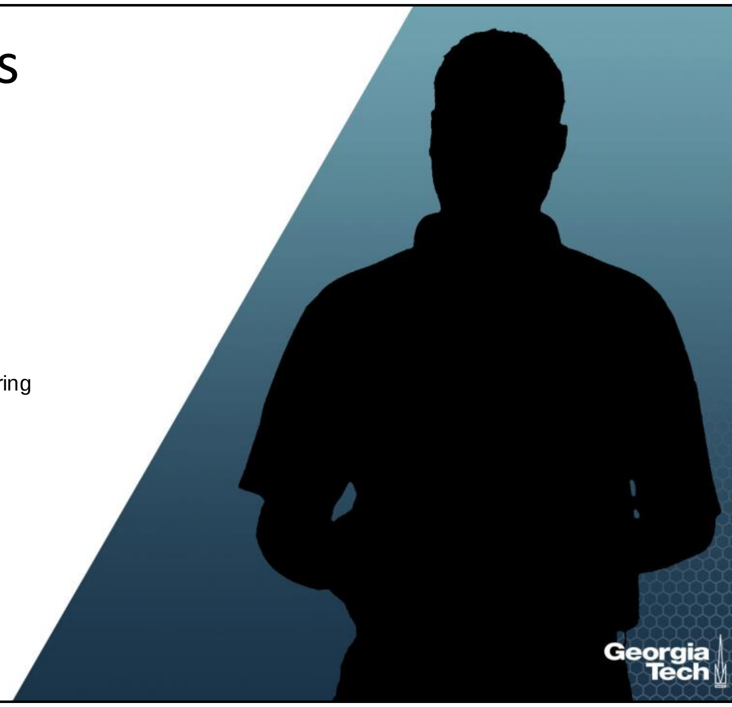
Regression Analysis

Poisson Regression



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

Model Estimation: Data Example

1

About This Lesson

2

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

3

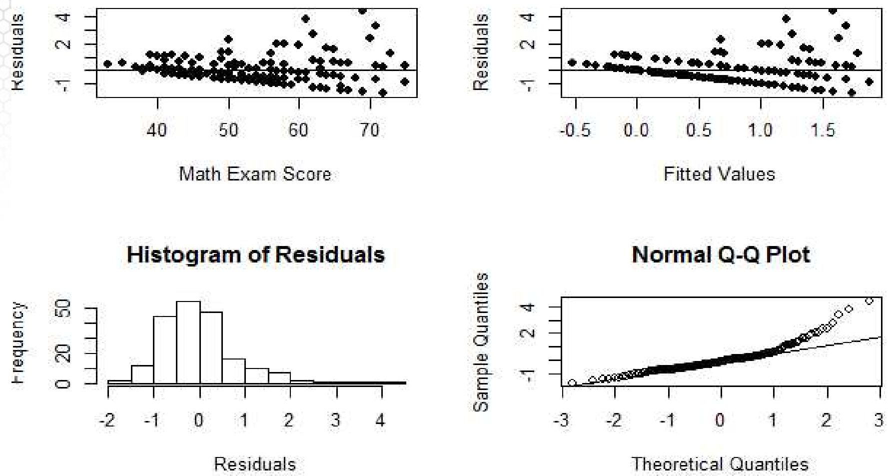
GOF: Standard Linear Regression

```
## 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)
```

Georgia Tech

4

GOF: Standard Linear Regression



5

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 ***

$\hat{\beta}_{math} = 0.07$: 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.

6

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 ***

$\boldsymbol{\beta_{academic}} = 1.084$: 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

7

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 <u>per</u> *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.

Georgia Tech

8

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:
  $\log(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \log(\text{exposure})$
- In this example, the number of policyholders is the exposure since the rate of claims is per policyholder (hence the unit).

Georgia Tech

9

Summary

Georgia Tech

10