribution appropriate through extensive simulation

Hosmer-Lemeshow Goodness of Fit

- R package generalhoslem provides HL test
  - Seeks $C = 10$ based on equal size but can alter it
- Lackfit option in SAS proc logistic
  - Loads data into 2000 bins and then combines bins seeking balanced groups and $C = 10$
- The daily # of cigarettes has many many 0's. Also, likely responses based on units in packs (e.g., 0.5, 1, 2 packs)
- Because of this, I will consider my own binning too

Software results

- In R

```
> logitgof(chd,fitted(model1))

Hosmer and Lemeshow test (binary model)

data:  chd, fitted(model1)
X-squared = 2.2092, df = 3, p-value = 0.5301

Warning message:
In logitgof(chd, fitted(model1)) :
  Not possible to compute 10 rows. There might be too few observations.
```

- In SAS

Partition for the Hosmer and Lemeshow Test

		chd1 = 1		chd1 = 0	
Group	Total	Observed	Expected	Observed	Expected
1	1652	98	100.00	1554	1552.00
2	322	20	24.31	302	297.69
3	483	50	44.76	433	438.24
4	405	50	45.00	355	360.00
5	292	39	42.94	253	249.06

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
2.5955	3	0.4583

My Binning

- Form 12 bins around 0, 5, 10, 15, 20,. . ., 50, $>$ 55

```
> E1 = numeric(length=12)       ###Expected # with CHD
> nj = numeric(length=12)       ###Total number of participants
> O1 = numeric(length=12)       ###Observed # with CHD

> E1[1] = sum(predict(model1,type="response")[cigs < 3])
> nj[1] = length(chd[cigs < 3])
> O1[1] = sum((as.numeric(chd)-1)[cigs < 3])
> for(i in 1:10){
>   nj[i+1] = length(chd[cigs > (2+(i-1)*5) & cigs < 8+(i-1)*5])
>   E1[i+1] = sum(predict(model1,type="response")[cigs > (2+(i-1)*5) & cigs < 8+(i-1)*5])
>   O1[i+1] = sum((as.numeric(chd)-1)[cigs > (2+(i-1)*5) & cigs < 8+(i-1)*5])
> }
> E1[12] = sum(predict(model1,type="response")[cigs > 52])
> nj[12] = length(chd[cigs > 52])
> O1[12] = sum((as.numeric(chd)-1)[cigs > 52])

> chisq = sum((1/E1 + 1/(nj-E1))*(O1-E1)^2)

> print(1-pchisq(chisq,10))
[1] 0.5774474
```

- In all cases, there is little evidence of a lack of fit

Model/Variable Selection

- So far have focused on a simple logistic regression
- Researchers may be interested in developing predictive model
- General approach similar to LM variable selection
- Given full model, use AIC to investigate all subsets
  - Minimize:

$$\begin{aligned} AIC_q &= -2\log L(\boldsymbol{\beta}) + 2q \\ &= \text{deviance} + 2q \end{aligned}$$

Fit a Multiple Logistic Model

```
> modelf = glm(chd~height+weight+sdp+dbp+age
               +cigs+dibep,family=binomial,wcgs)
> summary(modelf)

Call:
glm(formula = chd ~ height + weight + sdp + dbp + age + cigs +
    dibep, family = binomial, data = wcgs)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.9181130  2.1906762  -4.527 5.97e-06 ***
height      -0.0056924  0.0319100  -0.178   0.8584
weight       0.0088132  0.0037154   2.372   0.0177 *
sdp          0.0190428  0.0062184   3.062   0.0022 **
dbp          0.0009247  0.0104127   0.089   0.9292
age          0.0652009  0.0118924   5.483 4.19e-08 ***
cigs         0.0226468  0.0041545   5.451 5.00e-08 ***
dibepB       0.6804814  0.1437619   4.733 2.21e-06 ***
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1781.2  on 3153  degrees of freedom
Residual deviance: 1638.5  on 3146  degrees of freedom
AIC: 1654.5

Number of Fisher Scoring iterations: 6
```

Use Function to Search Subsets

```
> modelr = step(modelf,trace=0)
> summary(modelr)

Call:
glm(formula = chd ~ weight + sdp + age + cigs + dibep, family = binomial,
    data = wcgs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1880  -0.4452  -0.3360  -0.2512   2.7338

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.263018   0.867695 -11.828  < 2e-16 ***
weight        0.008526   0.003079   2.769  0.00563 **
sdp           0.019562   0.003946   4.957 7.16e-07 ***
age           0.065351   0.011866   5.507 3.64e-08 ***
cigs          0.022543   0.004117   5.476 4.36e-08 ***
dibepB        0.680252   0.143690   4.734 2.20e-06 ***
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1781.2  on 3153  degrees of freedom
Residual deviance: 1638.5  on 3148  degrees of freedom
AIC: 1650.5

Number of Fisher Scoring iterations: 6
```

Quality of Predictions

- Hosmer-Lemeshow requires some sort of binning
- Scoring methods do not require binning
  - Higher score means better overall prediction
  - Logarithmic scoring: $\sum (Y_i \log(\hat{p}_i) + (1 - Y_i) \log(1 - \hat{p}_i))$
- Can also assess model's ability to classify
  - Predict $Y_i = 1$ when say $\hat{p}_i > 0.50$
  - Can estimate overall misclassification rate
  - Or break down into sensitivity and specificity
- Finally, can look at measures of agreement

Measures of Agreement

- Have n observations
- Consider all pairs t of **distinct** responses
  - WCGS: 257 ($Y = 1$) $\times$ 2897 ($Y = 0$) = 744529 pairs
- Compare predicted probabilities
  - Concordant if $\hat{p}_{Y=1} > \hat{p}_{Y=0}$
  - Discordant if $\hat{p}_{Y=1} < \hat{p}_{Y=0}$
  - Tie if $\hat{p}_{Y=1} = \hat{p}_{Y=0}$
- Measures of agreement
  - Somers' D : (#C - #D)/t
  - Goodman-Kruskal Gamma : (#C - #D)/(#C + #D)
  - Kendall's Tau-a : (#C - #D)/(.5n(n-1))
  - c : (#C +.5(t-#C-#D))/t

Classification

- Choose a cutoff $c \in (0, 1)$

$$\hat{Y}_i = \begin{cases} 1, & \text{if } \hat{p}_i > c \\ 0, & \text{if } \hat{p}_i \leq c \end{cases}$$

  - Sensitivity: $\dfrac{\#\text{ predicted '1' \& true '1'}}{\#\text{ true '1'}} = \dfrac{\sum_{i=1}^{n} \hat{Y}_i Y_i}{\sum_{i=1}^{n} Y_i}$
  - Specificity: $\dfrac{\#\text{ predicted '0' \& true '0'}}{\#\text{ true '0'}} = \dfrac{\sum_{i=1}^{n} (1-\hat{Y}_i)(1-Y_i)}{\sum_{i=1}^{n} 1-Y_i}$

- Vary the cut-off $c \in (0, 1)$, and choose c to optimize sensitivity and specificity
- Results over multiple score cutoffs are summarized in a Receiver Operating Characteristic (ROC) curve.
- Best to apply to a new set of observations or perform cross-validation

Prediction Summaries

Hypothesized class		True class: p	True class: n
	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

Fawcett, "An introduction to ROC analysis". *Pattern Recognition Letters*, 2005

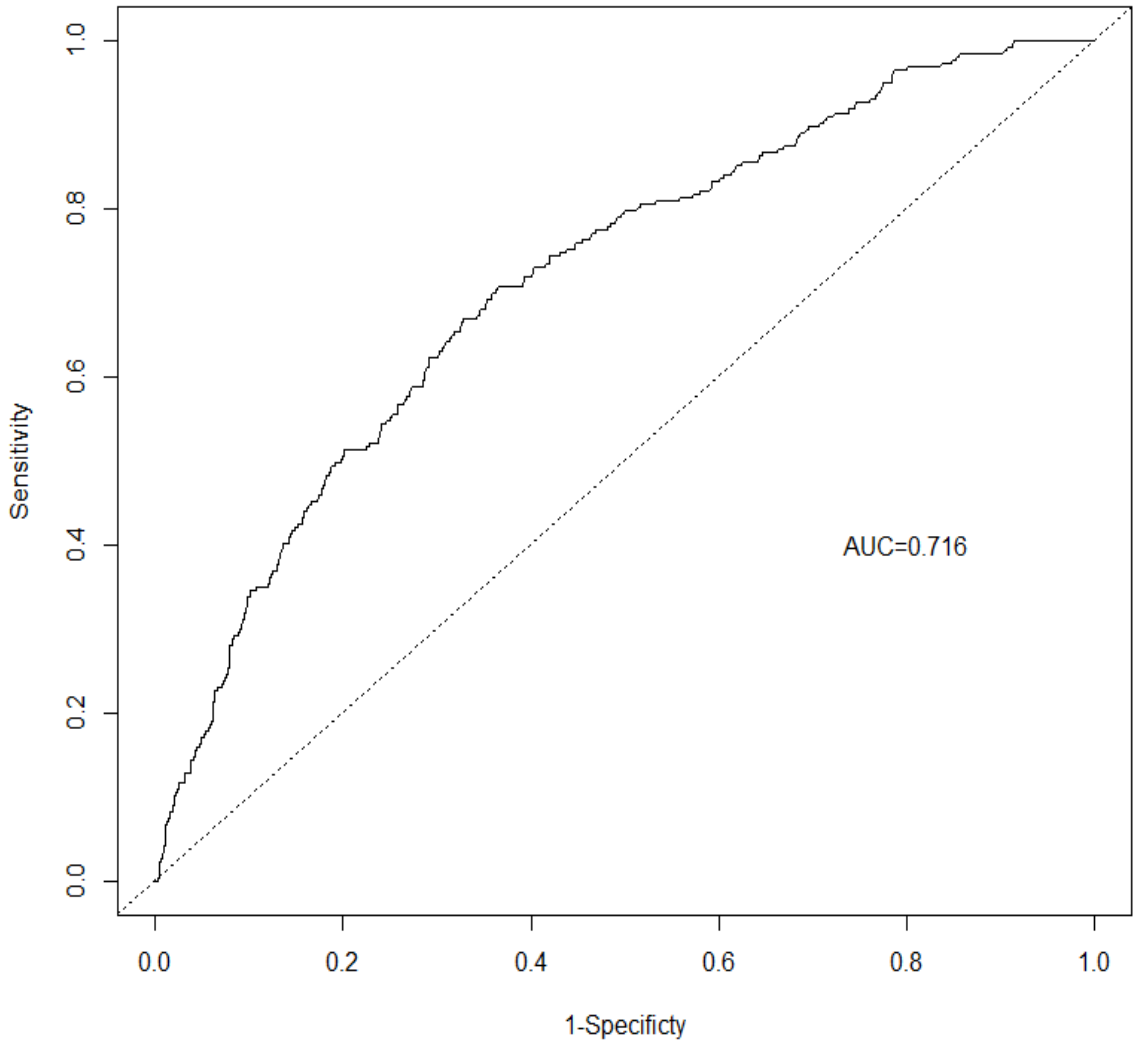
Return to Example

- Let's consider computing the ROC curve for our best multiple variable model

```
> cutpt = seq(0.0,1,.001)
> Sens = numeric(length(cutpt))
> Spec = numeric(length(cutpt))
> for(j in 1:length(cutpt)){
>   pp = factor(ifelse(modelr$fitted.values < cutpt[j],"no","yes"),
>        levels=c("no,"yes"))
>   xx = xtabs(~chd+pp,wcgs)
>   Spec[j] = xx[1,1]/(xx[1,1]+xx[1,2])
>   Sens[j] = xx[2,2]/(xx[2,1]+xx[2,2])
> }

> plot(1-Spec,Sens,type="l",xlab="1-Specificty",ylab="Sensitivity")
> abline(a=0,b=1,lty=3)
```

- Would be better to consider say a 5-fold cross-validation
  - Data set split into 5 equal parts
  - 80% of data used to classify other 20%

ROC Curve



Area under the ROC curve

- If no ties, area equal to Wilcoxon-Mann-Whitney statistic
  - $x_1, \ldots, x_m$: predictions on the $Y = 1$ cases
  - $y_1, \ldots, y_n$: predictions on the $Y = 0$ cases

$$\text{Area} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} I_{\{x_i > y_j\}}}{mn}$$

  - Represents estimate of the probability that the model ranks a randomly chosen $Y = 1$ case higher than $Y = 0$ case
- When ties, need to adjust. Recall c on Slide 48

What if Data Imbalanced?

- ROC will be the same regardless of the dist of Y because it's made up of probabilities that are conditioned on the true outcome
  - Specificity $= P(\hat{Y} = 0 | Y = 0)$
  - Sensitivity $= P(\hat{Y} = 1 | Y = 1)$
- When highly imbalanced, some researchers focus on assessing precision and recall
  - Precision $= P(Y = 1 | \hat{Y} = 1)$
  - Recall $= P(\hat{Y} = 1 | Y = 1)$
- Recall and sensitivity are the same probability
- Helpful when $Y = 1$ outcomes more interesting
- Often determine cutput using harmonic mean of metrics

$$F_i = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Estimation Issue

- Fitting algorithm can struggle when the groups of $Y = 1$ and $Y = 0$ responses are nearly linearly separated.
- Probabilities are tending towards 0 or 1, resulting in unstable estimates and SEs.
- Approaches to handle this issue
  - Bayesian analysis - prior on $\boldsymbol{\beta}$
  - Penalized regression methods - packages logistf, brglm
  - Exact logistic regression - packages elrm, logistiX
- Firth: maximize $l(\boldsymbol{\beta}) + \log |\mathrm{I}(\boldsymbol{\beta})| /2$