# STA303/1002 - Methods of Data Analysis II

(Week 07 lecture note - Extra Topics)

Wei (Becky) Lin

Feb 14/16, 2017

UNIVERSITY OF TORONTO

# Notes

- Assignment 2 is due 22:00, Sunday, Feb 26, 2017.
  - DO NOT submit it until last 10 minutes. 5% penalty is on for being just 1 minute.
  - OH for A2: 3-5pm, Friday, Feb 24, 2017.
- OH during reading week: 3-5pm Friday,Feb 24, 2017. (no more OH on Wednesday in that week)
- Midterm
  - Time and Location:
    - L0101 section: 10-12, Thursday, Mar 2nd. Rooms TBA.
    - L0201 section: 11-13, Thursday, Mar 2nd at **EX100**.
  - Coverage: Week 1 to Week 7.
  - My OH for midterm: Wednesday 12-2pm at SS6011

# Course and Book Recommendation about Data Science

- An introduction book, only application in R, no scary theory.
  - *ISL: An Introduction to Statistical Learning (with applications in R)*
  - Book is free: `http://www-bcf.usc.edu/~gareth/ISL/`

## An Introduction to Statistical Learning

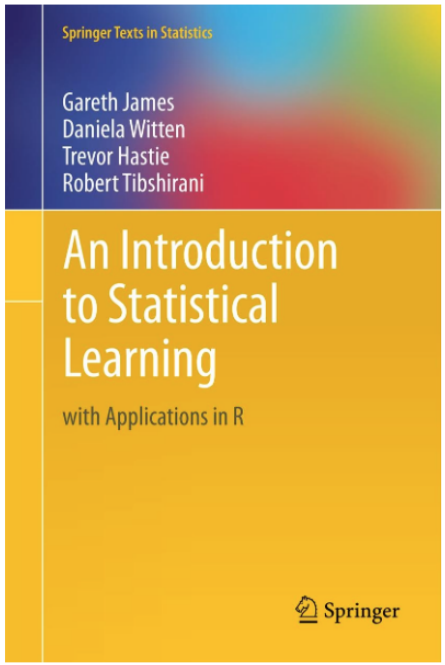
**with Applications in R**

**Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani**

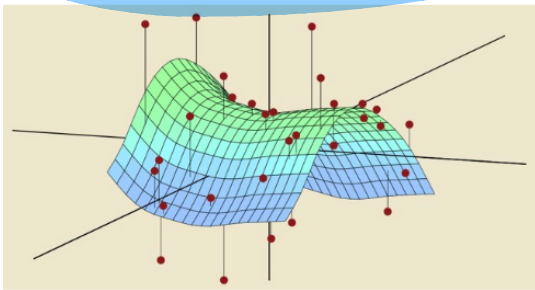
free ^_^

**Home**
**About this Book**
**R Code for Labs**
**Data Sets and Figures**
**ISLR Package**
**Get the Book**
**Author Bios**

**Errata**




**Download the book PDF**
**(corrected 6th printing)**



*Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Start anytime in self-paced mode.*

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students, masters students and Ph.D. students in the non-mathematical sciences. The book also contains a number of R labs with detailed explanations on how to implement the various methods in real life settings, and should be a valuable resource for a practicing data scientist.

# Course and Book Recommendation about Data Science

- Machine learning, one of th hotest courses on CourseRa.
  - Instructor: Andrew Ng, is also the co-founder and chairman of Coursera.
  - Link: `https://www.coursera.org/learn/machine-learning`

**Created by:** Stanford University

Stanford University

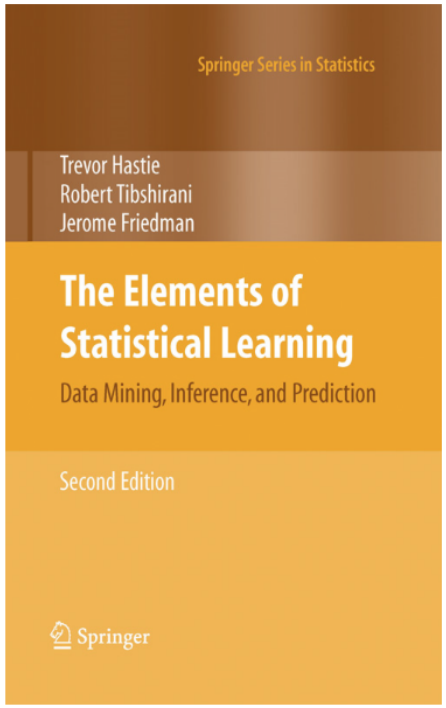
Strongly Recommend it to you!

**Taught by:** Andrew Ng, Associate Professor, Stanford University; Chief Scientist, Baidu; Chairman and Co-founder, Coursera

| | |
|---|---|
| **Language** | English, **Subtitles:** Spanish, Japanese, Chinese (Simplified) |
| **How To Pass** | Pass all graded assignments to complete the course. |
| **User Ratings** | ★★★★☆ Average User Rating 4.9 See what learners said |

# Course and Book Recommendation about Data Science

- Go deeper on data mining: a bit more theory.
  - *ESL: The Elements of Statistical Learning: data mining, inference and prediction*
  - Bood is free: `http://web.stanford.edu/~hastie/ElemStatLearn/`




***The Elements of Statistical Learning:***

**Data Mining, Inference, and Prediction.**

**Second Edition**

**February 2009**

Trevor Hastie

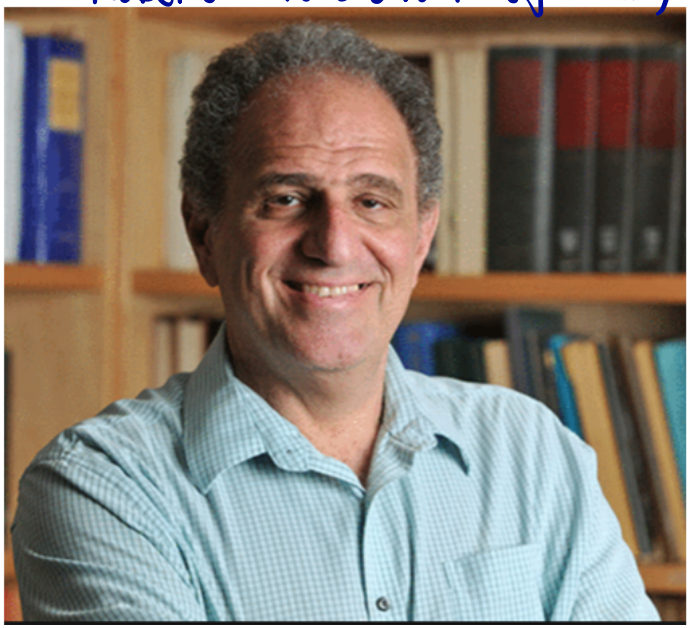
Robert Tibshirani → LASSO, 1996

Jerome Friedman

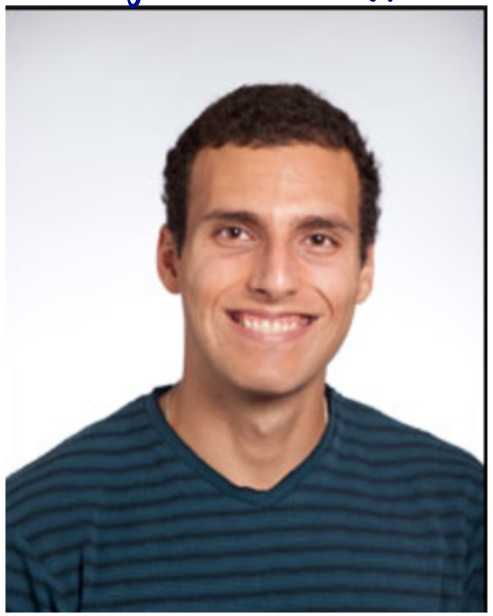
**What's new in the 2nd edition?**

# Course and Book Recommendation about Data Science

Robert Tibshirani (father)



Ryan Tibshirani (son)



- Ryan Tibshirani's website at CMU
  - Nice slides from ISL, ESL: `http://www.stat.cmu.edu/~ryantibs/datamining/`
  - Video for lectures: `http://www.stat.cmu.edu/~ryantibs/teaching.html`
- Want to learn from real data? `http://www.kaggle.com`

# Review: Topics learned last week

- Weighted Least squared to LOESS and LOWESS

WLS
LOESS / LOWESS
different def'n on w

$$Q_{wls} = \sum_{i}^{n} w_i (y_i - \vec{x}_i \cdot \beta)^2 = (Y - X\beta)' W (Y - X\beta)$$

- Bias-variance trade off
- Ridge regression (L2 penalty)

$$\min_{\beta} \sum_{i}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2, \quad \text{Subject to } \sum_{j=1}^{p} \beta_j^2 < C$$

L2 penalty

$$\rightarrow \min_{\beta} \sum_{i}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda_C \sum_{j=1}^{p} \beta_j^2$$

- LASSO (L1 penalty)

$$\min_{\beta} \sum_{i}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2, \quad \text{Subject to } \sum_{j=1}^{p} |\beta_j| < C$$

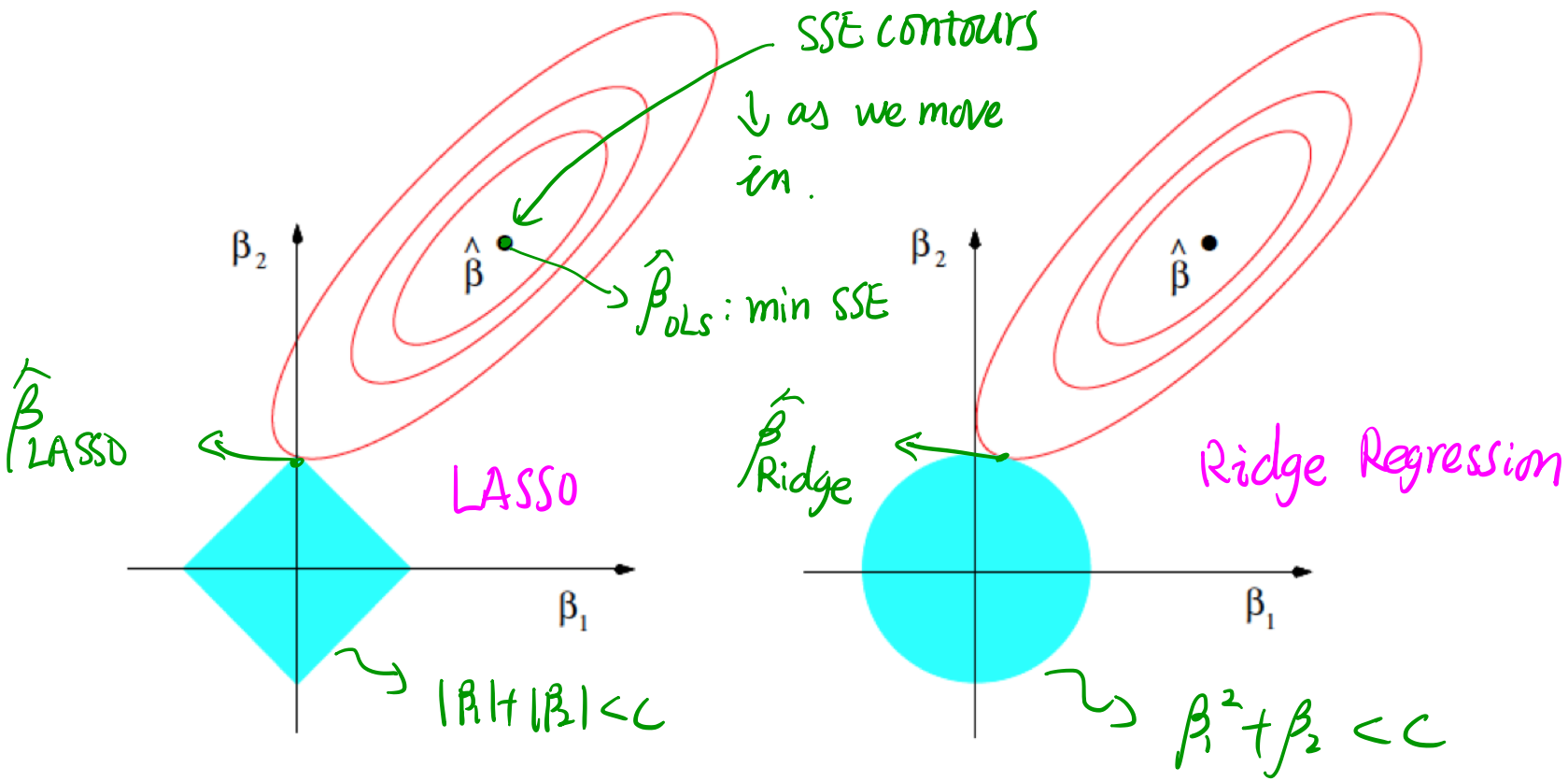
L1 penalty

$$\rightarrow \min_{\beta} \sum_{i}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda_C \sum_{j=1}^{p} |\beta_j|$$

# Review: Topics learned last week

- Ridge vs LASSO

  The lasso combines some of the shrinking advantages of ridge with variable selection




(From ESL page 71)

# Review: Topics learned last week

- Recall in linear regression with normal error model

$$Y_i = X_i\beta + \epsilon_i, E(Y_i|X_i) = \mu_i = X_i\beta$$

  - continuos response.
  - errors assumed to be normal, so $Y_i \sim N(X_i\beta, \sigma^2)$
  - variance of errors don't dependent on mean.

- Binary outcome
  - Response (Y) is either 0 or 1. $Y \sim Bern(\pi), \pi = P(Y = 1)$ $= E(Y)$
  - E(Y)=$\pi$ is restricted in [0,1] interval. Var(Y)= $\pi(1 - \pi)$
  - The idea of "normal error" does not apply here.

- Logistic regression for binary response variable
  - Apply logit transformation on the mean, $\mu_i = E(Y_i) = \pi_i$

$\omega$ = odds
$\eta = X\beta$

$$g(\mu_i) = logit(\pi) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

$$\omega = \frac{\pi_i}{1 - \pi_i}, \eta = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

- The inverse of the logit function (sigmoid function) gives the $\pi$,

$\pi = sigmoid(\eta) = \frac{e^\eta}{1+e^\eta}$

$$\pi = \frac{\omega}{1 + \omega} = sigmoid(\eta) = \frac{e^\eta}{1 + e^\eta}$$

# Review: Topics learned last week

- This is called a Generalized Linear Model with link function g(·)

$$E(Y) = \mu, g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

$* \boxed{g(\mu) = X\beta}$

- Coefficient interpretation in logistic regression

$$g(\mu) = logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

$\hat{p} = \hat{P}(Y=1) = \frac{\#(Y=1)}{n}$

$\Downarrow$

$\hat{\beta}_0 = \log\frac{\hat{p}}{1-\hat{p}}$

  - $\beta_0 = \log\left(\frac{\pi}{1-\pi}\right)$ is a logit function of the marginal probability that $Y = 1$, ignoring all the predictors or when all predictors X=0.
  - Similar to LR model, holding all the other predictors fixed, changing $X_k$ by one unit leads to a change in log odds, $\eta$, of $\beta_k$.
  - Or, holding all the other predictors fixed, changing $X_k$ by one unit leads to a change in the odds ratio of $\beta_k$.

$$\omega(X_k + 1) = \omega(X_k)e^{\beta_k} \approx \omega(X_k)(1 + \beta_k) \rightarrow \underbrace{\frac{\omega(X_k + 1)}{\omega(X_k)}}_{= OR} - 1 \approx \beta_k$$

  - The sign of $\beta_k$ indicates whether $\eta$ or $\omega$ increases ($\beta_k > 0$) or decreases ($\beta_k < 0$) as $X_k$ increases.

# Learning objectives this week

what is canonical parameter?
slide 16

- Generalized linear model
  - Link function: function that links the mean to the canonical parameter in the exponential model form
  - Bernoulli model
  - Conditional distribution of $Y|X$: exponential family
- Data example: interpretation of coefficient in logistic regression
- How to estimate regression coefficients in logistic regression: IRLS algorithm
- Inference on individual $\beta$
- Deviance analysis
- Goodness-of-fit

# Generalized Linear Model

$$\mu = E(Y), \quad g(\mu) = X\beta$$

- $g(\mu)$ is called a Link Function.
  - $g(\mu) = \mu$ is vanilla regression - Identity link function
  - $g(\mu) = \log(\mu)$ is Poisson regression - Log Link
  - $g(\mu) = \log(\frac{\mu}{1-\mu})$ is Logistic regression - Logit link
  - $g(\mu) = \phi^{-1}(\mu)$ is Probit regression - Probit Link
- Use MLE to get parameter estimates (coming soon for logistic regression)

↳ to estimate $\hat{\beta}$ in $g(\mu) = X\beta$, use MLE.

# GLM: Bernoulli model

- Y is binary. Assume $Y|X_1, \ldots, X_p \sim$ Bernoulli distribution with

$$P(Y = 1|X_1, \ldots, X_p) = p(X_1, \ldots, X_p) \in [0, 1]$$

- Suppose there is an unobserved latent variable $Z$ follows the linear model

$$\boxed{Z = -\beta_1 X_1 - \ldots - \beta_p X_p + \epsilon, \quad \epsilon \sim F}$$

F is distm for ε

  - For linear regression model, we have $Y = Z, F = \Phi(\cdot)$
  - For binary response variable, we observed $Y = I(Z \leq \tau)$ according to certain thresholding parameter $\tau$. Then

$$Y|X_1, \ldots, X_p \sim \text{Bernoulli}\Big(p(X_1, \ldots, X_p)\Big), p(X_1, \ldots, X_p) = E(Y|X_1, \ldots, X_p)$$

$$\begin{aligned}
E(Y) &= E\{I(Z \leq \tau)\} = P(Z \leq \tau) \\
&= P(Z + \beta_1 X_1 + \ldots + \beta_p X_p \leq \tau + \beta_1 X_1 + \ldots + \beta_p X_p) \\
&= P(\epsilon \leq \tau + \beta_1 X_1 + \ldots + \beta_p X_p) \\
&= F(\tau + \beta_1 X_1 + \ldots + \beta_p X_p) \\
&= p(X_1, \ldots, X_p)
\end{aligned}$$

  - If there is an intercept term in the linear model, let $\tau = 0$. This gives

$$E(Y|X) = p(Y|X_1, \ldots, X_p) = F(\beta_1 X_1 + \ldots + \beta_p X_p)$$

# GLM: Bernoulli model

$$E(Y|X) = p(Y|X_1, \ldots, X_p) = F(\beta_1 X_1 + \ldots + \beta_p X_p)$$

- If $\epsilon$ is distributed a logistic distribution (receives this name from its CDF),

PDF →
$$f(\epsilon; \mu, s) = \frac{e^{-(\epsilon-\mu)/s}}{s(1 + e^{-(\epsilon-\mu)/s})^2}$$

CDF →
$$F(x) = P(\epsilon \leq x) = \frac{e^{(x-\mu)/s}}{1 + e^{(x-\mu)/s}}$$

- let $\mu = 0, s = 1$ and $x = \beta_1 X_1 + \ldots + \beta_p X_p$, we then have $p(X_1, \ldots, X_p) = \mu$, it gives us a logit link

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \beta_1 X_1 + \ldots + \beta_p X_p$$

$$\leftrightarrow \quad \text{logit}(\mu) = \eta = X\beta$$

# GLM: Bernoulli model

$$p(Y|X_1, \ldots, X_p) = F(\beta_1 X_1 + \ldots + \beta_p X_p)$$

- If $\epsilon \sim N(0, \sigma^2)$, then

$$p(X_1, \ldots, X_p) = \Phi(\frac{\beta_1 X_1 + \ldots + \beta_p X_p}{\sigma}) = \Phi(\alpha_1 X_1 + \ldots + \alpha_p X_p)$$

$$\eta = \mu_Y = p(Y|X_1, \ldots, X_p) = F(\beta_1 X_1 + \ldots + \beta_p X_p)$$

  - This leads to probit link $g(\mu) = \Phi^{-1}(\alpha_1 X_1 + \ldots + \alpha_p X_p)$

- If $\epsilon \sim \text{Gumbel}(\mu, s)$ then

PDF $\longrightarrow$ $f(\epsilon; \mu, s) = \frac{1}{s} e^{-(z+e^{-z})}, \quad z = \frac{\epsilon - \mu}{s}$

CDF $\longrightarrow$ $F(x) = P(\epsilon \leq x) = e^{-e^{-(x-\mu)/s}}$

$$\mu_Y = p(Y|X_1, \ldots, X_p) = F(\beta_1 X_1 + \ldots + \beta_p X_p)$$

$$\eta = g(\mu_Y) = \log\big(-\log(1 - \mu_Y)\big) = \beta_1 X_1 + \ldots + \beta_p X_p$$

# GLM: conditional distribution

*

- In GLM, it is assumed that $Y|\vec{X}$ belongs to a canonical exponential family which p.d.f has the following form

$$\boxed{f(y|x) = \exp\left\{\frac{\theta(x)y - b\big(\theta(x)\big)}{a(\phi)} + c(y;\phi)\right\}}$$

free of $\theta$ (no $\theta$)

for some known function $a(\cdot), b(\cdot), c(\cdot,\cdot)$.

  - $b\big(\theta(x)\big)$ is called a canonical parameter
  - $\phi$ is called a dispersion parameter

- Can be shown that (proof, refer to PP116-117 in the PDF by Prof Fan)

Exercise {
  - $\mu = E(Y) = b'(\theta) = \partial b(\theta)/\partial\theta$
  - $Var(Y) = a(\phi)b''(\theta)$

- Definition: The function g that links the mean to the canonical parameter is called the canonical link

$$g(\mu) = \theta, \text{ since } \mu = b'(\theta) \rightarrow g(\mu) = (b')^{-1}(\mu)$$

$$\Rightarrow \boxed{g(\cdot) = (b')^{-1}(\cdot)} \leftarrow \text{Canonical Link}$$

# GLM: conditional distribution

- Lin. Reg. model example: $Y|x \sim N(\mu(x), \sigma^2)$

$$f(y;\mu,\sigma) = \exp\left\{\frac{\mu(x)y - \mu(x)^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right\}$$

  - $\theta = \mu$, $a(\phi) = \sigma^2$, $b(\theta) = \theta^2/2$, $c(y,\phi) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$
  - $g(\mu) = \mu$ $\quad \hookrightarrow b'(\theta) = \theta = \mu \rightarrow g(\mu) = \mu.$

- Binomial distribution: $Y|x \sim B(m, p(x)), 0 < p(x) < 1$

$$P(Y=y|X=x) = \binom{m}{y} p^y (1-p)^y = \exp\{y\log(\frac{p}{1-p}) + m\log(1-p) + \log\binom{m}{y}\}$$

$\rightarrow p = \frac{e^\theta}{1+e^\theta}$

  - $\theta = \log(\frac{p}{1-p})$, $b(\theta) = -m\log(1-p)$, $c(y;\phi) = \log\binom{m}{y}$ with $\phi = 1$
  - $g(\mu) = g(p) = \log(\frac{p}{1-p})$

$\Downarrow$

$b(\theta) = -m\log(1 - \frac{e^\theta}{1+e^\theta}) = +m\log(1+e^\theta) \rightarrow b'(\theta) = m\frac{e^\theta}{1+e^\theta} = m\mu$

$\rightarrow \boxed{\theta = \log\frac{\mu}{1-\mu}}$

- Exercise: for $Y|x \sim Poisson(\lambda(x)$, find $\theta, b(\theta), \phi, c(y;\phi)$

# GLM: summarizes canonical link functions for common distributions

Summary Table

| Distribution | $b(\theta)$ | Canonical Link | Range |
|---|---|---|---|
| Normal | $\theta^2/2$ | $g(\mu) = \mu$ | $(-\infty, \infty)$ |
| Poisson | $\exp(\theta)$ | $g(\mu) = \log(\mu)$ | $(-\infty, \infty)$ |
| Binomial | $-\log(1 - e^{\theta})$ | $g(\mu) = \frac{\mu}{1-\mu}$ | $(-\infty, \infty)$ |
| Gamma | $-\log(-\theta)$ | $g(\mu) = \frac{1}{\mu}$ | $(0, \infty)$ |

* STA303

Theorem: If the conditional distribution of $Y|X$ belongs to the canonical exponential family, then

- 1. The canonical link function is strictly increasing.
- 2. The likelihood function $\ell(\theta) = \log L(\theta) = \log(\prod_i^n f(y_i; \theta))$ using canonical parameter strictly concave.

Proof:

- 1. Show $dg(\mu)/d\theta = 1/b''(\theta) > 0$.
- 2. Show $\ell''(\theta) = -b''(\theta)/a(\phi) < 0$

Data Example
Interpretation of regression coefficients

# Data: graduate school admission

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

- Admit: binary response (outcome, dependent) variable
- gre: predictor, continuos variable.
- gpa: predictor, continuos variable.
- rank: factor variable, takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.
- Reference of this data example: `http://www.ats.ucla.edu/stat/r/dae/logit.htm`

```r
mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
str(mydata)
```

```
## 'data.frame':    400 obs. of  4 variables:
##  $ admit: int  0 1 1 1 0 1 1 0 1 0 ...
##  $ gre  : int  380 660 800 640 520 760 560 400 540 700 ...
##  $ gpa  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
##  $ rank : int  3 3 1 4 4 2 1 2 3 2 ...
```

# Data: graduate school admission

```r
mydata$rank=as.factor(mydata$rank)
head(mydata)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

```r
## two-way contingency table of categorical outcome and predictors
## we want to make sure there are not 0 cells
xtabs(~ admit + rank, data = mydata)
```

```
##      rank
## admit  1  2  3  4
##     0 28 97 93 55
##     1 33 54 28 12
```

in total, 55 observations in category: rank=4, admit=0

similarly, for the remaining #s.

# Interpretation of the intercept

- $\beta_0 = \log(\frac{\pi}{1-\pi})$ or, $e^{\beta_0} = \frac{\pi}{1-\pi}$ $\pi = p(Y=1)$

```r
p=mean(mydata$admit); cbind( p,b0=log(p/(1-p)) ) # p=(# of 1)/n, b0
```

```
##           p         b0
## [1,] 0.3175 -0.7652847
```

$\boxed{b_0 = \log \hat{\pi}/(1-\hat{\pi})}$

```r
fit0 <- glm(admit~1, family=binomial,data=mydata)
summary(fit0)
```

```
##
## Call:
## glm(formula = admit ~ 1, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8741  -0.8741  -0.8741   1.5148   1.5148
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7653     0.1074  -7.125 1.04e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 499.98  on 399  degrees of freedom
## AIC: 501.98
##
```

no predictors
only have intercept.

# Interpretation of regression slope of a Categorical Predictor

$$logit(\pi) = \eta = \beta_0 + \beta_1 Lev_2 + \beta_2 Lev_3 + \beta_3 Lev_4$$

- In SLR, $\beta_1$ is the difference in means between the level 2 group and the reference group (level 1).
- In logistic regression, $\beta_1 = \log \frac{p_2}{1-p_2} - \log \frac{p_1}{1-p_1} = \log \frac{p_2/(1-p_2)}{p_1/(1-p_1)}$ is the increases in the log-odds that $Y = 1$, comparing the level 2 group to the reference group (level 1).
- $e^{\beta_1} = \frac{p_2/(1-p_2)}{p_1/(1-p_1)}$ is the Odds Ratio (OR)

# Interpretation of regression slope of a Categorical Predictor

```r
fit1 <- glm(admit~rank, family=binomial,data=mydata)
summary(fit1)
```

```
##
## Call:
## glm(formula = admit ~ rank, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2479  -0.9408  -0.7255   1.1085   1.8546
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1643     0.2569   0.639   0.5225
## rank2        -0.7500     0.3080  -2.435   0.0149 *
## rank3        -1.3647     0.3354  -4.069 4.72e-05 ***
## rank4        -1.6867     0.4093  -4.121 3.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 474.97  on 396  degrees of freedom
## AIC: 482.97
##
## Number of Fisher Scoring iterations: 4
```

$\hat{\beta}$ $se(\hat{\beta})$ $\hat{\beta}-0/se(\hat{\beta})$

$$0.1643 = \log \frac{P_{Rank1}}{1-P_{Rank1}} = \log \frac{P_1}{1-P_1}$$

$$-0.75 = \log \frac{P_2/(1-P_2)}{P_1/(1-P_1)}$$

$\vdots$

# Interpretation of regression slope of a Categorical Predictor

```r
(pvec = with(mydata,tapply(admit,rank, mean)) )
```

```
##         1         2         3         4
## 0.5409836 0.3576159 0.2314050 0.1791045
```

$p_1$ $p_2$ $p_3$ $p_4$

```r
p1 = pvec[1]/(1-pvec[1])
log(pvec[-1]/(1-pvec[-1])/p1 ) # log(Odds)
```

$\hat{\beta}[2,3,4]$

```
##         2         3         4
## -0.750030 -1.364698 -1.686730
```

$\Leftarrow \log \frac{p_i/(1-p_i)}{p_1/(1-p_1)}, \; i=2,3,4$

```r
pvec[-1]/(1-pvec[-1])/p1 # OR
```

```
##         2         3         4
## 0.4723524 0.2554578 0.1851240
```

$\Leftarrow \frac{p_i/(1-p_i)}{p_1/(1-p_1)} = OR_i, \; i=2,3,4$

```r
exp(coef(fit))
```

```
## (Intercept)       rank2       rank3       rank4
##   1.1785714   0.4723524   0.2554578   0.1851240
```

# Interpretation of $\beta$ with mixed predictors

Prob of survival

$$logit(\pi) = \beta_0 + \beta_1 \text{ age} + \beta_2 I_{female}$$

- In LR model
  - $\beta_1$ is the increase in survival for a unit change in age, for each level of sex.
  - $\beta_2$ is the difference between group means, controlling for age.
- In logistic regression
  - $\beta_1$ is the increase in the log-odds of survival, for a unit change in age, for each level of sex (must be the same)
  - $\beta_2$ is the increase in the log-odds of survival for females compared to males, controlling for age.

# Interpretation of $\beta$ with mixed predictors

```r
fit3 <- glm(admit~gpa+gre+rank, family=binomial,data=mydata)
summary(fit3)
```

```
##
## Call:
## glm(formula = admit ~ gpa + gre + rank, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## gpa          0.804038   0.331819   2.423 0.015388 *
## gre          0.002264   0.001094   2.070 0.038465 *
## rank2       -0.675443   0.316490  -2.134 0.032829 *
## rank3       -1.340204   0.345306  -3.881 0.000104 ***
## rank4       -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

? Later.

# Interpretation of $\beta$ with mixed predictors

- For every one unit change in gre, the log odds of admission (versus non-admission) increases by 0.002.
- For a one unit increase in gpa, the log odds of being admitted to graduate school increases by 0.804.
- The indicator variables for rank have a slightly different interpretation. For example, having attended an undergraduate institution with rank of 2, versus an institution with a rank of 1, decreases the log odds of admission by 0.675.

$\hat{\beta}$ for rank2 = $-0.675$

- The estimated curve $\hat{\pi}$ is

$$\eta = -3.9899 + 0.8040gpa + 0.0023gre - 0.6754I_{rank2} - 1.3402I_{rank3} - 1.5515I_{rank4}$$

$$\hat{\pi} = \frac{e^{\eta}}{1 + e^{\eta}}$$

# Fitting logistic regression to binominal data

```r
ucb.df = data.frame(gender=rep(c("Male","Female"),c(6,6)),
                    dept=rep(LETTERS[1:6],2),
                    yes=c(512,353,120,138,53,22,89,17,202,131,94,24),
                    no=c(313,207,205,279,138,351,19,8,391,244,299,317))
ucb.df
```

```
##    gender dept yes  no
## 1    Male    A 512 313
## 2    Male    B 353 207
## 3    Male    C 120 205
## 4    Male    D 138 279
## 5    Male    E  53 138
## 6    Male    F  22 351
## 7  Female    A  89  19
## 8  Female    B  17   8
## 9  Female    C 202 391
## 10 Female    D 131 244
## 11 Female    E  94 299
## 12 Female    F  24 317
```

$Y$

$Y_1 \sim Bin(512+313, p)$, we observed $Y_1 = 512$

# Fitting logistic regression to binominal data

```r
glm.ucb = glm(cbind(yes,no)~gender+dept+gender*dept,
              family=binomial(link="logit"), data=ucb.df)
summary(glm.ucb)
```

```
##
## Call:
## glm(formula = cbind(yes, no) ~ gender + dept + gender * dept,
##     family = binomial(link = "logit"), data = ucb.df)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.5442     0.2527   6.110 9.94e-10 ***
## genderMale         -1.0521     0.2627  -4.005 6.21e-05 ***
## deptB              -0.7904     0.4977  -1.588  0.11224
## deptC              -2.2046     0.2672  -8.252  < 2e-16 ***
## deptD              -2.1662     0.2750  -7.878 3.32e-15 ***
## deptE              -2.7013     0.2790  -9.682  < 2e-16 ***
## deptF              -4.1250     0.3297 -12.512  < 2e-16 ***
## genderMale:deptB    0.8321     0.5104   1.630  0.10306
## genderMale:deptC    1.1770     0.2996   3.929 8.53e-05 ***
## genderMale:deptD    0.9701     0.3026   3.206  0.00135 **
## genderMale:deptE    1.2523     0.3303   3.791  0.00015 ***
## genderMale:deptF    0.8632     0.4027   2.144  0.03206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

⇒ This is the so called "saturated model". We have 12 obs but we have 12 $\beta_s$ to estimate $(\beta_0, \beta_1 \cdots \beta_{11})$.

How to estimate regression coefficients
in a logistic regression

# IRLS algorithm

- To estimate the unknown $\beta$ in $g(\mu) = \vec{X}\vec{\beta}$, it relies on the Iteratively Reweighted Least Squares algorithm.
- Denote the model log-likelihood function as $\ell(\beta) = \log(\prod_i f(y_i; \beta))$
  1. $\ell'(\hat{\beta}) = 0$, no analytical/ explicit solution. Need find solution iteratively. Given an initial value $\beta_0$ of $\hat{\beta}$, by Taylor's expansion
  $$0 = \ell'(\hat{\beta}) \approx \ell'(\beta_0) + \ell''(\beta_0)(\hat{\beta} - \beta_0)$$
  2. We then have
  $$\hat{\beta} \approx \beta_0 - [\ell''(\beta_0)]^{-1}\ell'(\beta_0)$$
- The Newton-Raphson method: iteratively use the following equation
  $$\boxed{\hat{\beta}_{t+1} = \hat{\beta}_t - [\ell''(\hat{\beta}_t)]^{-1}\ell'(\hat{\beta}_t)} \quad *$$
  until converges, say $|\hat{\beta}_{t+1} - \hat{\beta}_t| \leq \epsilon$.
- If we replace $\ell''(\hat{\beta}_t)$ by any non-degenerate matrix, $A(\hat{\beta}_t)$, then we have
  $$\boxed{\hat{\beta}_{t+1} = \hat{\beta}_t - [A(\hat{\beta}_t)]^{-1}\ell'(\hat{\beta}_t)}$$
  - If $A(\hat{\beta}_t) = I$, the identity matrix, it leads to the deepest decent method. Advantage: don't need to compute the second derivative matrix but convergence is slow.
  - If $A(\hat{\beta}_t) = E\{\ell''(\hat{\beta}_t)\} = -I_n(\hat{\beta}_t)$, we have Fisher Score method:
  $$\boxed{\hat{\beta}_{t+1} = \hat{\beta}_t + I_n(\hat{\beta}_t)^{-1}\ell'(\hat{\beta}_t)} \leftarrow \text{GLM uses it}$$

# IRLS: logistic regression

- The logistic regression function uses $g(\cdot)$ equal to the logit function to describe a distribution with $y \in \{0, 1\}$

$$\mu(x) = \pi(x) = E(Y|x) = logit^{-1}(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

$$P(Y_i = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} = \exp\{y_i \log(\frac{p_i}{1 - p_i}) + \log(1 - p_i)\}$$

- Show $\ell(\beta; \overset{i=1,\cdots n}{y_i|x_i}) = y'X\beta - \log(1 + e^{X\beta})$

Proof: $P(Y_i = y_i) = \exp\{ y_i \log\frac{p_i}{1-p_i} + \log(1-p_i) \}$, $\log\frac{p_i}{1-p_i} = x_i\beta$, $p_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

$$\Rightarrow \ell(\beta; Y_i) = y_i x_i\beta + \log(1 - \frac{e^{x_i\beta}}{1+e^{x_i\beta}})$$

$$= y_i x_i\beta - \log(1+e^{x_i\beta})$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\Rightarrow \ell(\beta; y_i|x_i, i=1,\cdots n) = \sum_{i=1}^n \{ y_i x_i\beta - \log(1+e^{x_i\beta}) \}$$

$$= (y_1, y_2, \cdots y_n)\begin{pmatrix} x_1\beta \\ \vdots \\ x_n\beta \end{pmatrix} - \boxed{\sum_{i=1}^n \log(1+e^{x_i\beta})}$$

$$= Y'X\beta - \boxed{\log(1+e^{X\beta})}$$

# IRLS: logistic regression

Show $\nabla_\beta \ell(\beta; Y|X) = X'(Y - p)$

Proof: From previous slide, we have $\ell(\beta; \vec{Y}|\vec{X}) = Y'X\beta - \sum_{i=1}^n \log(1+e^{x_i\beta})$

① $\frac{\partial}{\partial \beta} Y'X\beta = (Y'X)' = X'Y$

② $\frac{\partial}{\partial \beta_k} \sum_{i=1}^n \log(1+e^{x_i\beta}) = \sum_{i=1}^n x_{ik} \underbrace{\frac{e^{x_i\beta}}{1+e^{x_i\beta}}}$

$$= \sum_{i=1}^n x_{ik} p_i$$

$$= (x_{1k} \ x_{2k} \ \cdots \ x_{nk}) \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}$$

$$= x_k' p$$

$$\Rightarrow \frac{\partial}{\partial \beta} \sum_{i=1}^n \log(1+e^{x_i\beta}) = X'p$$

① + ② $\Rightarrow \frac{\partial}{\partial \beta} \ell(\beta; \vec{Y}|\vec{X}) = \nabla_\beta \ell(\beta; Y|X) = X'Y - X'p = X'(Y-p)$

# IRLS: Logistic regression

Show $H(\beta; Y|X) = -\nabla^2_\beta \ell(\beta; Y|X) = -X'DX$,

where $D = diag\{p_i(1-p_i)\}$, $p = (p_1, \ldots, p_n)'$

Proof: $\ell(\beta; Y|X) = Y'X\beta - \sum_{i=1}^n e^{X_i\beta}/(1+e^{X_i\beta})$

From last derivation we have

$$\frac{\partial}{\partial \beta_k} \sum_{i=1}^n \log(1+e^{X_i\beta}) = \sum_{i=1}^n X_{ik} p_i, \qquad p_i = \frac{e^{X_i\beta}}{1+e^{X_i\beta}} = \frac{1}{1+e^{-X_i\beta}}$$

$$\Rightarrow \frac{\partial}{\partial \beta_j \partial \beta_k} \sum_{i=1}^n \log(1+e^{X_i\beta}) = \frac{\partial}{\partial \beta_j} \sum_{i=1}^n X_{ik} p_i = \sum_{i=1}^n X_{ik} \frac{\partial p_i}{\partial \beta_j}$$

$$\frac{\partial p_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} (1+e^{-X_i\beta})^{-1} = -(1+e^{-X_i\beta})^{-2}(-1)X_{ij}e^{-X_i\beta}$$

$$= X_{ij}\left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)\left(\frac{1}{1+e^{X_i\beta}}\right)$$

$$= X_{ij} p_i(1-p_i) = X_{ij} V(Y_i)$$

$$\Rightarrow \frac{\partial}{\partial \beta_j \partial \beta_k} \sum_{i=1}^n \log(1+e^{X_i\beta}) = \sum_{i=1}^n X_{ik} V(Y_i) X_{ij}$$

$$\Rightarrow H(\beta; Y|X) = -\nabla^2_\beta \ell(\beta; Y|X) = -X'DX \quad \text{where } D = \begin{pmatrix} p_1(1-p_1) & & 0 \\ & \ddots & \\ 0 & & p_n(1-p_n) \end{pmatrix}$$

# IRLS (Newton-Raphson): Logistic regression

$$\hat{\beta}_{t+1} = \hat{\beta}_t - H^{-1}(\hat{\beta}_t)\nabla_\beta \ell(\beta_t; Y|X) = (X'DX)^{-1}X'Dz$$

where $D = diag\{p_i(1-p_i)\}, z = X\hat{\beta}_t + D^{-1}(y-p)$ and $p = (p_1, \ldots, p_n)'$.

$$H(\hat{\beta}_t) = -X'DX, \quad \nabla_\beta \ell(\beta_t; Y|X) = X'(Y-P)$$

$$\Rightarrow \hat{\beta}_{t+1} = \hat{\beta}_t + (X'DX)^{-1}X'(Y-P)$$

$$= \underline{(X'DX)^{-1}(X'DX)}\hat{\beta}_t + \underline{(X'DX)^{-1}X'D}D^{-1}(Y-P)$$

$$= (X'DX)^{-1}X'D[\underbrace{X\hat{\beta}_t + D^{-1}(Y-P)}_{=Z}]$$

$$= (X'DX)^{-1}X'DZ$$

Q.E.D.

# Why is this algorithm important

- This method generalizes to generalized linear models with exponential families, and has some fairly deep connections to Fisher information.
- Helps explain and address convergence properties in generalized linear models.
- The normal equations of logistic regression version can be used to establish large sample theory convergence results.
- Gives us an idea of how the geometric and analytic concepts underlying ridge regression, principal component analysis, and lasso regression connect to generalized linear models.

A generalized IRLS for GLM based on exponential family of distribution is provided by J.Q.Fan (PP124-PP127 in the PDF posted)

Take a break, and see you on Thursday

# Variance estimate of $\hat{\beta}$ in Logistic regression

- $\hat{\beta}$ is the MLE that maximizes the $L(\beta; y|x)$ or the corresonding $\ell(\beta; y|x)$.
- For large n, $\hat{\beta}$ is nearly unbiased. (MLE is consistent estimator)
- $Var(\hat{\beta})$ can be estimated
  - The information matrix $I_{p\times p}(\beta)$

$$\boxed{I_{ij}(\beta) = -\frac{\partial \ell^2(\beta)}{\partial \beta_i \partial \beta_j}}$$

$\beta$ by $\hat{\beta}$

  - The inverse of the observed information approximates the Var($\hat{\beta}$)

$$\boxed{\widehat{Var(\hat{\beta})} = I^{-1}(\hat{\beta})}$$

Note that $\widehat{Var(\hat{\beta})}$ is the var-covariance matrix. The elements on diagonal gives the $var(\hat{\beta}_j)$ and the off diagonal entry is the $cov(\hat{\beta}_i, \hat{\beta}_j)$.

- Among approximately unbiased estimators, the MLE has a variance smaller than any other estimator.
- For large n, the sampling distribution of $\hat{\theta}$ is approximately normal. So the CI can be construsted. $\hat{\theta}-\theta \longrightarrow N(0, I^{-1}(\hat{\theta}))$, $I(\theta)=n i(\theta)$. $i(\theta)$ for one Y
- For a montonic function $h(\cdot)$, $h(\hat{\beta})$ is the MLE of $h(\beta)$. Transformations of MLEs are MLEs: invariance property.

# Wald test and Wald CI

## Inference for individual $\beta$ in logistic regression

- From large sample theory, we have

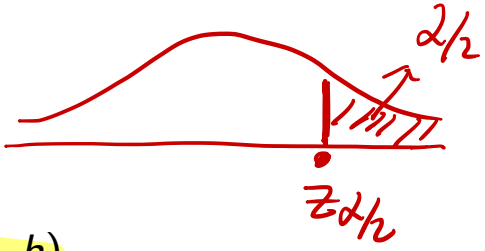
$$\boxed{Z = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim_{approx} N(0,1)}$$

- An approximate confidence interval for $\beta_j$

$$\hat{\beta}_j \pm Z_{\alpha/2}\, se(\hat{\beta}_j) \quad *$$

- Two sided test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$: reject $H_0$ if

$$\left| \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} \right| > Z_{\alpha/2} \quad *$$




- Test of constraint $H_0 : C_{j\times p}\beta_{p\times 1} = C_{j\times 1}$
  - reject if

$$(C\hat{\beta} - h)'\{C' avar(\hat{\beta}) C\}^{-1}(C\hat{\beta} - h)$$

  is larger than $\chi^2_{j,1-\alpha}$ where $Z_{\alpha/2}$ is the usual normal critical value.

R doesn't give you these CIs, but it gives the information needed to calculate it.

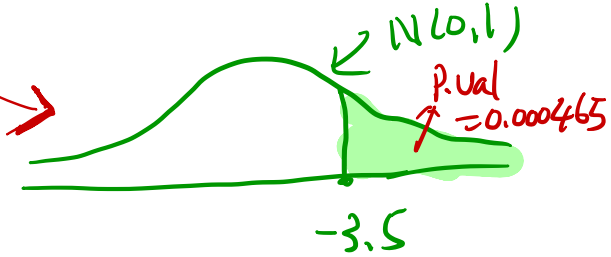
# Revisit the Graduate school admission data

```r
mod <- glm(admit~gpa+gre+rank, family=binomial,data=mydata)
summary(mod)
```

```
##
## Call:
## glm(formula = admit ~ gpa + gre + rank, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## gpa          0.804038   0.331819   2.423 0.015388 *
## gre          0.002264   0.001094   2.070 0.038465 *
## rank2       -0.675443   0.316490  -2.134 0.032829 *
## rank3       -1.340204   0.345306  -3.881 0.000104 ***
## rank4       -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

$\hat{\beta}$ $\quad$ $s.e.(\hat{\beta})$

$\frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$ $\quad$ eg. $-3.5 = \frac{-3.989979}{1.139951}$



# ANOVA table and Estimate of Var-covariance of $\beta$

```r
options(show.signif.stars=F)
anova(mod,test="Chisq")
```

→ disable "** ..."

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: admit
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    399     499.98
## gpa   1  13.0089       398     486.97   0.00031
## gre   1   6.6236       397     480.34   0.01006
## rank  3  21.8265       394     458.52 7.088e-05
```

```r
print(vcov(mod),digits=3)
```

→ give var-cov($\hat{\beta}$)

```
##             (Intercept)       gpa       gre     rank2     rank3     rank4
## (Intercept)    1.299488 -0.303660 -3.01e-04 -8.45e-02 -4.86e-02 -8.94e-02
## gpa           -0.303660  0.110104 -1.24e-04  4.52e-03 -9.47e-03  3.57e-03
## gre           -0.000301 -0.000124  1.20e-06 -1.69e-06  1.86e-05  1.18e-05
## rank2         -0.084476  0.004521 -1.69e-06  1.00e-01  6.96e-02  7.01e-02
## rank3         -0.048644 -0.009469  1.86e-05  6.96e-02  1.19e-01  6.97e-02
## rank4         -0.089431  0.003568  1.18e-05  7.01e-02  6.97e-02  1.75e-01
```

$cov(\hat{\beta}_i, \hat{\beta}_j)$

$var(\hat{\beta}_j)$

```r
head( cbind(y=mydata$admit, fitted=fitted(mod)) )
```

# Confidence interval for $\beta$

```r
## CIs using se, the formula we provided
cbind(b=coef(mod), confint.default(mod) )
```

$\hat{\beta}_j \pm z_{\frac{\alpha}{2}} se(\hat{\beta}_j)$

agree with S.E. column in summary output.

```
##                        b         2.5 %       97.5 %
## (Intercept) -3.989979073 -6.2242418514 -1.755716295
## gpa          0.804037549  0.1536836760  1.454391423
## gre          0.002264426  0.0001202298  0.004408622
## rank2       -0.675442928 -1.2957512650 -0.055134591
## rank3       -1.340203916 -2.0169920597 -0.663415773
## rank4       -1.551463677 -2.3703986294 -0.732528724
```

```r
# CIs using profiled log-likelihood
cbind(b=coef(mod), confint(mod) )
```

```
##                        b         2.5 %       97.5 %
## (Intercept) -3.989979073 -6.2716202334 -1.792547080
## gpa          0.804037549  0.1602959439  1.464142727
## gre          0.002264426  0.0001375921  0.004435874
## rank2       -0.675442928 -1.3008888002 -0.056745722
## rank3       -1.340203916 -2.0276713127 -0.670372346
## rank4       -1.551463677 -2.4000265384 -0.753542605
```

# Confidence interval for $\beta$

You can also exponentiate the coefficients and the corresponding CI and interpret them as odds-ratios.

$$\frac{p_2/(1-p_2)}{p_1/(1-p_1)} = OR \Rightarrow \frac{1}{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

```r
# OR only
exp( coef(mod) )
```

```
## (Intercept)         gpa         gre       rank2       rank3       rank4
##   0.0185001   2.2345448   1.0022670   0.5089310   0.2617923   0.2119375
```

```r
## OR and 95% for OR
exp(  cbind( OR=coef(mod), confint.default(mod) ) )
```

```
##                    OR       2.5 %    97.5 %
## (Intercept) 0.0185001 0.001980825 0.1727834
## gpa         2.2345448 1.166121956 4.2818768
## gre         1.0022670 1.000120237 1.0044184
## rank2       0.5089310 0.273692172 0.9463578
## rank3       0.2617923 0.133055086 0.5150889
## rank4       0.2119375 0.093443470 0.4806919
```

CI for $OR_i$

CI for $OR = \frac{p_2/(1-p_2)}{p_1/(1-p_2)}$

$\shortparallel$

$(L, U)$

$\Downarrow$

$$\boxed{\left(\frac{1}{U}, \frac{1}{L}\right) = \text{CI for } \frac{1}{OR}}$$

# Testing of constraint $H_0$

Using Wald.test() in R to test
$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$

- b: supplies the coefficients
- Sima: supplies the variance covariance matrix of the error terms,
- Terms: tells R which terms in the model are to be tested, in this case, terms 4, 5, and 6, are the three terms for the levels of rank
- L: tells R that we wish to base the test on the specified vector

```r
# Testing b1=b2=0
library(aod)
wald.test(b = coef(mod), Sigma = vcov(mod), Terms = 4:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 20.9, df = 3, P(> X2) = 0.00011
```

↳ Reject $H_0$. We have strong evidence that at least one of $\beta_4, \beta_5, \beta_6$ is nonzero.

# Testing of constraint $H_0$

$$\beta_4 - \beta_5 = 0 \Rightarrow \begin{pmatrix} 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_4 \\ \beta_5 \end{pmatrix} = 0$$

Using Wald.test() in R to test
$H_0 : \beta_3 - \beta_4 = 0$ this is same as $H_0 : C\beta = 0$ where C=(0,0,0,1,-1,0)

```
# Testing b1=b2=0
C=cbind(0,0,0,1,-1,0)
wald.test(b = coef(mod), Sigma = vcov(mod), L = C)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 5.5, df = 1, P(> X2) = 0.019
```

$\rightarrow$ we have evidence that $\beta_3 \neq \beta_4$.

# Predict probability of admission

create X for predicting

```
newdata1 <- with( mydata, data.frame(gre = mean(gre), gpa = mean(gpa), rank = factor(1:4))
newdata1$rankP <- predict(mod, newdata = newdata1, type = "response")
newdata1
```

asking for $\hat{\pi}$

$\hat{\pi}$

```
##     gre    gpa rank     rankP
## 1 587.7 3.3899    1 0.5166016
## 2 587.7 3.3899    2 0.3522846
## 3 587.7 3.3899    3 0.2186120
## 4 587.7 3.3899    4 0.1846684
```

X information

# Predict probability of admission

Prediction + corresponding CI.

```
newdata2 <- with(mydata,
  data.frame(gre = rep(seq(from = 200, to = 800, length.out = 100), 4),
             gpa = mean(gpa), rank = factor(rep(1:4, each = 100))))

newdata3 <- cbind(newdata2, predict(mod, newdata = newdata2, type="link", se=TRUE))
newdata3 <- within(newdata3, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})

## view first few rows of final dataset
head(newdata3)
```

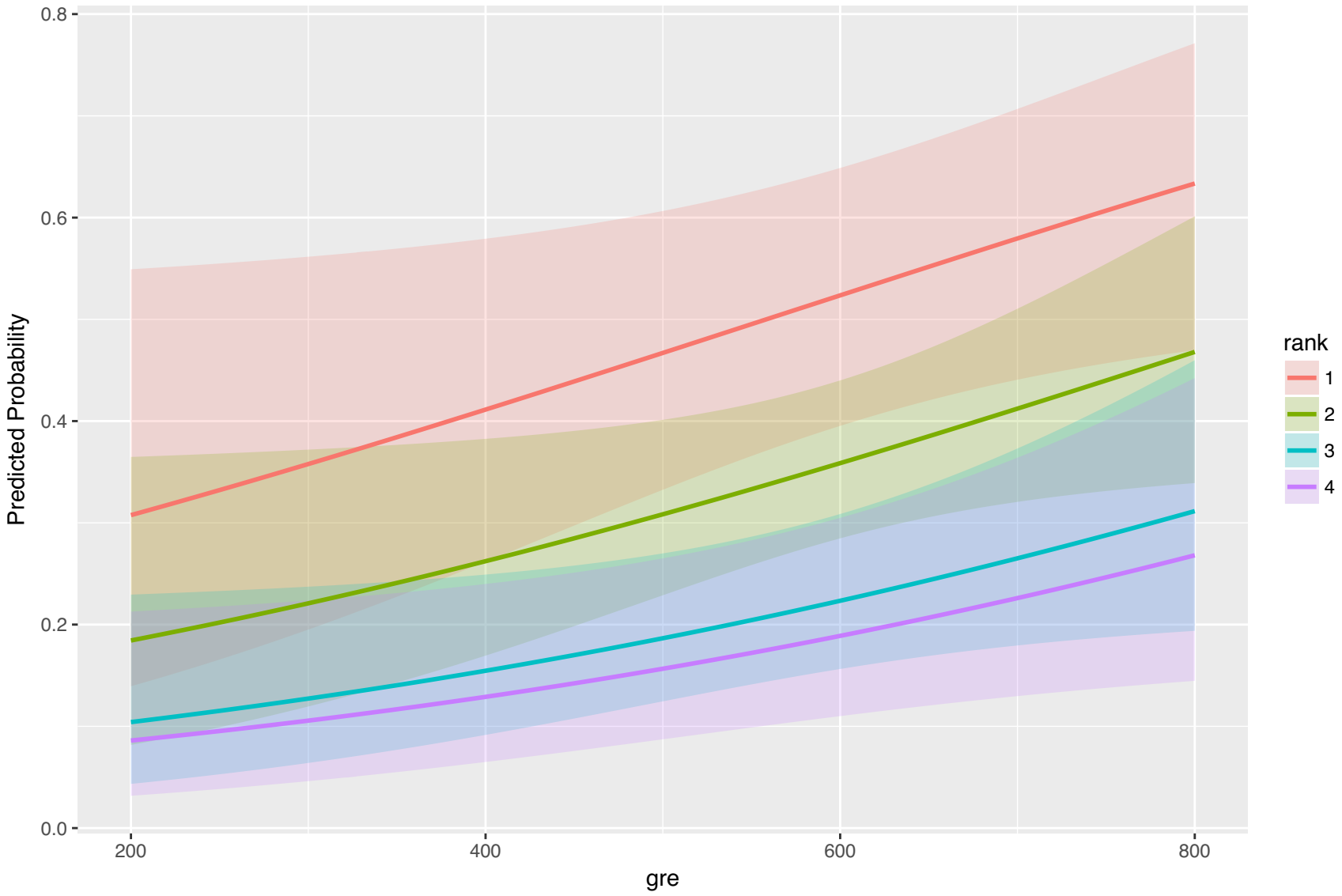
$\eta$ = eta

```
##        gre    gpa rank        fit    se.fit residual.scale        UL
## 1 200.0000 3.3899    1 -0.8114870 0.5147714              1 0.5492064
## 2 206.0606 3.3899    1 -0.7977632 0.5090986              1 0.5498513
## 3 212.1212 3.3899    1 -0.7840394 0.5034491              1 0.5505074
## 4 218.1818 3.3899    1 -0.7703156 0.4978239              1 0.5511750
## 5 224.2424 3.3899    1 -0.7565919 0.4922237              1 0.5518545
## 6 230.3030 3.3899    1 -0.7428681 0.4866494              1 0.5525464
##          LL PredictedProb
## 1 0.1393812     0.3075737
## 2 0.1423880     0.3105042
## 3 0.1454429     0.3134499
## 4 0.1485460     0.3164108
## 5 0.1516973     0.3193867
## 6 0.1548966     0.3223773
```

$$\pi = \frac{e^{\eta}}{1+e^{\eta}}$$

# Predict probability of admission

```
library(ggplot2)
ggplot(newdata3, aes(x = gre, y = PredictedProb)) +ylab("Predicted Probability")+
      geom_ribbon(aes(ymin = LL, ymax = UL, fill = rank), alpha = .2) +
      geom_line(aes(colour = rank), size=1)
```



# Assessment of model fit

- In linear regression, we use the F-test

$$F = \frac{(SSE(RM) - SSE(FM))/\triangle_{df}}{SSE(FM)/df(FM)} \sim F_{\triangle_{df}, df(FM)}$$

- In logistic regression, the analogous quantity of SSE for non-linear models is deviance:

$$D(\beta) = -2\ln\left(\frac{\text{Likelihood of Fitted model}}{\text{Likelihood of saturated model}}\right)$$

$$\text{Likelihood of saturated model} = 1$$

$$D(\beta) = -2\{\ell(\hat{\beta}) - \ell(\tilde{\beta})\}, \ell(\cdot) \text{ is the log-likelihood}$$

where $\tilde{\beta}$ is the fit of data using saturated model (n predictor).

# Assessment of model fit

- The deviance residual is useful for determining if individual points are not well fit by the model. The deviance residual for the ith observation

$$d_i = \text{sign}(Y_i - \hat{\pi}_i)\sqrt{-2[Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)]}$$

  - Obtain deviance residuals in R : residuals() in R

- Deviance analysis The deviance of a fitted model compares the log-likelihood of the fitted model to the log-likelihood of a model with n parameters that fits the n observations perfectly. It can be shown that the likelihood of this **saturated model** is equal to 1 so $\ell(\beta) = 0$. Therefore, the deviance for the logistic regression model is

$$D = \sum_i^n d_i^2 = -2[Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)]$$

$Y_i = 0/1$

$$D = \sum_i^n d_i^2 = 2[Y_i \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i)\log \frac{1 - Y_i}{1 - \hat{\pi}_i}] = 2\sum \text{observed} \times \log \frac{\text{observed}}{\text{fitted}}$$

  - where $\hat{\pi}_i$ is the fitted value for the i-th observation.
  - convention: $0 \times log(0) = 0$
  - The smaller the deviance, the closer the fitted value is to the saturated model. The larger the deviance, the poorer the fit.

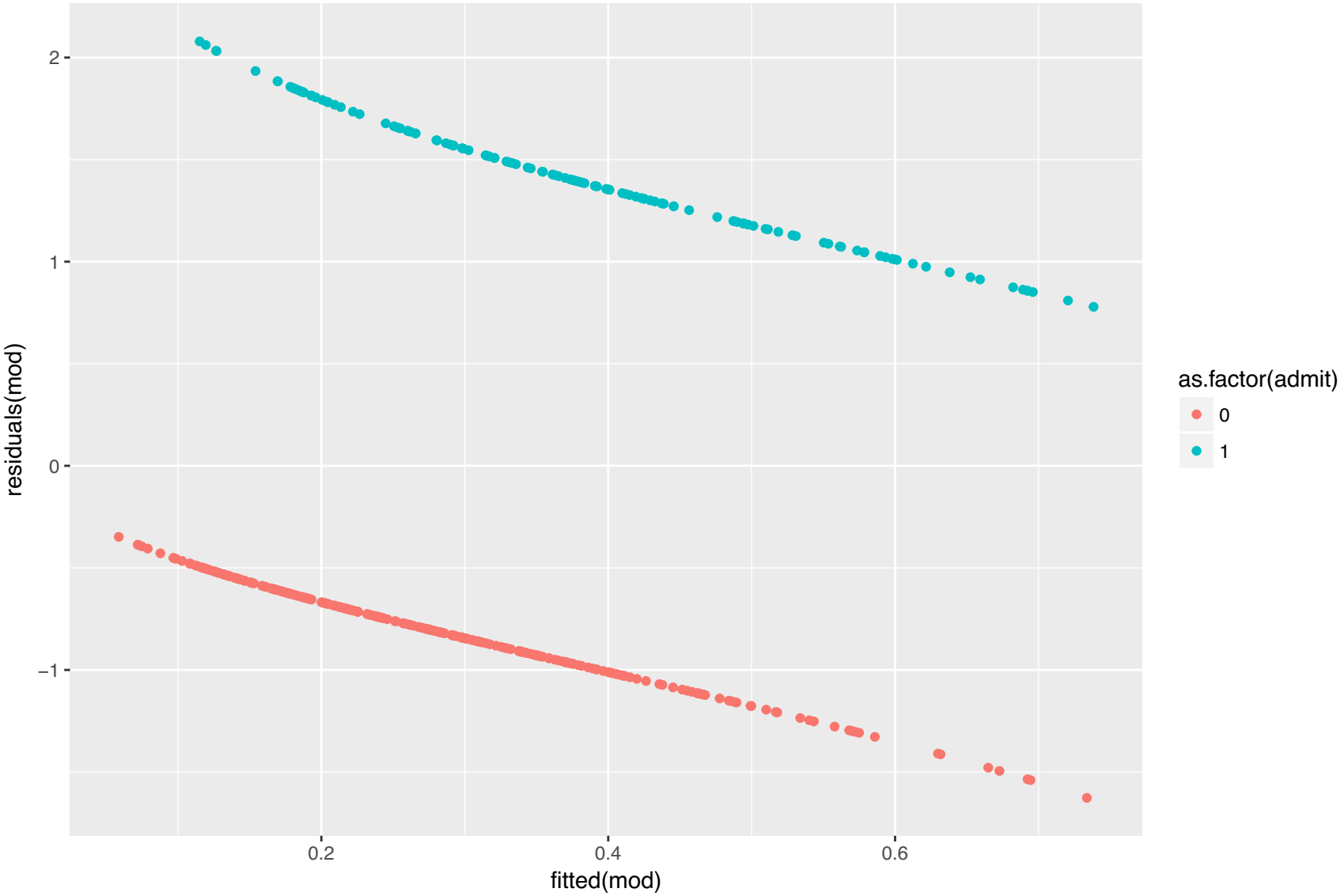
# Assessment of model fit : Saturated model (simulation)

```r
# saturated model
y = c(0,0,0,1,1,1)
a <- factor(1:length(y))
fit <- glm(y~a,family=binomial)
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ a, family = binomial)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.457e+01  1.310e+05       0        1
## a2          -2.628e-14  1.853e+05       0        1
## a3          -5.360e-14  1.853e+05       0        1
## a4           4.913e+01  1.853e+05       0        1
## a5           4.913e+01  1.853e+05       0        1
## a6           4.913e+01  1.853e+05       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.3178e+00  on 5  degrees of freedom
## Residual deviance: 2.5720e-10  on 0  degrees of freedom
## AIC: 12
##
## Number of Fisher Scoring iterations: 23
```

# Assessment of model fit : deviance residual vs fitted values

```
ggplot(mydata, aes(x=fitted(mod),y=residuals(mod),color=as.factor(admit)))+geom_point()
```



# Nested Chi-squared tests of model fit

As for SSE, the greater the deviance, the poorer the fit. If reduced model (RM) were true, then

$$Deviance(RM) - Deviance(FM) \rightarrow \chi^2_{\text{df(RM)-df(FM)}}$$

Thus, reject Reduced Model(RM) at asymptotic level $\alpha$ if

$$Deviance(RM) - Deviance(FM) > \chi^2_{\text{df(RM)-df(FM)},1-\alpha}$$

where

- RM = Reduced model ( with less predictors compared to FM)
- FM = Full model (includes more predictors to RM)

# Hosmer-Lemeshow goodness of fit test

- $H_0$: Actual and predicted event rates are similar across 10 deciles
- $H_a$: they are not the same

$$\Leftrightarrow \quad H_0 : E(Y) = \frac{\exp\{X\beta\}}{1+\exp\{X\beta\}}, \quad H_a : E(Y) \neq \frac{\exp\{X\beta\}}{1+\exp\{X\beta\}}$$

To calculate the test statistics

- Order the fitted values
- Group the fitted values in to g groups (g is between 6 and 10) of roughly equal size
- Calculate the observed and expected number in each group
- Perform a $\chi^2$ goodness of fit (GOF) test

$$X^2 = \sum_{j=1}^{g} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \sim \chi^2_{g-2}$$

# Hosmer-Lemeshow goodness of fit test

Implementation :

```
hosmlemeGOF = function(y, yhat, g=10) {
   cutyhat = cut(yhat,breaks = quantile(yhat, probs=seq(0,1, 1/g)),
                 include.lowest=TRUE)
  obs = xtabs(cbind(1 - y, y) ~ cutyhat)
  expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
  chisq = sum( (obs - expect)^2/expect )
  P = 1 - pchisq(chisq, g - 2)
  return(list(chisq=chisq,p.value=P))
}

hosmlemeGOF(y=mydata$admit, yhat=fitted(mod))
```

```
## $chisq
## [1] 11.08547
##
## $p.value
## [1] 0.1969031
```

```
library(ResourceSelection)
hoslem.test(mydata$admit, fitted(mod),g=10)
```

← using Package

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  mydata$admit, fitted(mod)
## X-squared = 11.085, df = 8, p-value = 0.1969
```

# After Lecture This Week

## Practice problems

- Review all the slides
- Try all the R example in slides.
- Try posted practice problem

## Topics for next 2 weeks:

- Reading week (no class, 3-5pm Friday,Feb 24, 201 at SS6011, no more OH in that week)
- The following week after reading week (Monday, Mar 1st)
  - 1-hour Review for midterm
  - 1-hour OH before midterm in class room
- Topic in the following week after midterm
  - classification with logistic regression
  - overdispersion in logistic regression
  - poisson regression