Generalized Linear Models 1 (GLM)

SYS 4021/6021

Laura Barnes and Julianne Quinn

Organization of lecture

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

1. Review of linear regression assumptions and what we can do if they are not met.
   a) Non-linear relationships
   b) Non-Gaussian residuals

2. What if we have a categorical response variable?

3. Introduction to Generalized Linear Models

Assumptions of Multivariate Linear Regression

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

$$Y = X\beta + \varepsilon$$

Assumptions:

1. X relates linearly to Y
   - Transform X if relationship is non-linear
2. ε are independent and Gaussian-distributed
   - Transform Y if residuals are not Gaussian
3. Y is quantitative

What can we do if Y is not quantitative? If we don't change anything, what are the consequences of using linear regression?

Example of Categorical Y Response

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

On Jan 28, 1986, the space shuttle Challenger exploded shortly after liftoff. All astronauts on board were lost, including a grade school teacher.

Before the launch, the engineers from Morton Thiokol, the makers of the solid rocket boosters on the shuttle, gave NASA managers a briefing on the association between temperature and O-ring failure.

The analysis had no visual displays and no statistical models. NASA managers were unconvinced and decided to launch. What if they had built a statistical model?

Challenger Example

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

$$Y = X\beta + \varepsilon$$

What would be the response variable in this model?

$$Y = \begin{cases} 0, & \text{O} - \text{ring success} \\ 1, & \text{O} - \text{ring failure} \end{cases}$$

What would be the predictor variable(s)?

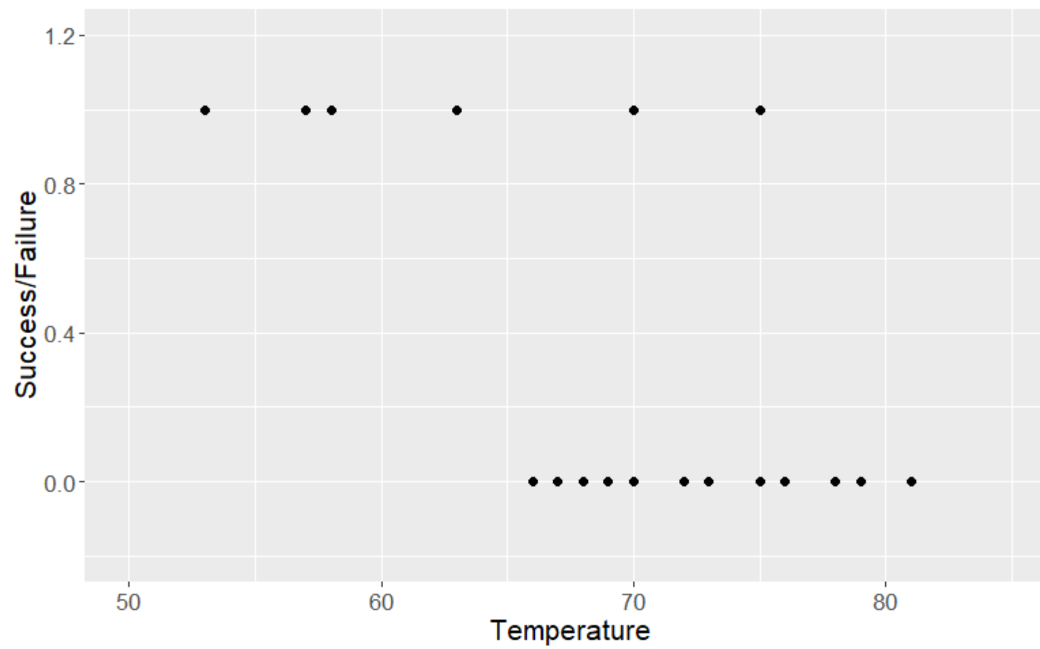
$$X_1 = \text{temperature}$$

Challenger Example

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

```r
library(mcsm)
library(ggplot2)
library(ggfortify)

# challenger plots
data(challenger)

ggplot(challenger, aes(x=temp, y=oring)) + geom_point(size=2) +
  xlab("Temperature") + ylab("Success/Failure") +
  theme(text = element_text(size=16)) + ylim(-0.2,1.2) + xlim(50,85)
```



Challenger Example

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

```
> linear.fit = lm(oring~temp, challenger)
> summary(linear.fit)

Call:
lm(formula = oring ~ temp, data = challenger)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43762 -0.30679 -0.06381  0.17452  0.89881 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.90476    0.84208   3.450  0.00240 **
temp        -0.03738    0.01205  -3.103  0.00538 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3987 on 21 degrees of freedom
Multiple R-squared:  0.3144,	Adjusted R-squared:  0.2818 
F-statistic:  9.63 on 1 and 21 DF,  p-value: 0.005383
```

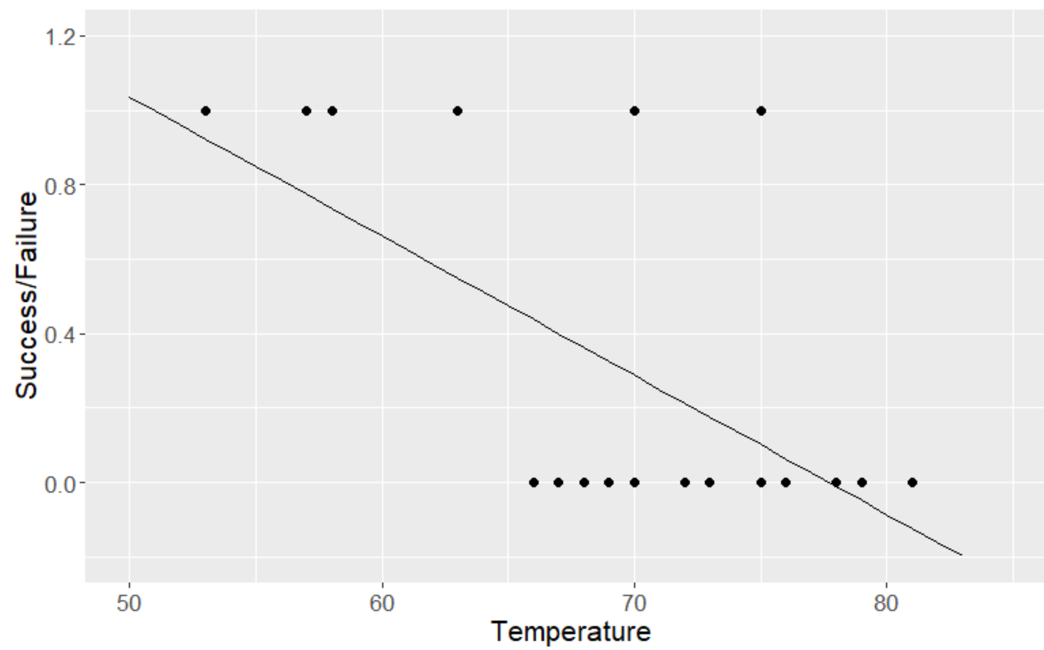
Even though the response is categorical, we can still fit a model and find temperature is significant. But is it a reasonable model?

Challenger Example

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

```
newdata = data.frame(temp=c(seq(50,85,1)))
linear.predict = predict(linear.fit,newdata)

ggplot(challenger, aes(x=temp, y=oring)) + geom_point(size=2) +
  geom_line(data=newdata, aes(y=linear.predict)) +
  xlab("Temperature") + ylab("Success/Failure") +
  theme(text = element_text(size=16)) + ylim(-0.2,1.2) + xlim(50,85)
```

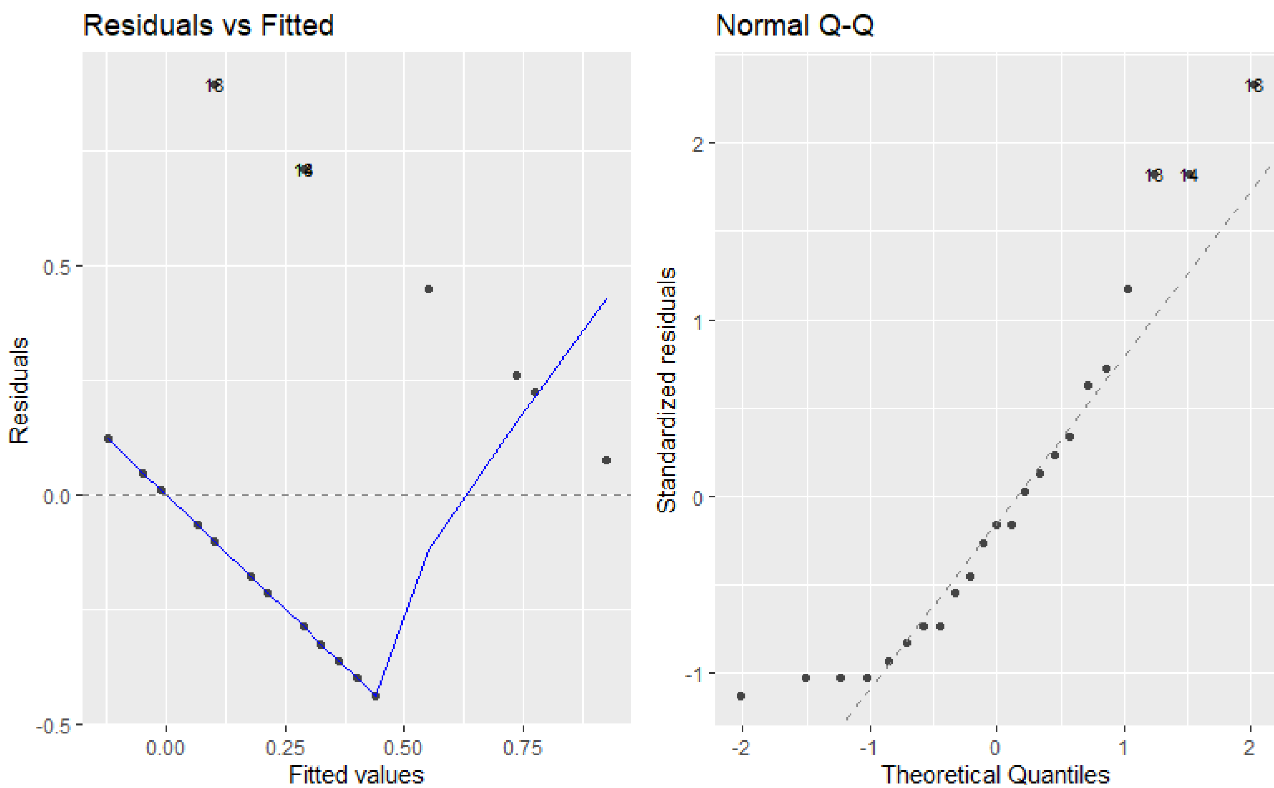


Challenger Example

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

```
> autoplot(linear.fit, which=1:2, label.size=3)
```

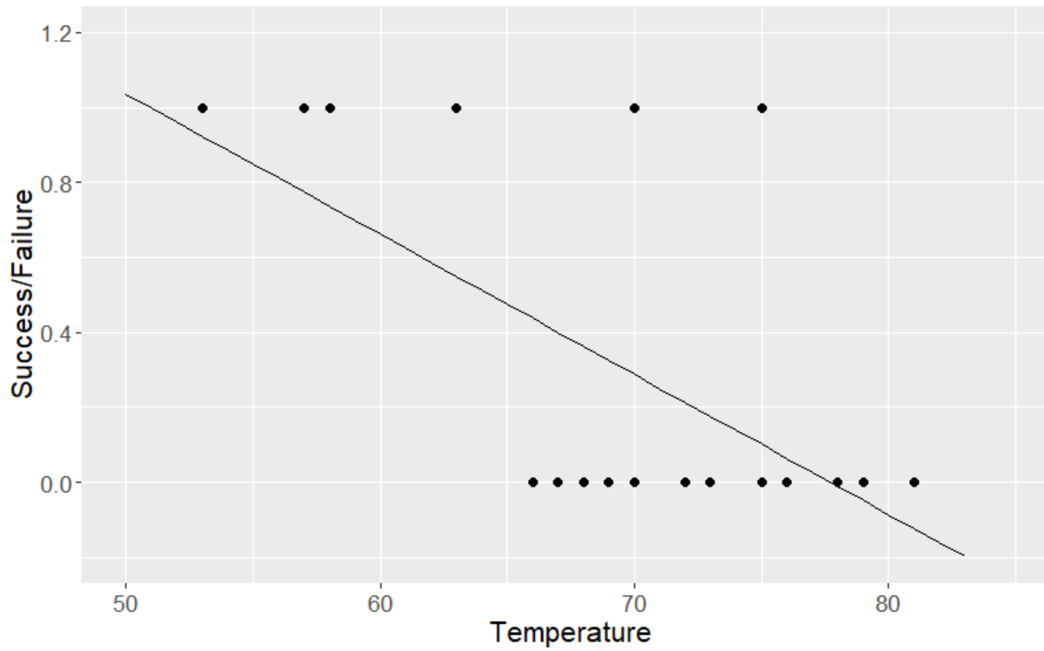


Problems with this regression model

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

1. Predicts values of the response outside of its range, and between discrete values.



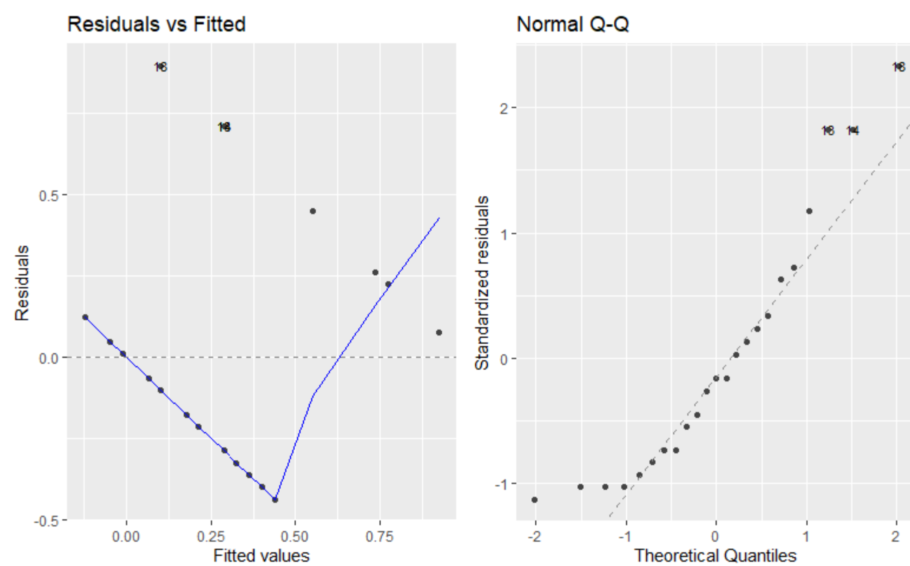
Problems with this regression model

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

1. Predicts values of the response outside of its range, and between discrete values.

2. Non-Gaussian distribution of Y results in non-Gaussian distribution of ε, and binary values of Y result in a pattern in the residuals vs. fitted plot.



Problems with this regression model

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

These problems will arise whenever we have a non-continuous response variable. For example:

1. Binary response (e.g. success/failure)
2. Categorical response (e.g. species of flowers)
3. Countable response (e.g. casualties)

Are there ways around this?

If Y is skewed, we can use a log or Box-Cox transformation to make it normal. Are there other transformations that can make categorical variables continuous?

Generalized Linear Models (GLMs)

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

Generalized linear models (GLMs) extend regression modeling to include non-Gaussian-distributed response variables.

GLMs use a link function to relate the mean of the response to a linear function of the predictors (where those predictors may be non-linear transformations of the original variables):

$$g(E[Y]) = X\beta$$

Technically, we assume $g(\cdot)$ has a distribution from the exponential family. The Gaussian, binomial, and Poisson distributions satisfy this condition, among others.

Generalized Linear Models

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

GLM families include:

Family	Link Function	Variance Function	Response
Gaussian	$E[Y]$	1	Real
Poisson	$\log(E[Y])$	$E[Y]$	Count
Binomial	$\log\left(\frac{E[Y]}{1-E[Y]}\right)$	$E[Y](1-E[Y])$	Binary

$E[Y] = X\beta + \varepsilon$ where $\mathrm{Var}(\varepsilon)$ is constant and independent of $E[Y]$

Generalized Linear Models

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

GLM families include:

Family	Link Function	Variance Function	Response
Gaussian	$E[Y]$	1	Real
Poisson	$\log(E[Y])$	$E[Y]$	Count
Binomial	$\log\left(\frac{E[Y]}{1-E[Y]}\right)$	$E[Y](1-E[Y])$	Binary

$\log(E[Y]) = X\beta \rightarrow E[Y] = \exp(X\beta) + \varepsilon$
where $\mathrm{Var}(\varepsilon) = E[Y]$

This is distinct from linear regression with a log transformation of Y because it assumes a different distribution of the residuals.

Generalized Linear Models

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

GLM families include:

Family	Link Function	Variance Function	Response
Gaussian	$E[Y]$	1	Real
Poisson	$\log(E[Y])$	$E[Y]$	Count
Binomial	$\log\left(\frac{E[Y]}{1-E[Y]}\right)$	$E[Y](1-E[Y])$	Binary

$$\log\left(\frac{E[Y]}{1-E[Y]}\right) = X\beta \rightarrow E[Y] = \frac{\exp(X\beta)}{1+\exp(X\beta)} + \varepsilon$$

where $\mathrm{Var}(\varepsilon) = E[Y](1-E[Y])$

What is $E[Y]$ for a binary variable? E[Y] = P(Y=1)

Generalized Linear Models

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

GLM families include:

Family	Link Function	Variance Function	Response
Gaussian	$E[Y]$	1	Real
Poisson	$\log(E[Y])$	$E[Y]$	Count
Binomial	$\log\left(\frac{E[Y]}{1-E[Y]}\right)$	$E[Y](1-E[Y])$	Binary

$$\log\left(\frac{E[Y]}{1-E[Y]}\right) = X\beta \rightarrow E[Y] = \frac{\exp(X\beta)}{1+\exp(X\beta)} + \varepsilon$$

where $\text{Var}(\varepsilon) = E[Y](1-E[Y])$

Therefore $\log\left(\frac{E[Y]}{1-E[Y]}\right) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right)$.

This is called the logit or logistic function.

Generalized Linear Models

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

GLM families include:

Family	Link Function	Variance Function	Response
Gaussian	$E[Y]$	1	Real
Poisson	$\log(E[Y])$	$E[Y]$	Count
Binomial	$\log\left(\frac{E[Y]}{1-E[Y]}\right)$	$E[Y](1-E[Y])$	Binary

$$\log\left(\frac{E[Y]}{1-E[Y]}\right) = X\beta \rightarrow E[Y] = \frac{\exp(X\beta)}{1+\exp(X\beta)} + \varepsilon$$

where $\mathrm{Var}(\varepsilon) = E[Y](1-E[Y])$

Therefore $\log\left(\frac{E[Y]}{1-E[Y]}\right) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right)$.

It is the log of the odds of the event Y=1

Generalized Linear Models

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

Key Insight: Rather than predicting the event Y=0 or 1 with linear regression, we predict the log of the odds Y=1 instead of Y=0. We can convert this to the P(Y=1) or P(Y=0).

Let $p = P(Y = 1)$ and $P(Y = 0) = 1 - p$:

$$\log\left(\frac{p}{1-p}\right) = X\beta$$

$$\frac{p}{1-p} = \exp(X\beta)$$

$$p = \exp(X\beta)[1-p]$$

$$p = \exp(X\beta) - \exp(X\beta)p$$

$$p + p\exp(X\beta) = \exp(X\beta)$$

$$p[1 + \exp(X\beta)] = \exp(X\beta)$$

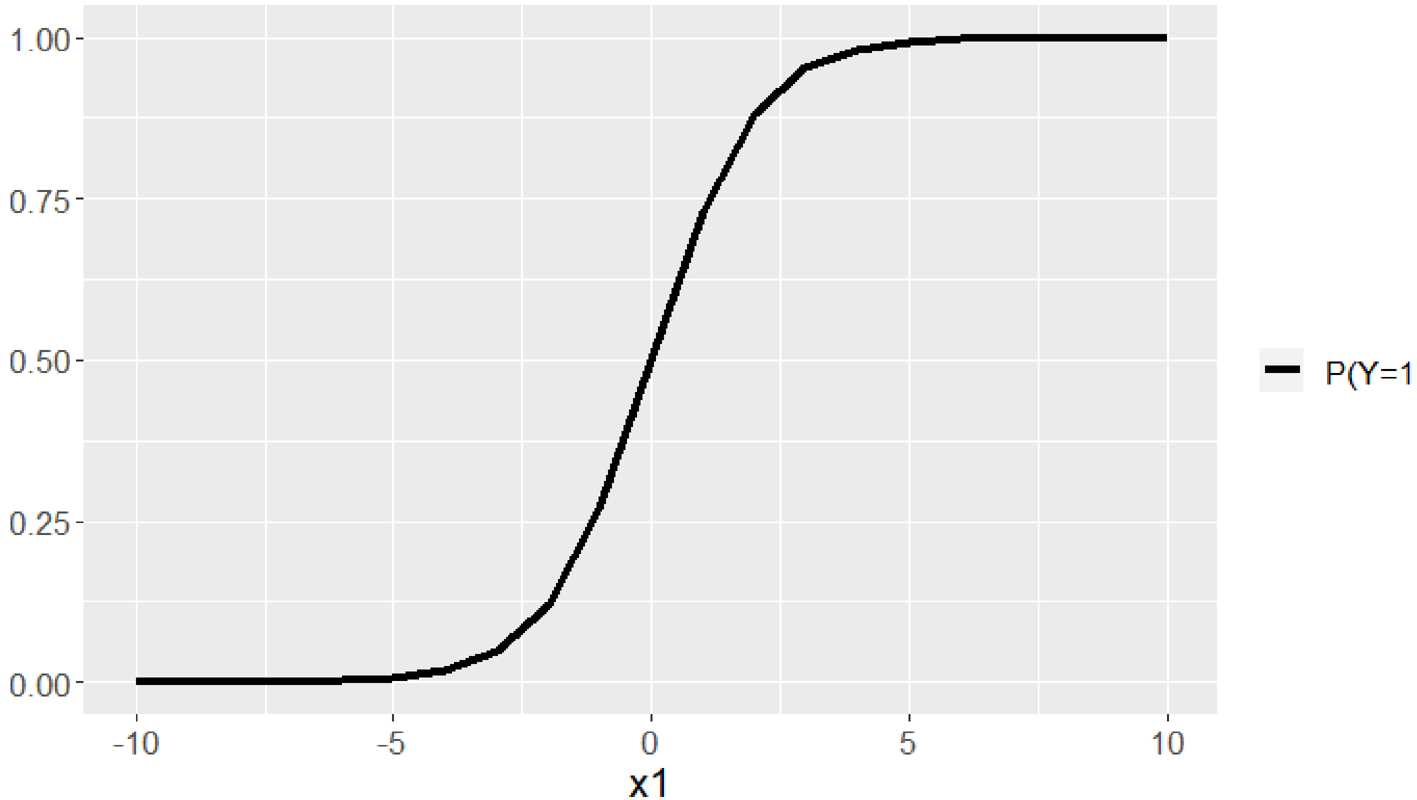
$$p = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$

What does this function look like?

Logistic Function for P(Y=1)

Agenda

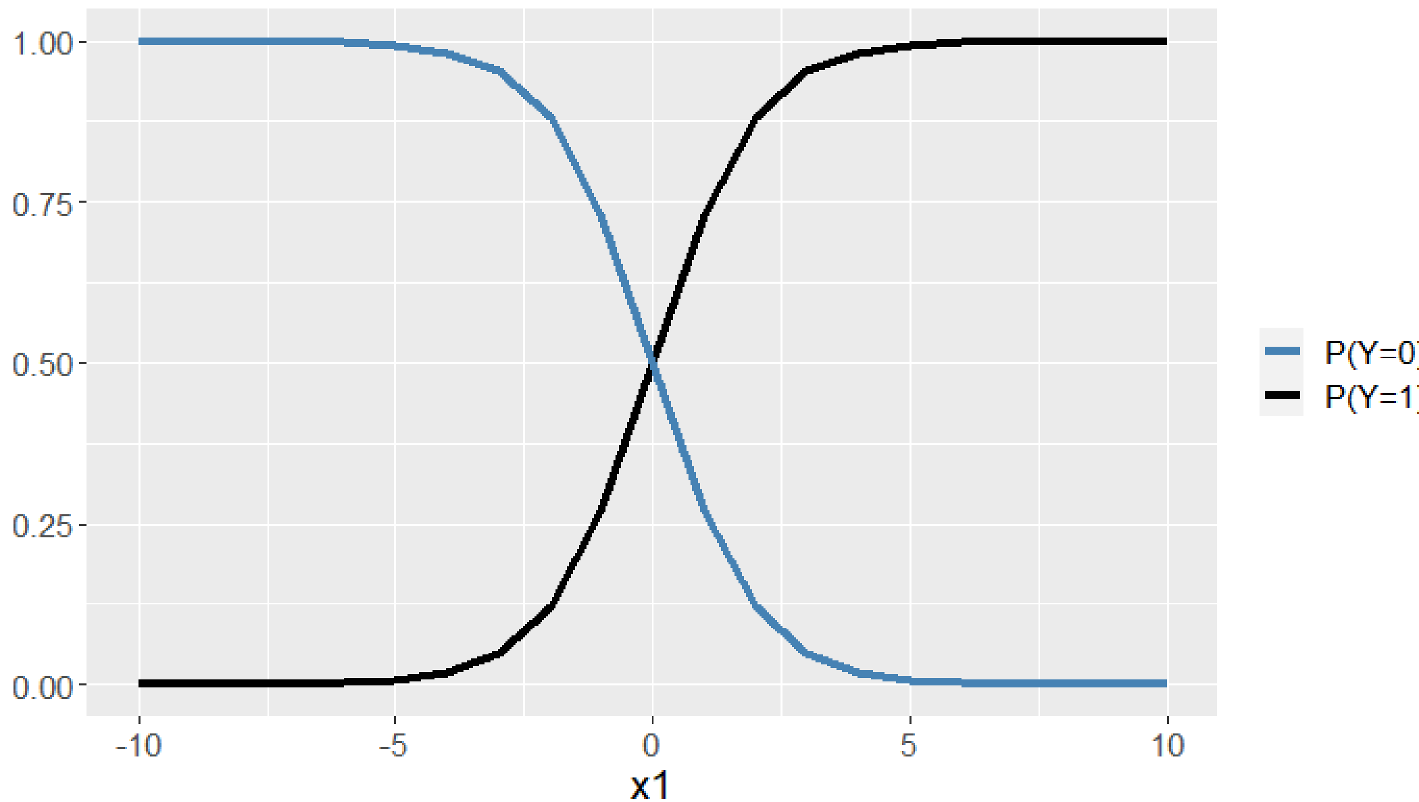
- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM



Logistic Function for P(Y=1) and P(Y=0)

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM



Properties of Logistic Function

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

The S-shape of the logistic function has many convenient properties over a linear function:

1. The domain of the predictions for Y is (0,1), the same as the observations of Y.

2. The response function is smooth unlike a piecewise linear function.

3. The predicted values between 0 and 1 can be interpreted as probabilities, so we don't need a binary classification.

   But we can choose a probability threshold (such as $P(Y=1)>0.5$) to create a binary classification if we want.

Why does this work?

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

Linear Regression: $Y = X\beta + \varepsilon$

Logistic Regression: $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = X\beta$

Domain of Y: (-∞,∞)

Domain of P(Y=1): (0,1)

Domain of $\frac{P(Y=1)}{1-P(Y=1)}$: (0,∞)

Domain of $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$: (-∞,∞)

Logistic regression on the Challenger dataset

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

```
glm.fit = glm(oring~temp, family="binomial", challenger)
```

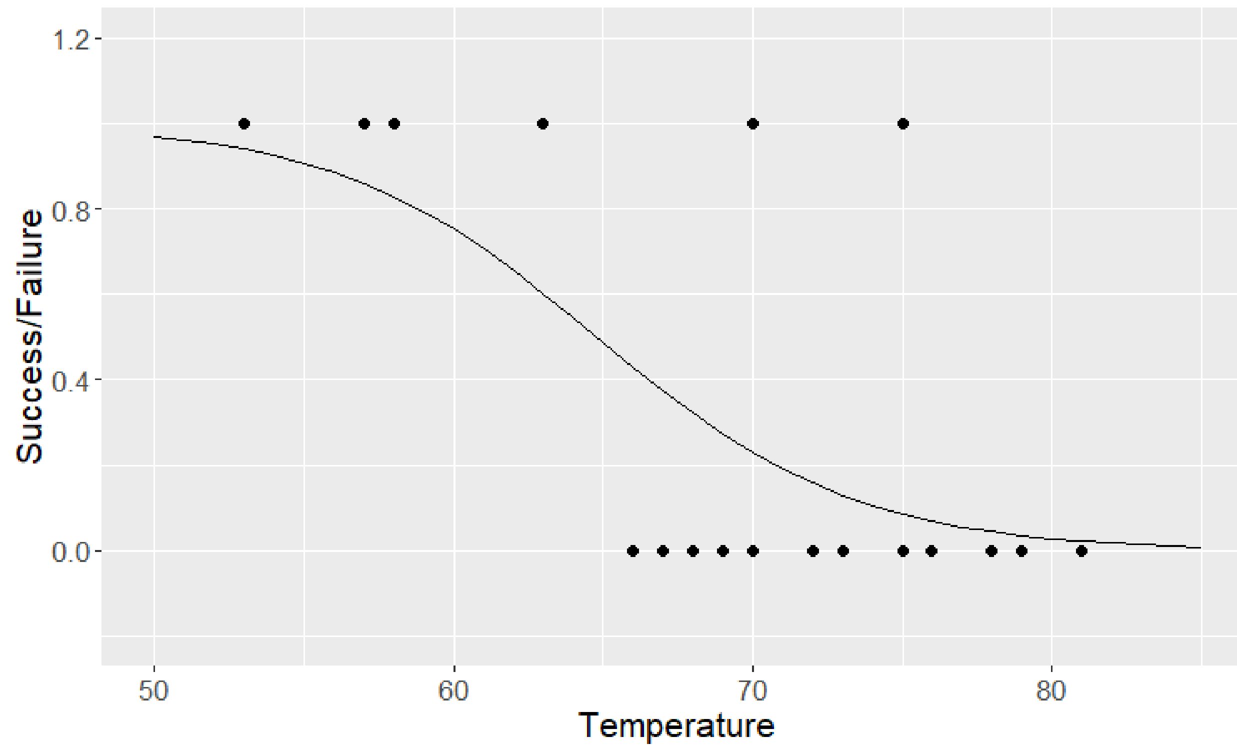
```
> summary(glm.fit)

Call:
glm(formula = oring ~ temp, family = "binomial", data = challenger)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
temp         -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

Logistic regression on the Challenger dataset

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

```
glm.predict = predict(glm.fit, type="response", newdata)

ggplot(challenger, aes(x=temp, y=oring)) + geom_point(size=2) +
  geom_line(data=newdata, aes(y=glm.predict)) +
  xlab("Temperature") + ylab("Success/Failure") +
  theme(text = element_text(size=16)) + ylim(-0.2,1.2) + xlim(50,85)
```

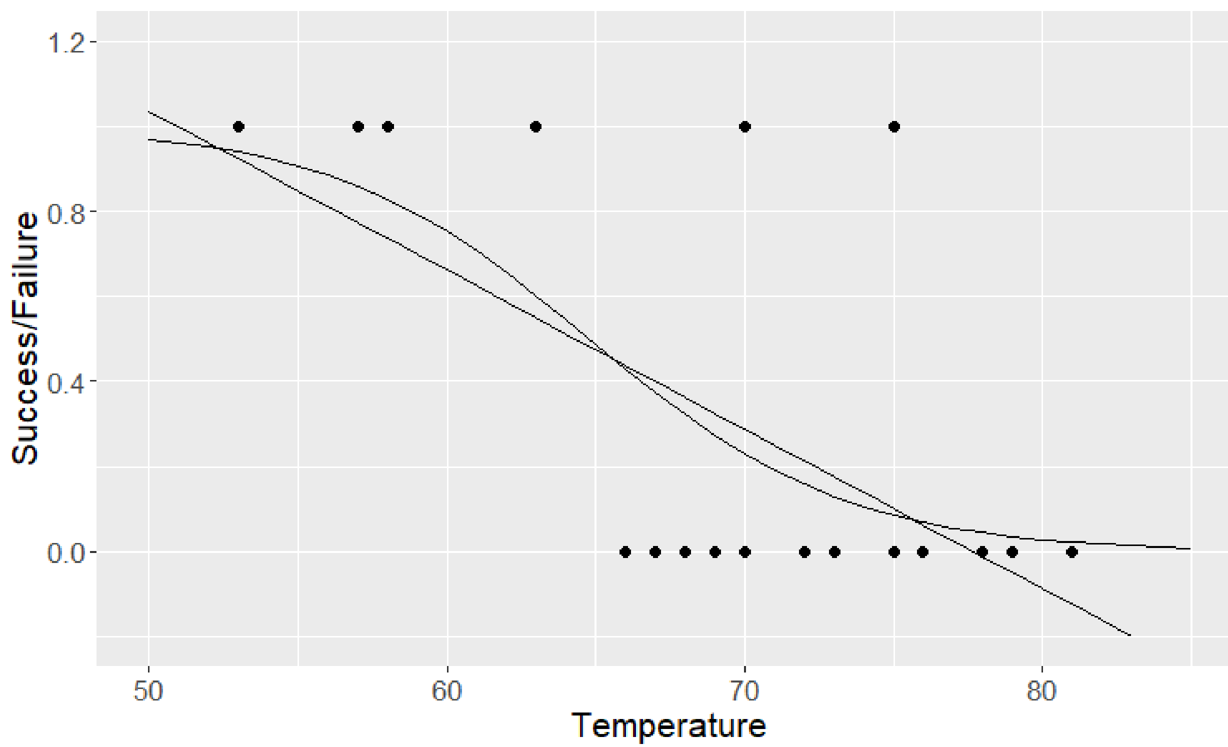


Logistic regression on the Challenger dataset

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

```
ggplot(challenger, aes(x=temp, y=oring)) + geom_point(size=2) +
  geom_line(data=newdata, aes(y=glm.predict)) +
  geom_line(data=newdata, aes(y=linear.predict)) +
  xlab("Temperature") + ylab("Success/Failure") +
  theme(text = element_text(size=16)) + ylim(-0.2,1.2) + xlim(50,85)
```



Summary

Agenda

- Linear regression assumptions
- Problems with linear regression for categorical response
- Generalized Linear Models (GLM)
- Example GLM

1. Linear regression is inappropriate for predicting binary or categorical response variables.

2. Generalized Linear Models allow us to build linear models of a function of the response, transforming a binary or categorical variable onto the real numbers.

3. For binary responses, a logit or logistic function of the log of the odds of the event is modeled as a linear function of the predictors. This can be re-arranged to predict the probability of the event's occurrence.

Generalized Linear Models 2 (GLM)

SYS 4021/6021

Laura Barnes and Julianne Quinn

Organization of lecture

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

1. Review of generalized linear models and logistic regression.

2. Parameter estimation for logistic regression models.

3. Interpreting coefficients in logistic regression models and testing their statistical significance.

Review of Generalized Linear Models (GLMs)

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

Generalized linear models (GLMs) extend regression modeling to include non-Gaussian-distributed response variables.

GLMs use a link function to relate the mean of the response to a linear function of the predictors (where those predictors may be non-linear transformations of the original variables):

$$g(E[Y]) = X\beta$$

Technically, we assume $g(\cdot)$ has a distribution from the exponential family. The Gaussian, binomial, and Poisson distributions satisfy this condition, among others.

Logistic Regression

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

When we have a binary response variable, the link function is the logit or logistic function, and E[Y] represents p=P(Y=1):

$$g(E[Y]) = X\beta$$

$$\log\left(\frac{E[Y]}{1 - E[Y]}\right) = X\beta$$

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{P(Y=1)}{1 - P(Y=1)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) = X\beta$$

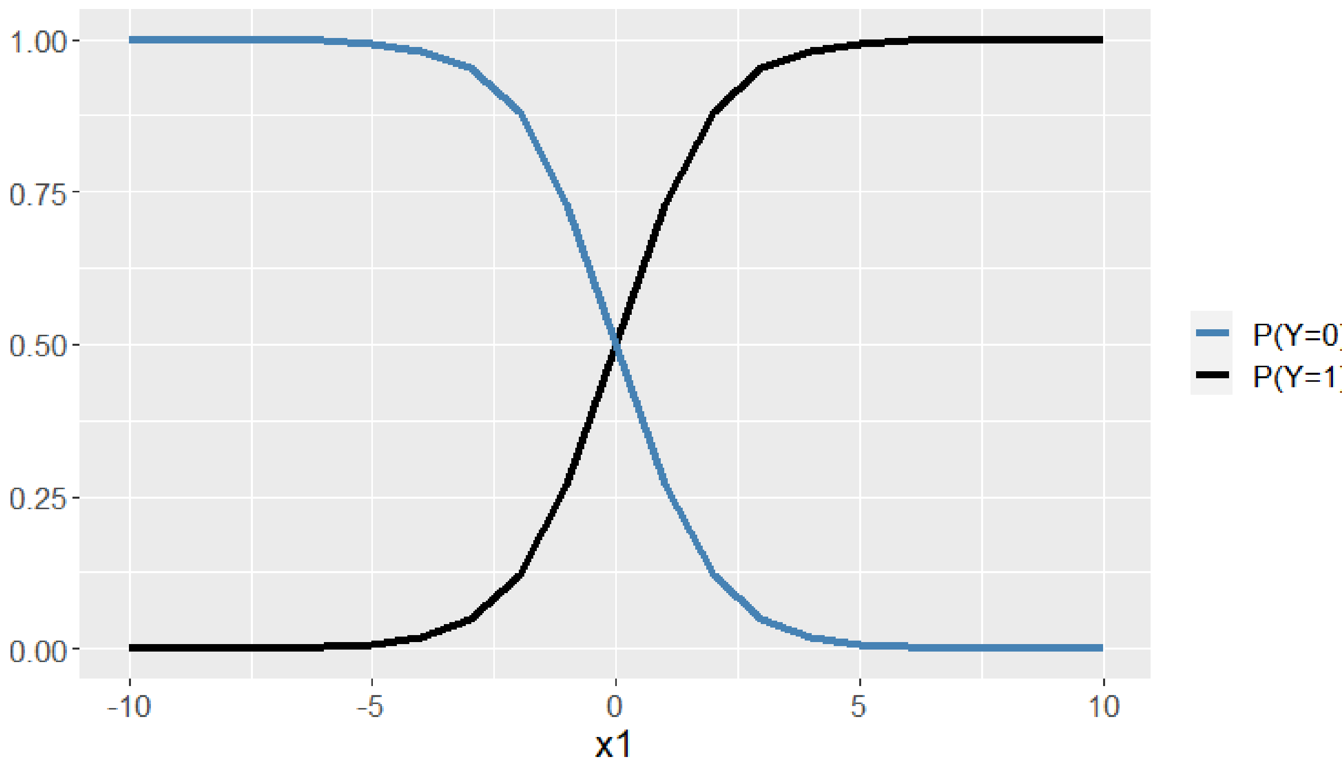
$\frac{P(Y=1)}{P(Y=0)}$ is called the odds, so we are predicting the log-odds as a linear function of our predictors (which may be non-linear).

This is the same as predicting p as a logistic function of our predictors: $p = \frac{\exp(X\beta)}{1+\exp(X\beta)}$

Logistic Function for P(Y=1) and P(Y=0)

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance



Parameter estimation

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

How do we estimate the parameters β for logistic regression?

How do we estimate them for linear regression?

Find the values which minimize the sum of squared errors

What are our errors in logistic regression?

We're predicting a probability of occurrence, but we don't observe the probability, we only observe a binary 0/1 response.

If we don't know the true probability, we can't compute a residual, and therefore can't minimize the sum of squared errors.

Parameter estimation

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

In logistic regression we need to use a different objective function for parameter estimation.

Rather than finding the parameters that minimize the sum of squared errors, we find the parameters under which we would be most likely to observe the binary responses we did.

This is called maximum likelihood estimation (MLE).

Maximum likelihood estimation (MLE)

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

Maximum likelihood estimation finds the parameters that maximize the likelihood function

The likelihood function is the probability of observing exactly what was observed as a function of different parameter values:

$$\mathcal{L}(\beta) = P(Y|\beta) = P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|\beta)$$

If all our observations are independent, an assumption of both linear and logistic regression, then

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|\beta)$$
$$= \mathrm{P}(Y_1 = y_1|\beta)\mathrm{P}(Y_2 = y_2|\beta) \cdots \mathrm{P}(Y_n = y_n|\beta)$$
$$= \prod_{i=1}^{n} P(Y_i = y_i|\beta)$$

Maximum likelihood estimation

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

We want to find β that maximizes:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P(Y_i = y_i|\beta)$$

So what is $P(Y_i = y_i|\beta)$? We know:

$$P(Y_i = 1) = p = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \text{ and}$$

$$P(Y_i = 0) = 1 - p = 1 - \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

We can represent this in the likelihood function as:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \left\{ \left[\frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}\right]^{y_i} \left[1 - \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}\right]^{1-y_i} \right\}$$

Maximum likelihood estimation

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

We want to find β that maximizes:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \left\{ \left[\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \right]^{y_i} \left[1 - \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \right]^{1-y_i} \right\}$$

Or alternatively, the log-likelihood:

$$\log(\mathcal{L}(\beta)) = \sum_{i=1}^{n} \left\{ y_i \log\left[\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \right] + (1-y_i)\log\left[1 - \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \right] \right\}$$

Finding the values of β that maximize this function requires an iterative search algorithm. It is common to use iteratively reweighted least squares. (We'll see why this is a least squares algorithm in GLM3.)

Example logistic regression problem

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

Companies compete for road projects in Florida. They submit a proposal with a project cost and the state department of transportation awards contracts based on these proposals.

Sometimes the companies cheat by colluding on the bids.

Can we build a system to help Florida detect collusion?

For each project, they know the number of bidders and the difference between the proposed cost and the DOT's estimate of the costs provided by one of their engineers.

Road bid collusion problem

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

Let p = P(Y=1) = probability of collusion.

$$\log\left(\frac{p}{1-p}\right) = X\beta$$

We have two predictors:

X_1 = number of bidders

X_2 = difference from engineer's estimate

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

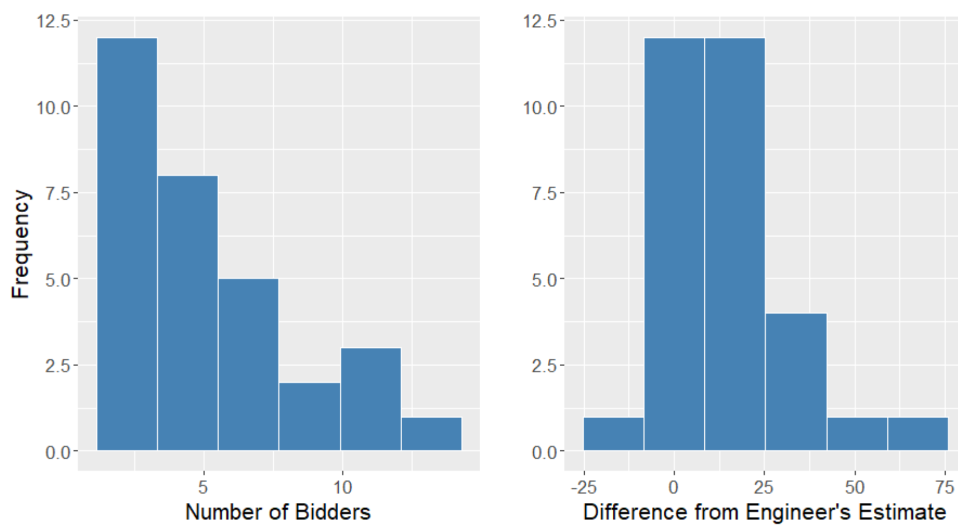
How would you expect X_1 and X_2 to relate to the odds of collusion?

Road bid collusion problem

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
# road bid plots
library(ggpubr)
bidders = read.table("bidders.csv", sep=",", header=TRUE)
bidders$Collusion = as.factor(bidders$Collusion)

bid.hist = ggplot(bidders, aes(x=Bidders)) +
  geom_histogram(fill="steelblue", color="white",
                 bins=nclass.Sturges(bidders$Bidders)) +
  labs(x="Number of Bidders", y="Frequency") +
  theme(text=element_text(size=16))

diff.hist = ggplot(bidders, aes(x=Difference)) +
  geom_histogram(fill="steelblue", color="white",
                 bins=nclass.Sturges(bidders$Bidders)) +
  labs(x="Difference from Engineer's Estimate",y="") +
  theme(text=element_text(size=16))

ggarrange(bid.hist, diff.hist, nrow=1, ncol=2)
```

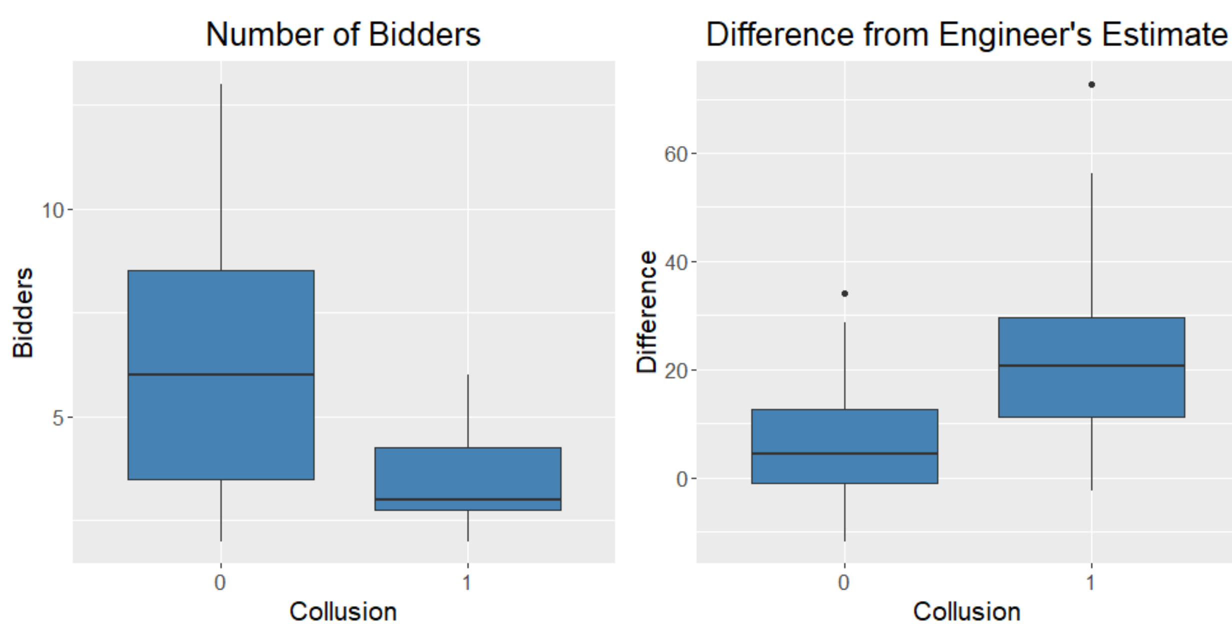


Road bid collusion problem

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
bid.boxplot = ggplot(bidders, aes(x=Collusion, y=Bidders)) +
  geom_boxplot(fill= "steelblue") +
  ggtitle("Number of Bidders") +
  theme(text=element_text(size=16), plot.title = element_text(hjust=0.5))

diff.boxplot = ggplot(bidders, aes(x=Collusion, y=Difference)) +
  geom_boxplot(fill= "steelblue") +
  ggtitle("Difference from Engineer's Estimate") +
  theme(text=element_text(size=16), plot.title = element_text(hjust=0.5))

ggarrange(bid.boxplot, diff.boxplot, nrow=1, ncol=2)
```

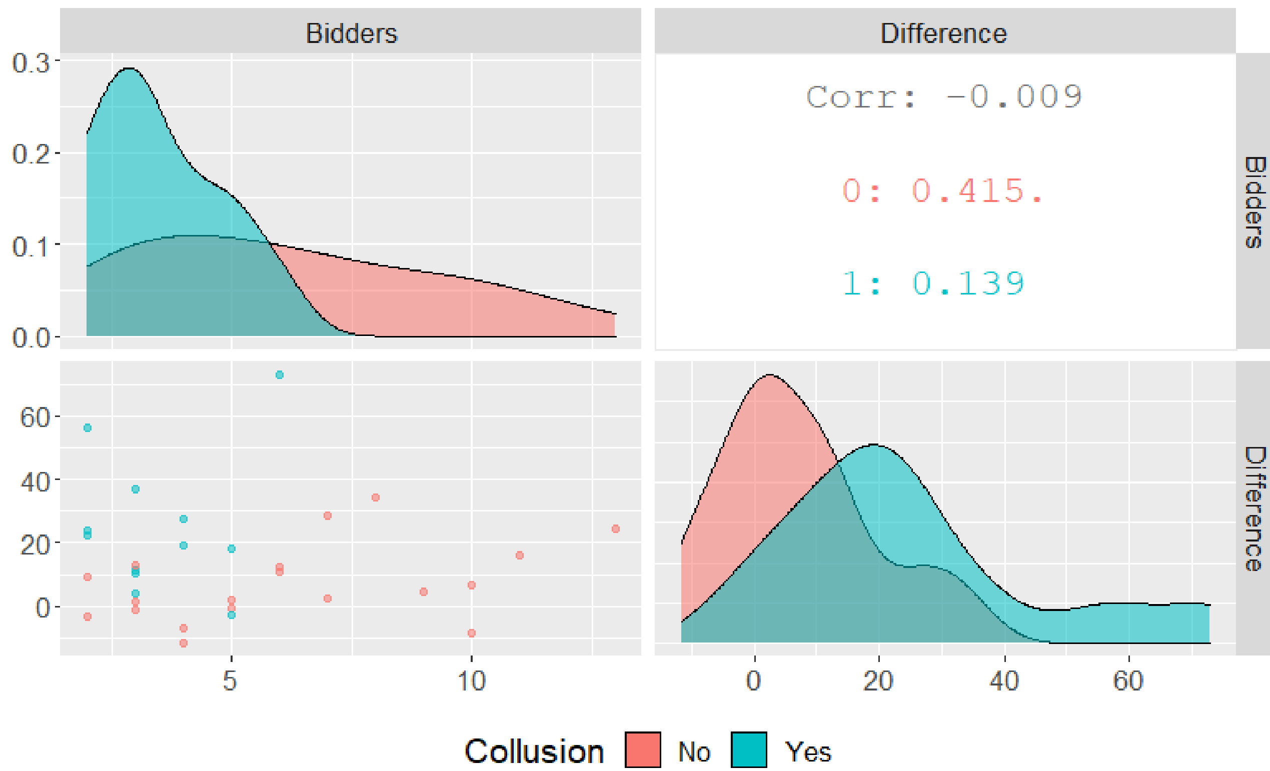


Road bid collusion problem

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
library(GGally)
ggpairs(bidders[,c("Bidders","Difference")],
        aes(color=bidders$Collusion, alpha=0.5),
        legend=1, upper = list(continuous=wrap("cor", size=6))) +
  theme(text=element_text(size=16), legend.position="bottom") +
  scale_fill_discrete(name="Collusion", labels=c("No","Yes")) +
  scale_alpha_continuous(guide="none")
```



Fitting the GLM

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
> bidder.glm = glm(Collusion~Bidders+Difference, family="binomial",
+                  bidders)
> summary(bidder.glm)

Call:
glm(formula = Collusion ~ Bidders + Difference, family = "binomial",
    data = bidders)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5898  -0.5514  -0.1119   0.3973   2.3260

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.42120    1.28677   1.104   0.2694
Bidders     -0.75534    0.33880  -2.229   0.0258 *
Difference   0.11220    0.05139   2.183   0.0290 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.381  on 30  degrees of freedom
Residual deviance: 22.843  on 28  degrees of freedom
AIC: 28.843

Number of Fisher Scoring iterations: 6
```

Interpreting the logistic regression model

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
Call:
glm(formula = Collusion ~ Bidders + Difference, family = "binomial",
    data = bidders)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.42120    1.28677   1.104   0.2694
Bidders     -0.75534    0.33880  -2.229   0.0258 *
Difference   0.11220    0.05139   2.183   0.0290 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Write an expression for the probability of cheating.

$$\log\left(\frac{P(Collusion)}{P(No\ Collusion)}\right)$$
$$= 1.42120 - 0.75534\ Bidders + 0.11220\ Difference$$

$$p = \frac{\exp(1.42120 - 0.75534\ Bidders + 0.11220\ Difference)}{1 + \exp(1.42120 - 0.75534\ Bidders + 0.11220\ Difference)}$$

What is the probability of cheating if there are 3 bidders and the difference from the engineer's estimate is 20%?

Interpreting the logistic regression model

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
Call:
glm(formula = Collusion ~ Bidders + Difference, family = "binomial",
    data = bidders)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.42120    1.28677   1.104   0.2694
Bidders     -0.75534    0.33880  -2.229   0.0258 *
Difference   0.11220    0.05139   2.183   0.0290 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Write an expression for the probability of cheating.

$$\log\left(\frac{P(Collusion)}{P(No\ Collusion)}\right) = 1.42120 - 0.75534\ Bidders + 0.11220\ Difference$$

$$p = \frac{\exp(1.42120 - 0.75534\ Bidders + 0.11220\ Difference)}{1 + \exp(1.42120 - 0.75534\ Bidders + 0.11220\ Difference)}$$

What is the probability of cheating if there are 3 bidders and the difference from the engineer’s estimate is 20%? 0.80

Interpreting the logistic regression model

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

In linear regression, β_i represents the change in Y per unit change in X_i, while holding the other predictors constant.

In logistic regression, our response is the log(odds). So β_i tells us the change in the log(odds) for a unit change in X_i.

This is more interpretable if we exponentiate both sides. Then e^{β_i} tells us the factor change in the odds for a unit change in X_i.

Even better, $100(e^{\beta_i}-1)$ gives the % change in the odds per unit change in X_i, while holding the other predictors constant.

Interpreting logistic regression coefficients

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
Call:
glm(formula = Collusion ~ Bidders + Difference, family = "binomial",
    data = bidders)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.42120    1.28677   1.104   0.2694
Bidders     -0.75534    0.33880  -2.229   0.0258 *
Difference   0.11220    0.05139   2.183   0.0290 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\log\left(\frac{P(Collusion)}{P(No\ Collusion)}\right) = 1.42120 - 0.75534\ Bidders + 0.11220\ Difference$$

Bidders: 100(exp(-0.75534)–1) = -53%

Difference: 100(exp(-0.11220)–1) = +12%

Model and Coefficient Significance

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

We now know how to interpret the model coefficients, but how do we determine if they are significant, and likewise, if the model has predictive value?

In linear least squares regression, we use the model utility test to assess model significance.

Recall the test statistic for the model utility test is the F statistic:

$$F = \frac{MSS/p}{RSS/(n-p-1)} \text{ where}$$

$MSS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$ is the model sum of squares (variance in Y explained by the model),

$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares (variance of the residuals, i.e. unexplained variance in Y)

and p is the number of predictors, n the number of observations.

Model utility test for logistic regression

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

We can't use an F test for logistic regression for the same reason we can't estimate parameters by minimizing the sum of squared errors: we don't know the true probabilities.

But we can test the same hypothesis another way:

H0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

Ha: $\beta_i \neq 0$ for some *i*

If we accept H0, we choose a model with only an intercept, a "null model": $Y = \beta_0 + \varepsilon$

If we reject H0, there is evidence the larger model has utility beyond the null model.

Model utility for logistic regression

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

What does a null model predict?

$$Y = \beta_0 + \varepsilon$$

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

In both cases it is predicting the mean of the response.

What is the mean of the response in logistic regression?

The average log odds, or re-formatted for probability, the average probability of observing the event: k/n, where k is the number of times Y=1 out of n observations.

Model utility for logistic regression

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

Both parameter estimation and model utility are based on sum of squared errors for linear regression.

Parameter estimation for logistic regression is based on the likelihood function. Guess what? So is model utility.

In logistic regression we compare the likelihood of the null model to the likelihood of the larger model to see if we can reject H0 in favor of the more complex model.

The test statistic is not an F statistic, but the deviance.

Deviance test statistic

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

Deviance, D, is defined as:

$$D = -2\log(L(\beta))$$

We want high likelihoods, or low deviances.

When comparing two models, we compare their deviances, $D_0 - D_1$, where D_0 is the deviance of the null model and D_1 is the deviance of the larger model. The greater this difference, the better the larger model.

But like RSS, adding predictors will always reduce the deviance, so we need to penalize the addition of parameters.

$D_0 - D_1 \sim \chi_p^2$ as $n \to \infty$ where p is the number of predictors in the complex model. This parameter will control for overfitting.

Deviance test for model utility

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

To implement the deviance test (aka likelihood test) for model utility in R, we need to build a null model.

```
> bidder.null.glm = glm(Collusion~1, family="binomial", bidders)
```

We can then compare the null and complex models using the anova function with the argument test="Chi" to indicate our test-statistic has a Chi-Squared distribution, not an F distribution like in a partial F test.

```
> anova(bidder.null.glm, bidder.glm, test='Chi')
Analysis of Deviance Table

Model 1: Collusion ~ 1
Model 2: Collusion ~ Bidders + Difference
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        30     41.381
2        28     22.843  2   18.538 9.431e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Deviance to compare nested models

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

We can also use deviance to compare nested models (in fact, the null model is nested within the complex model).

Let D_1 be the deviance of the smaller model with p predictors and D_2 be the deviance of the larger model with $p+q$ predictors. Then $D_1 - D_2 \sim \chi_q^2$.

The null and hypotheses are the same as for the F test in linear regression:

H0: $\beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$

Ha: $\beta_{p+i} \neq 0$ for some i

Comparing a main effects and interaction model

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
> bidder.interact.glm = glm(Collusion~(Bidders+Difference)^2,
+                           family="binomial", bidders)
> summary(bidder.interact.glm)

Call:
glm(formula = Collusion ~ (Bidders + Difference)^2, family = "binomial",
    data = bidders)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6783  -0.5037  -0.2122   0.2323   2.2081

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.9835     1.7674  -0.556   0.5779
Bidders             -0.2127     0.3351  -0.635   0.5256
Difference           0.3533     0.1696   2.083   0.0372 *
Bidders:Difference  -0.0467     0.0281  -1.662   0.0965 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.381  on 30  degrees of freedom
Residual deviance: 19.625  on 27  degrees of freedom
AIC: 27.625

Number of Fisher Scoring iterations: 6
```

Comparing a main effects and interaction model

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

```
> anova(bidder.glm, bidder.interact.glm, test='Chi')
Analysis of Deviance Table

Model 1: Collusion ~ Bidders + Difference
Model 2: Collusion ~ (Bidders + Difference)^2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        28     22.843
2        27     19.625  1   3.2176  0.07285 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, we reject H0 at the 10% level, but not at the 5% level.

Deviance test for coefficient significance

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

Since we can use the deviance test to compare nested models, we can therefore use it to test the significance of particular coefficients by adding or dropping a single predictor from one model to the next and comparing their deviances.

If the model with one additional predictor has a statistically lower deviance, that predictor is significant.

This approach is called a partial deviance, or partial likelihood test.

Parameter significance

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

If we use the drop1 function with the argument test="Chi", it will tell us the significance of each coefficient based on the deviance of the full model compared to the model with all but that predictor.

```
> drop1(bidder.glm, response~., test = "Chi", data = bidders)
Single term deletions

Model:
Collusion ~ Bidders + Difference
           Df Deviance    AIC    LRT Pr(>Chi)
<none>          22.843 28.843
Bidders     1   33.252 37.252 10.409 0.001254 **
Difference  1   33.296 37.296 10.453 0.001224 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

Agenda

- Review of GLM
- Logistic regression
- Maximum likelihood
- Example logistic regression
- Interpreting logistic regression
- Model and Coefficient Significance
- Model utility of logistic regression
- Deviance test
- Parameter significance

1. Logistic regression allows us to predict the probability of a binary response using a logistic function of the predictors.

2. Parameter estimation for logistic regression models must be performed using maximum likelihood estimation, since the true probability is unknown to compute a residual.

3. The likelihood, or deviance, can also be used to test model utility against a null model. Partial likelihood can be used to assess parameter significance.

Generalized Linear Models 3 (GLM)

SYS 4021/6021

Laura Barnes and Julianne Quinn

Organization of lecture

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

1. Model evaluation for generalized linear models
   a) Diagnostics
   b) Variable selection

2. Decision functions to map probabilities to classifications
   a) Types of classification error
   b) Receiver Operator Characteristic (ROC) curves of classification errors for different decision functions

3. Principal Components Regression (PCR)

Review of Generalized Linear Models (GLMs)

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

Generalized linear models (GLMs) extend regression modeling to include non-Gaussian-distributed response variables.

GLMs use a link function to relate the mean of the response to a linear function of the predictors (where those predictors may be non-linear transformations of the original variables):

$$g(E[Y]) = X\beta$$

Technically, we assume $g(\cdot)$ has a distribution from the exponential family. The Gaussian, binomial, and Poisson distributions satisfy this condition, among others.

Diagnostics in logistic regression

Agenda

- Review of GLM
- Logistic Regression
    - Diagnostics
    - Residual
    - Example
    - Variable selection
    - Test set
    - Type of error
    - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

In linear regression, we assume the residuals are independent and Gaussian-distributed with a constant mean of 0 and constant variance.

In generalized linear models, and specifically logistic regression models, do we make these same assumptions?

What are the residuals in a logistic regression model?

Residuals in logistic regression

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

In logistic regression we predict the log-odds or probability (p) of an event Y=1:

$$\log\left(\frac{p}{1-p}\right) = X\beta, \qquad p = \frac{\exp(X\beta)}{1+\exp(X\beta)}$$

But we don't know the true probability, so we can't calculate the true residual.

But we can calculate the response residual $y_i - \hat{p}_i$, where y_i is the binary response and $\hat{p}_i$ is the predicted probability Y=1.

Another residual computed in logistic regression is the deviance residual. This relates to the Deviance statistic for model utility.

Deviance as an analogy to least squares

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

Because logistic regression uses maximum likelihood to estimate the parameters β, it also minimizes the deviance $D = -2\log(L(\beta))$:

$$\text{Max } \log(\mathcal{L}(\beta)) = \sum_{i=1}^{n}\left\{y_i \log\left[\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}\right] + (1-y_i)\log\left[1 - \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}\right]\right\}$$

$$\text{Max } \log(\mathcal{L}(\beta)) = \sum_{i=1}^{n}\{y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\}$$

$$\text{Min } D = -2\log\big(L(\beta)\big) = \sum_{i=1}^{n} -2\{y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\}$$

In linear regression we try to minimize the sum of squared errors. We can rewrite the deviance in an analogous fashion:

$$\text{Min } D = \sum_{i=1}^{n} \sqrt{-2\{y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\}}^{\,2} = \sum_{i=1}^{n} d_i^2$$

These d_i values are called deviance residuals.

Residuals in logistic regression

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

We will use both response and deviance residuals in diagnostics.

Unlike in linear regression, the response residuals $y_i - \widehat{p}_i$ do not have constant variance; their variance depends on the mean:

$$\text{Var}(\text{E}[\text{Y}]) = \text{E}[\text{Y}](1\text{-E}[\text{Y}]).$$

These residuals will also exhibit a pattern. So we are less concerned about the residuals vs. fitted plot.

But if we standardize $y_i - \widehat{p}_i$ by dividing by its standard deviation, this Pearson residual should be approximately Gaussian:

$$\frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1 - \widehat{p}_i)}}$$

Standardized deviance residuals should also be Gaussian. And we will still be concerned about potential influential points.

Consider again the road bid collusion problem

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```
> bidder.glm = glm(Collusion~Bidders+Difference, family="binomial",
+                  bidders)
> summary(bidder.glm)

Call:
glm(formula = Collusion ~ Bidders + Difference, family = "binomial",
    data = bidders)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5898  -0.5514  -0.1119   0.3973   2.3260

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.42120    1.28677   1.104   0.2694
Bidders     -0.75534    0.33880  -2.229   0.0258 *
Difference   0.11220    0.05139   2.183   0.0290 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.381  on 30  degrees of freedom
Residual deviance: 22.843  on 28  degrees of freedom
AIC: 28.843

Number of Fisher Scoring iterations: 6
```

Regression diagnostics

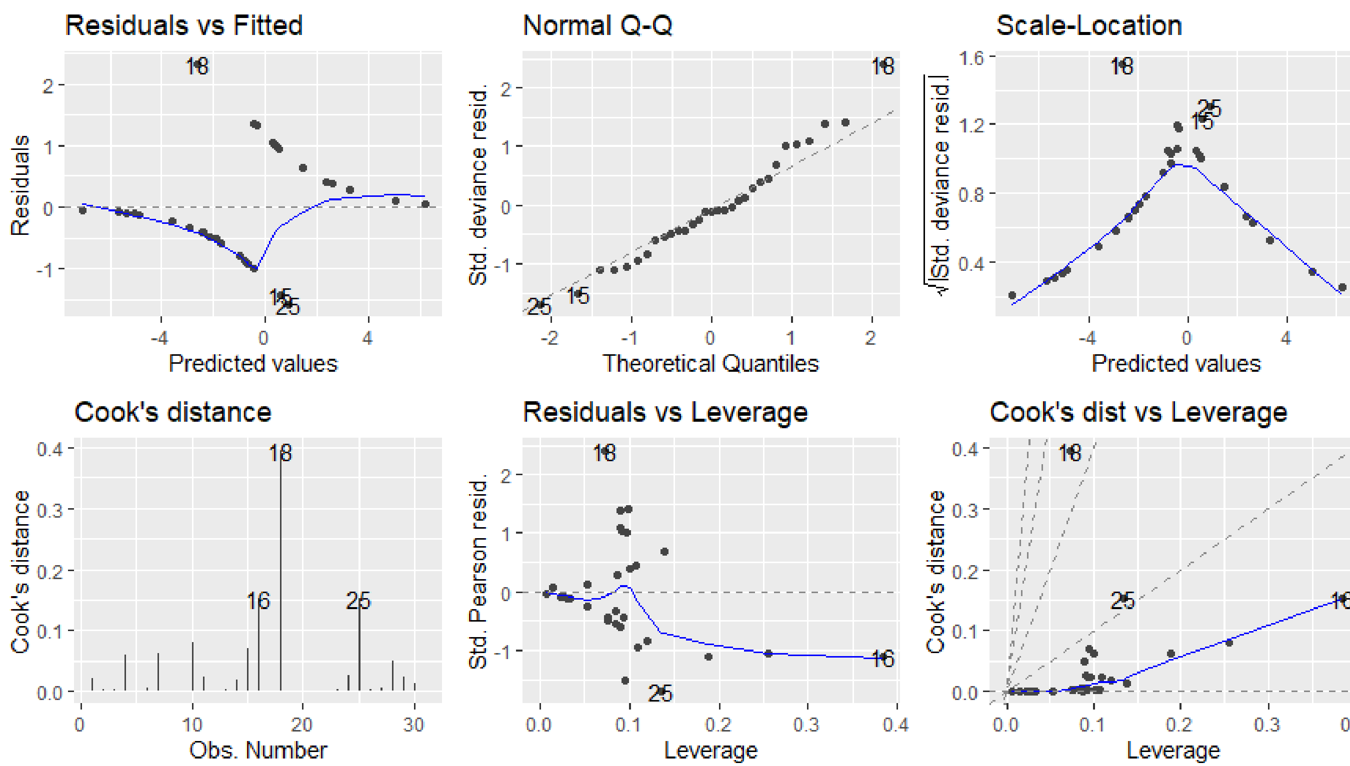
Agenda

- Review of GLM
- Logistic Regression
    - Diagnostics
    - Residual
    - Example
    - Variable selection
    - Test set
    - Type of error
    - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```
> autoplot(bidder.glm, which=1:6, nrow=2, ncol=3)
```

We see an expected pattern

The residuals are fairly Gaussian



Some points approaching influential

Variable selection in logistic regression

Agenda

- Review of GLM
- Logistic Regression
    - Diagnostics
    - Residual
    - Example
    - Variable selection
    - Test set
    - Type of error
    - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

Variable selection in logistic regression is similar to variable selection in linear regression.

We can use a partial likelihood test to determine significance, but this isn't ideal for the same reason a t-test isn't: test multiplicity.

Criterion based metrics still apply, and we can still use automated selection methods.

But test sets and cross-validation are even more important for logistic regression because we care more about prediction out of sample than inference on model coefficients.

Test sets for logistic regression

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

Similar to linear regression, we split our data into ~2/3 for training and ~1/3 for testing.

In linear regression, we use pMSE on the test set as our evaluation metric. What should it be for logistic regression?

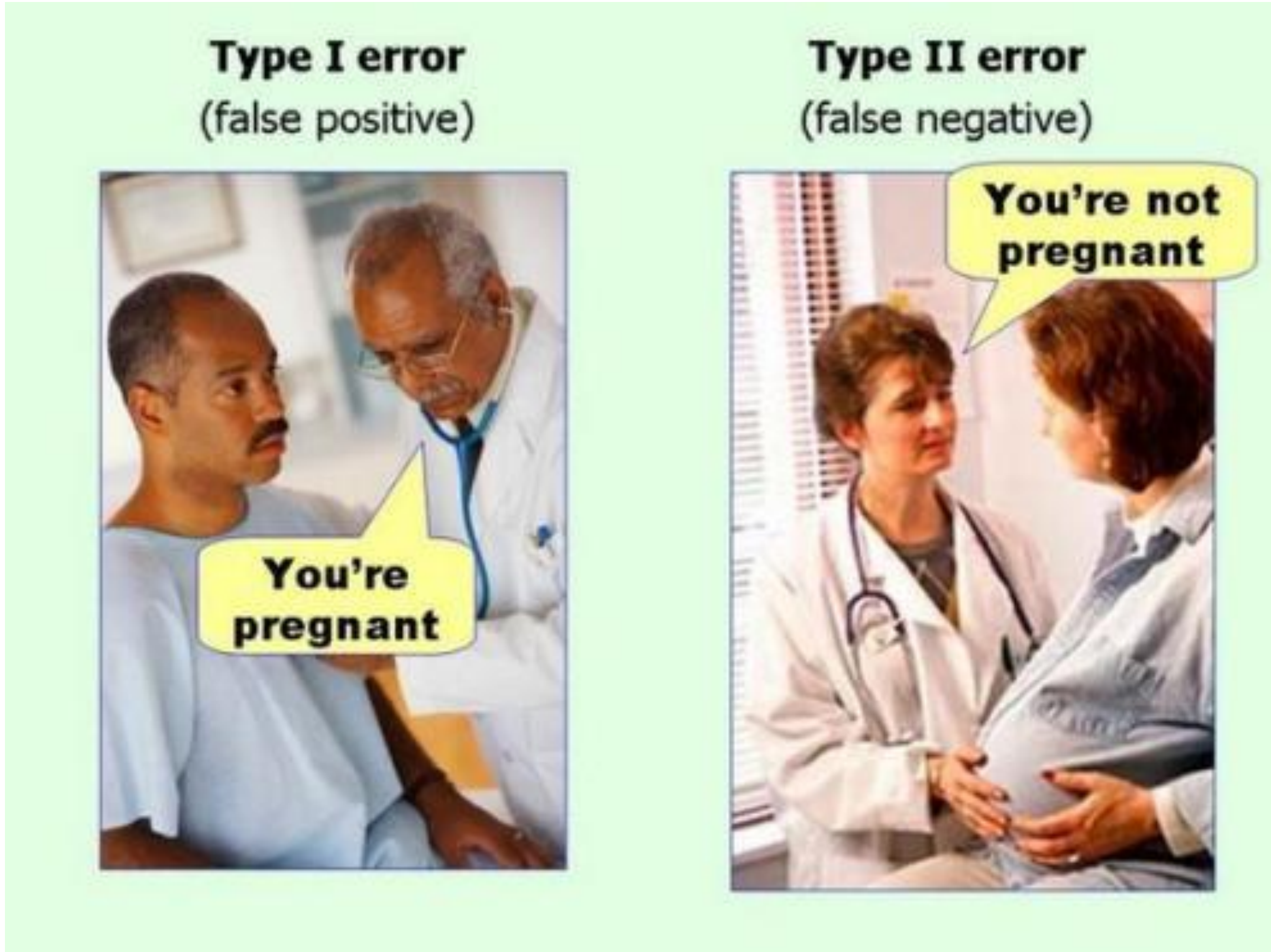
Since we care about classification into binary categories, we need a decision function to map probabilities to categories.

Based on the decision function, we can count our total misclassification errors: false positives + false negatives

Types of errors

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix



Decision function

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

A decision function maps the probability/odds of an event to a finite set of categories.

For logistic regression, we choose a threshold, T, for the probability of the event. If $P(y_i=1) > T$, we predict $\hat{y}_i = 1$, otherwise $\hat{y}_i = 0$.

A threshold T=0.5 is "risk neutral"; false positives and false negatives are weighted equally.

If T>0.5, we prefer false negatives to false positives and vice versa.

Example classification problem

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

The R package bestglm has a dataset on heart disease in a heart disease-high region of South Africa.

```
# heart data plots
library(bestglm)
data(SAheart)
SAheart$famhist=as.factor(SAheart$famhist)
```

We want to predict whether an individual will have coronary heart disease (chd) as a function of their personal and family health history.

```
> head(SAheart)
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45   0
```

Comparing heart disease models in testing

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

We will compare the performance of 3 models in testing:

1) Main effects
2) Main effects + interactions
3) Stepwise from main effects + interactions

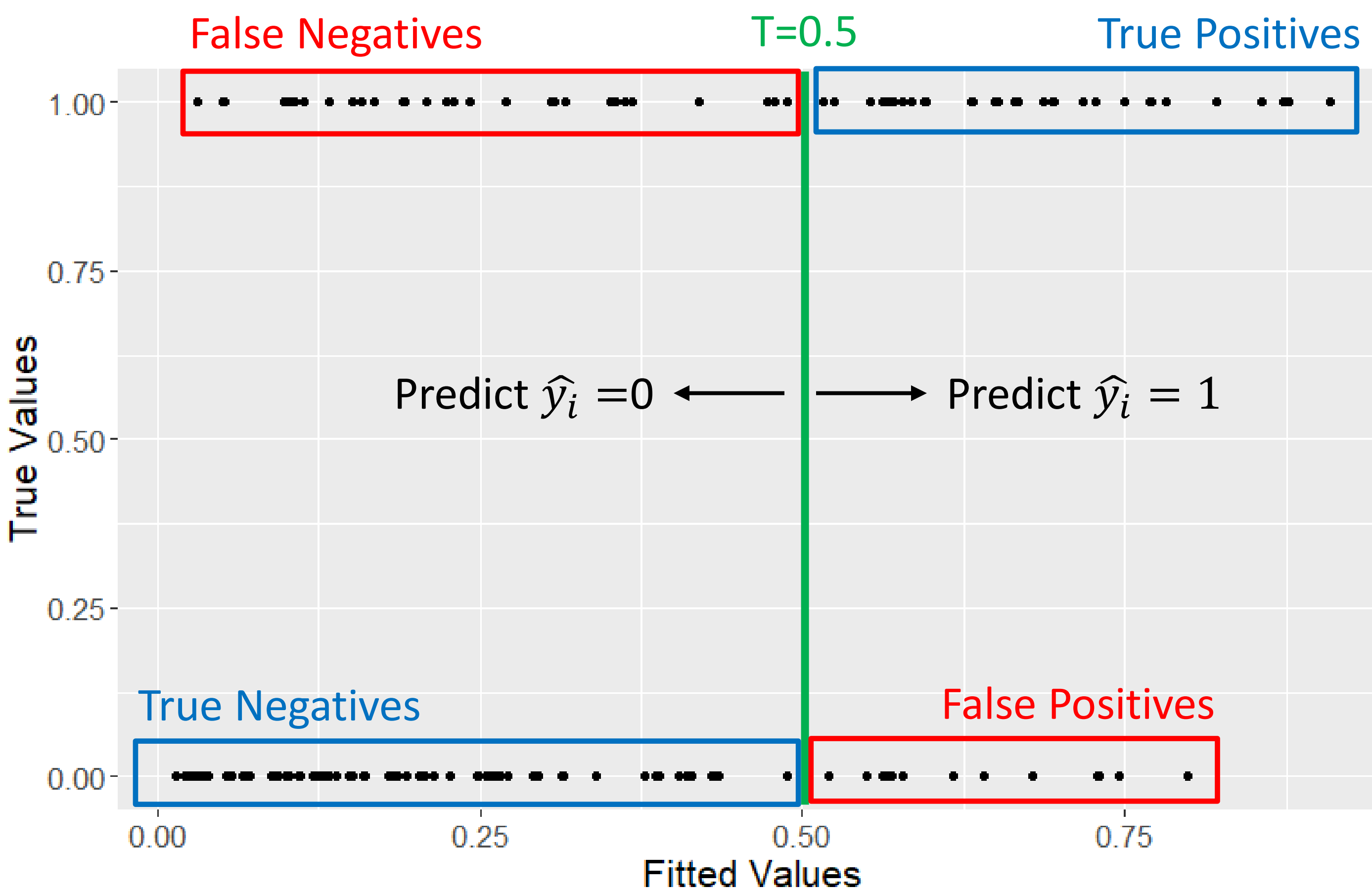
```
# split into training and test
source("TestSet.R")
Heart <- test.set(SAheart, .33)

# build models to training
main.glm = glm(chd~., family="binomial", Heart$train)
interact.glm = glm(chd~.^2, family="binomial", Heart$train)
step.glm = step(interact.glm, trace=0)

# predict on testing
main.glm.pred = predict(main.glm, type = "response",
                        newdata = Heart$test)
interact.glm.pred = predict(interact.glm, type = "response",
                            newdata = Heart$test)
step.glm.pred = predict(step.glm, type = "response",
                        newdata = Heart$test)
```

Main effects true vs. fitted plot

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```
ggplot() + geom_point(aes(x=main.glm.pred, y=Heart$test$chd)) +
  xlab("Fitted Values") + ylab("True Values") +
  theme(text = element_text(size=16))
```

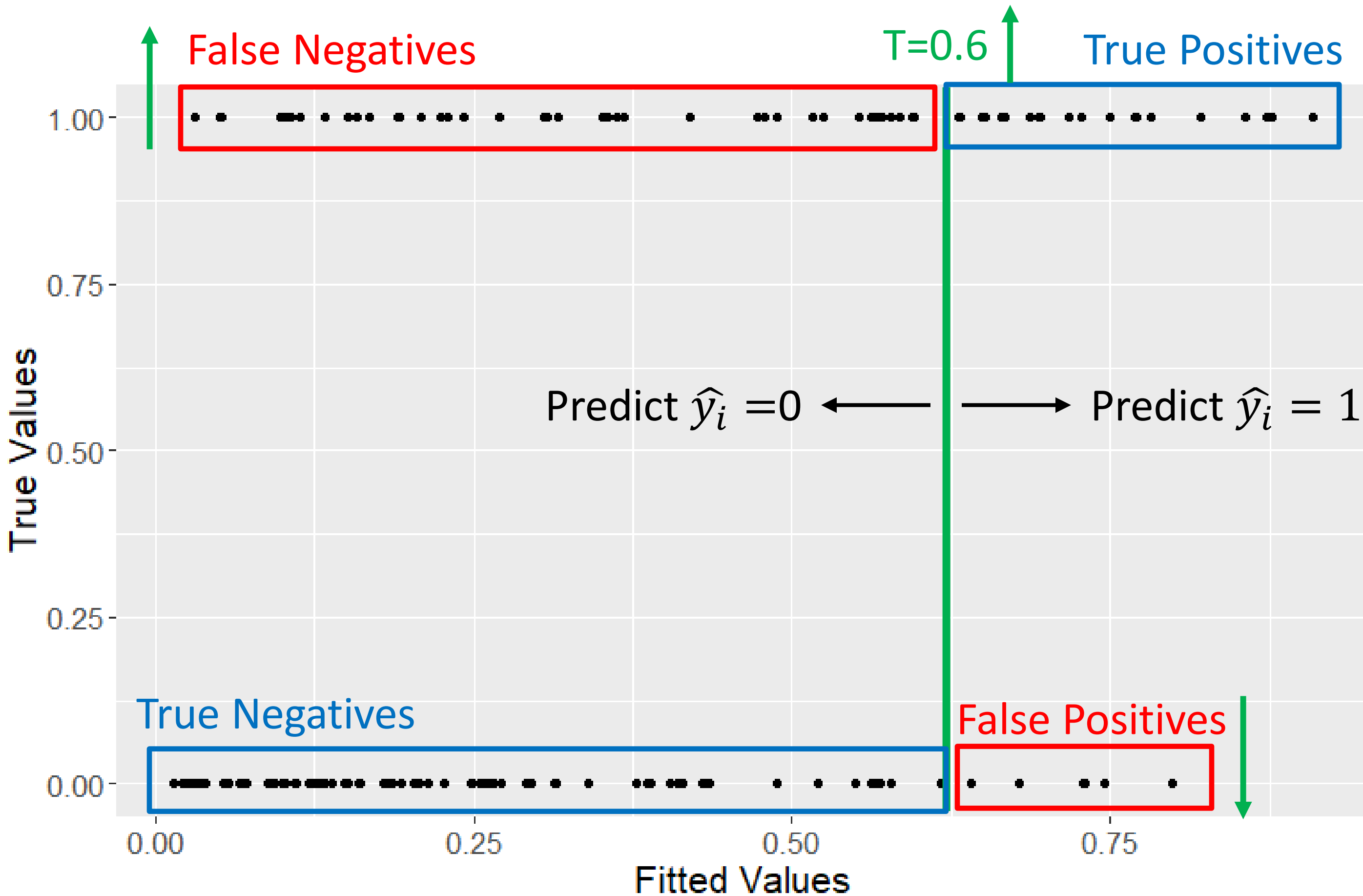


Main effects true vs. fitted plot

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```
ggplot() + geom_point(aes(x=main.glm.pred, y=Heart$test$chd)) +
  xlab("Fitted Values") + ylab("True Values") +
  theme(text = element_text(size=16))
```

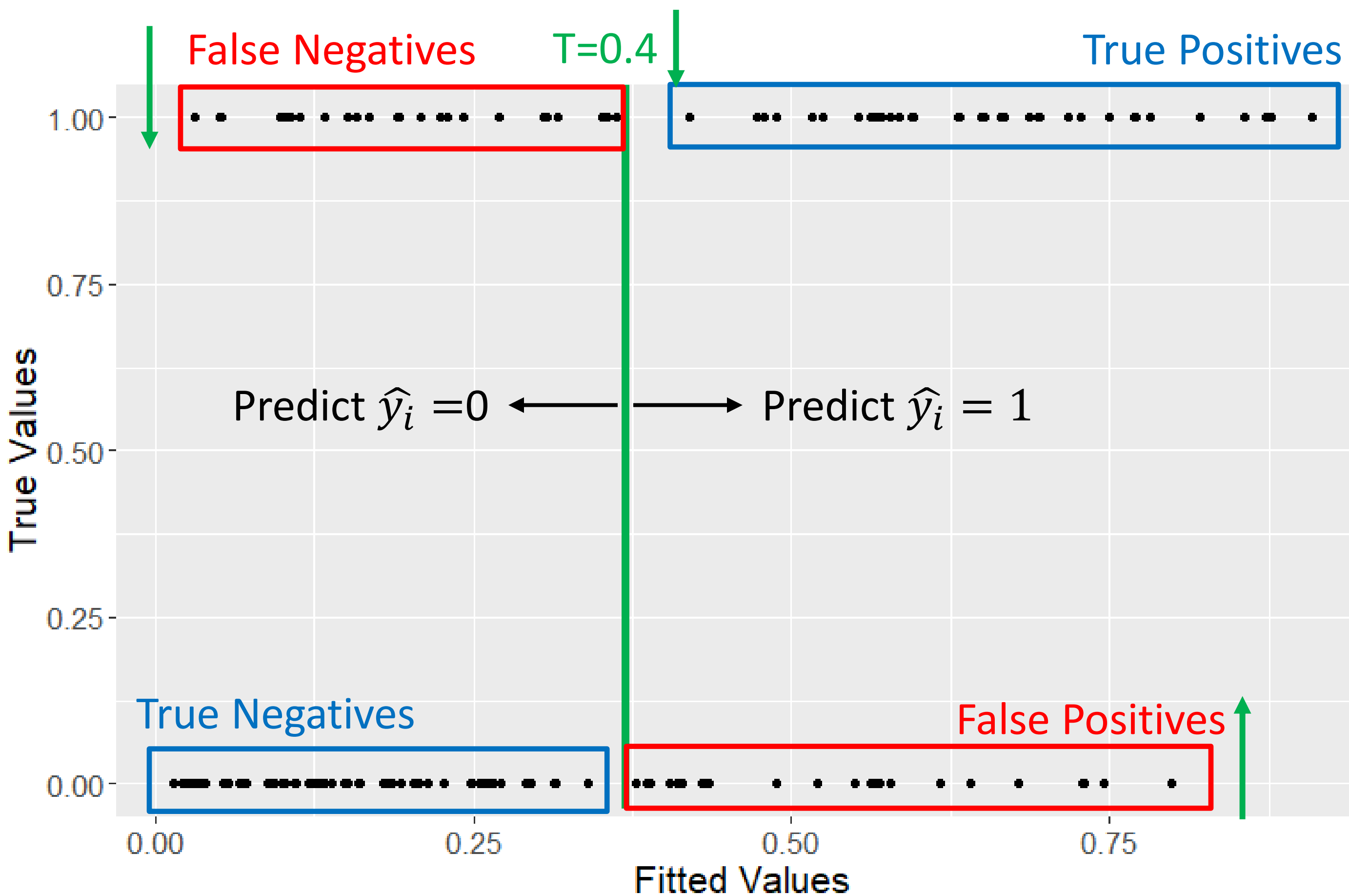


Main effects true vs. fitted plot

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```
ggplot() + geom_point(aes(x=main.glm.pred, y=Heart$test$chd)) +
  xlab("Fitted Values") + ylab("True Values") +
  theme(text = element_text(size=16))
```



Comparing models

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

We can compare the true and false positives and negatives across models in a confusion matrix or score table.

```
> score.table(main.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    77   15
     1    30   30
> score.table(interact.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    74   18
     1    29   31
> score.table(step.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    78   14
     1    28   32
```

Threshold T

Comparing models

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

We can compare the true and false positives and negatives across models in a confusion matrix or score table.

```
> score.table(main.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    77   15
     1    30   30
> score.table(interact.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    74   18
     1    29   31
> score.table(step.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    78   14
     1    28   32
```

False positives

Comparing models

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

We can compare the true and false positives and negatives across models in a confusion matrix or score table.

```
> score.table(main.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    77   15
     1    30   30
> score.table(interact.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    74   18
     1    29   31
> score.table(step.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    78   14
     1    28   32
```

False negatives

What if we changed the threshold?

Agenda

- Review of GLM
- Logistic Regression
    - Diagnostics
    - Residual
    - Example
    - Variable selection
    - Test set
    - Type of error
    - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

A Receiver Operator Characteristic (ROC) curve allows us to compare models at different thresholds of the decision function

It plots the True Positive Rate (TPR) vs. the False Positive Rate (FPR)

$$TPR = P(\hat{y}_i = 1 | y_i = 1) = \frac{TP}{TP + FN}$$

$$FPR = P(\hat{y}_i = 1 | y_i = 0) = \frac{FP}{FP + TN}$$

What is 1-TPR? 1-FPR?

Understanding True/False Positive/Negative Rates

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

$$TPR = P(\hat{y}_i = 1 | y_i = 1) = \frac{TP}{TP + FN}$$

What is 1-TPR?

$$1 - TPR = 1 - P(\hat{y}_i = 1 | y_i = 1)$$

$$1 - TPR \ = P(\hat{y}_i = 0 | y_i = 1)$$

$$1 - TPR = \frac{FN}{TP + FN}$$

Understanding True/False Positive/Negative Rates

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

$$TPR = P(\hat{y}_i = 1 | y_i = 1) = \frac{TP}{TP + FN}$$

What is 1-TPR? False Negative Rate

$$1 - TPR = 1 - P(\hat{y}_i = 1 | y_i = 1)$$

$$1 - TPR = P(\hat{y}_i = 0 | y_i = 1)$$

$$1 - TPR = \frac{FN}{TP + FN}$$

$$1 - TPR = FNR$$

Understanding True/False Positive/Negative Rates

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

$$FPR = P(\hat{y}_i = 1 | y_i = 0) = \frac{FP}{FP + TN}$$

What is 1-FPR?

$$1 - FPR = 1 - P(\hat{y}_i = 1 | y_i = 0)$$
$$1 - FPR = P(\hat{y}_i = 0 | y_i = 0)$$
$$1 - FPR = \frac{TN}{FP + TN}$$

Understanding True/False Positive/Negative Rates

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

$$FPR = P(\hat{y}_i = 1 | y_i = 0) = \frac{FP}{FP + TN}$$

What is 1-FPR? True Negative Rate

$$1 - FPR = 1 - P(\hat{y}_i = 1 | y_i = 0)$$

$$1 - FPR = P(\hat{y}_i = 0 | y_i = 0)$$

$$1 - FPR = \frac{TN}{FP + TN}$$

$$1 - FPR = TNR$$

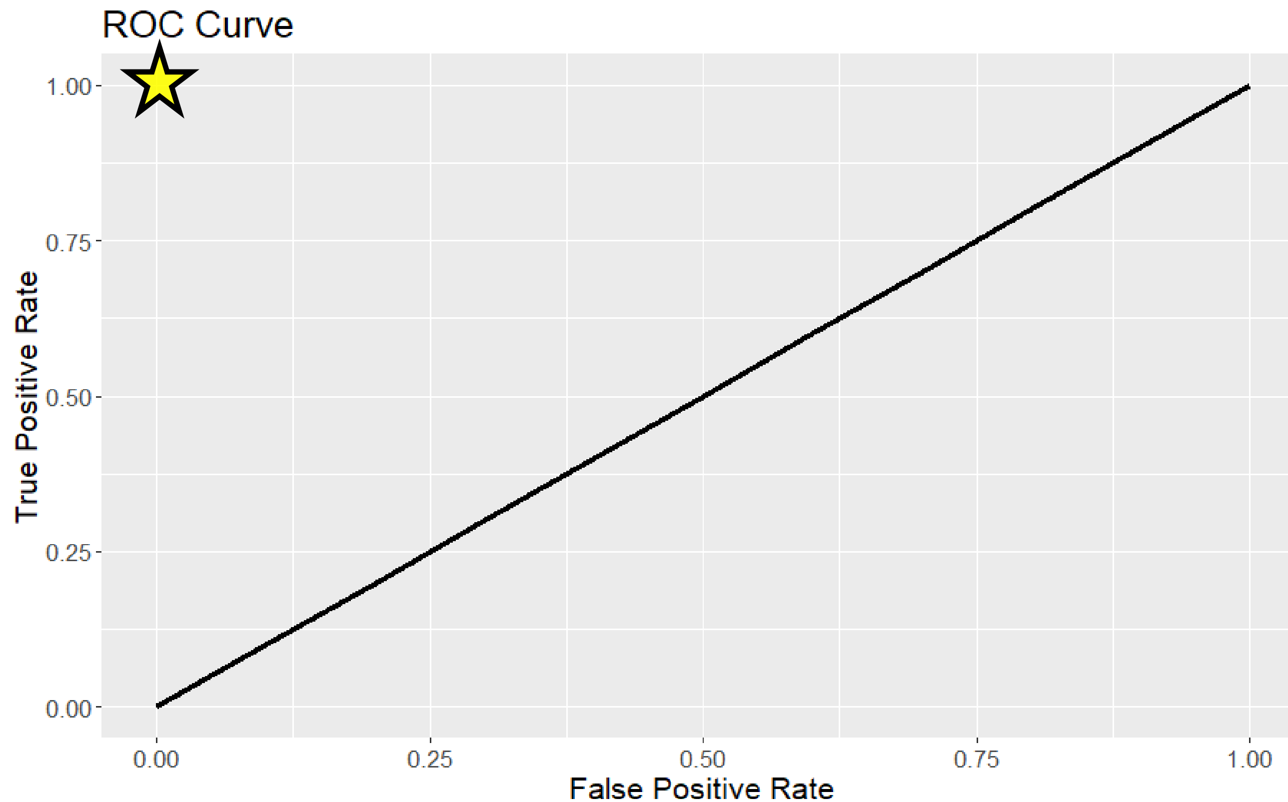
ROC Curve

Agenda

- Review of GLM
- Logistic Regression
    - Diagnostics
    - Residual
    - Example
    - Variable selection
    - Test set
    - Type of error
    - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

The ROC curve plots TPR vs. FPR

What's the ideal point?



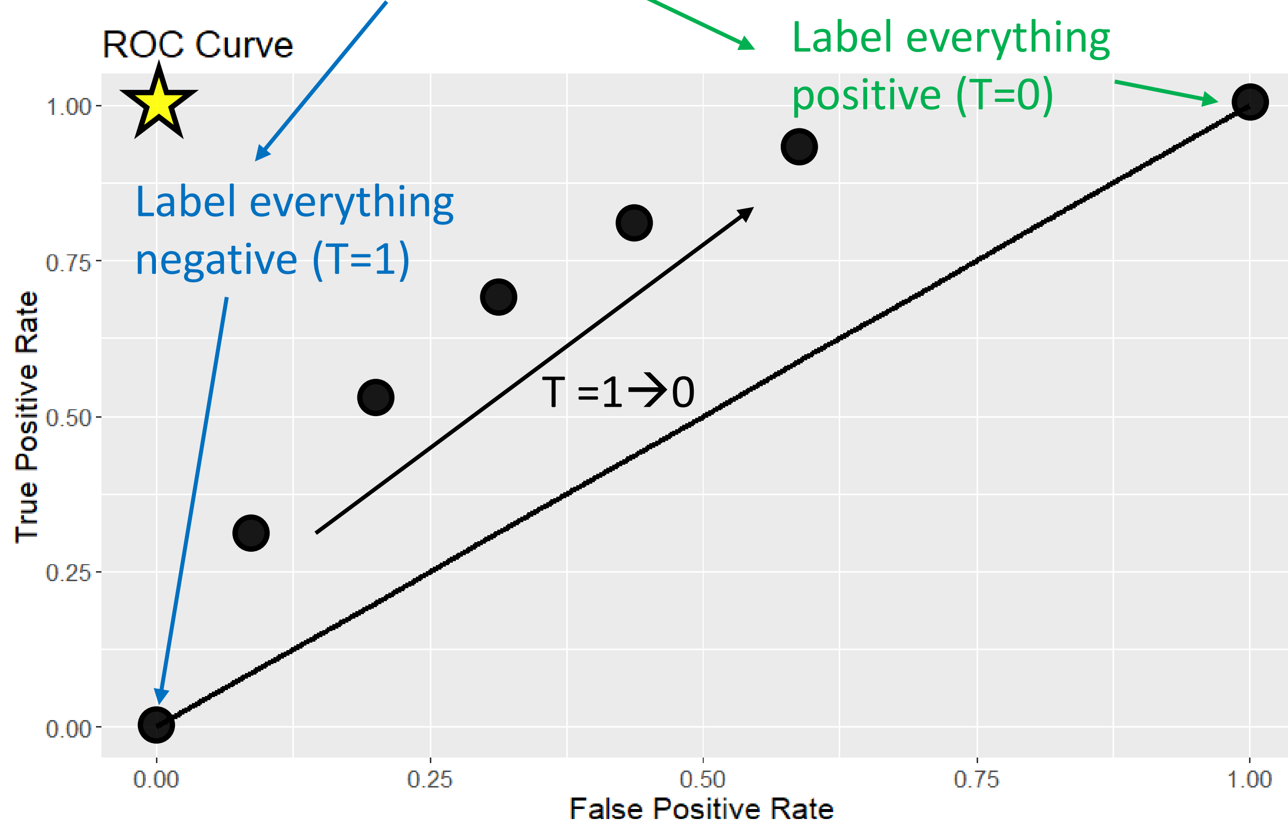
ROC Curve

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

The ROC curve plots TPR vs. FPR

How do we get FPR=0? TPR=1?

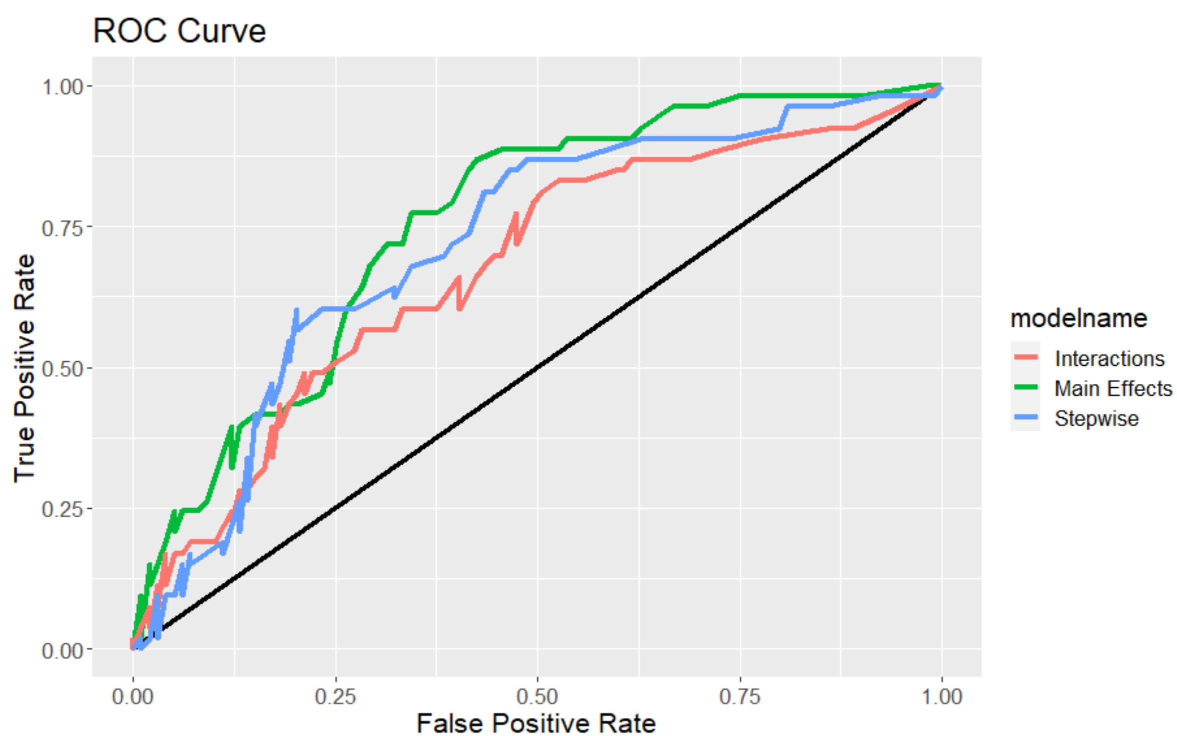


ROC curve for coronary heart disease models

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```
# make ROC curves
classColors = c("#f8766d","#00ba38","#619cff")
SAheart.roc.gg = lines.roc.gg(SAheart.roc.gg, step.glm.pred,
                              Heart$test$chd, "Stepwise")
SAheart.roc.gg + scale_color_manual(name="Model", values = classColors)
```



Principal Components Regression

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

Another way to select variables to try to improve classification is to combine them. This can also reduce multicollinearity. We can use PCs for this.

The first step is to perform PCA using all quantitative predictor variables.

We then select a subset of the PCs that explain p% of the variance. The modeler can use variable selection methods to decide on p.

The PCs then become the predictors instead of the variables themselves. While this decreases interpretability, we're more concerned about prediction than inference for classification.

Principal Components Regression

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

$$[U] = [X][E]$$

$$\begin{array}{c} \begin{array}{ccc} \overrightarrow{u_1} & \overrightarrow{u_2} & \cdots\ \overrightarrow{u_k} \end{array} \\ \boxed{\begin{bmatrix} u_{11} & u_{12} & \ldots\ u_{1m} \\ u_{21} & u_{22} & \ldots\ u_{2m} \\ u_{31} & u_{32} & \ldots\ u_{3m} \\ \vdots & \vdots & \ddots \quad \vdots \\ u_{n1} & u_{n2} & \ldots\ u_{nm} \end{bmatrix}} \end{array} = \begin{bmatrix} x_{11} & x_{12} & \ldots\ x_{1m} \\ x_{21} & x_{22} & \ldots\ x_{2m} \\ x_{31} & x_{32} & \ldots\ x_{3m} \\ \vdots & \vdots & \ddots \quad \vdots \\ x_{n1} & x_{n2} & \ldots\ x_{nm} \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} & \ldots\ e_{1m} \\ e_{21} & e_{22} & \ldots\ e_{2m} \\ e_{31} & e_{32} & \ldots\ e_{3m} \\ \vdots & \vdots & \ddots \quad \vdots \\ e_{m1} & e_{m2} & \ldots\ e_{mm} \end{bmatrix}$$

Instead of fitting the model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \varepsilon$$

We fit the model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_k u_k + \varepsilon$$

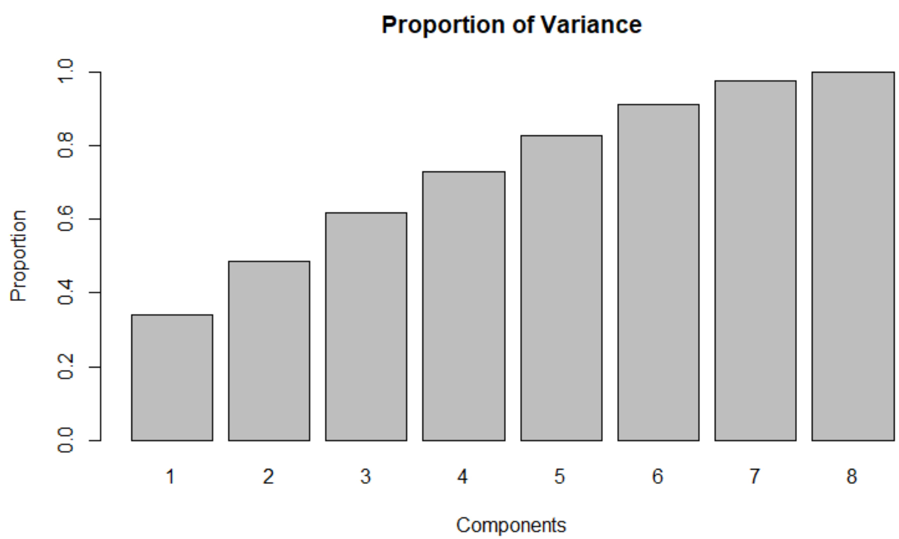
Where k PCs explain $p\%$ of the variance of [X].

Principal Components Regression

Agenda

- Review of GLM
- Logistic Regression
    - Diagnostics
    - Residual
    - Example
    - Variable selection
    - Test set
    - Type of error
    - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```r
# perform PC regression
source("pc.glm.R")
source("PCAplots.R")

# remove response variable (column 10) and categorical predictor (column 5)
heart.pca = princomp(Heart$train[,-c(5,10)], cor = T)

cumplot = cumplot(heart.pca)
```




```
> cumplot
   Component Proportion
1:    Comp.1  0.3410677
2:    Comp.2  0.4852510
3:    Comp.3  0.6189930
4:    Comp.4  0.7311980
5:    Comp.5  0.8270689
6:    Comp.6  0.9140724
7:    Comp.7  0.9754025
8:    Comp.8  1.0000000
```

Principal Components Regression

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

We test three logistic regression models using the PCs that explain:

1. 80% of the variability
2. 90% of the variability
3. 95% of the variability

```
# use number of PCs to explain 80% of the data (5 PCs)
heart.pc.glm80 <- pc.glm(heart.pca, 80, Heart$train$chd)

# use number of PCs to explain 90% of the data (6 PCs)
heart.pc.glm90 <- pc.glm(heart.pca, 90, Heart$train$chd)

# use number of PCs to explain 95% of the data (7 PCs)
heart.pc.glm95 <- pc.glm(heart.pca, 95, Heart$train$chd)

# predict on testing
heart.pc.glm80.pred = predict.pc.glm(heart.pc.glm80, heart.pca,
                                     Heart$test[,-c(5,10)])
heart.pc.glm90.pred = predict.pc.glm(heart.pc.glm90, heart.pca,
                                     Heart$test[,-c(5,10)])
heart.pc.glm95.pred = predict.pc.glm(heart.pc.glm95, heart.pca,
                                     Heart$test[,-c(5,10)])
```

Confusion Matrices

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

Main effects vs. three PC regression models

```
> score.table(main.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    77   15
     1    30   30
> score.table(heart.pc.glm80.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    79   13
     1    35   25
> score.table(heart.pc.glm90.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    78   14
     1    34   26
> score.table(heart.pc.glm95.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    73   19
     1    35   25
```

Using 80% of the PCs results in the lowest False Positives

Confusion Matrices

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

Main effects vs. three PC regression models

```
> score.table(main.glm.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    77   15
     1    30   30
> score.table(heart.pc.glm80.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    79   13
     1    35   25
> score.table(heart.pc.glm90.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    78   14
     1    34   26
> score.table(heart.pc.glm95.pred, Heart$test$chd, .5)
Actual vs. Predicted
      Pred
Actual FALSE TRUE
     0    73   19
     1    35   25
```

The main effects model has the lowest False Negatives

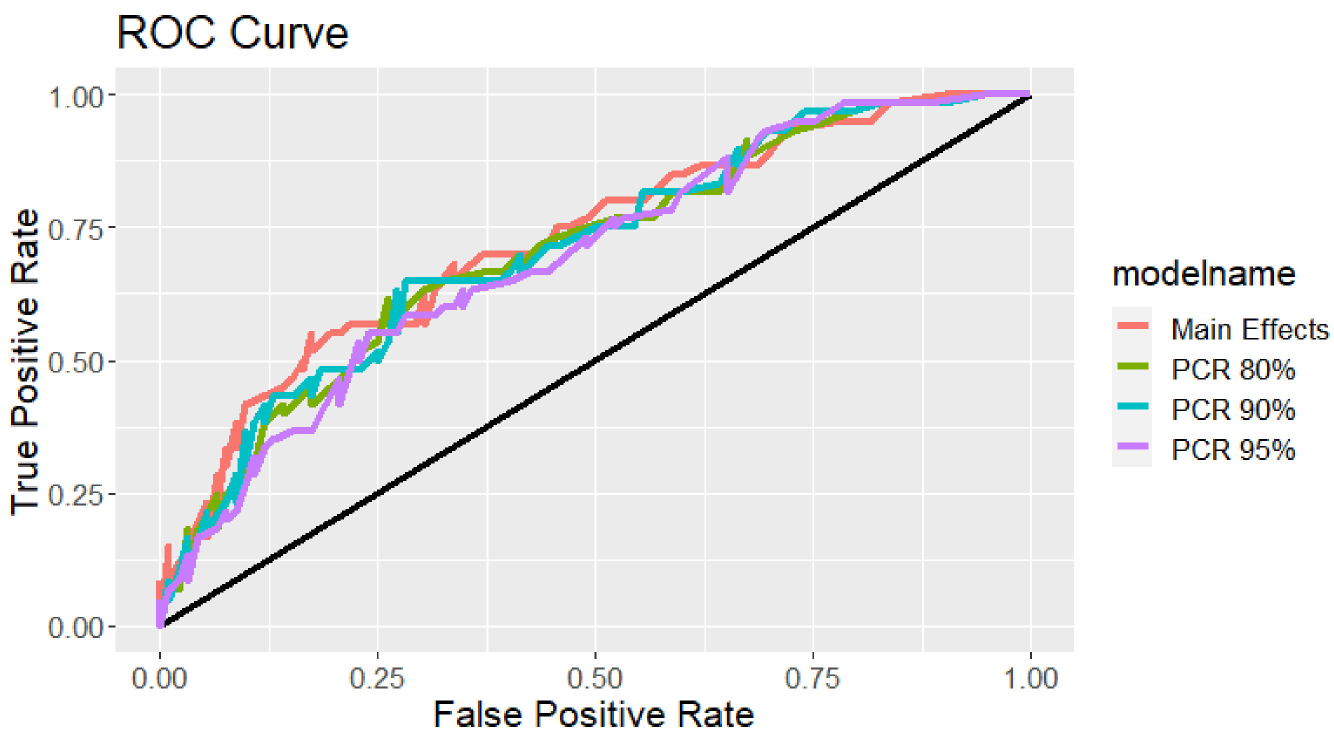
But the highest False Negatives

ROC Curves of PC regression models

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

```
# make ROC curves
plot.roc2.gg = plot.roc.gg(main.glm.pred, Heart$test$chd, "Main Effects")
plot.roc2.gg = lines.roc.gg(plot.roc2.gg, heart.pc.glm80.pred, Heart$test$chd,
                            "PCR 80%")
plot.roc2.gg = lines.roc.gg(plot.roc2.gg, heart.pc.glm90.pred, Heart$test$chd,
                            "PCR 90%")
plot.roc2.gg = lines.roc.gg(plot.roc2.gg, heart.pc.glm95.pred, Heart$test$chd,
                            "PCR 95%")
plot.roc2.gg
```



Summary

Agenda

- Review of GLM
- Logistic Regression
  - Diagnostics
  - Residual
  - Example
  - Variable selection
  - Test set
  - Type of error
  - Decision function
- Example Classification problem
- Threshold
- True/False negative/positive
- ROC curve
- Principal components regression
- Confusion matrix

1. Model diagnostics are less crucial for GLMs.
   a) We expect to see patterns and non-constant variance in the residuals vs. fitted plot.
   b) Pearson and deviance residuals should be normal.
   c) We don't want influential points.

2. Decision functions allow us to map probabilities to classifications to compute false positives and false negatives.

3. Receiver Operator Characteristic (ROC) curves allow us to compare models' errors across decision function thresholds.

4. Principal Components Regression (PCR) can be useful for prediction in logistic regression models.