THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Data Science Generalized Linear Models: Logistic Regression

Model Selection Process

- Selection of the appropriate model is a critical process to any data analysis problem. Below are a few simple steps you can use to help with this process.

  1. Objective of the Model
  2. Model Structure
  3. Model Assumptions
  4. Parameter Estimates and Assumptions
  5. Model Fit (goodness of fit tests)
  6. Model Selection

Example for Simple Linear Regression

1. Objective of the Model – Model the expected (predication) value of a continuous variable y
2. Model Structure : $\hat{y} = b_0 + b_1 x + \ldots.$
3. Model Assumptions –
   - y is normal distributed
   - Linear relationship between dep and independent
   - Homoscedasticity (errors normal)
   - Multicollinearity (Independent variables not correlated)
4. Parameter Estimates and Interpretation b_0 estimate of the intercept and b_1 is the estimate of the slope
5. Model Fit: R^2 , residual, goodness of fit, F Statistics
6. Model Selection: From Model Fit, which variables to include?

Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

General Linear model-GLM vs Generalized Linear model-GLiM

- We've focused largely on General Linear Models (GLM)
  - Ordinary Least Squares (OLS) with fairly strict assumptions
  - Residuals are normally distributed with mean 0
  - GLM, uses built-in function lm( ) in R
  - $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$
- Generalized Linear Models GLiM (or GLzM) less restrictive
  - includes a "link" function f that connects the linear predictor βX to the mean of distribution of the response Y
  - Use maximum likelihood estimation (MLE) to determine parameters
  - GLiM, uses built-in function glm( ) in R
  - $f(\hat{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$, where f is the link function
- Different variable types → different models

Generalized Linear Model (GLiM) Assumptions

- Dependent variable does not need to be normally distributed, but assumes a distribution exponential family such as binomial or Poisson
- Does not assume a linear relationship between dependent and independent but between the transformed response, i.e. link function and the explanatory variables
- Homogeneity of variance does not need to be satisfied
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate parameters, and thus relies on large scale approximations
- Goodness of fit relies on sufficiently large samples

Examples of When to Use Certain GLiM

Model	Random (Y)	Link function	Systematic (X)
Linear Regression	Normal	Identity	Mixed
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Log-linear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

- Random: Is the probability dist. of the response variable Y
- Systematic: Explanatory variables in the model
- Link Function: Connects the random to the systematic

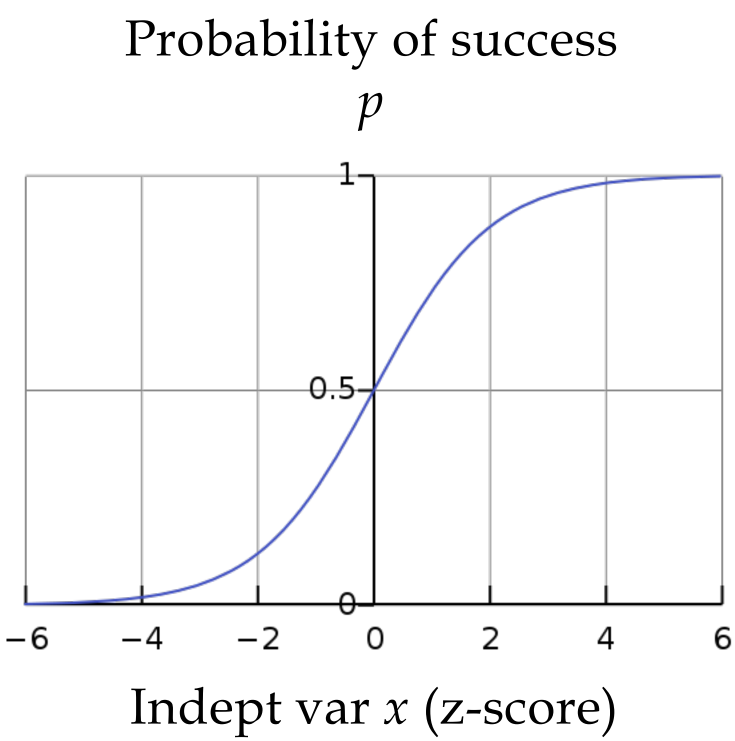
Logistic Regression : Intro

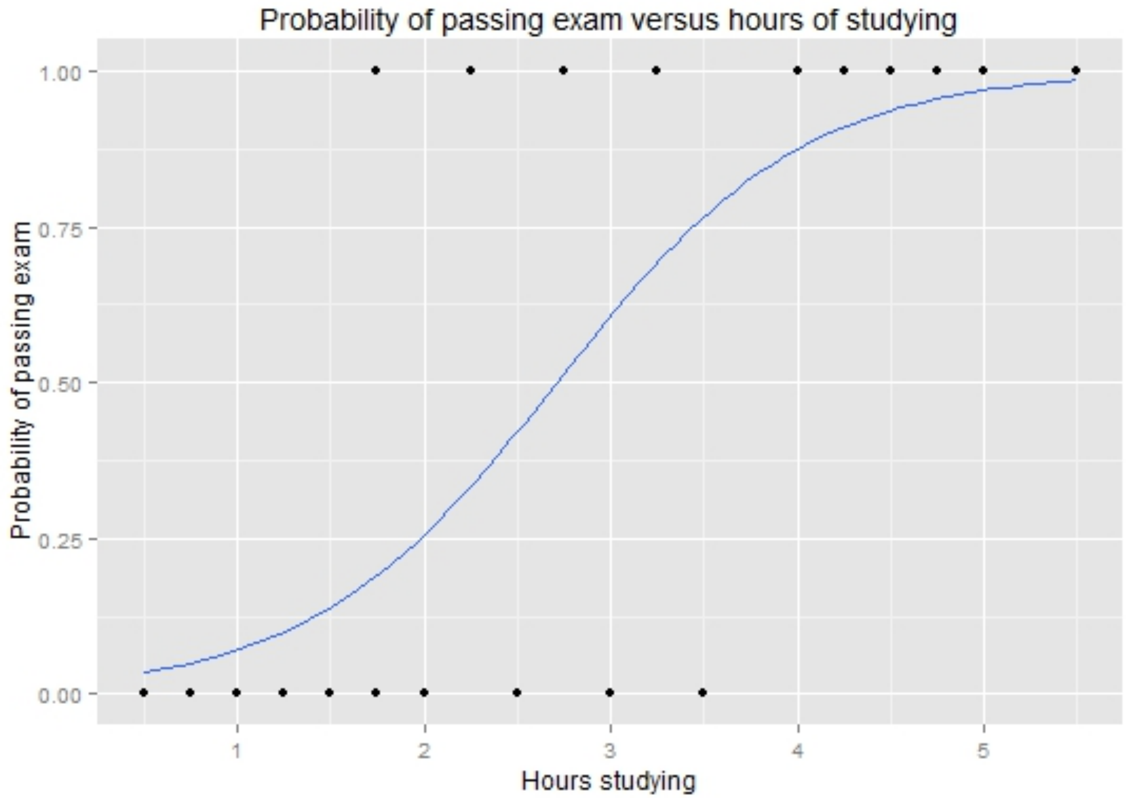
- Definition: A type of generalized linear model (GLiM) that uses statistical analysis to predict an event based on known factors when using a **dichotomous** dependent variable.

- Classification algorithm used to predict a binary outcome (1/0 or Yes/No or True/False), predicts the probability of a event occurring by fitting data to a logit fit function.

- What Can We Use Logit Regression to Answer?
  - make predictions about whether a customer will buy a product based on age, gender, geography, and other demographics. Also to assess likelihood of loan defaults, or political success/failure, …

Logistic Regression : Intro





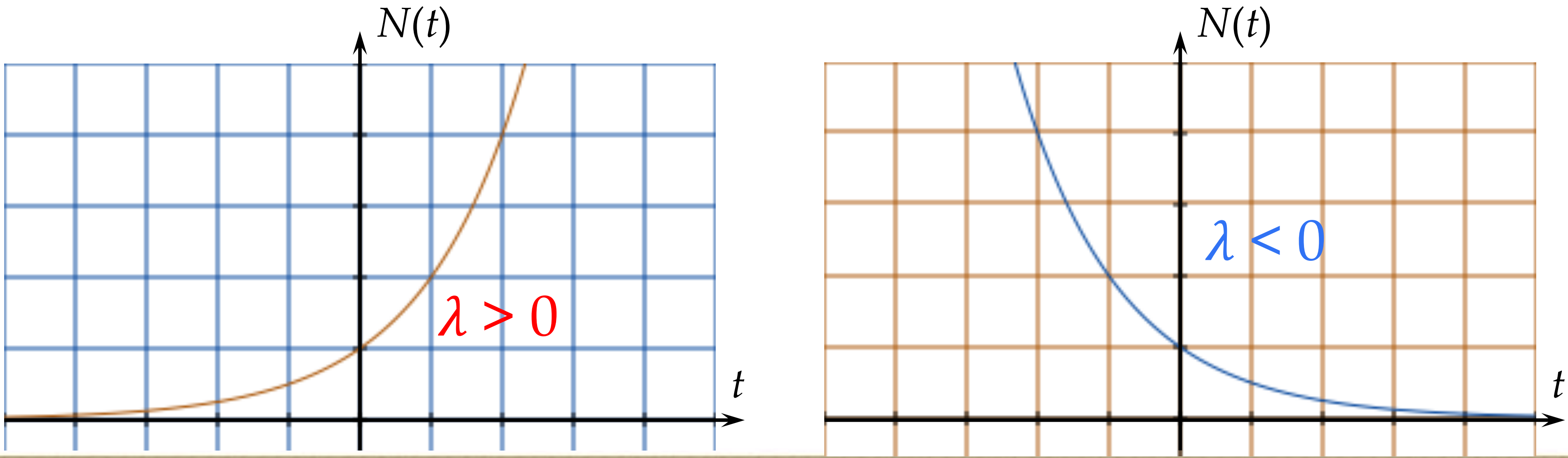
Wikipedia

By Qef (talk) - Created from scratch with gnuplot,
Public Domain, https://commons.wikimedia.org/w/index.php?curid=4310325

Logistic Regression : Calculus approach

- A lot of systems can be argued the growth rate (or decay rate) of a population should be proportional to population itself. Hence we set up the population (differential) equation:

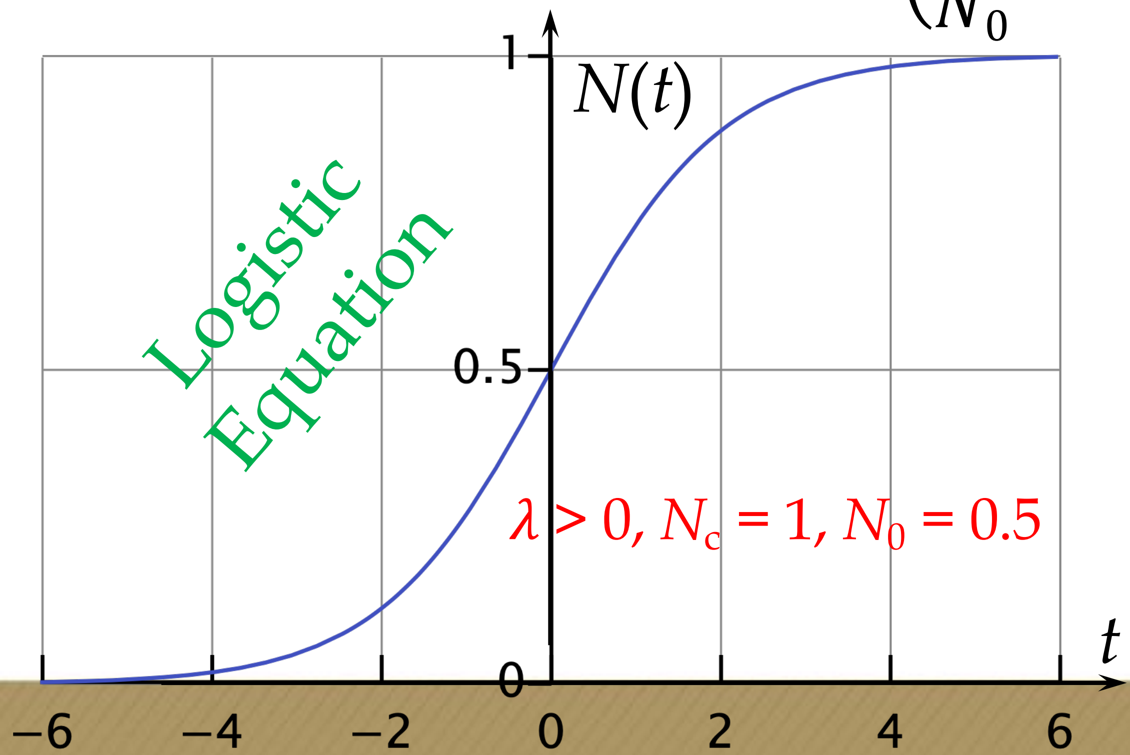
$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda\, N(t) \implies N(t) = N_0\, e^{\lambda t}$$



Logistic Regression : Calculus approach

- More realistic models will have a limit on the growth, by a "carrying capacity" N_c (dropping the explicit t on the LHS):

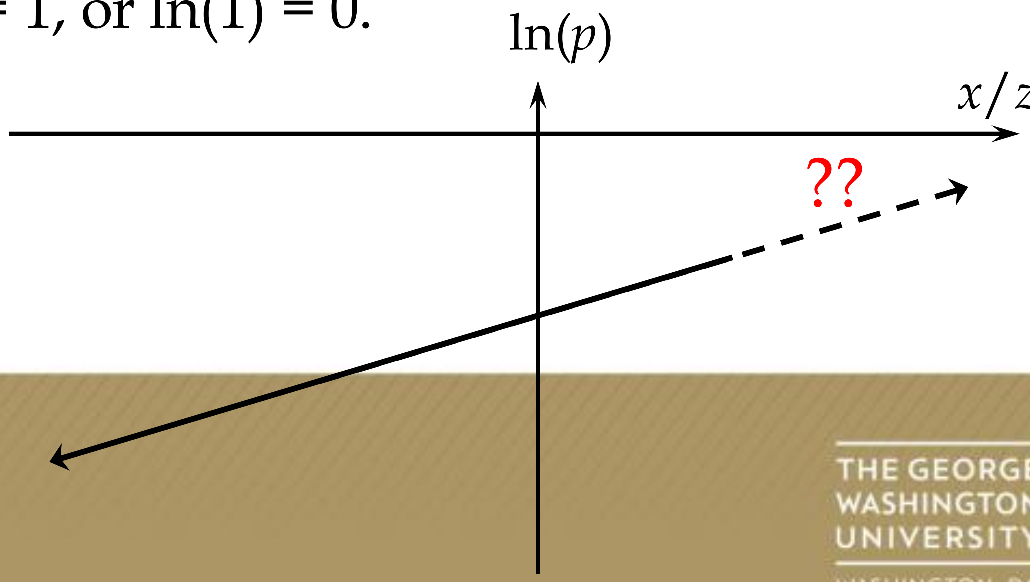
$$\frac{\mathrm{d}N}{\mathrm{d}t} = \lambda\, N\left(1 - \frac{N}{N_c}\right) \implies N(t) = N_c \frac{1}{1 + \left(\frac{N_c}{N_0} - 1\right) e^{-\lambda t}}$$



Logistic Regression : Data approach

- The probability of success p (between 0 and 1) should NOT be linearly dependent on x_i.
  - Typically, x_i can be unbounded. It is almost impossible to restrict a linear fit to have a range of 0 and 1.
  - Residuals are normally distributed with mean 0

- Try to model $\ln(p)$? [natural log ln( ) as link function?]
  - Say as $x \rightarrow -\infty$, we want $p = 0$, or $\ln(0) = -\infty$.
  - And as $x \rightarrow +\infty$, we want $p = 1$, or $\ln(1) = 0$.
  - <u>Doesn't work</u>



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Logistic Regression : Odds Ratio

Before we actually introduce what "logit" means, we need to learn about "odds ratio".

Simple Example:

	Show A	Show B
Male	200	100
Female	50	150

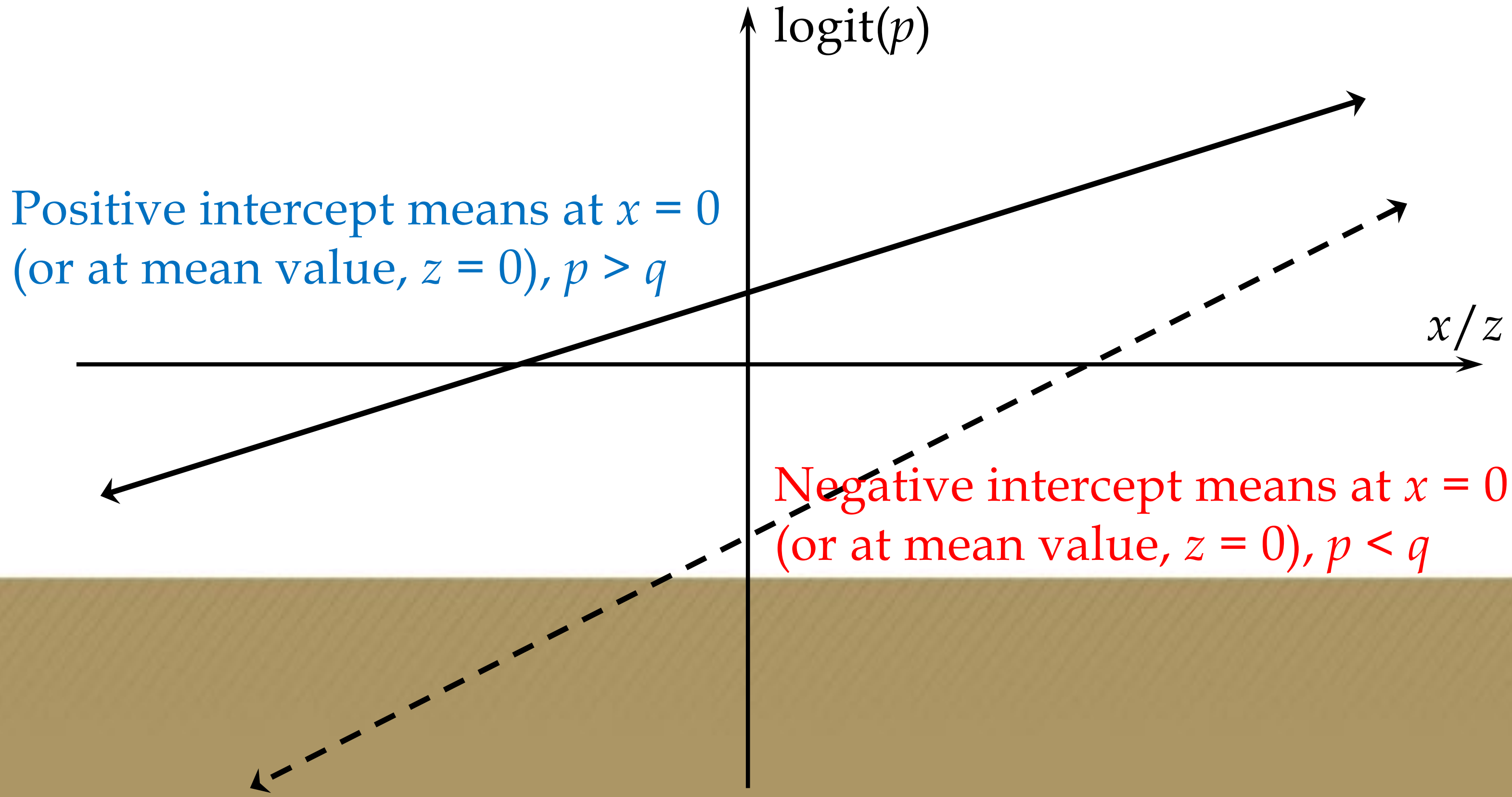
- Odds of Males watching show A are 200/100 or 2 to 1
- Odds Female watching show A is 50/150 or 1 to 3
- Divide results by each other to get the ratio of odds for gender of show A = (2/1)/(1/3) = 6
- So males are 6 times more likely to be watching Show A than females

Logistic Regression : Odds Ratio

- The odds ratio success-to-failure (or win-to-lose) is $p/q = p/(1 - p)$
- The odds ratio ranges from 0 ($p = 0$) to $+\infty$ ($p = 1$).
- The ratio equals 1 means 50-50 chance.
- Now let's model $\ln(p/q) = \ln(p/(1 - p))$.
- We call this the logit function $\text{logit}(p) = \ln(p/(1 - p))$.

Logistic Regression

- Now $\text{logit}(p) = \ln(p/(1 - p))$. [logit( ) as link function]
  - when $p = 0$, or $\ln(0/1) = -\infty$.
  - when $p = 1$, or $\ln(1/0) = \ln(\infty) = \infty$.
  - In general, $\text{logit}(p) > 0$ means $p > q$ and vice versa.
  - Positive slope: odds (and odds-ratio) increases with x.



Logistic Regression

- Note that there is no obvious reason why logit(p) would be linear to x_i. It could very well be $x_i^{\,2}$, $x_i^{\,3}$, $x_i^{\,1/2}$, $\sin(x_i)$, or interaction terms, etc. The same is true for ordinary linear regression. We can use other info to decide what transformed variable(s) should we include in our model.

- Remember that logit regression does not give us directly p, nor $\ln(p)$. It gives us logit(p) instead.

- $\text{logit}(\hat{p}) = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

$$\frac{\hat{p}}{1-\hat{p}} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots} = e^{\beta_0}\, e^{\beta_1 x_1}\, e^{\beta_2 x_2} \ldots = e^{\beta_0}(e^{\beta_1})^{x_1}(e^{\beta_2})^{x_2} \ldots$$

$$\hat{p} = \frac{1}{1+e^{-\beta_0-\beta_1 x_1-\beta_2 x_2-\cdots}} = \frac{1}{1+e^{-\beta_0}(e^{-\beta_1})^{x_1}(e^{-\beta_2})^{x_2}\ \ldots}$$

Logistic Regression : useful results

- Another way to express the probability p is

$$\hat{p} = \frac{1}{1 + e^{-\beta_0}(e^{-\beta_1})^{x_1}(e^{-\beta_2})^{x_2}\ \ldots} = \frac{\text{odds-ratio}}{\text{odds-ratio} + 1}$$

- When p is small (less than 0.1), $p \approx p/q$.
  In other words, calculating the odds and the odds-ratio are about the same

- When the coefficient is small ($|\beta| \ll 1$), the growth/decay factor $e^{\beta} \approx 1 + \beta$

Logistic Regression: Interpret Results

```
Call:
glm(formula = admit ~ gre + gpa + rank, family = binomial(link = "logit"),
    data = data)

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 ***
gre          0.002264   0.001094   2.070 0.038465 *
gpa          0.804038   0.331819   2.423 0.015388 *
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
```

Gives us the log odds-ratio

- For every one unit gain in GRE, ln(odds-ratio) of admit gain by .0023
- Similar for GPA
- For rank: use rank1 as baseline, ln(odds-ratio) deceases by .67 when changing from 1 to 2

Logistic Regression: Interpret Results

```
> exp(coef(admitlogit))
(Intercept)         gre         gpa       rank2       rank3       rank4 
  0.0185001   1.0022670   2.2345448   0.5089310   0.2617923   0.2119375 
> 
```

- Produces Log(odds-ratio) – usually needs to be converted to be useful (unless the numerical value is close to zero)

- We converted our coefficients to be more usable so what we can now say is that for every unit increase in GPA the <u>**odds-ratio**</u> of being admitted is multiplied by 2.23. Or we conclude the <u>**odds-ratio**</u> is 123% higher (subtract 1, then times 100).

Model Evaluation

- For regular Linear Regression models, we use:
  - ✓ Coefficients' p-values
  - ✓ F-statistics for overall model significance
  - ✓ R^2 for percentage explained
  - ✓ Feature Selection / Model comparison: Adj R^2, BIC, C_p
- For Logit Regression models, we use:
  - ✓ Coefficients' p-values
  - ✓ Hosmer and Lemeshow Goodness of Fit (GOF) test (c.f. χ^2)
  - ✓ Receiver Operating Characteristic/Area-Under-Curve, value > 0.8
  - ✓ McFadden (c.f. R^2)
  - ✓ Feature Selection / Model comparison: Akaike Info Criterion (AIC), lower is better

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Model Evaluation: Goodness of Fit

- Hosmer Lemeshow goodness of fit test: Above .05 better

```
> hoslem.test(data$admit, fitted(admitlogit))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  data$admit, fitted(admitlogit)
X-squared = 11.085, df = 8, p-value = 0.1969
```

- p-value > 0.05, fail to reject the null
- no significant difference between the model and the observed data
- Similar to comparing the actual frequency distribution to the model predicted frequency distribution.

χ^2 : Goodness of Fit (refresher)

We use Goodness of Fit to test hypotheses for the distribution of a variable (usually categorical).

Example: National crime distribution vs Local data

Type of violent crime	Relative frequency
Murder	0.011
Forcible rape	0.063
Robbery	0.286
Agg. assault	0.640

Type of violent crime	Frequency
Murder	3
Forcible rape	37
Robbery	154
Agg. assault	306
	500

Model Evaluation: ROC and AUC

- Use library "pROC" for logit regressions
- Receiver Operating Characteristic: Measure of the model Sensitivity (true positive rate) vs Specificity (false positive)
- Area Under Curve is a byproduct of ROC:
  - ✓ Range from 0.5 – 1.0, higher better
  - ✓ Measure Discrimination
  - ✓ Criteria is used to accept / decline model (AUC = 0.8)

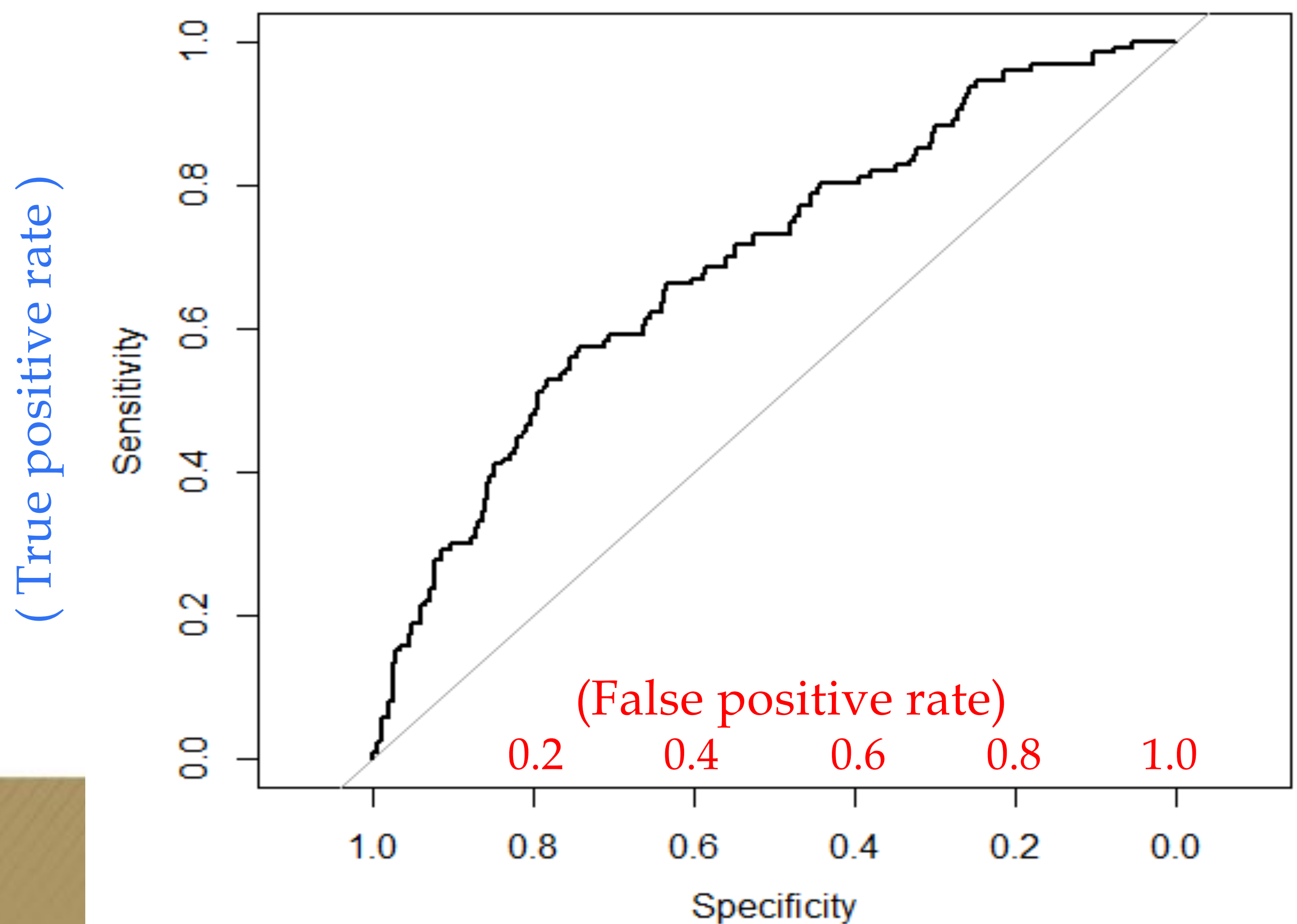
```
> prob=predict(admitlogit, type = c("response"))
> data$prob=prob
> library(pROC)
> h <- roc(admit~prob, data=data)
> plot(h)

Call:
roc.formula(formula = admit ~ prob, data = data)
```

Model Evaluation: ROC and AUC

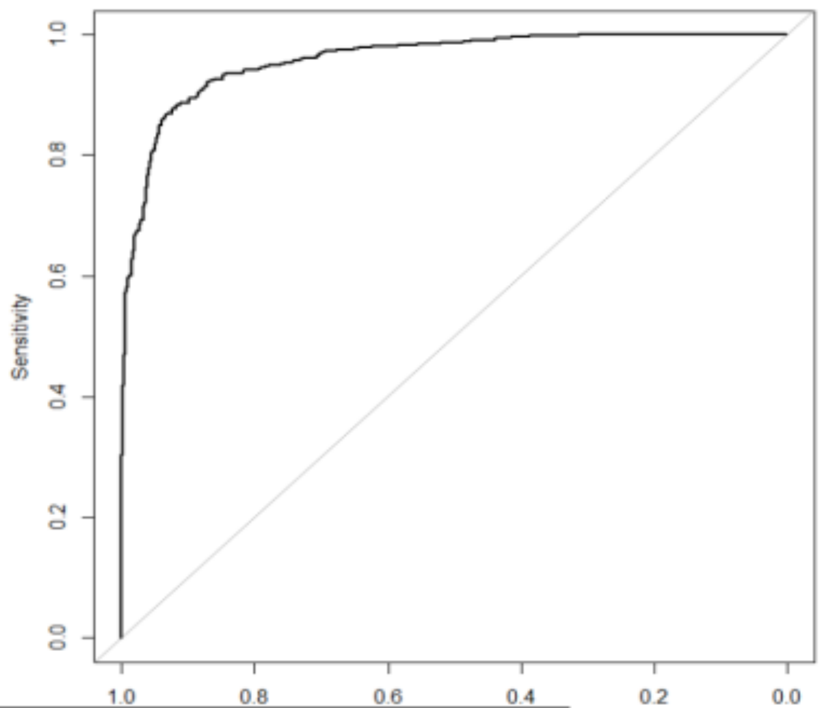
```
Data: prob in 273 controls (admit 0) < 127 cases (admit 1)
Area under the curve: 0.6928
```

← Not so great, want this output to be closer to 0.8



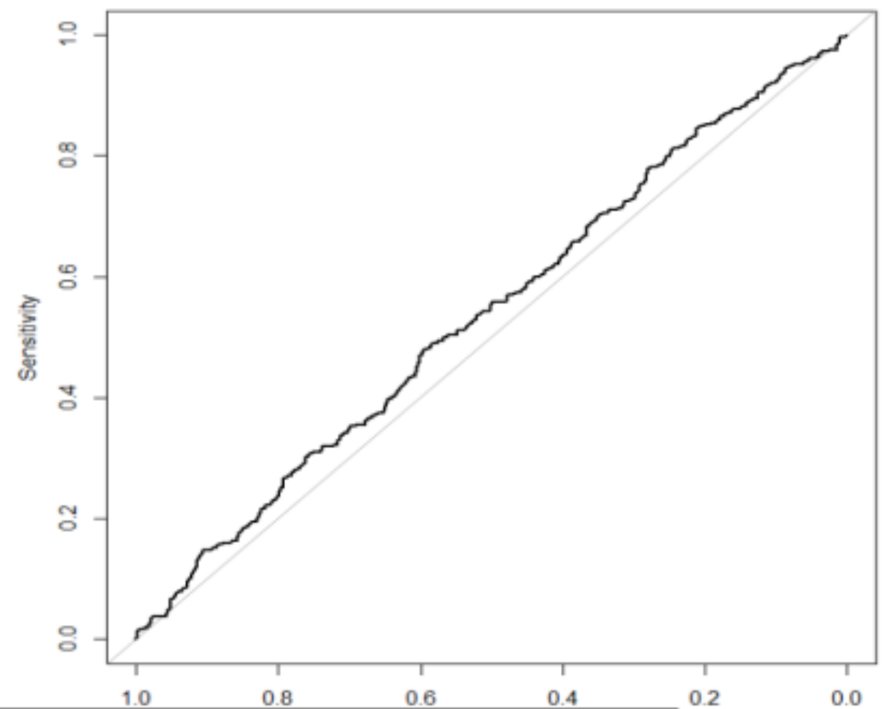
Model Evaluation: ROC and AUC

- Good Indicator / good model



ROC curve from a model with a very strong predictor

- Bad Indicator / bad model



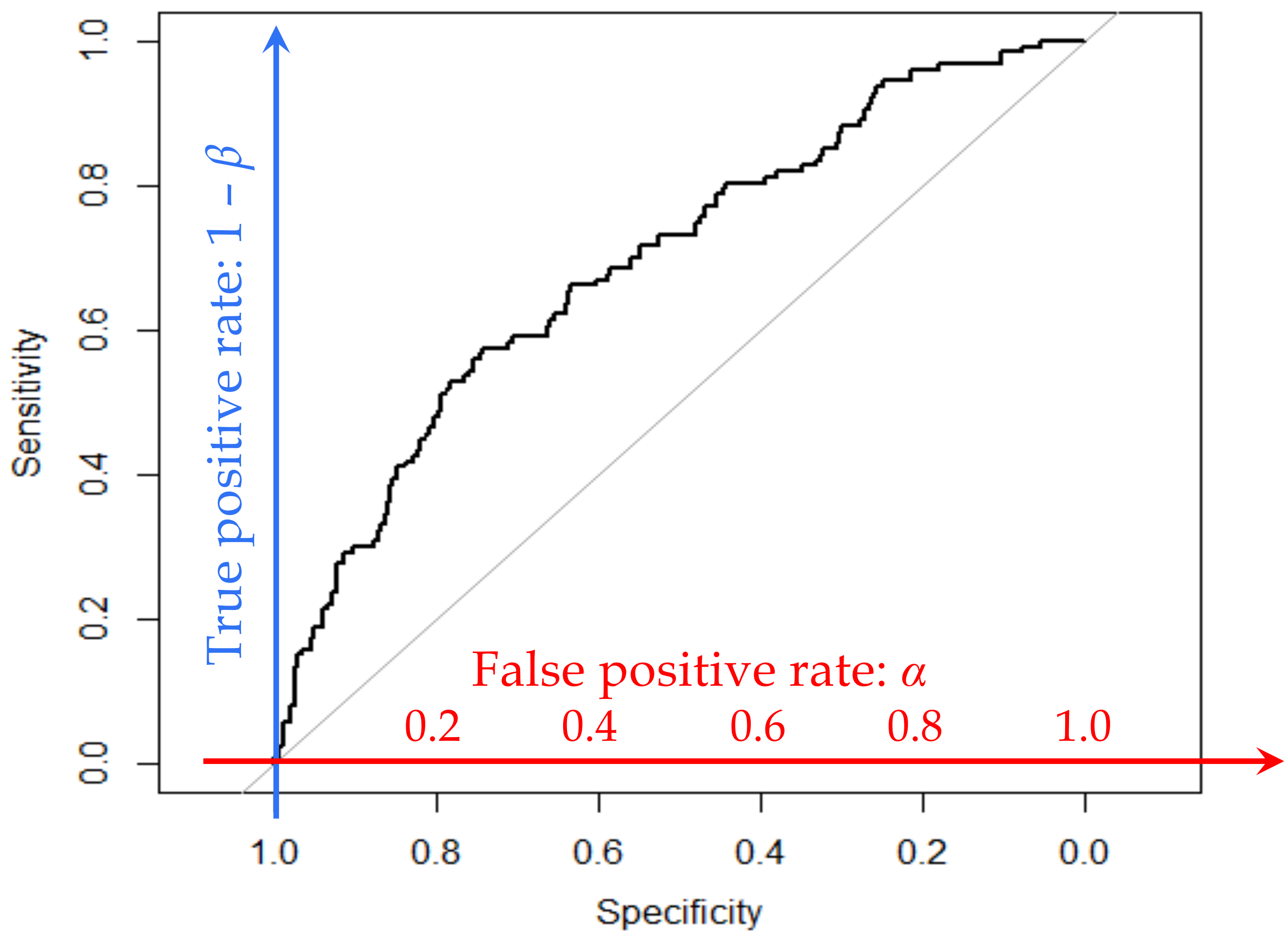
ROC curve where the predictor is independent of outcome

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Model Evaluation: ROC and AUC

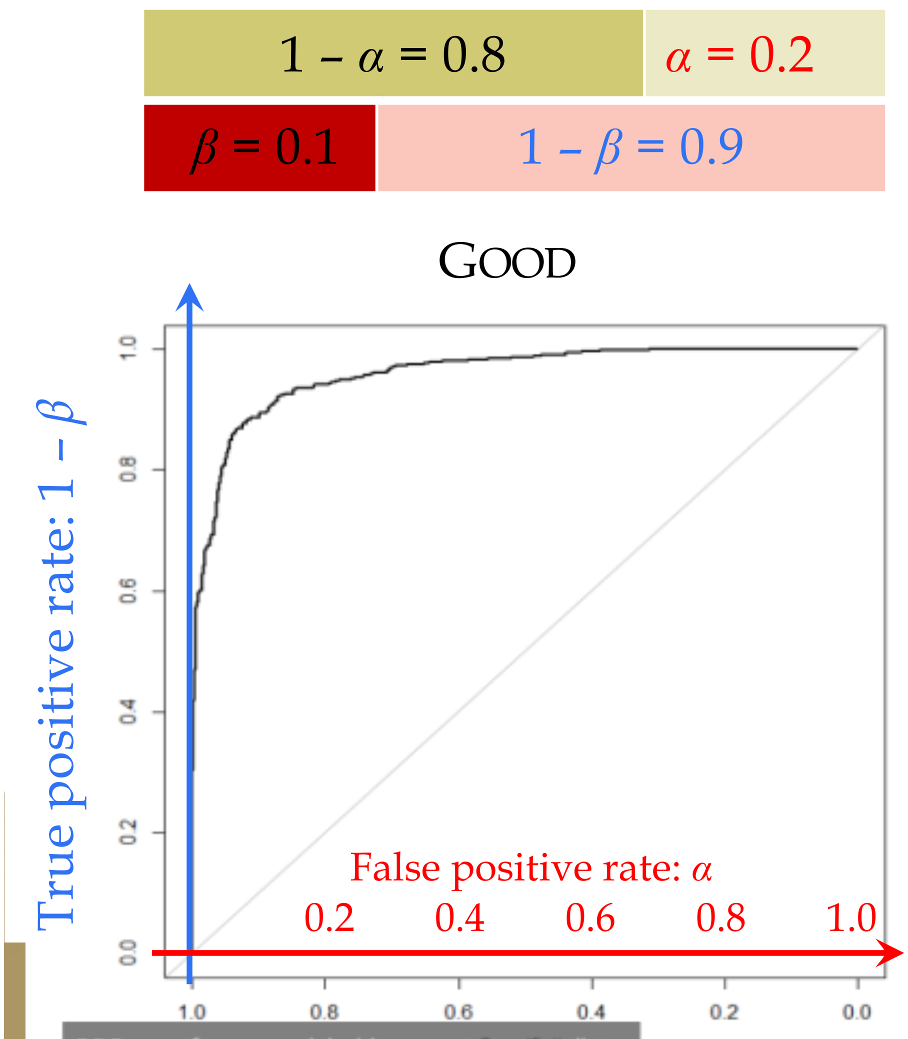
Truth of H_0	Negative, fail reject H_0	Positive, reject H_0
H_0 is true	True negative: $1 - \alpha$	False positive: α
H_0 is false	False negative: β	True positive: $1 - \beta$

- Recall that $\alpha\uparrow \implies \beta\downarrow$ (and vice versa) except to increase sample size n.

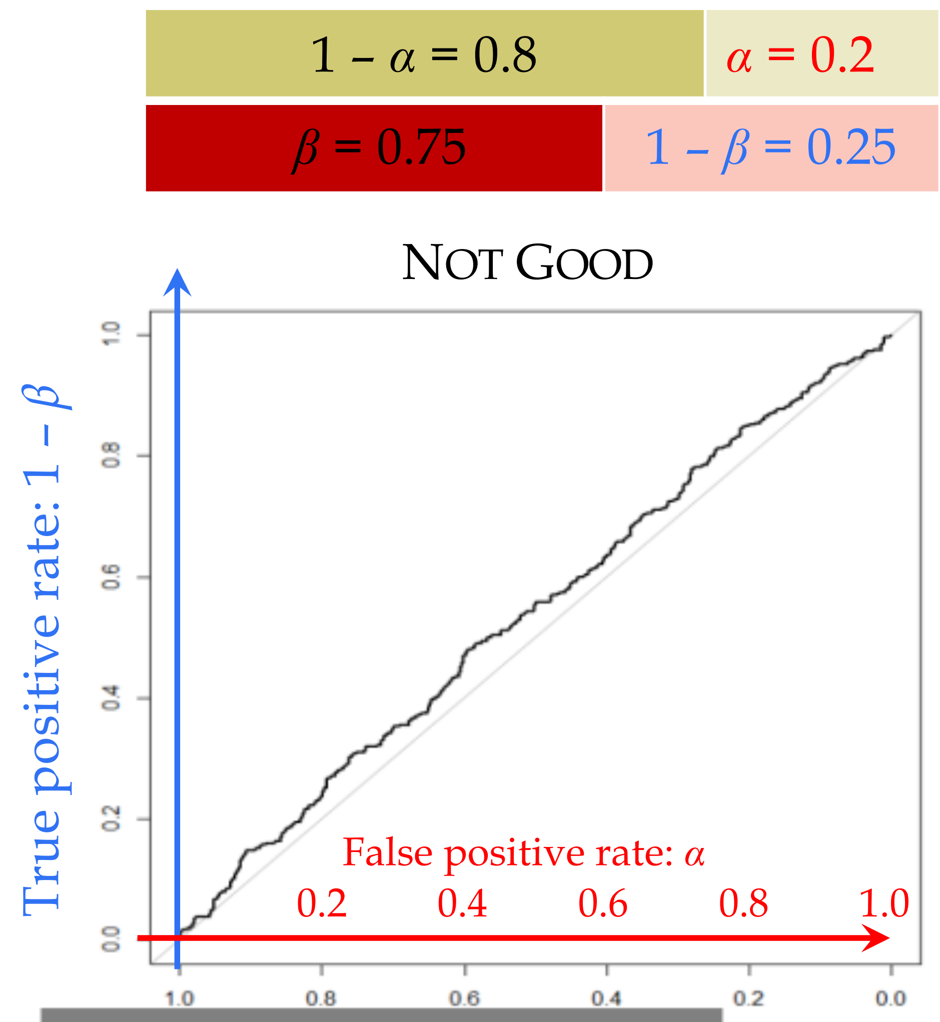


Model Evaluation: ROC and AUC

Truth of H_0	Negative, fail reject H_0	Positive, reject H_0
H_0 is true	True negative: $1 - \alpha$	False positive: α
H_0 is false	False negative: β	True positive: $1 - \beta$



ROC curve from a model with a very strong predictor



ROC curve where the predictor is independent of outcome

Model Evaluation: McFadden

- Calculate McFadden directly, compared to null model

  > mcFadden = 1 - logLik(admitLogit)/logLik(nullModel)

  'log Lik.' 0.08292194 (df=6)

- Calculate the Pseudo R^2 values via the “pscl” library (pscl – political science computational lab).

- Use pR2( ) function

```
> pR2(admitlogit)
         llh      llhNull           G2     McFadden         r2ML         r2CU
-229.25874624 -249.98825878  41.45902508   0.08292194   0.09845702   0.13799580
```

- Not so great with only 8% of variation explained

Model Comparison: AIC

```
(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
```

- AIC: Akaike information criterion – Compare between models, lower AIC is better
- Function AIC( ) is also in the R core base library
  - ✓ `> AIC(admitlogit)`
- Residual Deviance: From 499.98 to 458.52, not great

Feature Selections on Logit Models

- Leaps package (regsubsets) does not work with logistic regression model
- Bestglm is a package that will handle logistic regression feature selection, but with some limitations.
- Explore different packages that you might find useful.