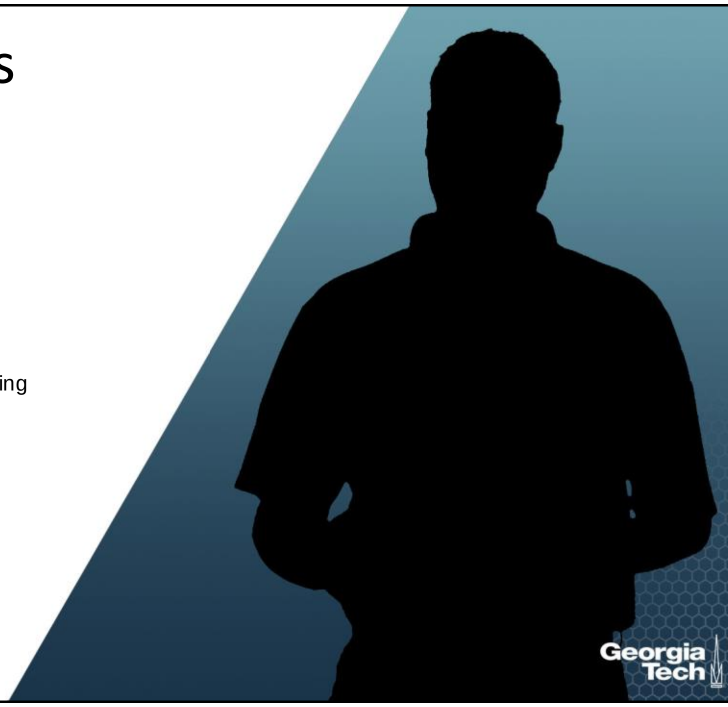
Regression Analysis

Poisson Regression



Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Statistical Inference: 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

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

4

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

5

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

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

Test for significance
$\boldsymbol{\beta}_{age.L}$ = -0.3944 or exp($\boldsymbol{\beta}_{age.L}$) =
$\boldsymbol{\beta}_{age.Q}$& $\boldsymbol{\beta}_{age.C}$: p-value>0.1 thus not statistically significant

Georgia Tech

6

Data Example 2: Statistical Inference (cont'd)

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

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

Test for overall regression: p-value ≈ 0 thus at least one predicting variables significantly explains the variability in the number of awards

7

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: p-value $= 0.002$ reject the null hypothesis that the District variable does have significant explanatory power

8

Summary

Georgia Tech

9