Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Introduction

About This Lesson

Yes/No Questions

- How likely is it that users will like a new layout of our website?
- Will my customers leave my wireless service at the end of their subscription?
- What financial characteristics can be used to predict whether or not a business will go bankrupt?

➔ **Model the probability of ‘Yes’**

Linear Regression

Model: $\boldsymbol{Y}_i = \beta_0 + \beta_1 \boldsymbol{X}_{i,1} + \beta_2 \boldsymbol{X}_{i,2} + \cdots + \beta_p \boldsymbol{X}_{i,p} + \varepsilon_i \; i = 1, \cdots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent random variables
- *Normality Assumption*: $\varepsilon_i \sim \mathrm{Normal}$

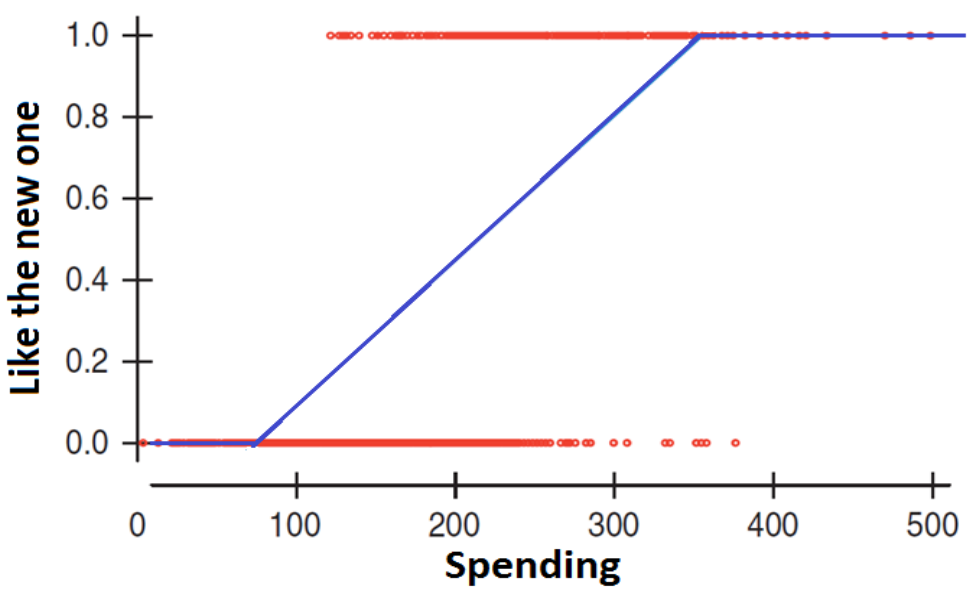
Georgia Tech

Linear Regression for Yes/No Question?

- Uber recently changed their logo.

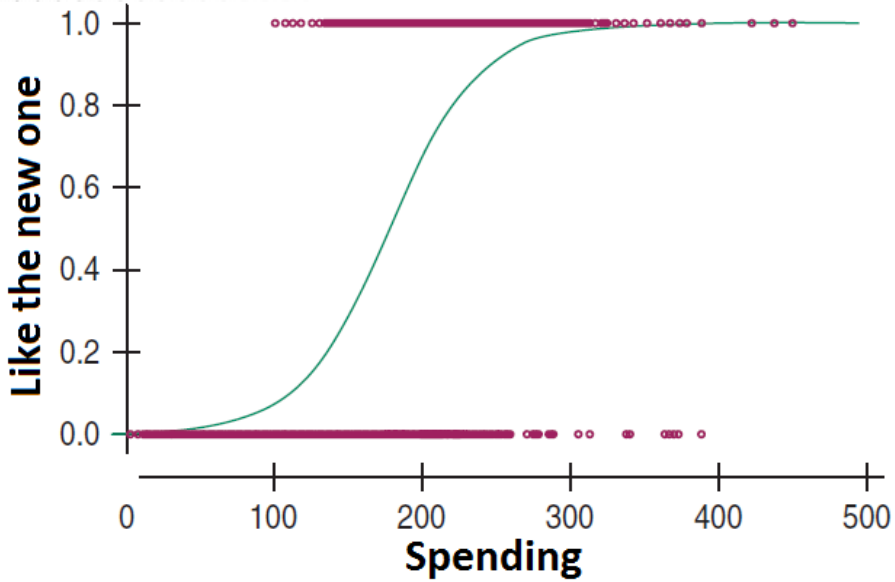
- You are asked to model whether Uber users will like the new logo based on how much they spent in the last 3 months using Uber.

What Is Wrong with Linear Regression?



Customers will not behave like this!

S-shaped Curve



Georgia Tech

Logistic Regression Model

Data: $\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Model: We model the *probability of success given the predictor(s)*

$$\mathrm{p} = \mathrm{p}(X_1, \cdots, X_p) = \Pr(Y = 1 \mid X_1, \cdots, X_p)$$

by linking p to the predicting variables through a nonlinear *link function* g:

$$\mathrm{g}(\mathrm{p}) = +\mathrm{g}\left(\mathrm{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

There is no error term!
What are the model assumptions?

Georgia Tech

Logistic Regression Model

Data: $\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Assumptions:

- *Linearity Assumption*: $\mathrm{g}\left(\mathrm{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- *Independence Assumption*: $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are independent random variables
- *Logit Link Function:*

$$\mathrm{g(p)} = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right)$$

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Data Example

About This Lesson

Data Example: Smoking

- Between 1972 and 1974, a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.
  - Among the information obtained originally was whether a person was a smoker or not.
- Twenty years later a follow-up study was conducted.
  - 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers!
Call Philip Morris, smoking leads to a longer life span!

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

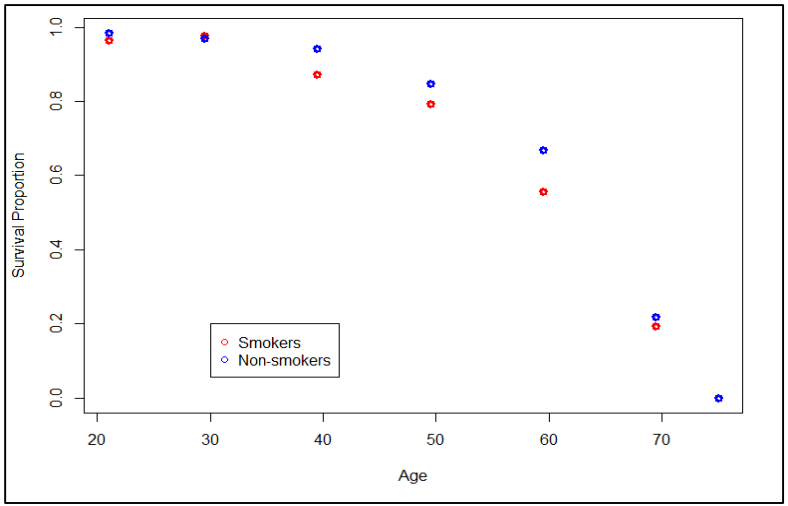
Georgia Tech

Data Example in R

```
## Read data in R
smoking <- read.table("CIGARETT.dat",sep="",row.names=NULL)
names(smoking) <- c("Age","Smoker","Survived","At.risk")
attach(smoking)
## Plot proportion of survival
plot(Age,Survived/At.risk, xlab="Age", ylab="Survival Proportion", col=c("red","blue"),lwd=3)
legend(30,0.2, legend=c("Smokers","Non-smokers"), pch=1, col=c("red","blue"))
```

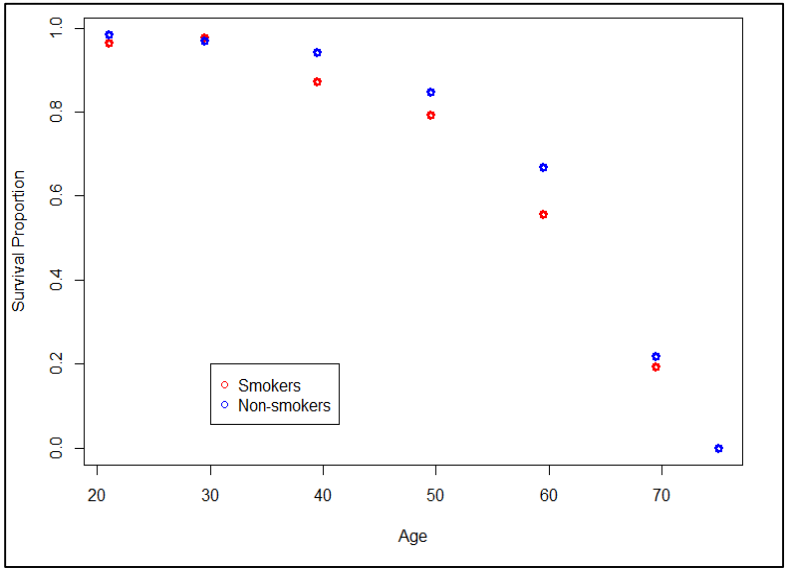
Data Example in R

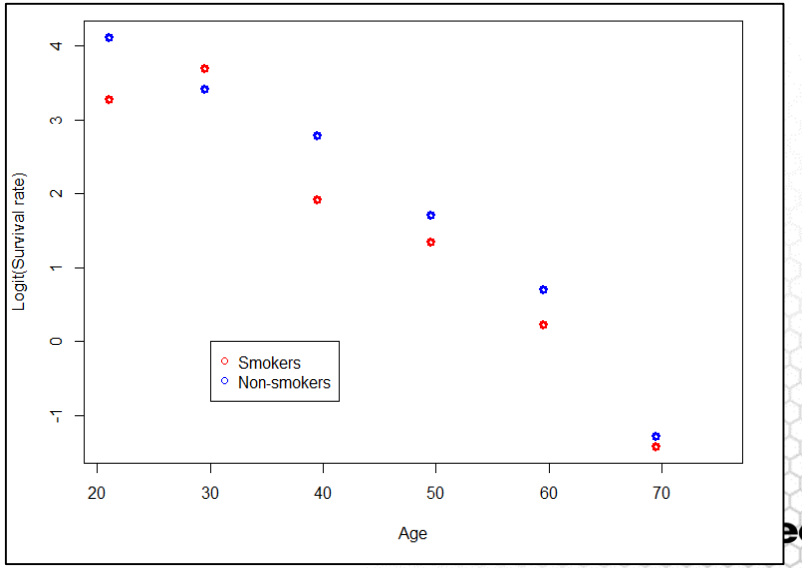
```
## Read data in R
smoking <- read.table("CIGARETT.dat",sep="",row.names=NULL)
names(smoking) <- c("Age","Smoker","Survived","At.risk")
attach(smoking)
## Plot proportion of survival
plot(Age,Survived/At.risk, xlab="Age", ylab="Survival Proportion", col=c("red","blue"),lwd=3)
legend(30,0.2, legend=c("Smokers","Non-smokers"),pch=1, col=c("red","blue"))
```



Data Example in R (cont'd)

```
## Plot of logit transformation of the proportion survival
prop.survival <-Survived/At.risk
plot(Age,log(prop.survival/(1-prop.survival)), col=c("red","blue"), xlab="Age", ylab="Logit(Survival Proportion)", lwd=3)
legend(30,0, legend=c("Smokers","Non-smokers"), pch=1, col=c("red","blue"))
```





Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Model Description and Estimation

Georgia Tech

About This Lesson

Logistic Regression Model

Data: $\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Model: We model the *probability of success given the predictor(s)*

$$\mathrm{p} = \mathrm{p}(X_1, \cdots, X_p) = \Pr(Y = 1 \mid X_1, \cdots, X_p)$$

by linking p to the predicting variables through the ***logit*** *link function* g:

$$\mathrm{g}(\mathrm{p}) = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

OR

$$\mathrm{p}(X_1, \cdots, X_p) = \frac{\mathrm{e}^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + \mathrm{e}^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Georgia Tech

Model Interpretation

- The probability of success given one predicting variable $X = x$ is
$$\mathrm{p} = \mathrm{p}(x) = \Pr(Y = 1 \mid x)$$
- The logit function $\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1 x$ is the ***log odds*** function.
- The exponential of the logit function $\frac{\mathrm{p}(X)}{1-\mathrm{p}(X)} = \mathrm{e}^{\beta_0+\beta_1 x}$ is the ***odds*** of $Y = 1$ at $X = x$
- The odds at $X = a$ versus $X = b$ is equal to the ***odds ratio***:
$$\frac{\mathrm{e}^{\beta_0+\beta_1 a}}{\mathrm{e}^{\beta_0+\beta_1 b}} = \mathrm{e}^{\beta_1(a-b)}$$

Georgia Tech

Model Interpretation

If we calculate the odds ratio of the odds at $X = b + 1$ versus $X = b$, we have

$$\frac{e^{\beta_0+\beta_1(b+1)}}{e^{\beta_0+\beta_1 b}} = e^{\beta_1}$$

- The regression coefficient β_1 can be interpreted as the log of the odds ratio for an increase of one unit in the predicting variable.
- If X a dummy variable of a categorical factor, interpret as the log of odds ratio of one category versus baseline.
- Interpret β with respect to the odds of success, not directly with respect to the response variable.

Model Estimation

Model the probability of success given predictor(s):

$$\text{Logit}\Big(\Pr(Y=1 \mid X_1,\cdots,X_p)\Big) = \text{Logit}\Big(\text{p}(X_1,\cdots,X_p)\Big) = \beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

Parameters: $\beta_0,\beta_1,\cdots,\beta_p$

Approach: Maximum Likelihood Estimation

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \mathcal{L}(\beta_0,\beta_1,\cdots,\beta_p) = \prod_{i=1}^{n} \text{p}\big(\boldsymbol{X}_{i,1},\boldsymbol{X}_{i,2},\cdots,\boldsymbol{X}_{i,p}\big)^{\boldsymbol{Y}_i}\Big(1-\text{p}\big(\boldsymbol{X}_{i,1},\boldsymbol{X}_{i,2},\cdots,\boldsymbol{X}_{i,p}\big)\Big)^{1-\boldsymbol{Y}_i}$$

or

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \ell\big(\beta_0,\beta_1,\cdots,\beta_p\big) = \max_{\beta_0,\beta_1,\cdots,\beta_p} \log\big(\mathcal{L}(\beta_0,\beta_1,\cdots,\beta_p)\big)$$

$$= \max_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n}\left(\boldsymbol{Y}_i\log\left(\frac{\text{e}^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+\text{e}^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}\right)+(1-\boldsymbol{Y}_i)\log\left(\frac{1}{1+\text{e}^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}\right)\right)$$

Georgia Tech

Model Estimation (cont'd)

Approach: Maximum Likelihood Estimation

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} \left(\boldsymbol{Y}_i \log\left(\frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \right) + (1-\boldsymbol{Y}_i)\log\left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \right) \right)$$

- Maximizing the (log-)likelihood function with respect to $\beta_0, \beta_1, \cdots, \beta_p$ in closed form expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters.

- Use numerical algorithm to estimate $\beta_0, \beta_1, \cdots, \beta_p \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$

Upshot: The estimated parameters and their standard errors are approximate estimates. Do not attempt to do it yourself! Use statistical software to derive the estimated regression coefficients.

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Model Estimation: Data Example

Georgia Tech

About This Lesson

Data Example: Smoking

- Between 1972 and 1974, a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.
  - Among the information obtained originally was whether a person was a smoker or not.
- Twenty years later a follow-up study was conducted.
  - 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers!
Call Philip Morris, smoking leads to a longer life span!

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

Data Example in R

Data: $\boldsymbol{Y_i}$ binary responses ~ Binomial($\mathbf{p}_i$, $\boldsymbol{n_i}$)

- $\boldsymbol{Y_i}$ number of people at risk who survived (Survived)
- $\boldsymbol{n_i}$ number of people at risk (At.risk)

Fit a logistic regression model

smoke1 = glm(Survived/At.risk ~ Smoker, weights=At.risk, family=binomial)
summary(smoke1)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.78052	0.07962	9.803	< 2e-16 ***
Smoker	0.37858	0.12566	3.013	0.00259 **

$\widehat{\boldsymbol{\beta}}_{smoker} = \mathbf{0.378}$: The log odds of survival increases by 0.378 for smokers versus non-smokers OR the odds of survival are 46% higher for smokers than for non-smokers (the odds ratio is 1.459).

Georgia Tech

Data Example in R

Fit a logistic regression model

smoke2 = glm(Survived/At.risk ~ Smoker + Age, weights=At.risk, family=binomial)
summary(smoke2)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.785001	0.454999	17.110	<2e-16 ***
Smoker	-0.240831	0.167885	-1.435	0.151
Age	-0.127419	0.007397	-17.227	<2e-16 ***

$\widehat{\boldsymbol{\beta}}_{smoker} = -\mathbf{0.24}$ The odds of survival is 27.2% higher for non-smokers than for smokers (odds ratio for non-smokers versus smokers is 1/exp(-0.24) =1.272).

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Statistical Inference

Georgia Tech

About This Lesson

Model Estimation

Model the probability of success given predictor(s):

$$\text{Logit}\Big(\Pr\big(Y=1 \mid X_1,\cdots,X_p\big)\Big) = \text{Logit}\Big(\text{p}\big(X_1,\cdots,X_p\big)\Big) = \beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

Parameters: $\beta_0,\beta_1,\cdots,\beta_p$

Approach: Maximum Likelihood Estimation

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \mathcal{L}(\beta_0,\beta_1,\cdots,\beta_p) = \prod_{i=1}^{n} \text{p}\big(\boldsymbol{X}_{i,1},\boldsymbol{X}_{i,2},\cdots,\boldsymbol{X}_{i,p}\big)^{\boldsymbol{Y}_i}\Big(1-\text{p}\big(\boldsymbol{X}_{i,1},\boldsymbol{X}_{i,2},\cdots,\boldsymbol{X}_{i,p}\big)\Big)^{1-\boldsymbol{Y}_i}$$

or

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \ell\big(\beta_0,\beta_1,\cdots,\beta_p\big) = \max_{\beta_0,\beta_1,\cdots,\beta_p} \log\big(\mathcal{L}(\beta_0,\beta_1,\cdots,\beta_p)\big)$$

$$= \max_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n}\left(\boldsymbol{Y}_i\log\left(\frac{\text{e}^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+\text{e}^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}\right)+(1-\boldsymbol{Y}_i)\log\left(\frac{1}{1+\text{e}^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}\right)\right)$$

Georgia Tech

Statistical Inference

Maximum Likelihood Estimators (MLEs):

$$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$$

Statistical Properties of MLEs:

- Approximate Sampling Distribution: $\widehat{\boldsymbol{\beta}} \approx \boldsymbol{N}(\boldsymbol{\beta}, \boldsymbol{V})$
- The normal approximation relies on the assumption of large sample size
- Statistical inference is not reliable for small sample data

1-α Approximate Confidence interval

$$\hat{\beta}_j \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\beta}_j)}$$

Georgia Tech

Statistical Inference (cont'd)

- Hypothesis testing and Confidence Intervals rely on the approximately normal distribution of large sample sizes
- Use the z-test (**Wald test**)
  - Test is for the statistical significance of $\hat{\beta}_j$ *given all other predicting variables in the model*
  - Null hypothesis is that β_j is not significant

    $$\mathrm{H}_0: \beta_j = 0 \text{ vs. } \mathrm{H}_\mathrm{a}: \beta_j \neq 0$$

  - $$\mathrm{z\text{-}value} = \frac{\hat{\beta}_j - 0}{\mathrm{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$$
  - Reject H_0 if $|\mathrm{z\text{-}value}|$ is too large
    - Implies that β_j is statistically significant

Georgia Tech

Statistical Inference (cont'd)

For $\mathbf{H_0}$: $\boldsymbol{\beta_j = b}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j \neq b}$ (to test if the coefficient equals constant b)

- $\mathrm{z{-}value} = \dfrac{\widehat{\beta}_j - b}{\mathrm{se}(\widehat{\beta}_j)}$
- Reject H_0 if $|\mathrm{z{-}value}| > \mathrm{z}_{\alpha/2}$ for significance level α
- Alternatively, compute P-value
  - P−value = $2\Pr(Z > |\mathrm{z{-}value}|)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j \leq 0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j} > 0$ (to test for a significantly positive coefficient)

- P−value = $\Pr(Z > z - value)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j \geq 0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j < 0}$ (to test for a significantly negative coefficient)

- P−value = $\Pr(Z < z - value)$

Georgia Tech

Statistical Inference (cont'd)

For $\mathbf{H_0}$: $\boldsymbol{\beta_j = b}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j \neq b}$ (to test if the coefficient equals constant b)

- $\mathrm{z{-}value} = \dfrac{\widehat{\beta}_j - b}{\mathrm{se}(\widehat{\beta}_j)}$
- Reject H_0 if $|\mathrm{z{-}value}| > \mathrm{z}_{\alpha/2}$ for significance level α
- Alternatively, compute P-value
  - P−value = $2\Pr(Z > |\mathrm{z{-}value}|)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j \leq 0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j > 0}$ (to test for a significantly positive coefficient)

  - P−value = $\Pr(Z > z - value)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j \geq 0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j < 0}$ (to test for a significantly negative coefficient)

  - P−value = $\Pr(Z < z - value)$

- Because the approximation of the normal distribution relies on large sample size, so do the hypothesis testing procedures.
- What if *n* is small?
  - The hypothesis testing procedure will have a probability of type I error larger than the significance level.
  - In other words, there will likely be more type I errors than expected.

Georgia Tech

Testing for Subsets of Coefficients

Full model:

$$\text{Logit}\left(\text{p}\left(X_1, \cdots, X_p, Z_1, \cdots, Z_q\right)\right)$$
$$= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q$$

Reduced model:

$$\text{Logit}\left(\text{p}\left(X_1, \cdots, X_p\right)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The hypothesis test:

$$\mathbf{H_0}: \boldsymbol{\alpha_1 = \alpha_2 = \cdots = \alpha_q = 0}$$

vs.

$$\mathbf{H_a}: \boldsymbol{\alpha_i \neq 0 \text{ for at least one } \alpha_i, i = 1, \cdots, q}$$

- Maximize the likelihood function under reduced model: $\mathcal{L}\,(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)\right) - \log\left(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)\right) \approx \chi^2_q$
  - P-value = $\Pr(\chi^2_q > \text{Deviance})$

Georgia Tech

Testing for Subsets of Coefficients

Full model:

$$\text{Logit}\left(\text{p}(X_1, \cdots, X_p, Z_1, \cdots, Z_q)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q$$

Reduced model:

$$\text{Logit}\left(\text{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The hypothesis test:

$$\mathbf{H_0}: \boldsymbol{\alpha_1 = \alpha_2 = \cdots = \alpha_q = 0}$$

vs.

$$\mathbf{H_a}: \boldsymbol{\alpha_i \neq 0 \text{ for at least one } \alpha_i, i = 1, \cdots, q}$$

- The hypothesis test for subsets of coefficients is approximate
  - It relies on large sample size
- This is not a test for goodness of fit!
  - It only compares two models

- Maximize the likelihood function under reduced model: $\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p))$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)\right) - \log\left(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)\right) \approx \chi^2_q$
  - P-value = $\Pr(\chi^2_q > \text{Deviance})$

Georgia Tech

Testing for Overall Regression

Full model:

$$\text{Logit}\left(\text{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Reduced model:

$$\text{Logit}\left(\text{p}(X_1, \cdots, X_p)\right) = \beta_0$$

The hypothesis test:

$$\mathbf{H_0}: \boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_p = 0}$$

vs.

$$\mathbf{H_a}: \boldsymbol{\beta_i \neq 0 \text{ for at least one } \beta_i, i = 1, \cdots, p}$$

- Maximize the likelihood function under reduced model: $\mathcal{L}(\bar{\beta}_0)$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0)\right) - \log\left(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)\right) \approx \chi^2_p$
  - P-value = $\Pr(\chi^2_p > \text{Deviance})$

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Statistical Inference:
Data Example

Georgia Tech

About This Lesson

Data Example: Smoking

- Between 1972 and 1974, a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.
  - Among the information obtained originally was whether a person was a smoker or not.
- Twenty years later a follow-up study was conducted.
  - 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers!
Call Philip Morris, smoking leads to a longer life span!

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

Data Example in R

```
## Fit a logistic regression model
smoke1 = glm(Survived/At.risk ~ Smoker, weights=At.risk, family=binomial)
summary(smoke1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.78052	0.07962	9.803	< 2e-16 ***
Smoker	0.37858	0.12566	3.013	0.00259 **

Null deviance: 641.5 on 13 degrees of freedom
Residual deviance: 632.3 on 12 degrees of freedom

```
1 - pchisq(smoke1$null.deviance-smoke1$deviance, 1)
[1] 0.002419817
```

Test for significance: β_{smoker} P-value = **0.0025**, thus statistically significant
Test for overall regression: Null deviance - Residual Deviance = 9.2
P-value = $\Pr(\chi_1^2 > 9.2)$ = **0.0024**

Georgia Tech

Data Example in R (cont'd)

Fit a logistic regression model

smoke2 = glm(Survived/At.risk ~ Smoker + Age, weights=At.risk, family=binomial)
summary(smoke2)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.785001	0.454999	17.110	<2e-16 ***
Smoker	-0.240831	0.167885	-1.435	0.151
Age	-0.127419	0.007397	-17.227	<2e-16 ***

Null deviance: 641.496 on 13 degrees of freedom
Residual deviance: 43.459 on 11 degrees of freedom

Test for significance: β_{smoker} P-value = **0.151**, not statistically significant
Test for significance: β_{age} P-value $\approx$ **0**, statistically significant

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment

Georgia Tech

About This Lesson

Logistic Regression Model

Data:

$$\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$$

where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Assumptions:

- *Linearity Assumption*: $\mathrm{g}\left(\mathrm{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- *Independence Assumption*: $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are independent random variables
- *Logit Link Function:* $\mathrm{g}(\mathrm{p}) = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right)$

There is no error term!
How to check the model assumptions?

Georgia Tech

Residuals in Logistic Regression

Data:

$$\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$$

where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

- **Logistic Regression Without Replications**
  - One separate (possibly non-unique) set of predictors $(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p})$ for each individual observation $\boldsymbol{Y}_i$ (i=1,..,n)
  - $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim$ Bernoulli$\left(\mathrm{p}(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p})\right)$ or

    $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim$ Binomial$\left(1, \mathrm{p}(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p})\right)$
- **Logistic Regression With Replications**
  - Observe n_i repeated responses $\boldsymbol{Y}_i$ for each unique set of predictors $(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p})$ across all i=1,..,n
  - $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim$ Binomial$\left(n_i, \mathrm{p}(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p})\right), n_i > 1$

Georgia Tech

Residuals in Logistic Regression

Logistic Regression With Replications

- $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim \text{Binomial}\left(n_i, \text{p}\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right)\right),\ n_i > 1$
- Estimated probabilities are:

$$\hat{\text{p}}_i = \hat{\text{p}}_i\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right) = \frac{\text{e}^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}{1 + \text{e}^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}$$

- Pearson residuals:

$$r_i = \frac{Y_i - n_i\hat{\text{p}}_i}{\sqrt{n_i\hat{\text{p}}_i(1-\hat{\text{p}}_i)}}$$

- Deviance residuals:

$$\text{d}_i = \text{sign}\left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}_i\right)\sqrt{2\boldsymbol{Y}_i\log\left(\boldsymbol{Y}_i/\widehat{\boldsymbol{Y}}_i\right) + 2(n_i - \boldsymbol{Y}_i)\log\left((n_i - \boldsymbol{Y}_i)/\left(n_i - \widehat{\boldsymbol{Y}}_i\right)\right)}$$

Georgia Tech

Residuals in Logistic Regression

Logistic Regression With Replications

- $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim \text{Binomial}\left(n_i, \text{p}\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right)\right),\ n_i > 1$
- Estimated probabilities are:

$$\hat{\text{p}}_i = \hat{\text{p}}_i\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right) = \frac{\text{e}^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}{1 + \text{e}^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}$$

- Pearson residuals:

$$r_i = \frac{Y_i - n_i\hat{\text{p}}_i}{\sqrt{n_i\hat{\text{p}}_i(1-\hat{\text{p}}_i)}}$$

- Deviance residuals:

$$\text{d}_i = \text{sign}\left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}_i\right)\sqrt{2\boldsymbol{Y}_i\log\left(\boldsymbol{Y}_i/\widehat{\boldsymbol{Y}}_i\right) + 2(n_i - \boldsymbol{Y}_i)\log\left((n_i - \boldsymbol{Y}_i)/\left(n_i - \widehat{\boldsymbol{Y}}_i\right)\right)}$$

- Pearson's residuals follow directly a normal approximation to a binomial, hence approximately *N(0,1).*
- The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model vs. the fitted model, thus approximately *N(0,1)* if the model is a good fit.

Georgia Tech

Model Goodness of Fit

GOF Visual Analytics

- Normal probability plot & histogram of the residuals

Hypothesis Testing Procedure

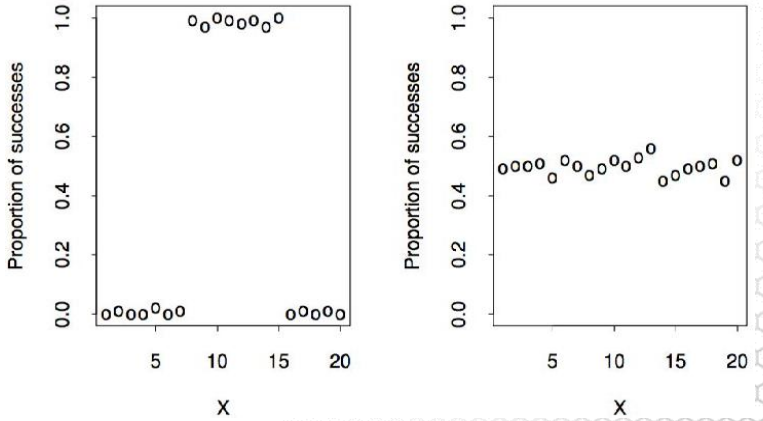
- $\mathbf{H_0}$: **the logistic model fits the data** vs.
  $\mathbf{H_a}$: **the logistic model does not fit the data**
- Deviance test statistic: $\mathrm{D} = \sum_{i=1}^{n} d_i^2$
  - Under $\mathbf{H_0}$, $\mathrm{D} \sim \chi^2_{n-p-1}$
- Reject $\mathbf{H_0}$ if $\mathrm{P-value} = \Pr(\chi^2_{df} > \mathrm{D})$ is small
- For this test we want large p-values!!!!

Georgia Tech

Goodness of Fit vs. Predictive Power

- **Goodness of fit**: Model assumptions hold
  - For example, does the S-shape logit function fit the data?
- **Predictive Power**: The predicting variables predict the data
  - Even if the one or more assumptions do not hold

While the logistic model is a sensible one for probabilities, it is not necessarily appropriate for any particular data set. That does not mean that the predicting variables are not good predictors of the probability of success.



Georgia Tech

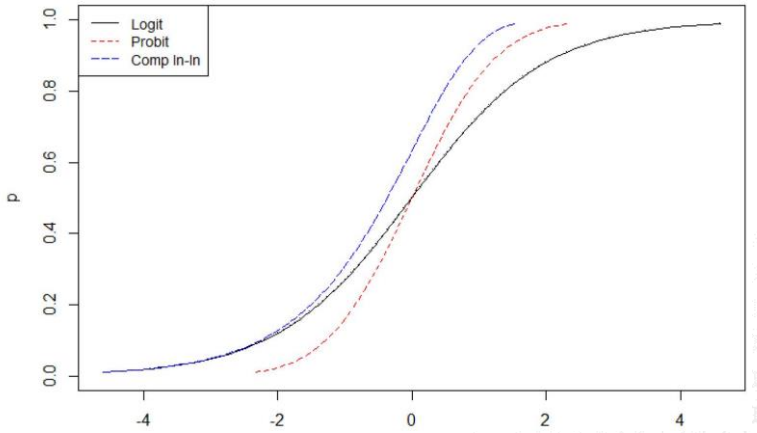
What If No Goodness of Fit?

- Add predicting variables
- Transform predicting variables to improve linearity
- Identify unusual observations (outliers, leverage points)
- The binomial distribution may not be appropriate
  - Overdispersion: The variability of the probability estimates is larger than would be implied by a binomial random variable
    - Correlation in the observed responses
    - Heterogeneity in the success probabilities that hasn't been modeled
- Logit function may not fit data
  - Other S-shape functions
    - probit, complementary log-log (c-log-log)

Georgia Tech

What If No Goodness of Fit?

- Add predicting variables
- Transform predicting variables to improve linearity
- Identify unusual observations (outliers, leverage points)
- The binomial distribution may not be appropriate
  - Overdispersion: The variability of the probability estimates is larger than would be implied by a binomial random variable
    - Correlation in the observed responses
    - Heterogeneity in the success probabilities that hasn't been modeled
- Logit function may not fit data
  - Other S-shape functions
    - probit, complementary log-log (c-log-log)



What If No Goodness of Fit?

- Add predicting variables
- Transform predicting variables to improve linearity
- Identify unusual observations (outliers, leverage points)
- The binomial distribution may not be appropriate
  - Overdispersion: The variability of the probability estimates is larger than would be implied by a binomial random variable
    - Correlation in the observed responses
    - Heterogeneity in the success probabilities that hasn't been modeled
- Logit function may not fit data
  - Other S-shape functions
    - probit, complementary log-log (c-log-log)

Why logistic regression?

- Logit link function is the canonical link function
- Ease of interpretation of the regression coefficients

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment: Data Examples

Georgia Tech

About This Lesson

Data Example: Smoking

- Between 1972 and 1974, a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.
  - Among the information obtained originally was whether a person was a smoker or not.
- Twenty years later a follow-up study was conducted.
  - 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers!
Call Philip Morris, smoking leads to a longer life span!

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

Georgia Tech

GOF Hypothesis Test

Deviance Test for GOF using deviance residuals

c(deviance(smoke2), 1-pchisq(deviance(smoke2),11))

[1] 4.345918e+01 9.033325e-06

Test for goodness-of-fit:

- Using deviance residuals: P-value ≈ 0
- Reject the null hypothesis of good fit (thus NOT a good fit)

GOF test using Pearson residuals

pearres2 = residuals(smoke2,type="pearson")

pearson.tvalue = sum(pearres2^2)

c(pearson.tvalue, 1-pchisq(pearson.tvalue,11))

[1] 36.751889370 0.000126796

Test for goodness-of-fit:

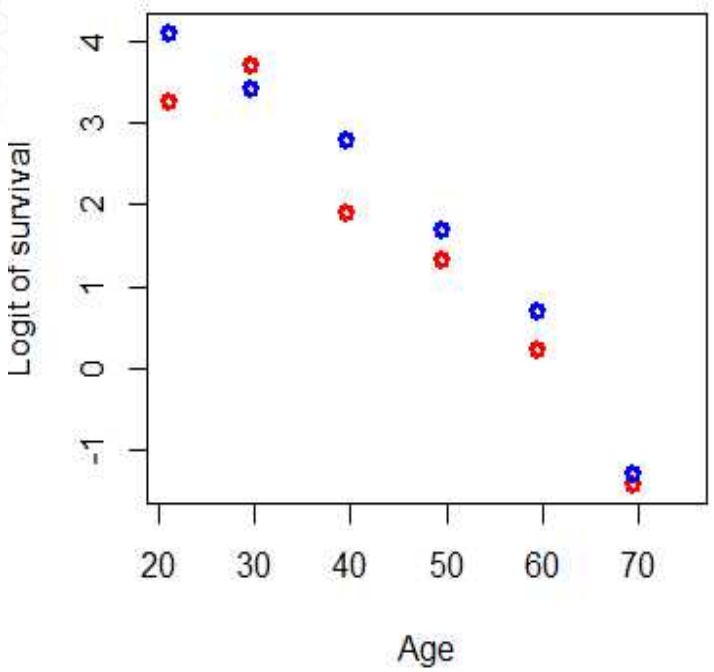
- Using Pearson residual: P-value ≈ 0.0001
- Reject the null hypothesis of good fit (thus NOT a good fit)

Georgia Tech

Linearity Assumption

Is it a linear fit?

plot(Age,log((Survived/At.risk)/(1-Survived/At.risk)), ylab="Logit of survival", main="Scatterplot of logit survival rate vs age", col=c("red","blue"), lwd=3)



The relationship between the logit of survival and age is more quadratic than linear.

Georgia Tech

Improve the Fit

Fit a logistic regression model

*Age.squared = Age*Age*

smoke3 = glm(Survived/At.risk ~ Smoker + Age + Age.squared, weights=At.risk, family=binomial)

summary(smoke3)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5190783	1.0248206	2.458	0.0140 *
Smoker	-0.4284561	0.1770581	-2.420	0.0155 *
Age	0.0951102	0.0430095	2.211	0.0270 *
Age.squared	-0.0021673	0.0004309	-5.030	4.91e-07 ***

Null deviance: 641.496 on 13 degrees of freedom

Residual deviance: 19.808 on 10 degrees of freedom

Test for significance: β_{smoker} P-value ≈ 0.015, statistically significant at 0.05

Test for significance: $\beta_{\text{Age.squared}}$ P-value ≈ 0, statistically significant

Georgia Tech

GOF Test for Improved Model

```
## Test for goodness of fit
round(c(deviance(smoke3), 1-pchisq(deviance(smoke3),10)),2)
[1] 19.81  0.03

pearres3 = residuals(smoke3,type="pearson")
pearson = sum(pearres3^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
[1] 14.79  0.14
```

Does the goodness of fit improve?

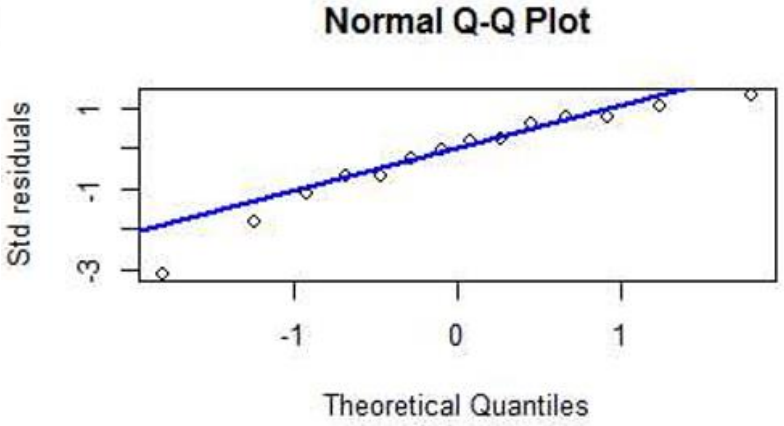
- Using deviance residuals: P-value = 0.03
- Using Pearson residual: P-value = 0.14
- Do not reject the null hypothesis of good fit using Pearson residuals, but do reject using Deviance residuals at the significance level 0.03 or higher.

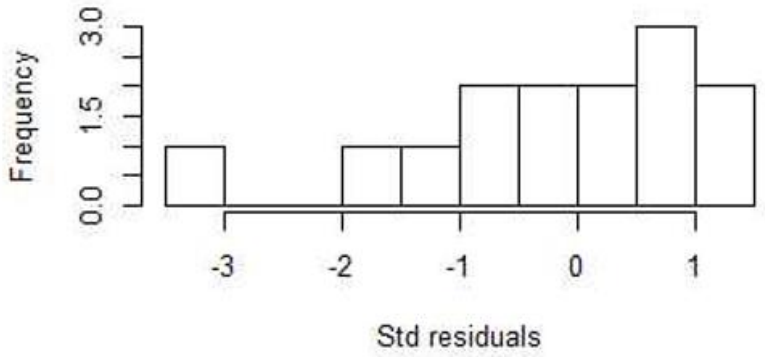
Georgia Tech

Residual Analysis

Residual Plots

res = resid(smoke3,type="deviance")
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")





Georgia Tech

Higher Order Nonlinearity

Fit a logistic regression model with Age as a factor

smoke4 = glm(Survived/At.risk ~ Smoker + factor(Age), weights=At.risk, family=binomial)

summary(smoke4)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8601	0.5939	6.500	8.05e-11 ***
Smoker	-0.4274	0.1770	-2.414	0.015762 *
factor(Age)29.5	-0.1201	0.6865	-0.175	0.861178
factor(Age)39.5	-1.3411	0.6286	-2.134	0.032874 *
factor(Age)49.5	-2.1134	0.6121	-3.453	0.000555 ***
factor(Age)59.5	-3.1808	0.6006	-5.296	1.18e-07 ***
factor(Age)69.5	-5.0880	0.6195	-8.213	< 2e-16 ***
factor(Age)75	-27.8073	11293.1437	-0.002	0.998035

Null deviance: 641.4963 on 13 degrees of freedom

Residual deviance: 2.3809 on 6 degrees of freedom

Georgia Tech

Higher Order Nonlinearity

Fit a logistic regression model with Age as a factor

smoke4 = glm(Survived/At.risk ~ Smoker + factor(Age), weights=At.risk, family=binomial)

summary(smoke4)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8601	0.5939	6.500	8.05e-11 ***
Smoker	-0.4274	0.1770	-2.414	0.015762 *
factor(Age)29.5	-0.1201	0.6865	-0.175	0.861178
factor(Age)39.5	-1.3411	0.6286	-2.134	0.032874 *
factor(Age)49.5	-2.1134	0.6121	-3.453	0.000555 ***
factor(Age)59.5	-3.1808	0.6006	-5.296	1.18e-07 ***
factor(Age)69.5	-5.0880	0.6195	-8.213	< 2e-16 ***
factor(Age)75	-27.8073	11293.1437	-0.002	0.998035

Null deviance: 641.4963 on 13 degrees of freedom

Residual deviance: 2.3809 on 6 degrees of freedom

Test for significance: β_{smoker} P-value ≈ 0.015, statistically significant at 0.05

Test for significance: Not all regression coefficients for the dummy variables for age are statistically significant.

Georgia Tech

Higher Order Nonlinearity: GOF

```
## Test for goodness of fit
round(c(deviance(smoke4), 1-pchisq(deviance(smoke4),6)),2)
[1] 2.38 0.88

pearres4 = residuals(smoke4,type="pearson")
pearson = sum(pearres4^2)
round(c(pearson, 1-pchisq(pearson,6)),2)
[1] 2.37 0.88
```

Does the goodness of fit improve?

- Using deviance residuals: P-value = 0.88
- Using Pearson residual: P-value = 0.88
- Do not reject the null hypothesis of good fit using either Pearson residuals or Deviance residuals.

Georgia Tech

Different Link Function

Use probit link function

smoke5 = glm(Survived/At.risk ~ Smoker + Age + Age.squared, weights=At.risk, family=binomial(link = probit))

summary(smoke5)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1033963	0.4904877	2.250	0.02447 *
Smoker	-0.2277451	0.0970191	-2.347	0.01890 *
Age	0.0681279	0.0213095	3.197	0.00139 **
Age.squared	-0.0013767	0.0002173	-6.335	2.37e-10 ***

Null deviance: 641.496 on 13 degrees of freedom
Residual deviance: 18.233 on 10 degrees of freedom

Test for significance: β_{smoker} P-value $\approx$ 0.018, statistically significant at 0.05
Test for significance: $\beta_{\text{Age.squared}}$ P-value $\approx$ 0, statistically significant

Georgia Tech

Different Link Function: GOF

```
## Test for goodness of fit
round(c(deviance(smoke5), 1-pchisq(deviance(smoke5),10)),2)
[1] 18.23  0.05

pearres5 = residuals(smoke5,type="pearson")
pearson = sum(pearres5^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
[1] 14.00  0.17
```

Does the goodness of fit improve?

- Using deviance residuals: P-value = 0.05
- Using Pearson residual: P-value = 0.17
- Do not reject the null hypothesis of good fit using Pearson residuals or using deviance residuals at the significance level 0.01.

Georgia Tech

Simpson's Paradox

Simpson's paradox: Reversal of an association when looking at a marginal relationship versus a conditional relationship.

- Smoking is statistically significant with a positive estimated coefficient under the marginal model.
- Smoking has a negative estimated coefficient under the conditional model.

Marginal versus Conditional Relationship

- ***Marginal***: Capturing the association of a predicting variable to the response variable without consideration of other factors
- ***Conditional***: Capturing the association of a predicting variable to the response variable conditional on other predicting variables in the model

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Classification

Georgia Tech

About This Lesson

Classification Objective

Data: $\{(x_{1,1}, x_{1,2}, \cdots, x_{1,p}), \boldsymbol{Y}_1\}, \cdots, \{(x_{n,1}, x_{n,2}, \cdots, x_{n,p}), \boldsymbol{Y}_n\}$, where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Model: Probability of success given predictor(s)

$$\mathrm{p} = (x_1, \cdots, x_p) = \Pr(Y = 1 \mid x_1, \cdots, x_p)$$

Objective: Classify (predict) a new binary response $\hat{Y}^*$ based on observed predicting variables $x^*_1,..,x^*_p$

- Predicted probability:

$$\hat{\mathrm{p}}(x^*_1, \cdots, x^*_p) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x^*_1 + \cdots + \hat{\beta}_p x^*_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x^*_1 + \cdots + \hat{\beta}_p x^*_p}}$$

- If the predicted probability is large, then classify $\hat{Y}^*$ as a success

Georgia Tech

Classification Objective

Data: $\{(x_{1,1}, x_{1,2}, \cdots, x_{1,p}), \boldsymbol{Y}_1\}, \cdots, \{(x_{n,1}, x_{n,2}, \cdots, x_{n,p}), \boldsymbol{Y}_n\}$,
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Model: Probability of success given predictor(s)

$$\mathrm{p} = (x_1, \cdots, x_p) = \Pr(Y = 1 \mid x_1, \cdots, x_p)$$

Objective: Classify (predict) a new binary response $\hat{Y}^*$ based on observed predicting variables $x^*_1,..,x^*_p$

- Predicted probability:

$$\hat{\mathrm{p}}(x^*_1, \cdots, x^*_p) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x^*_1 + \cdots + \hat{\beta}_p x^*_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x^*_1 + \cdots + \hat{\beta}_p x^*_p}}$$

- If the predicted probability is large, then classify $\hat{Y}^*$ as a success

How good is the classification or prediction?

- Goodness of fit doesn't guarantee good prediction;
- If we have many models for classification, how do we choose among them?

Georgia Tech

Classification Error Rate

- **Predicted probability** given $x_1, \cdots, x_p$:

$$\hat{\mathrm{p}}(x_1, \cdots, x_p) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}$$

- **Classifier:** $\mathrm{h}(x_1, \cdots, x_p) = \begin{cases} 1 \text{ if } \hat{\mathrm{p}}(x_1, \cdots, x_p) > \mathrm{r} \\ 0 \text{ otherwise} \end{cases}$,

  where r is a classification threshold between 0 and 1 (e.g., $\mathrm{r} = 1/2$)

- **Classification error rate:** $\mathrm{L}(\mathrm{h}) = 1 - \Pr\left(Y = \mathrm{h}(x_1, \cdots, x_p)\right)$
  - Training error
    - Use data to fit model, take proportion of responses misclassified
    - Biased downward as estimate of true classification error rate

Georgia Tech

Cross-Validation

Split the data $\{(x_{1,1}, x_{1,2}, \cdots, x_{1,p}), \boldsymbol{Y}_1\}, \cdots, \{(x_{n,1}, x_{n,2}, \cdots, x_{n,p}), \boldsymbol{Y}_n\}$, into:

- **Training Set**: Used to fit the model, i.e., to estimate $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$
- **Testing/Validation Set**: Used to estimate the classification error rate

$$\hat{\mathrm{L}}(\mathrm{h}) = \frac{1}{\mathrm{m}} \mathrm{count}\left(\left(1 - \mathrm{h}(x_{i,1}, x_{i,2}, \cdots, x_{i,p})\right) = \boldsymbol{Y}_i\right), i \in \text{Validation Set},$$

where *m* is the size of the validation set

How to split the data?

- Random subsampling
- *k*-fold cross-validation (KCV)
  - Leave-one-out cross-validation (LOOCV)

Cross-Validation: How to Split Data?

Random Subsampling

- Randomly split the data into two portions (training and validation sets)
- Train on training set and test on validation set
- Randomly split multiple times
- Average the classification error rate across all random splits

***k*-fold cross-validation (KCV)**

- Randomly divide the data into *k* chunks (folds) of approximately equal size
- For *i* = 1 to *k*:
  - The training data consist of data without the i^{th} fold of data
  - The testing data consist of the i^{th} fold
  - Compute classification error rate $\hat{\mathrm{L}}_i$ for the i^{th} fold testing data
  - Compute overall classification error: $\hat{\mathrm{L}}(\mathrm{h}) = \frac{1}{k}\sum_{i=1}^{k} \hat{\mathrm{L}}_i$

Georgia Tech

Cross-Validation: How to Split Data?

Random Subsampling

- Randomly split the data into two portions (training and validation sets)
- Train on training set and test on validation set
- Randomly split multiple times
- Average the classification error rate across all random splits

***k*-fold cross-validation (KCV)**

- Randomly divide the data into *k* chunks (folds) of approximately equal size
- For *i* = 1 to *k*:
  - The training data consist of data without the i^{th} fold of data
  - The testing data consist of the i^{th} fold
  - Compute classification error rate $\hat{\mathrm{L}}_i$ for the i^{th} fold testing data
  - Compute overall classification error: $\hat{\mathrm{L}}(\mathrm{h}) = \frac{1}{k}\sum_{i=1}^{k}\hat{\mathrm{L}}_i$

Random CV or *k*-fold CV?

- Random subsampling is computationally more expensive than *k*-fold CV

How to choose *k*?

- Leave-one-out CV is KCV with *k* = *n*
  - Less computationally efficient than KCV
- The larger the *k*, the less bias but the more variance

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

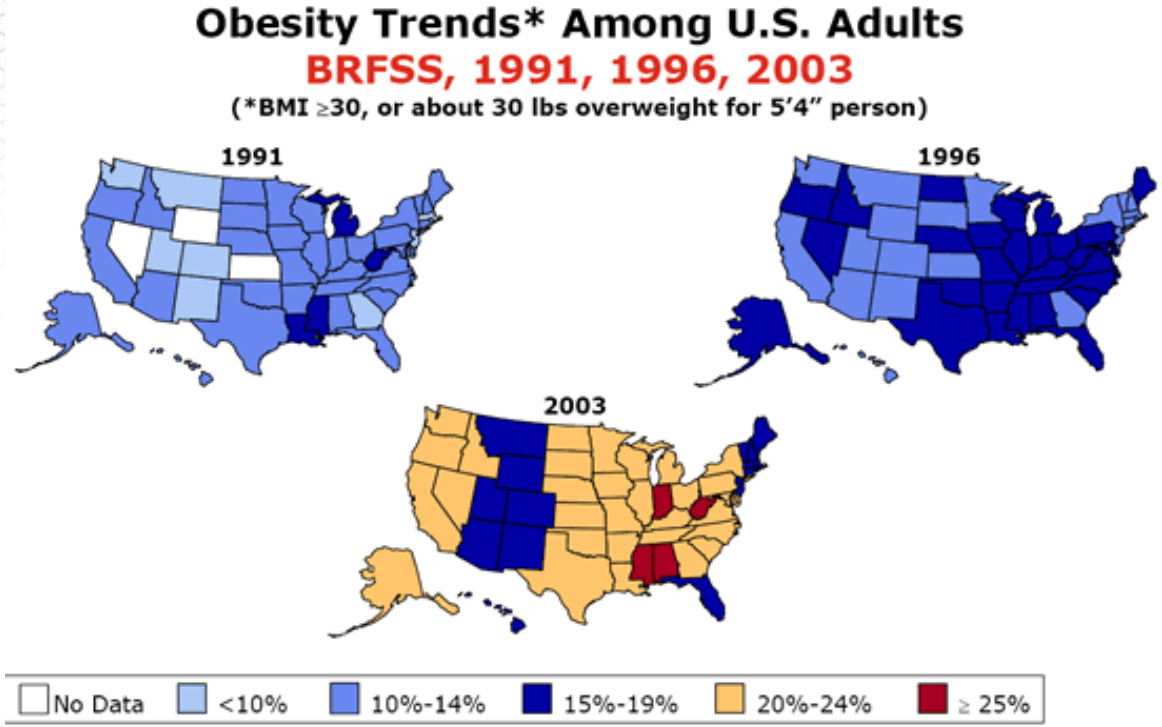
Stewart School of Industrial and Systems Engineering

Case Study: The Demographics of Obesity

Georgia Tech

About This Lesson

Obesity in the United States



Georgia Tech

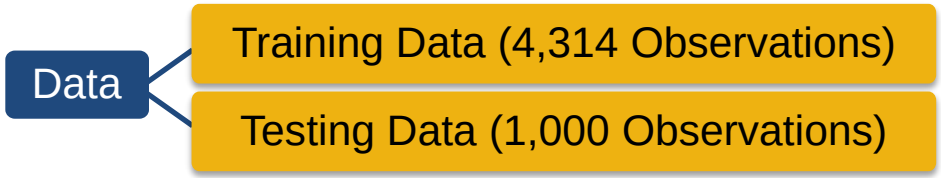
Case Study Overview

Objective:

- Use National Health and Nutrition Examination Survey (NHANES) to create a model used to predict likelihood of obesity.
- Identify predicting factors with predictive power

Variables:

- Response Variable: Whether an adult is classified as obese
- Predicting Variables: Age, Education Level, Gender



Georgia Tech

Case Study Overview

Objective:

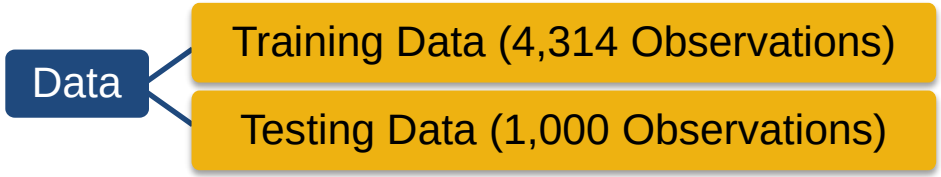
- Use National Health and Nutrition Examination Survey (NHANES) to create a model used to predict likelihood of obesity.
- Identify predicting factors with predictive power

Variables:

- Response Variable: Whether an adult is classified as obese
- Predicting Variables: Age, Education Level, Gender

Age variable

- Present as a continuous variable in the data
- Recoded into classes (or ranges) like
  - Class 1: 18-24 years
  - Class 2: 25-34 years
  - etc.



Georgia Tech

Obesity Data

```
## Read data in R
obdata = read.table("obesitydata.txt", h=T)
attach(obdata)

## Data before aggregation
obesityind = factor(Obesity, labels=c("NotObese", "Obese"))
agegr = factor(AgeGroup,
    labels=c("18to24", "25to34", "35to44", "45to64", "65+"))
gender = factor(Gender, labels=c("Male", "Female"))
edu = factor(Education,
    labels=c("<9thGrade", "9to11Grade", "HighSchool", "SomeCollege", "College+"))
```

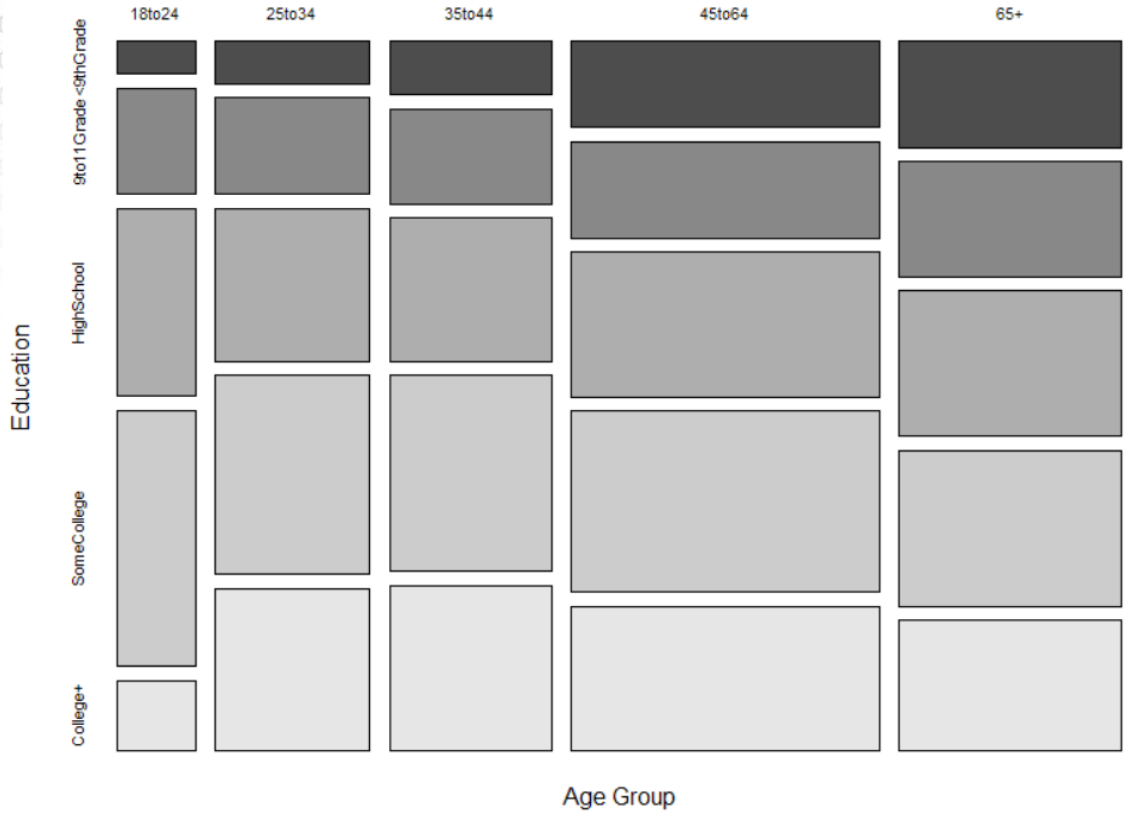
Georgia Tech

Exploratory Data Analysis

```
## Exploratory data analysis:  Categorical Predictors
tb_ageedu = xtabs(~agegr+edu)
library(vcd)
mosaicplot(tb_ageedu, xlab="Age Group", ylab="Education", color=TRUE, main="")
```

Georgia Tech

Exploratory Data Analysis

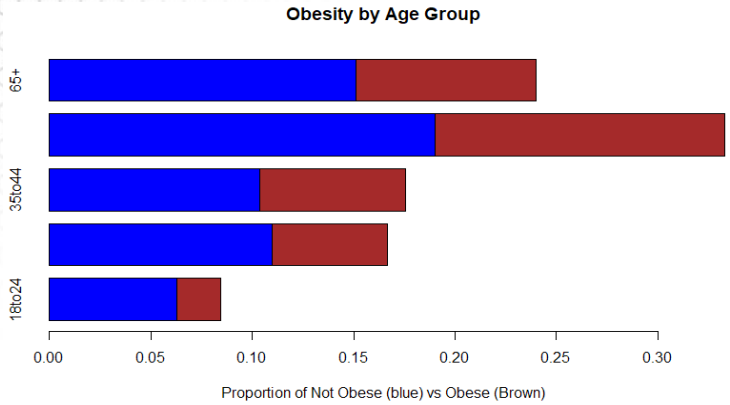


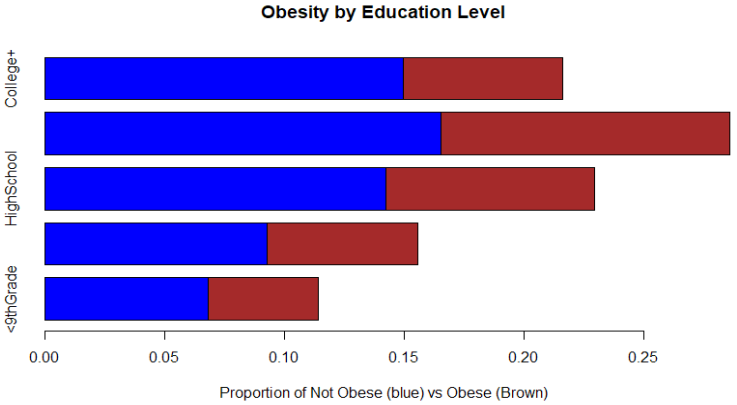
Exploratory Data Analysis

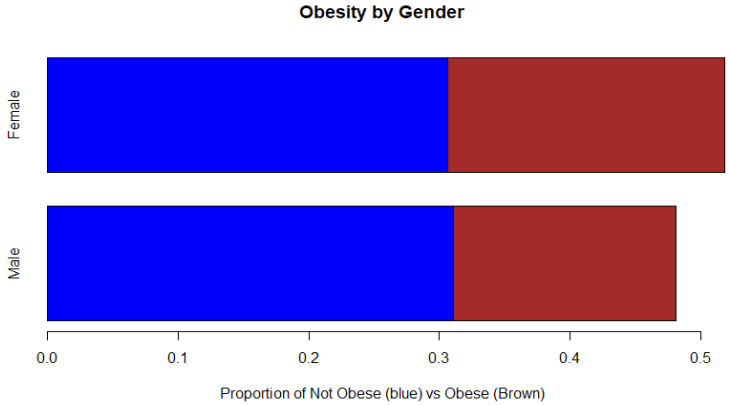
```
## Exploratory data analysis: Response vs Predictors
tb_obage = xtabs(~obesityind+agegr)
tb_obgender = xtabs(~obesityind+gender)
tb_obedu = xtabs(~obesityind+edu)

barplot(prop.table(tb_obage), axes=T, space=0.3, horiz=T,
    xlab="Proportion of Not Obese (blue) vs Obese (Brown)",
    col=c("blue","brown"), main="Obesity by Age Group")

barplot(prop.table(tb_obgender), axes=T, space=0.3, horiz=T,
    xlab="Proportion of Not Obese (blue) vs Obese (Brown)",
    col=c("blue","brown"), main="Obesity by Gender")

barplot(prop.table(tb_obedu), axes=T, space=0.3, horiz=T,
    xlab="Proportion of Not Obese (blue) vs Obese (Brown)",
    col=c("blue","brown"), main="Obesity by Education Level")
```

Georgia Tech

Exploratory Data Analysis







Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

The Demographics of Obesity: Modeling and Prediction

Georgia Tech

About This Lesson

Model Estimation

Fit a logistic regression model

model = glm(Obesity~agegr+gender+edu, family=binomial)
summary(model)

…

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.20581	0.15730	-7.666	1.78e-14 ***
agegr25to34	0.47271	0.14428	3.276	0.001052 **
agegr35to44	0.76486	0.14196	5.388	7.13e-08 ***
agegr45to64	0.84815	0.13240	6.406	1.49e-10 ***
agegr65+	0.60086	0.13751	4.370	1.24e-05 ***
genderFemale	0.23041	0.06363	3.621	0.000293 ***
edu9to11Grade	0.05632	0.12229	0.461	0.645110
eduHighSchool	-0.03440	0.11436	-0.301	0.763579
eduSomeCollege	0.13947	0.11036	1.264	0.206301
eduCollege+	-0.40077	0.11757	-3.409	0.000653 ***

…

Null deviance: 5739.9 on 4313 degrees of freedom
Residual deviance: 5641.3 on 4304 degrees of freedom

…

Georgia Tech

Model Estimation

Fit a logistic regression model

model = glm(Obesity~agegr+gender+edu, family=binomial)
summary(model)

…

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.20581	0.15730	-7.666	1.78e-14 ***
agegr25to34	0.47271	0.14428	3.276	0.001052 **
agegr35to44	0.76486	0.14196	5.388	7.13e-08 ***
agegr45to64	0.84815	0.13240	6.406	1.49e-10 ***
agegr65+	0.60086	0.13751	4.370	1.24e-05 ***
genderFemale	0.23041	0.06363	3.621	0.000293 ***
edu9to11Grade	0.05632	0.12229	0.461	0.645110
eduHighSchool	-0.03440	0.11436	-0.301	0.763579
eduSomeCollege	0.13947	0.11036	1.264	0.206301
eduCollege+	-0.40077	0.11757	-3.409	0.000653 ***

…

Null deviance: 5739.9 on 4313 degrees of freedom
Residual deviance: 5641.3 on 4304 degrees of freedom

…

$$\widehat{\beta}_{agegr25to34} = \mathbf{0.4727}$$

The ratio of the odds of obesity for age group 25-34 versus the age group 18-24 is 1.604 (or, equivalently the log odds ratio is 0.4727), holding all other predicting variables fixed. Odds of obesity for age group 25-34 are 60.4% higher than for age group 18-24 (baseline group).

$$\widehat{\beta}_{genderfemale} = \mathbf{0.2304}$$

The ratio of the odds of obesity for females versus males is 1.259, holding all other predicting variables fixed. Odds of obesity for females is 26% higher than for males.

Georgia Tech

Statistical Inference

```
## Test for overall regression
gstat = model$null.deviance - deviance(model)
cbind(gstat, 1-pchisq(gstat,length(coef(model))-1))
      gstat
[1,] 98.63672 0
```

Test for overall regression: ***p-value* ≈ 0 (< 0.01).** Reject the null hypothesis that all regression coefficients are zero. Conclude there are predicting variables that explain the variability in obesity.

```
round(coefficients(summary(model))[,4],4)
     (Intercept)   agegr25to34   agegr35to44   agegr45to64   agegr65+
          0.0000        0.0011        0.0000        0.0000     0.0000
     genderFemale  edu9to11Grade  eduHighSchool  eduSomeCollege  eduCollege+
          0.0003         0.6451         0.7636          0.2063       0.0007
```

Except for one, education regression coefficients are not statistically significant given that we account for age and gender.

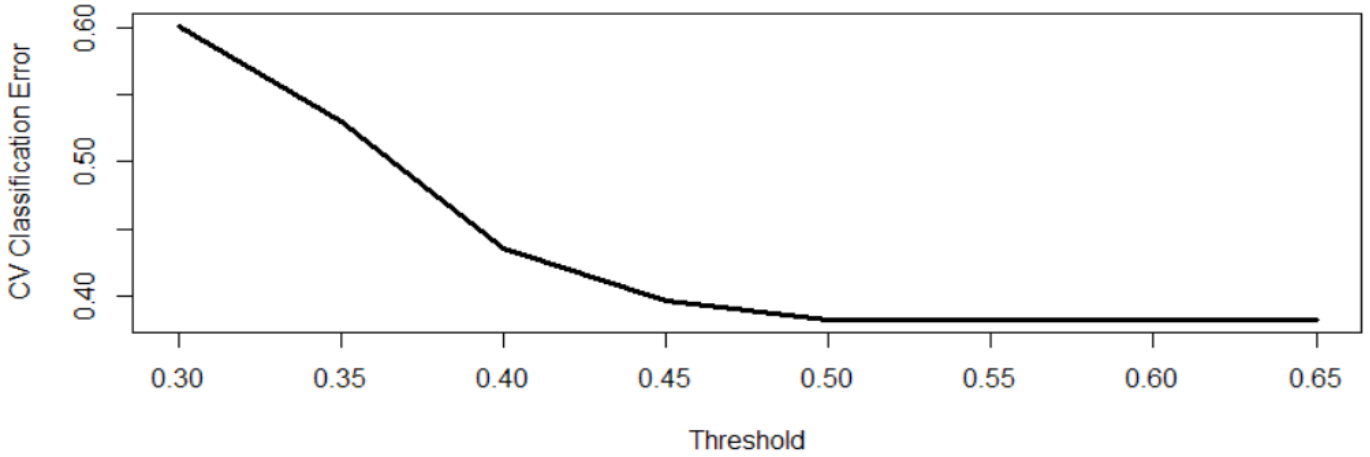
Georgia Tech

Predictive Power

```
## Prediction Accuracy
library(boot)
cost0.5 = function(y, pi){
      ypred=rep(0,length(y))
      ypred[pi>0.5] = 1
      err = mean(abs(y-ypred))
      return(err)}
obdata.fr = data.frame(cbind(Obesity, agegr, gender, edu))

## classification error for 10-fold cross-validation
cv.err0.5 = cv.glm(obdata.fr, model, cost=cost0.5, K=10)$delta[1]
:
cv.err = c(cv.err0.3, cv.err0.35, cv.err0.4, cv.err0.45, cv.err0.5,
      cv.err0.55, cv.err0.6, cv.err0.65)

## Smallest prediction error is 0.3824
plot(c(0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65), cv.err,
      type="l", lwd=3, xlab="Threshold", ylab="CV Classification Error")
```

Georgia Tech

Predictive Power



Prediction accuracy is highest and equal for thresholds above 0.5. ***Why?***

- It is the same prediction accuracy as if we were to replace all predictions with 0 (that is, predict everyone is not obese).
- The model has no predictive power since it performs worse than prediction without modeling.

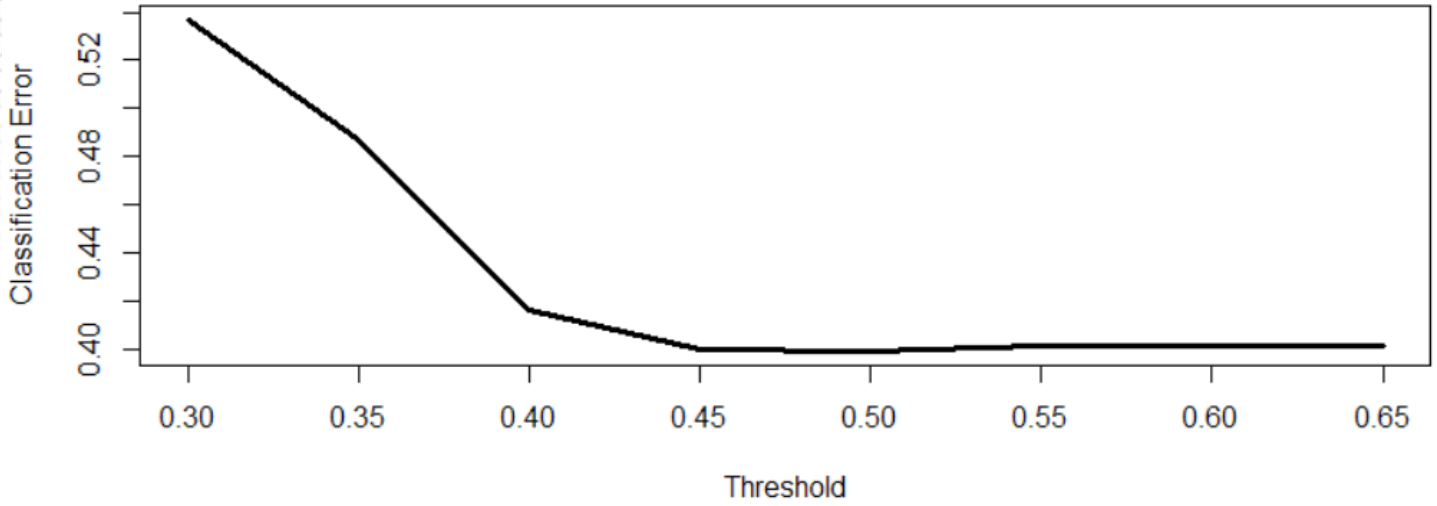
Georgia Tech

Prediction for Test Data

```
## Prediction given a set of new observations
## Prepare the test data
testobdata = read.table("testobesitydata.txt", h=T)
agegr.t = factor(testobdata$AgeGroup, labels=c("18to24", "25to34", "35to44",
    "45to64", "65+"))
gender.t = factor(testobdata$Gender, labels=c("Male","Female"))
edu.t = factor(testobdata$Education, labels=c("<9thGrade","9to11Grade",
    "HighSchool", "SomeCollege", "College+"))
pred.data = data.frame(agegr=agegr.t, gender=gender.t, edu=edu.t)
### Predict
pred.test = predict.glm(model,pred.data,type="response")
### Prediction Accuracy for multiple thresholds
err0.3 = cost0.3(testobdata$Obesity, pred.test)
:
err0.65 = cost0.65(testobdata$Obesity, pred.test)
err = c(err0.3, err0.35, err0.4, err0.45, err0.5, err0.55, err0.6, err0.65)
plot(c(0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65), err,
    type="l", lwd=3, xlab="Threshold", ylab="Classification Error")
```

Georgia Tech

Prediction for Test Data



- Prediction accuracy is highest at 0.5; it is similar as the prediction accuracy if we were to predict everyone is not obese.
- The prediction accuracy using the fitted model did not improve for the test data.

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

The Demographics of Obesity: Goodness of Fit

Georgia Tech

About This Lesson

Logistic Regression With Replications

```
### Aggregate data for Logistic Regression with repetitions
obdata.agg.n = aggregate(Obesity~agegr+gender+edu, FUN=length)
obdata.agg.y = aggregate(Obesity~agegr+gender+edu, FUN=sum)
obdata.agg = data.frame(Obesity = obdata.agg.y$Obesity,
                                  Total = obdata.agg.n$Obesity,
                                  agegr = obdata.agg.n$agegr,
                                  gender = obdata.agg.n$gender,
                                  edu = obdata.agg.n$edu)
## Fit a logistic regression model
model.agg = glm(cbind(Obesity,Total-Obesity)~agegr+gender+edu,
    data=obdata.agg, family=binomial)
## Test for GOF: Using deviance residuals
c(deviance(model.agg), 1-pchisq(deviance(model.agg),40))
[1] 29.0640209  0.8996714
```

Logistic Regression With Replications

```
### Aggregate data for Logistic Regression with repetitions
obdata.agg.n = aggregate(Obesity~agegr+gender+edu, FUN=length)
obdata.agg.y = aggregate(Obesity~agegr+gender+edu, FUN=sum)
obdata.agg = data.frame(Obesity = obdata.agg.y$Obesity,
                                  Total = obdata.agg.n$Obesity,
                                  agegr = obdata.agg.n$agegr,
                                  gender = obdata.agg.n$gender,
                                  edu = obdata.agg.n$edu)
## Fit a logistic regression model
model.agg = glm(cbind(Obesity,Total-Obesity)~agegr+gender+edu,
    data=obdata.agg, family=binomial)
## Test for GOF: Using deviance residuals
c(deviance(model.agg), 1-pchisq(deviance(model.agg),40))
[1] 29.0640209  0.8996714
```

With replications, we can perform a goodness of fit test.

***p-value* = 0.899** indicates a good fit.

Georgia Tech

Logistic Regression With Replications

summary(model.agg)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.20581	0.15730	-7.666	1.78e-14 ***
agegr25to34	0.47271	0.14428	3.276	0.001052 **
agegr35to44	0.76486	0.14196	5.388	7.13e-08 ***
agegr45to64	0.84815	0.13240	6.406	1.49e-10 ***
agegr65+	0.60086	0.13751	4.370	1.24e-05 ***
genderFemale	0.23041	0.06363	3.621	0.000293 ***
edu9to11Grade	0.05632	0.12229	0.461	0.645110
eduHighSchool	-0.03440	0.11436	-0.301	0.763579
eduSomeCollege	0.13947	0.11036	1.264	0.206301
eduCollege+	-0.40077	0.11757	-3.409	0.000653 ***

Null deviance: 127.701 on 49 degrees of freedom
Residual deviance: 29.064 on 40 degrees of freedom

- Regression coefficient output for estimation and statistical inference is the same with or without replications.
- Null and residual deviance output is different with replications. ***Why?***

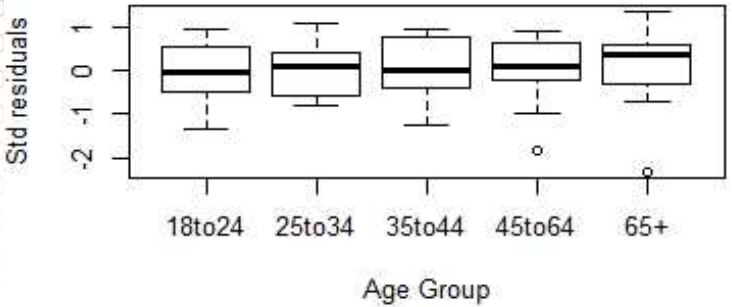
Georgia Tech

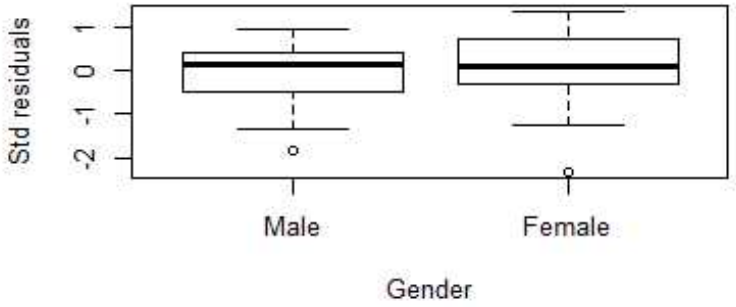
Residual Analysis

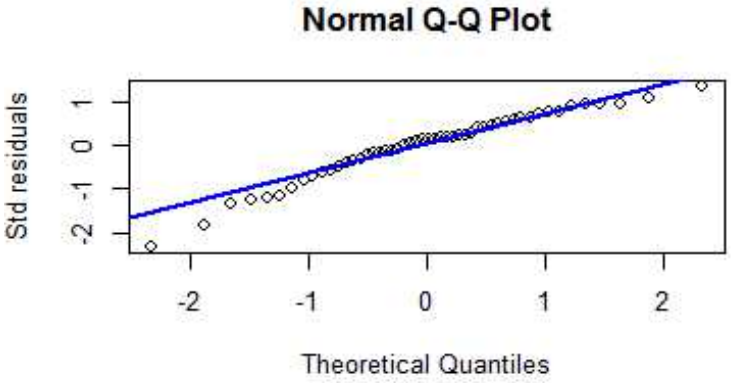
```
res = resid(model.agg, type="deviance")
par(mfrow=c(2,2))

boxplot(res~agegr,
        xlab="Age Group",
        ylab="Std residuals",
        data=obdata.agg)

boxplot(res~gender,
        xlab="Gender",
        ylab="Std residuals",
        data = obdata.agg)

qqnorm(res, ylab="Std residuals")
qqline(res, col="blue", lwd=2)

hist(res, 10, xlab="Std residuals", main="")
```

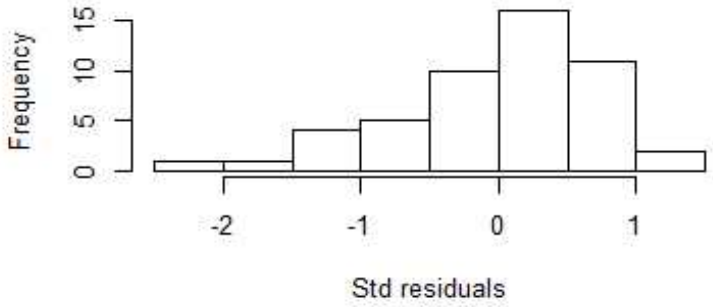
Georgia Tech

Residual Analysis









Georgia Tech

Prediction of Adult Obesity: Results

- Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity.
  - **But the fitted model with education, gender, and age group factors does not improve prediction.**
- After factor aggregation, goodness of fit can be performed.

- The *p-value* of the deviance test for goodness of fit is high, indicating good fit.
  - **But residual analysis suggests that there may be some departures from normality and thus from goodness of fit.**
- Models with different link functions or including interaction terms have not shown improvement. *(Results not shown in this lecture.)*
- The sample size is large enough for reliable statistical inference.

Georgia Tech

Prediction of Adult Obesity: Results

- Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity.
  - **But the fitted model with education, gender, and age group factors does not improve prediction.**
- After factor aggregation, goodness of fit can be performed.
- The *p-value* of the deviance test for goodness of fit is high, indicating good fit.
  - **But residual analysis suggests that there may be some departures from normality and thus from goodness of fit.**
- Models with different link functions or including interaction terms have not shown improvement. *(Results not shown in this lecture.)*
- The sample size is large enough for reliable statistical inference.

What can be done to improve the model fit and the predictive power?

- Include other factors in the model, such as income level, unemployment, race, and ethnicity, among others.
- Consider interaction terms between age, education, and gender groups and other factors.

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Introduction

About This Lesson

Other Distributions of the Response

➔ The response variable (e.g. rate) has a **Poisson distribution**

- What drives the rate of phone calls per day in a calling service center?
- What predicts the density per mile of trees in a forest?

➔ The response variable (e.g. wait time) has an **exponential distribution**

- What explains the wait time for a wellness visit at your physician offices?

➔ The response variable can have other distributions from the **exponential family of distributions**

Generalized Linear Model

Standard Linear Regression

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$, $i = 1,...,n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $E(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $Var(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1,..., \varepsilon_n\}$ are independent random variables
- ***Normality Assumption*: $\varepsilon_i \sim$ Normal**

Georgia Tech

Generalized Linear Model

Data: $\{(x_{11},..,x_{1p}),Y_1\},....,\{(x_{n1},..,x_{np}),Y_n\}$ where $Y_1, .., Y_n$ response variable with **a distribution from the exponential family**

Model: Model the conditional expectation**:**

$$g(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

OR

$$E(Y|x_1, \ldots, x_p) = g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$$

where g( ) is a *link function* and g^{-1}( ) the *inverse link function* depending on the distribution of Y.

Georgia Tech

Generalized Linear Model

Y~ *distribution in the exponential family* if its density function can be written as:

$$f(y;\theta) = h(y)e^{g(\theta)T(y)-B(\theta)}$$

where θ is the parameter of the distribution and g(θ) is the link function.

Distribution	Link	Regression Function
Normal	g(m) = m	m = $x^T\beta$
Poisson	g(m) = log(m)	m = $e^{x^T\beta}$
Bernoulli	g(m) = log(m/(1-m))	m = $\frac{e^{x^T\beta}}{1+ e^{x^T\beta}}$
Gamma	g(m) = 1/m	m = $\frac{1}{x^T\beta}$

Georgia Tech

Poisson Regression

Data: $\{(x_{11},.., x_{1p}),Y_1\},....,\{(x_{n1},.., x_{np}),Y_n\}$ where $Y_1, .., Y_n$ response variable with **a Poisson distribution**

Model: Model the conditional expectation**:**

$$\log(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

OR

$$E(Y|x_1, \ldots, x_p) = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

Georgia Tech

Linear Regression versus Poisson Regression

Standard Linear Regression with log-transformation:

- $E(\log(\mathrm{Y})|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = \beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}$
- $\mathrm{V}(\log(\mathrm{Y})|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})$ constant

Poisson Regression:

- $\log(\mathrm{E}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})) = \beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}$
- $\mathrm{V}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = e^{\beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}}$

OR

$$\log(\mathrm{V}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})) = \beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}$$

Georgia Tech

Linear Regression versus Poisson Regression

Standard Linear Regression with log-transformation:

- $E(\log(\mathrm{Y})|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = \beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}$
- $\mathrm{V}(\log(\mathrm{Y})|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})$ constant

Poisson Regression:

- $\log(\mathrm{E}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})) = \beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}$
- $\mathrm{V}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = e^{\beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}}$

OR

$$\log(\mathrm{V}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})) = \beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p}$$

- Using Standard Linear Regression with log-transformation instead of Poisson Regression will result in violations of the assumption of constant variance.
- Alternatively, Standard Linear Regression could be use if the number of counts are large and with the variance stabilizing transformation $\sqrt{\mu + 3/8}$.

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Data Examples

About This Lesson

Data Example 1: High School Awards

Objective: To model and predict the number of awards earned by students at one high school for multiple high schools.

Response Variable: The number of awards earned by students at a high school per *year.*

Predicting Variables:

- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 ="Vocational"; and
- The score on the final exam in math.

Acknowledgement: This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.

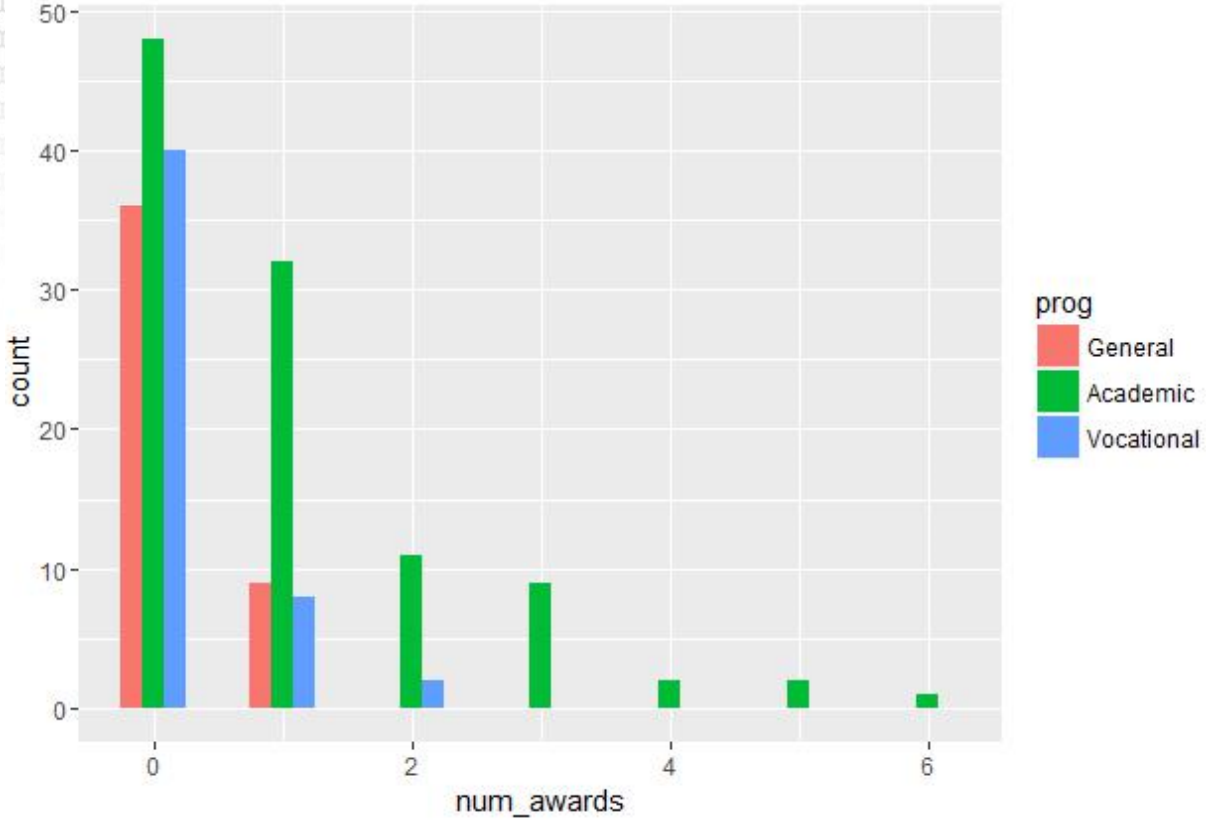
Georgia Tech

Exploratory Data Analysis

```
## Read data in R
awardsdata = read.csv("students_awards.csv",header=T)
## Convert qualitative variable in the data into factor in R
awardsdata = within(awardsdata, {
      prog = factor(prog, levels=1:3, labels=c("General", "Academic", "Vocational"))
      id = factor(id)})

## Conditional histograms
library(ggplot2)
ggplot(awardsdata, aes(num_awards, fill = prog)) + geom_histogram(binwidth=.5, position="dodge")
```

Georgia Tech

Exploratory Data Analysis



Data Example 2: Insurance Claims

Objective: To explain factors that are associated to car insurance claims due to accidents or other events leading to car damage.

Response Variable: The number of car insurance claims per *policyholder*:

- Holders: numbers of policyholders; and
- Claims: numbers of claims

Predicting Variables:

- District of residence of policyholder (1 to 4): 4 is major cities.
- Classification of cars with levels <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.
- Age group of the policyholder: <25, 25–29, 30–35, >35.

Exploratory Data Analysis

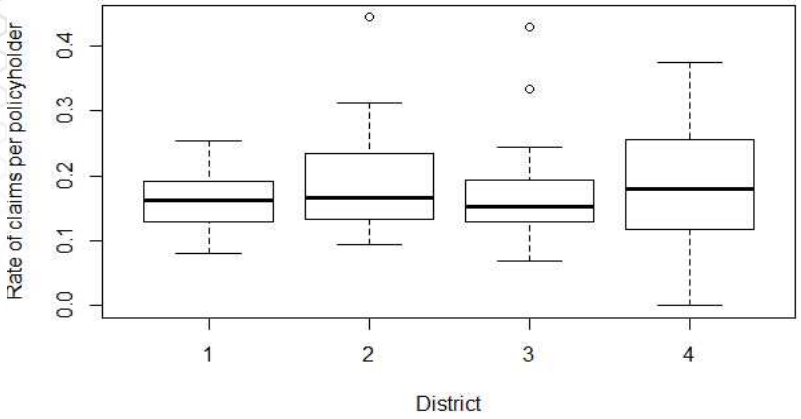
Data in the R library MASS

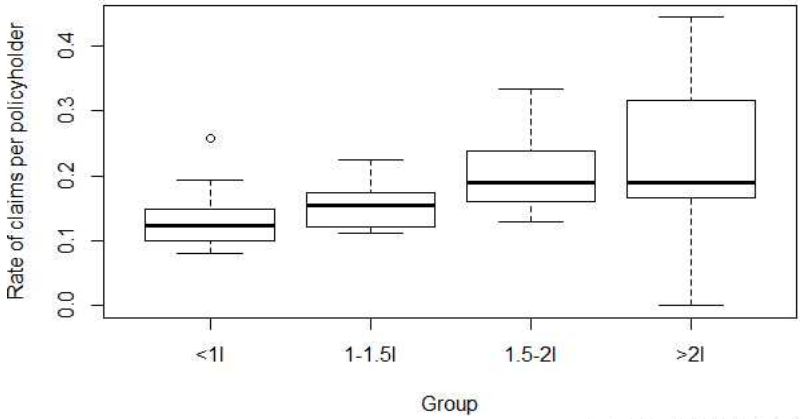
library(MASS)
summary(Insurance)

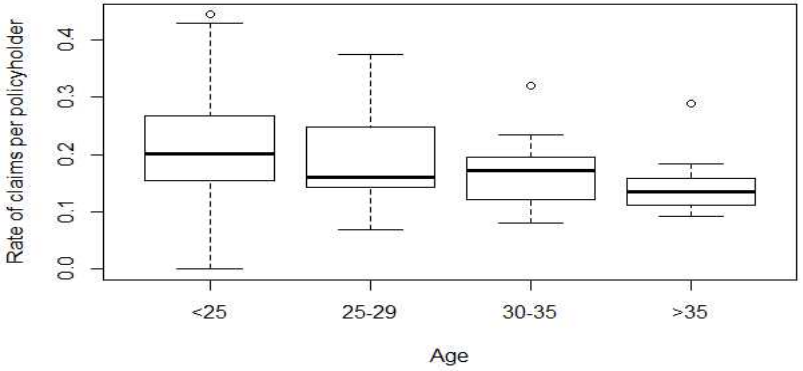
Relationship between rate of claims and predictors

boxplot(Claims/Holders~District, xlab = "District", ylab = "Rate of claims per policyholder",data=Insurance)
boxplot(Claims/Holders~Group, xlab = "Group", ylab = "Rate of claims per policyholder",data=Insurance)
boxplot(Claims/Holders~Age, xlab = "Age", ylab = "Rate of claims per policyholder",data=Insurance)

Exploratory Data Analysis







Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Model Description and Estimation

Georgia Tech

About This Lesson

Poisson Regression Model

Data: $\{(x_{11},..,x_{1p}),Y_1\},....,\{(x_{n1},..,x_{np}),Y_n\}$ where $Y_1,..,Y_n$ are event count data per observation unit with a Poisson distribution

Poisson Distribution: $Y \sim \text{Poisson}(\lambda)$: $P(Y=y) = \frac{e^{-\lambda}\lambda^y}{y!}$

$$E(Y) = V(Y) = \lambda$$

Model: Model the conditional expectation:

$Y_i|x_{i1}, \ldots, x_{ip} \sim \text{Poisson}(\lambda_i)$ with

$\lambda_i = E(Y|x_1, \ldots, x_p) = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$

OR

$\log(\lambda_i) = \log(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1+\ldots+\beta_p x_p$

Georgia Tech

Model Interpretation

The rate of event occurrence given predicting variable X = x:

$$\lambda = \lambda(\mathrm{x}) = E(Y|\mathrm{x}) = e^{\beta_0+\beta_1 x}$$

- The log function $\ln(\lambda(\mathrm{x})) = \beta_0 + \beta_1 \mathrm{x}$ is the *log rate*.
- With an increase with one unit in x (if quantitative): $\frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$
- If x categorical: $\frac{e^{\beta_0+\beta_1(x=1)}}{e^{\beta_0+\beta_1(x=0)}} = e^{\beta_1}$
- Interpretation of the regression coefficients in terms of log ratio of the rate.
- If other predicting variables are in the model, then we need to hold fixed all other predicting variables.

Georgia Tech

Model Estimation

Model the log rate given predictor(s):

$$\log(\lambda_i) = \log(\mathrm{E}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})) = \beta_0 + \beta_1 \mathrm{x}_1 + \ldots + \beta_p \mathrm{x}_\mathrm{p}$$

Parameters: $\beta_0, \beta_1, \ldots, \beta_p$

Approach: Maximum Likelihood Estimation:

$$\mathrm{L}(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

$$\max_{\beta_0, \beta_1, \ldots, \beta_p} \mathrm{l}(\beta_0, \beta_1, \ldots, \beta_p) = \log(\mathrm{L}(\beta_0, \beta_1, \ldots, \beta_p)) =$$

$$\sum_{i=1}^{n} \{y_i \log \lambda_i - \lambda_i\} = \sum_{i=1}^{n} \{y_i (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - e^{\beta_0 + \beta_1 \mathrm{x}_1 + \ldots + \beta_p \mathrm{x}_\mathrm{p}}\}$$

Georgia Tech

Model Estimation (cont'd)

Approach: Maximum Likelihood Estimation

$$\max_{\beta_0,\, \beta_1,\ldots,\, \beta_p} l(\beta_0, \beta_1,\ldots, \beta_p) = \log(L(\beta_0, \beta_1,\ldots, \beta_p)) =$$

$$\sum_{i=1}^{n} \{y_i \log \lambda_i - \lambda_i\} = \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}\}$$

- Maximizing the (log-)likelihood function with respect to $\beta_0, \beta_1,\ldots, \beta_p$ in close form expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters
- Use numerical algorithm to estimate $\beta_0, \beta_1,\ldots, \beta_p \Rightarrow \hat{\beta}_0, \hat{\beta}_1,\ldots, \hat{\beta}_p$

Upshot: The estimated parameters and their standard errors are approximate estimates. Do not attempt to do it yourself! Use a statistical software to derive the estimated regression coefficients.

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Model Estimation: Data Example

Georgia Tech

About This Lesson

Data Example 1: High School Awards

Objective: To model and predict the number of awards earned by students at one high school for multiple high schools.

Response Variable: The number of awards earned by students at a high school per *year*

Predicting Variables:

- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 ="Vocational"; and
- The score on the final exam in math.

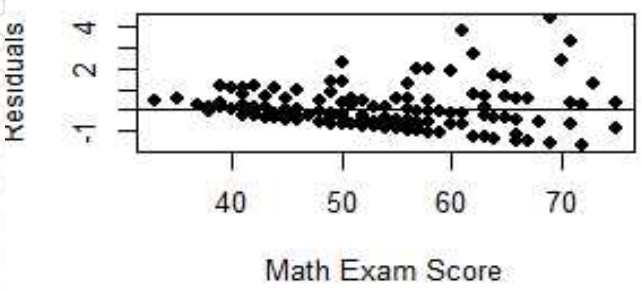
Acknowledgement: This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.

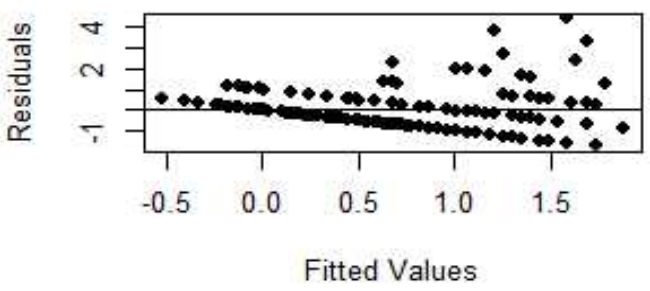
Georgia Tech

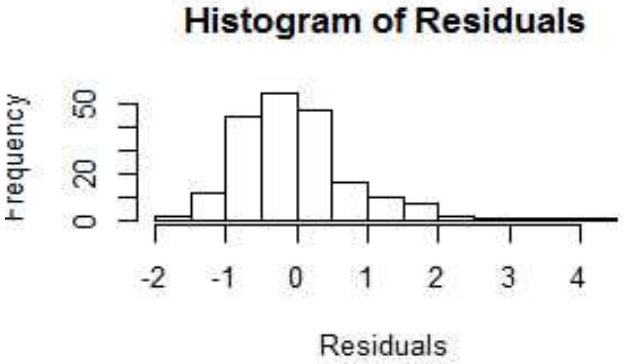
GOF: Standard Linear Regression

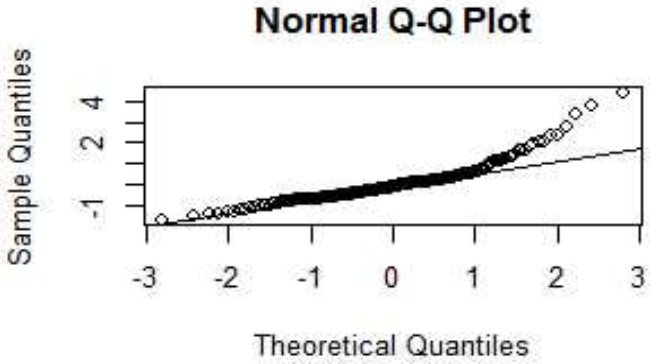
```
## Fit a standard regression model
m0 = lm(num_awards ~ prog + math, data=awardsdata)
## Residual Analysis for Goodness of Fit
par(mfrow = c(2,2))
plot(awardsdata$math, res, xlab = "Math Exam Score", ylab = "Residuals", pch = 19)
abline(h = 0)
plot(fitted(m0), res, xlab = "Fitted Values", ylab = "Residuals", pch = 19)
abline(h = 0)
hist(res, xlab="Residuals", main= "Histogram of Residuals")
qqnorm(res)
qqline(res)
```

GOF: Standard Linear Regression









Georgia Tech

Poisson Regression Estimation

m1 = glm(num_awards ~ prog + math, family="poisson", data=awardsdata)
summary(m1)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.24712	0.65845	-7.969	1.60e-15 ***
progAcademic	1.08386	0.35825	3.025	0.00248 **
progVocational	0.36981	0.44107	0.838	0.40179
math	0.07015	0.01060	6.619	3.63e-11 ***

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{math}} = 0.07$: For one unit increase in the math exam score,

- the log expected award count would be expected to increase by 0.07, holding the program fixed.
- the rate ratio for awards would be expected to increase by a factor of 1.07, holding the program fixed.

Georgia Tech

Poisson Regression Estimation

m1 = glm(num_awards ~ prog + math, family="poisson", data=awardsdata)
summary(m1)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.24712	0.65845	-7.969	1.60e-15 ***
progAcademic	1.08386	0.35825	3.025	0.00248 **
progVocational	0.36981	0.44107	0.838	0.40179
math	0.07015	0.01060	6.619	3.63e-11 ***

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{academic}} = 1.084$: While holding math score fixed, academic programs compared to general programs are expected to have

- The log of expected award counts 1.084 higher
- The rate for awards exp(1.084) = 2.956 times higher

Georgia Tech

Data Example 2: Insurance Claims

Objective: To explain factors that are associated to car insurance claims due to accidents or other events leading to car damage.

Response Variable: The number of car insurance claims per *policyholder*:

- Holders: numbers of policyholders; and
- Claims: numbers of claims

Predicting Variables:

- District of residence of policyholder (1 to 4): 4 is major cities.
- Classification of cars with levels <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.
- Age group of the policyholder: <25, 25–29, 30–35, >35.

Poisson Regression Estimation

m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)), data = Insurance, family = poisson)

Important to note!

- Event rates can be calculated as events per units of varying size; the unit size is called **exposure***;*
- In Poisson regression, exposure is accounted for using an **offset** -- the exposure variable enters in the linear combination of the predicting variables, but with the coefficient (for log(exposure)) constrained to 1:
  $\log(\mathrm{E}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})) = \beta_0 + \beta_1\mathrm{x}_1 + \ldots + \beta_p\mathrm{x}_\mathrm{p} + \log(\text{exposure})$
- In this example, the number of policyholders is the exposure since the rate of claims is per policyholder (hence the unit).

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Statistical Inference

Georgia Tech

About This Lesson

Model Estimation

Model the log rate given predictor(s):

$$\log(\lambda_i) = \log(\mathrm{E}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p})) = \beta_0 + \beta_1 \mathrm{x}_1 + \ldots + \beta_p \mathrm{x}_\mathrm{p}$$

Parameters: $\beta_0, \beta_1, \ldots, \beta_p$

Approach: Maximum Likelihood Estimation:

$$\mathrm{L}(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} {\lambda_i}^{y_i}}{y_i!}$$

$$\max_{\beta_0, \beta_1, \ldots, \beta_p} \mathrm{l}(\beta_0, \beta_1, \ldots, \beta_p) = \log(\mathrm{L}(\beta_0, \beta_1, \ldots, \beta_p)) =$$

$$\sum_{i=1}^{n} \{y_i \log \lambda_i - \lambda_i\} = \sum_{i=1}^{n} \{y_i (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - e^{\beta_0 + \beta_1 \mathrm{x}_1 + \ldots + \beta_p \mathrm{x}_\mathrm{p}}\}$$

Georgia Tech

Statistical Inference

Maximum Likelihood Estimators (MLEs): $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$

Statistical Properties of MLEs:

- Approximate Sampling Distribution: $\widehat{\boldsymbol{\beta}} \approx \boldsymbol{N}(\boldsymbol{\beta}, \boldsymbol{V})$
- The normal approximation relies on the assumption of large sample size ⇒ Statistical inference is not reliable for small sample data

1-α Approximate Confidence interval

$$\hat{\beta}_j \pm z_{\frac{\alpha}{2}} \sqrt{V\left(\hat{\beta}_j\right)}$$

Georgia Tech

Statistical Inference (cont'd)

- Hypothesis testing and Confidence Intervals rely on the approximately normal distribution of large sample sizes
- Use the z-test (Wald test)
  - Test is for the statistical significance of $\hat{\beta}_j$ *given all other predicting variables in the model*
  - Null hypothesis is that β_j is not significant
    $\mathrm{H}_0: \beta_j = 0$ vs. $\mathrm{H}_\mathrm{a}: \beta_j \neq 0$
  - $\mathrm{z-value} = \frac{\hat{\beta}_j - 0}{\mathrm{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$
  - Reject H_0 if $|\mathrm{z-value}|$ is too large
    - Implies that β_j is statistically significant

Georgia Tech

Statistical Inference (cont'd)

$$\text{z-value} = \frac{\widehat{\beta_j} - b}{se\left(\widehat{\beta_j}\right)}$$ how large to reject $H_0: \beta_j = b$?

For significance level α , Reject if $|\text{z-value}| > z_{\frac{\alpha}{2}}$

Alternatively, compute $\text{P-value} = 2P(Z > |\text{ z-value}|)$

What if we want to test for positive relationship?

$H_0: \beta_j \leq 0$ versus $H_A: \beta_j > 0$?

$\text{P-value} = P(Z > \text{z-value})$

What if we want to test for negative relationship?

$H_0: \beta_j \geq 0$ versus $H_A: \beta_j < 0$?

$\text{P-value} = P(Z < \text{z -value})$

Georgia Tech

Statistical Inference (cont'd)

$$\text{z-value} = \frac{\widehat{\beta_j} - b}{se(\widehat{\beta_j})}$$ how large to reject $H_0: \beta_j = b$?

For significance level α , Reject if $|\text{z-value}| > z_{\frac{\alpha}{2}}$

Alternatively, compute $\text{P-value} = 2P(Z > |\text{ z-value}|)$

What if we want to test for positive relationship?

$H_0: \beta_j \leq 0$ **versus** $H_A: \beta_j > 0$?

$\text{P-value} = P(Z > \text{z-value})$

What if we want to test for negative relationship?

$H_0: \beta_j \geq 0$ **versus** $H_A: \beta_j < 0$?

$\text{P-value} = P(Z < \text{z -value})$

- Because the approximation of the normal distribution relies on large sample size, so do the hypothesis testing procedures.
- What if *n* is small?
  - The hypothesis testing procedure will have a probability of type I error larger than the significance level.
  - In other words, there will likely be more type I errors than expected.

Georgia Tech

Testing for Subsets of Coefficients

Full model:

$$\text{Log}\left(\text{p}\left(X_1, \cdots, X_p, Z_1, \cdots, Z_q\right)\right) =$$

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q$$

Reduced model:

$$\text{Log}\left(\text{p}\left(X_1, \cdots, X_p\ \right)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The hypothesis test:

$\mathbf{H_0}$: $\boldsymbol{\alpha_1 = \alpha_2 = \cdots = \alpha_q = 0}$

vs.

$\mathbf{H_a}$: $\boldsymbol{\alpha_i \neq 0}$ **for at least one** $\boldsymbol{\alpha_i}, \mathbf{i = 1}, \cdots, \boldsymbol{q}$

- Maximize the likelihood function under reduced model: $\mathcal{L}\,(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)\right) - \log\left(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)\right) \approx \chi^2_q$
  - $\text{P-value} = \Pr(\chi^2_q > \text{Deviance})$

Georgia Tech

Testing for Subsets of Coefficients

Full model:

$$\text{Log}\left(\text{p}\left(X_1, \cdots, X_p, Z_1, \cdots, Z_q\right)\right) =$$

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q$$

Reduced model:

$$\text{Log}\left(\text{p}\left(X_1, \cdots, X_p\right)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The hypothesis test:

$$\mathbf{H_0}: \boldsymbol{\alpha_1 = \alpha_2 = \cdots = \alpha_q = 0}$$

vs.

$$\mathbf{H_a}: \boldsymbol{\alpha_i \neq 0 \text{ for at least one } \alpha_i, i = 1, \cdots, q}$$

- The hypothesis test for subsets of coefficients is approximate
- This is not a test for goodness of fit!
  - It only compares two models

- Maximize the likelihood function under reduced model: $\mathcal{L}\,(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)\right) - \log\left(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)\right) \approx \chi^2_q$
  - $\text{P-value} = \Pr(\chi^2_q > \text{Deviance})$

Georgia Tech

Testing for Overall Regression

Full model:

$$\text{Log}\left(\text{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Reduced model:

$$\text{Log}\left(\text{p}(X_1, \cdots, X_p)\right) = \beta_0$$

The hypothesis test:

$\mathbf{H_0}$: $\boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_p = 0}$

vs.

$\mathbf{H_a}$: $\boldsymbol{\beta_i \neq 0}$ **for at least one** $\boldsymbol{\beta_i}, \mathbf{i = 1, \cdots, p}$

- Maximize the likelihood function under reduced model: $\mathcal{L}(\bar{\beta}_0)$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0)\right) - \log(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)\right) \approx \chi^2_p$
  - $\text{P-value} = \Pr(\chi^2_p > \text{Deviance})$

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Statistical Inference: Data Example

Georgia Tech

About This Lesson

Data Example 1: High School Awards

Objective: To model and predict the number of awards earned by students at one high school for multiple high schools.

Response Variable: The number of awards earned by students at a high school per *year*

Predicting Variables:

- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 ="Vocational"; and
- The score on the final exam in math.

Acknowledgement: This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.

Georgia Tech

Data Example 1: Statistical Inference

m1 = glm(num_awards ~ prog + math, family="poisson", data=awardsdata)
summary(m1)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.24712	0.65845	-7.969	1.60e-15 ***
progAcademic	1.08386	0.35825	3.025	0.00248 **
progVocational	0.36981	0.44107	0.838	0.40179
math	0.07015	0.01060	6.619	3.63e-11 ***

Null deviance: 287.67 on 199 degrees of freedom
Residual deviance: 189.45 on 196 degrees of freedom

1-pchisq((287.67-189.45),(199-196))
[1] 0

Test for significance β_{math}: p-value≈0 thus statistically significant
Test for overall regression: p-value ≈0 thus at least one predicting variables significantly explains the variability in the number of awards

Georgia Tech

Data Example 2: Insurance Claims

Objective: To explain factors that are associated to car insurance claims due to accidents or other events leading to car damage.

Response Variable: The number of car insurance claims per *policyholder*:

- Holders: numbers of policyholders; and
- Claims: numbers of claims

Predicting Variables:

- District of residence of policyholder (1 to 4): 4 is major cities.
- Classification of cars with levels <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.
- Age group of the policyholder: <25, 25–29, 30–35, >35.

Data Example 2: Statistical Inference

m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)), data = Insurance, family = poisson)
summary(m.ins)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.810508	0.032972	-54.910	< 2e-16 ***
…….				
Age.L	-0.394432	0.049404	-7.984	1.42e-15 ***
Age.Q	-0.000355	0.048918	-0.007	0.994210
Age.C	-0.016737	0.048478	-0.345	0.729910

What are these Age variables?

Null deviance: 236.26 on 63 degrees of freedom
Residual deviance: 51.42 on 54 degrees of freedom

Georgia Tech

Data Example 2: Statistical Inference

```
library(MASS)
summary(Insurance)

Ins.dat = within(Insurance, {
  Age = factor(Age, ordered = F)
  Group = factor(Group, ordered = F)
  District = factor(District, ordered = F)})

m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)), data = Ins.dat, family = poisson)
summary(m.ins)
```

Georgia Tech

Data Example 2: Statistical Inference

m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)), data = Ins.dat, family = poisson)
summary(m.ins)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.82174	0.07679	-23.724	< 2e-16 ***
…….				
Age25-29	-0.19101	0.08286	-2.305	0.021149 *
Age30-35	-0.34495	0.08137	-4.239	2.24e-05 ***
Age>35	-0.53667	0.06996	-7.672	1.70e-04 ***

Null deviance: 236.26 on 63 degrees of freedom
Residual deviance: 51.42 on 54 degrees of freedom

Test for significance

$\hat{\beta}_{age25-29}$ = -0.191 or
exp($\hat{\beta}_{age25-29}$) = 0.826

$\hat{\beta}_{age25-29}$, $\hat{\beta}_{age30-35}$ & $\hat{\beta}_{age>35}$: p-value<0.05, thus statistically significant.

Georgia Tech

Data Example 2: Statistical Inference (cont'd)

```
m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)),
data = Ins.dat, family = poisson)
summary(m.ins)

# test for overall regression
1-pchisq((236.26-51.42),(63-54))
```

Test for overall regression: $\text{p-value} \approx 0$ thus at least one predicting variables significantly explains the variability in the number of claims

Georgia Tech

Data Example 2: Statistical Inference (cont'd)

 # Is the district of residence of policyholder a statistically significant variable given all other predicting variables in the model?

Full model: District + Group + Age
Reduced model: Group + Age

```
library(aod)
wald.test(b=coef(m.ins), Sigma=vcov(m.ins), Terms=2:4)
```

Wald test:

Chi-squared test:
X2 = 14.6, df = 3, P(> X2) = 0.0022

Test for subsets of coefficients: $\text{p-value} = 0.002$ reject the null hypothesis and conclude that the District variable does have significant explanatory power

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment

About This Lesson

Poisson Regression Model

Data: $\{(x_{11},..,x_{1p}),Y_1\},....,\{(x_{n1},..,x_{np}),Y_n\}$ where $Y_1, .., Y_n$ are event count data per observation unit with a Poisson distribution

Assumptions:

- *Linearity Assumption*: $\log(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- *Independence Assumption*: $Y_1, .., Y_n$ are independent random variables
- *Variance Assumption*: $E(Y|x_1, \ldots, x_p) = V(Y|x_1, \ldots, x_p)$

There is no error term! How to check the assumptions?

Georgia Tech

Residuals in Poisson Regression

Poisson Regression:

$$Y_i|(x_{i1},..,x_{ip}) \sim \text{Poisson}(\lambda(x_{i1},\dots,x_{ip}))$$

- Estimated rates are:

$$\hat{\lambda}_i = \hat{\lambda}(x_{i1,..,}x_{ip}) = e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\dots+\hat{\beta}_p x_{ip}}$$

- Pearson Residuals: $r_i = \frac{Y_i-\hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$

- Deviance Residuals:

$$d_i = sgn(Y_i-\hat{\lambda}_i)\sqrt{2\left\{Y_i \log\left(\frac{Y_i}{\hat{\lambda}_i}\right)-(Y_i-\hat{\lambda}_i)\right\}}$$

- Pearson's residuals follow directly a normal approximation to a binomial. Hence approximately N(0,1)
- The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model vs. the fitted model. Thus approximately N(0,1) if the model is a good fit.
- Deviances play the role of sum of squares in a linear model.

Georgia Tech

Goodness of Fit

GOF Visual Analytics:

- Normal Probability plot & Histogram of the Residuals
- Log of the event rate vs predictors

Hypothesis Testing Procedure:

H_0: *the Poisson model fits the* data

H_A: *the Poisson model does not fit the data*

Deviance test statistic: $D = \sum_{i=1}^{n} d_i^2$

Under null hypothesis, $D \sim \chi^2_{df}$ with $df = n\text{-}p\text{-}1$

Reject the null that the model is correct if $p\text{-value} = P(\chi^2_{df} > D)$ small.

Note that for this test, we want large p-values!!!!

Georgia Tech

What if No Goodness of Fit?

- Add predicting variables, consider interaction terms, or/and transform predicting variables to improve linearity;

- Identify unusual observations (outliers, leverage points);

- The Poisson distribution isn’t appropriate:
  - Overdispersion: the variability of the estimated rates is larger than would be implied by a Poisson model
    - Correlation in the observed responses
    - Heterogeneity in the rates that hasn’t been modeled

Overdispersion

Overdispersion: the variability of the response variable is larger than would be implied by the model

Binomial regression model:

- $\mathrm{V}(Y_i|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = n_i p(\mathrm{x}_{\mathrm{i}1}, \ldots, \mathrm{x}_{ip})(1\text{-}p(\mathrm{x}_{\mathrm{i}1}, \ldots, \mathrm{x}_{ip}))$
- Overdispersed Binomial: $\mathrm{V}(Y_i|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = \phi\, n_i p(\mathrm{x}_{\mathrm{i}1}, \ldots, \mathrm{x}_{ip})(1\text{-}p(\mathrm{x}_{\mathrm{i}1}, \ldots, \mathrm{x}_{ip}))$

Poisson regression model:

- $\mathrm{V}(Y_i|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = \lambda(\mathrm{x}_{\mathrm{i}1}, \ldots, \mathrm{x}_{ip})$
- Overdispersed Poisson: $\mathrm{V}(Y_i|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = \phi\lambda(\mathrm{x}_{\mathrm{i}1}, \ldots, \mathrm{x}_{ip})$

Overdispertion Parameter: ϕ

- Estimate: $\hat{\phi} = \frac{D}{n-p-1}$ where D is the sum of the squared deviances
- If $\hat{\phi} > 2$ then overdispersed model

Georgia Tech

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment: Data Examples

Georgia Tech

About This Lesson

Data Example 1: High School Awards

Objective: To model and predict the number of awards earned by students at one high school for multiple high schools.

Response Variable: The number of awards earned by students at a high school per *year*

Predicting Variables:

- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 ="Vocational"; and
- The score on the final exam in math.

Acknowledgement: This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.

Georgia Tech

Goodness-Of-Fit

```
## Deviance Test for GOF
with(m1, cbind(res.deviance = deviance, df = df.residual,
               p = 1- pchisq(deviance, df.residual)))
      res.deviance  df     p
[1,]  189.4496      196  0.6182274
```

Test for goodness-of-fit:

- Using deviance residuals: p-value = 0.61
- Do not reject the null hypothesis of good fit.

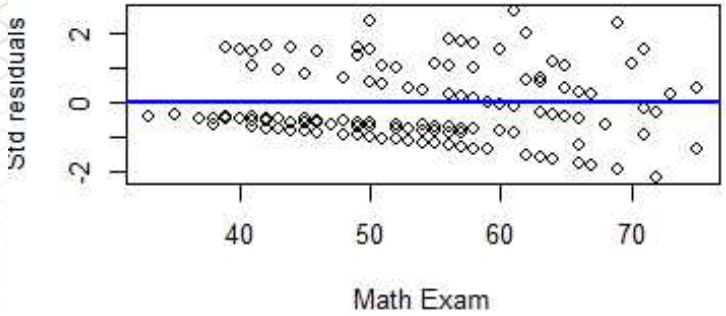
Georgia Tech

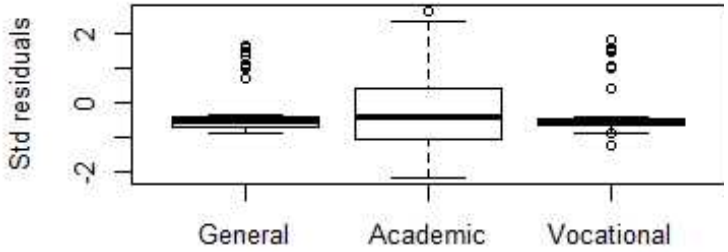
Residual Analysis

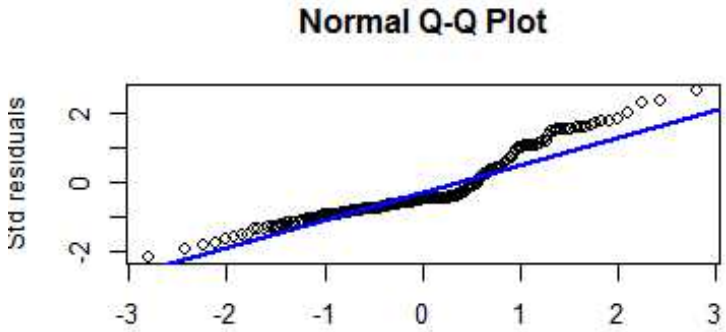
```
## Residual Plots
res = resid(m1,type="deviance")
par(mfrow=c(2,2))
plot(awardsdata$math,res,ylab="Std residuals",xlab="Math Exam")
abline(0,0,col="blue",lwd=2)
boxplot(res~prog,ylab = "Std residuals")
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")
```

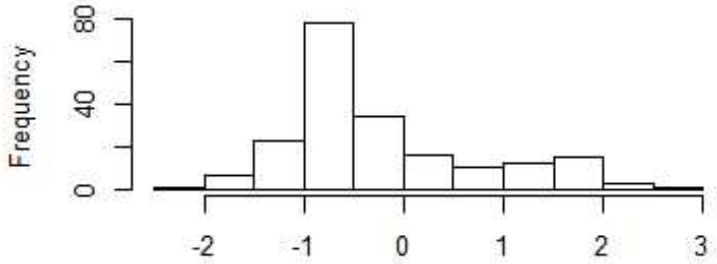
Georgia Tech

Residual Analysis









Georgia Tech

Modeling Nonlinear Relationships

```
## Fit a logistic regression model with math nonlinearly associated to awards count
library(mgcv)
m2 = gam(num_awards ~ prog + s(math), family="poisson", data=awardsdata)
```

- The residuals vs math: downward trend: Consider a **non-parametric** transformation of ‘math’ predicting variable
- *Nonparametric association*: not specifying the transformation but allowing the data to best identify/fit the transformation
- For this example, we do not see an improvement in the fit.

Summary

Georgia Tech

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Predicting Demand for Rental Bikes: Poisson Regression

Georgia Tech

About This Lesson

Predicting Demand for Rental Bikes

Bike sharing systems are of great interest due to their important role in traffic management.

Dataset: Historical data for years 2011-2012 for the bike sharing system in Washington D.C.

Data Source: UCI Machine Learning Repository

Acknowledgement*: This example was prepared with support from students in the Masters of Analytics program, including Naman Arora, Puneeth Banisetti, Mani Chandana Chalasani, Joseph (Mike) Tritchler and Kevin West*

Georgia Tech

Response & Predicting Variables

The response variable is:

Y (*Cnt*): Total bikes rented by both casual & registered users together

The qualitative predicting variables are:

Season: Season which the observation is made (1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall)

Yr: Year on which the observation is made

Mnth: Month on which the observation is made

Hr: Day on which the observation is made (0 through 23)

Holiday: Indictor of a public holiday or not (1 = public holiday, 0 = not a public holiday)

Weekday: Day of week (0 through 6)

Weathersit: Weather condition (1 = Clear, Few clouds, Partly cloudy, Partly cloudy, 2 = Mist & Cloudy, Mist & Broken clouds, Mist & Few clouds, Mist, 3 = Snow, Rain, Thunderstorm & Scattered clouds, Ice Pallets & Fog)

The quantitative predicting variables are:

Temp: Normalized temperature in Celsius

Atemp: Normalized feeling temperature in Celsius

Hum: Normalized humidity

Windspeed: Normalized wind speed

Georgia Tech

Poisson Regression Analysis in R

 # Applying Poisson regression model

*model1 = **glm**(cnt ~ ., data=train, family='poisson')*

***summary**(model1)*

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	2.93659	0.007629	384.941	<2e-16
season2	0.265486	0.004129	64.298	<2e-16
season3	0.255689	0.00473	54.059	<2e-16
season4	0.448706	0.004582	97.918	<2e-16
yr1	0.4684	0.001289	363.518	<2e-16
mnth2	0.115282	0.004247	27.143	<2e-16
mnth3	0.235149	0.004422	53.179	<2e-16
mnth4	0.210302	0.005857	35.909	<2e-16
mnth5	0.271895	0.006138	44.295	<2e-16
mnth6	0.2239	0.006247	35.84	<2e-16

⋮

Deviance Residuals:

Min	1Q	Median	3Q	Max
-24.6089	-3.7805	-0.8685	3.0436	22.6553

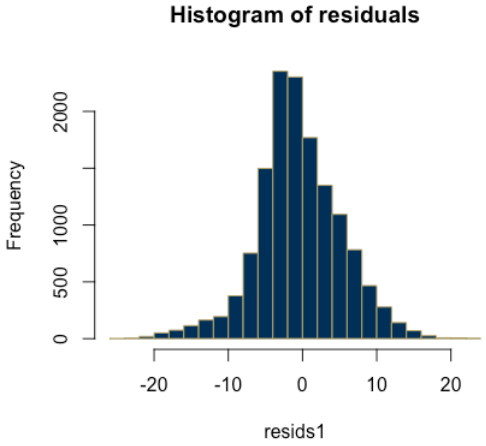
In the full output there are 51 predictor rows in addition to the intercept.

- All predicting variables are statistically significantly explaining the variability in the response (all p-values are small)
- Inflated statistical significance is also an issue in Poisson regression when the sample size is large

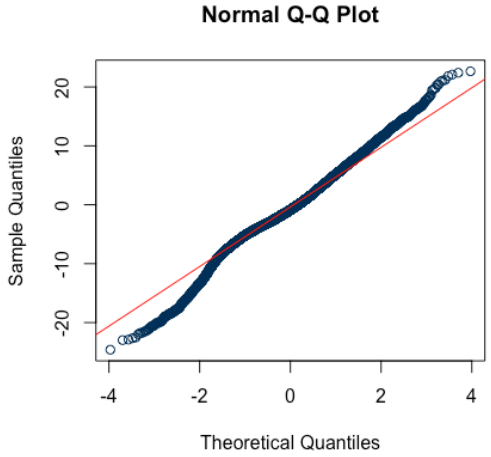
Georgia Tech

Goodness of Fit

```
## Checking normality
# histogram
hist(resids1,
     nclass=20,
     col=gtblue,
     border=techgold,
     main="Histogram of residuals")
```




```
# q-q plot
qqnorm(resids1,
       col="gtblue")
qqline(resids1,
       col="red")
```




```
# GOF Test
with(model1, cbind(res.deviance = deviance, df = df.residual,
                   p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

res.deviance	df	p
1458653.4	13851	0

Prediction

```
## Read New Data (Test Data)
test=data[-picked,]
test <- test[-c(1,2,9,15,16)]

## Prepare the test data the same as the training data
## Convert the numerical categorical variables to
predictors in the test data
test$season = as.factor(test$season)
test$yr = as.factor(test$yr)
test$mnth = as.factor(test$mnth)
test$hr = as.factor(test$hr)
test$holiday = as.factor(test$holiday)
test$weekday = as.factor(test$weekday)
test$weathersit = as.factor(test$weathersit)

## Build a prediction for model1 with the test data
# Specify whether a confidence or prediction interval
pred = predict(model1, test, interval = 'prediction')
```

Georgia Tech

Prediction Accuracy: Model 1

Save Predictions to compare with observed data

test.pred1 <- predict(model1, test, type='response')

 # Mean Squared Prediction Error (MSPE)

mean((test.pred1-test$cnt)^2)

[1] 8060.083

 # Mean Absolute Prediction Error (MAE)

mean(abs(test.pred1-test$cnt))

[1] 59.96461

 # Mean Absolute Percentage Error (MAPE)

mean(abs(test.pred1-test$cnt)/test$cnt)

[1] 0.8214892

 # Precision Measure (PM)

sum((test.pred1-test$cnt)^2)/sum((test$cnt-mean(test$cnt))^2)

[1] 0.2425596

Accuracy Measures

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i - \hat{Y}_i\right|}{|Y_i|}$$

$$\text{PM} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Prediction Accuracy

MSPE = 8060.08

MAE = 59.96

MAPE = 0.82

PM = 0.243

Georgia Tech

Model Comparison

Model	MSPE	MAE	MAPE	PM
Full MLR	10304.95	74.52	2.72	0.310
MLR Transformed	8955.41	62.69	0.80	0.271
Poisson Reg	8060.08	59.96	0.82	0.243

- **The Poisson regression models outperform the multi-variable linear regression models in terms of predictive power across most prediction measures except MAPE.**
- **While the GOF test rejects the null of good fit, the deviance residuals seem approximately normally distributed.**

Georgia Tech

Summary