

Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.

Professor
Stewart School of Industrial and Systems Engineering

Data Examples



1

About This Lesson



2

Data Example 1: High School Awards

Objective: To model and predict the number of awards earned by students 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.



3

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)})
```

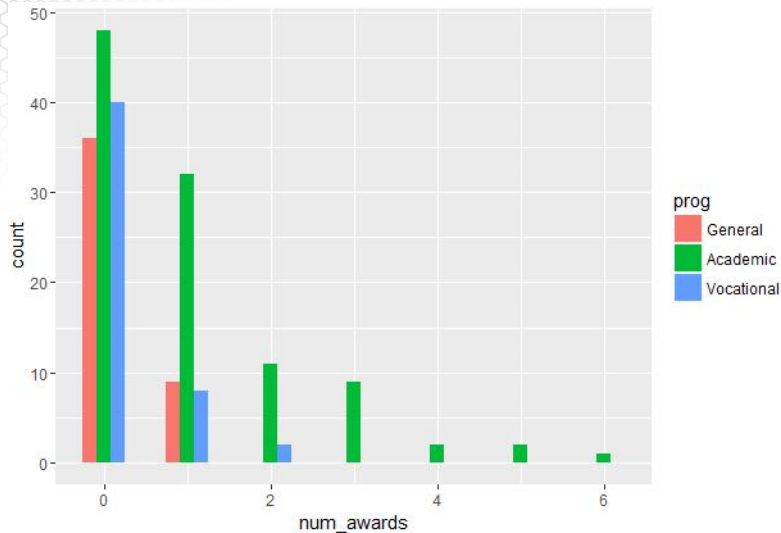
Conditional histograms

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



4

Exploratory Data Analysis



5

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.

6

Exploratory Data Analysis

Data in the R library MASS

```
library(MASS)
summary(Insurance)
```

Relationship between rate of claims and predictors

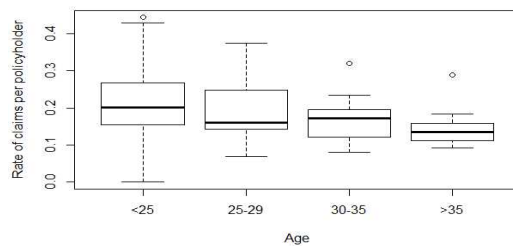
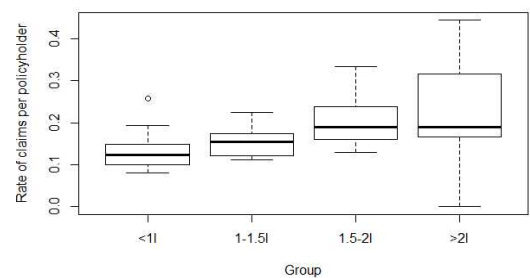
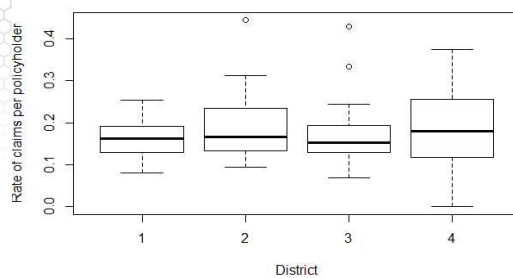
```
boxplot(Claims/Holders~District, xlab = "District", ylab = "Rate of claims per
policyholder", data=Insurance)
```

```
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)
```

7

Exploratory Data Analysis



8

Summary

