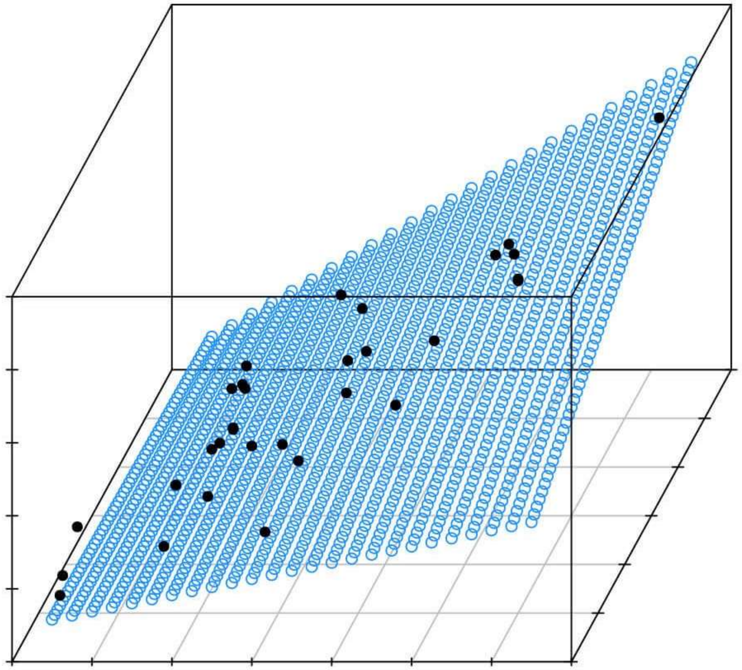
Regression Models with R Applications

by

Olga Korosteleva, Ph.D.
CSULB



October 13, 2020, OCRUG

ABOUT ME

- ❑ *BS in Mathematics, Wayne State University, Detroit, MI, 1996*
- ❑ *MS in Statistics, Purdue University, West Lafayette, IN, 1998*
- ❑ *Ph.D. in Statistics, Purdue University, West Lafayette, IN, 2002*
- ❑ *Professor of Statistics, CSU, Long Beach, 2002-present*

OUTLINE

- ❑ *Normal Linear Regression*
- ❑ *Gamma Regression*
- ❑ *Binary Logistic Regression*
- ❑ *Poisson Regression*

Greek Letters

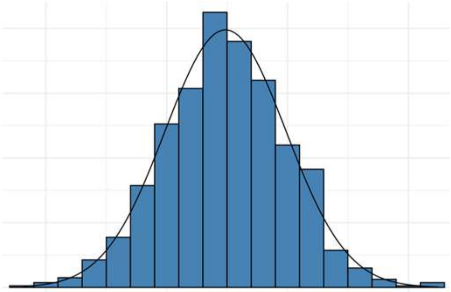
□ Alpha	α
□ Beta	β
□ Epsilon	ε
□ Lambda /lam-da/	λ
□ Pi	π
□ Sigma	σ

GENERAL LINEAR REGRESSION: OVERVIEW

- ❑ *General Linear Regression* model is $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$ where ε is a $N\left(0, \sigma^2\right)$ *random error*. Equivalently, y is a normally distributed random variable with mean $Ey = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and variance σ^2.

- ❑ Parameters are $\beta_0, \beta_1, \ldots \beta_k$, and σ^2.

- ❑ Fitted model is $\hat{E}y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$.



- ❑ Terminology: y is *response* (or *dependent variable*), $x_1, \ldots, x_k$ are *predictors* (or *independent variables*), β_0 is *intercept*, and $\beta_1, \ldots \beta_k$ are *slopes*.

GENERAL LINEAR REGRESSION: OVERVIEW (CONT.)

❑ Interpretation of fitted coefficients:

- If x_1 is continuous, $\hat{\beta}_1$ represents the change in the estimated mean of y for a one-unit increase in x_1, provided all the other variables are unchanged.

  Indeed,

  $\hat{E}y|_{x_1+1} - \hat{E}y|_{x_1} = \hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \cdots + \hat{\beta}_k x_k - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k) = \hat{\beta}_1$.
- If x_1 is a 0 - 1 variable, $\hat{\beta}_1$is interpreted as the difference of the estimated means of y for $x_1 = 1$ and $x_1 = 0$, controlling for the other predictors. Indeed,

  $\hat{E}y|_{x_1=1} - \hat{E}y|_{x_1=0} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \cdots + \hat{\beta}_k x_k - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \cdots + \hat{\beta}_k x_k) = \hat{\beta}_1$.

GENERAL LINEAR REGRESSION: OVERVIEW (CONT.)

- Prediction: For a given set of predictors $x_1^0, x_2^0, \dots, x_k^0$, the predicted response y^0 is computed as:

$$y^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0.$$

GENERAL LINEAR REGRESSION: EXAMPLE

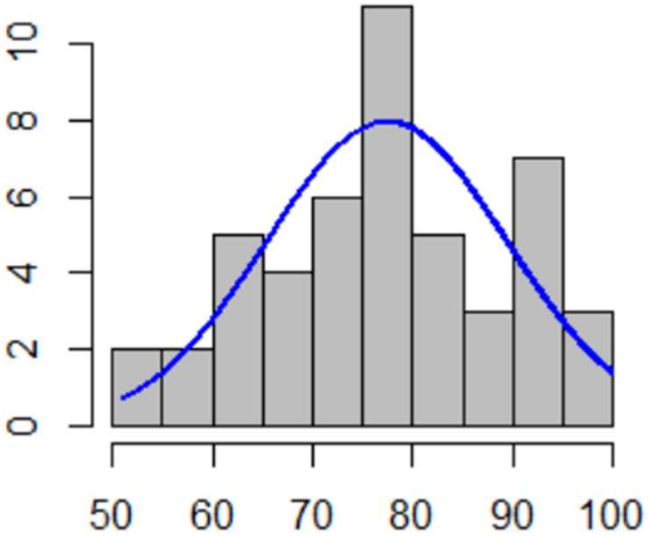
- A survey of 48 employees of a large company was conducted with the purpose of determining how satisfied they are with their jobs. Such demographic variables as gender, age, and education (Bachelor, Master, or Doctoral degree) were recorded. The total satisfaction score was calculated as a sum of scores on 20 questions on a 5-point Likert scale. We use these **data** to develop a regression model that relates the job satisfaction score to the other variables.

GENERAL LINEAR REGRESSION: EXAMPLE

- First, we plot the histogram for the scores.

```
job.satisfaction.data<- read.csv(file="./NormalExampleData.csv", header=TRUE, sep=",")
```

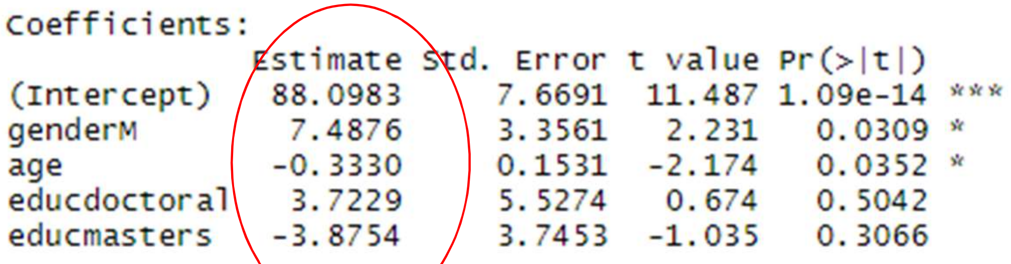
```
install.packages("rcompanion")
library(rcompanion)
plotNormalHistogram(job.satisfaction.data$score)
```



GENERAL LINEAR REGRESSION: EXAMPLE (CONT.)

- Next, we run the model.

```
summary(fitted.model<- glm(score ~ gender+ age + educ, data=job.satisfaction.data,
family=gaussian(link=identity))
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   88.0983     7.6691  11.487 1.09e-14 ***
genderM        7.4876     3.3561   2.231   0.0309 *  
age           -0.3330     0.1531  -2.174   0.0352 *  
educdoctoral   3.7229     5.5274   0.674   0.5042    
educmasters   -3.8754     3.7453  -1.035   0.3066    
```

```
print(sigma.hat<- sigma(fitted.model))
10.97801
```

- Then we write the fitted model.

$$\hat{E}(score) = 88.0983 + 7.4876 \cdot male - 0.3330 \cdot age + 3.7229 \cdot doctoral - 3.8754 \cdot masters$$

and $\hat{\sigma} = 10.97801.$

GENERAL LINEAR REGRESSION: EXAMPLE (CONT.)

$$\hat{E}(score) = 88.0983 + 7.4876 \cdot male - 0.3330 \cdot age + 3.7229 \cdot doctoral - 3.8754 \cdot masters$$

❑ Then we interpret the estimated regression coefficients.

- Gender: The estimated mean job satisfaction score for men is 7.4876 points larger than that for women.
- Age: With a one-year increase in age, the estimated average job satisfaction score is reduced by 0.333 points.
- Edu: For employees with doctoral degree, the estimated mean job satisfaction score is 3.7229 points larger than that for those with bachelor’s degree. For employees with Master’s degree, the estimated mean job satisfaction score is 3.8742 points lower than that for those with bachelor’s degree.

GENERAL LINEAR REGRESSION: EXAMPLE (CONT.)

- Finally, we use the fitted model for prediction of the job satisfaction score for a new female employee of this company who is 40 years of age and has a bachelor’s degree.

$$\mathit{predicted\ score}\ =\ 88.0983 -\ 0.3330 \cdot 40 = 74.7783.$$

```
print(predict(fitted.model, data.frame(gender="F", age=40, educ="bachelor")))
```

```
74.78019
```

GENERAL LINEAR REGRESSION: EXERCISE

- ❑ A cardiologist conducts a study to find out what factors are good predictors of elevated heart rate (HR) in her patients. She measures heart rate at rest in 30 patients on their next visit, and obtains from the medical charts additional **data** on their age, gender, ethnicity, body mass index (BMI), and the number of currently taken heart medications. She also obtains the air quality index (AQI) for the area of residence of her patients.

GENERAL LINEAR REGRESSION: EXERCISE (CONT.)

(1) Check normality of the heart rate measurements.

(2) Fit the general linear regression model. Write down the fitted model.

(3) Give interpretation of the estimated regression coefficients.

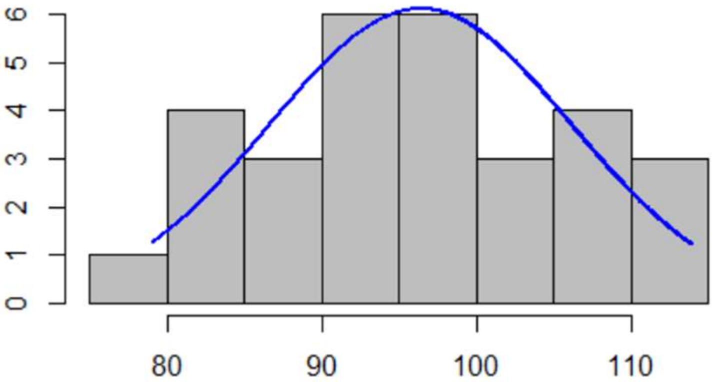
(4) Compute the predicted heart rate of a 50-year-old Hispanic male who has a BMI of 20, is not taking any heart medications, and resides in an area with a moderate air quality.

GENERAL LINEAR REGRESSION: EXERCISE SOLUTION

```
HR.data<- read.csv(file="./NormalExerciseData.csv", header=TRUE, sep=",")
```

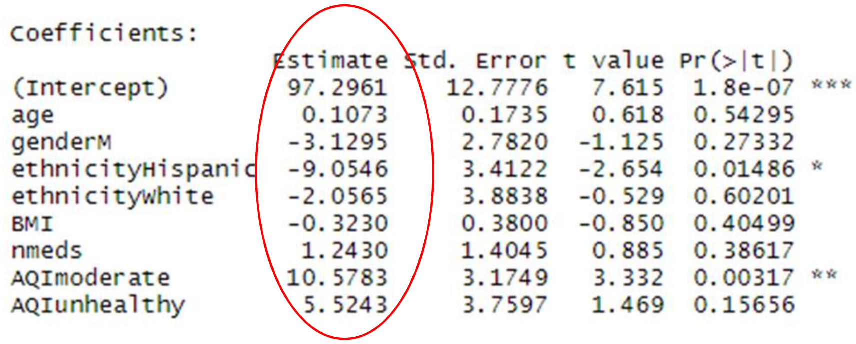
❑ Construct histogram

```
install.packages("rcompanion")
library(rcompanion)
plotNormalHistogram(HR.data$HR)
```



❑ Fit the model

```
summary(fitted.model<- glm(HR ~ age + gender + ethnicity + BMI + nmeds+AQI, data=HR.data,
family=gaussian(link=identity)))
```



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        97.2961    12.7776   7.615   1.8e-07 ***
age                 0.1073     0.1735   0.618   0.54295    
genderM            -3.1295     2.7820  -1.125   0.27332    
ethnicityHispanic  -9.0546     3.4122  -2.654   0.01486 *  
ethnicitywhite     -2.0565     3.8838  -0.529   0.60201    
BMI                -0.3230     0.3800  -0.850   0.40499    
nmeds               1.2430     1.4045   0.885   0.38617    
AQImoderate        10.5783     3.1749   3.332   0.00317 ** 
AQIunhealthy        5.5243     3.7597   1.469   0.15656    
```

```
print(sigma.hat<- sigma(fitted.model))
```

```
7.060314
```

GENERAL LINEAR REGRESSION: EXERCISE SOLUTION (CONT.)

❑ Write down the fitted model.

$$\hat{E}(HR) = 97.2961 + 0.1073 \cdot age - 3.1295 \cdot male - 9.0546 \cdot Hispanic \; - \; 2.0565 \cdot White \\ - 0.3230 \cdot BMI + 1.2430 \cdot nmeds + 10.5783 \cdot AQImoderate + 5.5243 \cdot AQIunhealthy$$

and $\hat{\sigma} = 7.060314$.

❑ Give interpretation of estimated regression coefficients. For example,

- Age: as age increases by one year, the estimated mean heart rate increases by 0.1073 units.
- Gender: the estimated average heart rate for males is 3.1295 points below that for females.

GENERAL LINEAR REGRESSION: EXERCISE SOLUTION (CONT.)

$$\hat{E}(HR) = 97.2961 + 0.1073 \cdot age - 3.1295 \cdot male - 9.0546 \cdot Hispanic \ - \ 2.0565 \cdot White \\ - 0.3230 \cdot BMI + 1.2430 \cdot nmeds + 10.5783 \cdot AQImoderate + 5.5243 \cdot AQIunhealthy$$

- Compute the predicted heart rate of a 50-year-old Hispanic male who has a BMI of 20, is not taking any heart medications, and resides in an area with a moderate air quality.

$$Predicted\ HR \ = 97.2961 + 0.1073 \cdot 50 \ - \ 3.1295 \ - \ 9.0546 \ - \ 0.3230 \cdot 20 + 10.5783 = 94.5953$$

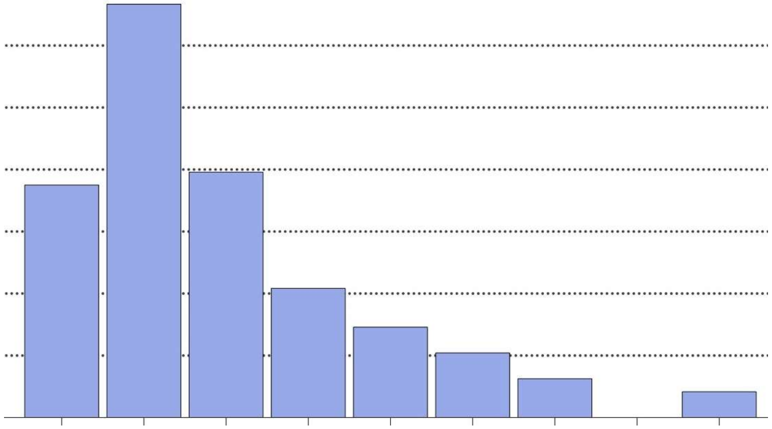
```
print(predict(fitted.model, data.frame(age=50, gender="M", ethnicity="Hispanic",
 BMI=20, nmeds=0,AQI="moderate")))

 94.59647
```

GENERALIZED LINEAR REGRESSION MODELS: THEORY

- ❑ Model response y as having a certain distribution defined by the setting.
- ❑ Model the mean Ey as a certain function of the linear regression term $g(Ey) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, where $g(.)$ is called a *link function*.
- ❑ Fitted model looks like this: $g\left(\hat{E}y\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$.
- ❑ Interpret the estimated regression coefficients as
  - If x_1 is continuous, $\hat{\beta}_1 = g\ (\hat{E}y)|_{x_1+1} - g\ (\hat{E}y)|_{x_1}$.
  - If x_1 is a 0 - 1 variable, $\hat{\beta}_1 = g\ (\hat{E}y)|_{x_1=1} - g\ (\hat{E}y)|_{x_1=0}$.
- ❑ Predict as $y^0 = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0\ )$.

GAMMA REGRESSION MODEL: THEORY

❑ If distribution of y is skewed to the right (has a long right tail), a *gamma regression* is appropriate.



- $f(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-y/\beta}, y > 0, \alpha, \beta > 0.$
- $Ey = \alpha\beta = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}.$
- Generalized linear model with a *log link* function: $ln(Ey) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$
- Parameters of the model are $\alpha, \beta_0, \beta_1, \ldots, \beta_k,$ where α is called a *scale* or *dispersion* parameter.

GAMMA REGRESSION MODEL: THEORY (CONT.)

❑ Interpretation of the estimated regression coefficients:

- If x_1 is continuous, $\left(e^{\widehat{\beta}_1}-1\right)\cdot 100\%$ represents percent change in the estimated mean response for a one-unit increase in x_1, provided all the other variables stay intact. Indeed,

$$\frac{\hat{E}y|_{x_1+1}-\hat{E}y|_{x_1}}{\hat{E}y|_{x_1}}\cdot 100\% = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1(x_1+1)+\cdots+\hat{\beta}_k x_k)-\exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k)}{\exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k)}\cdot 100\%$$

$$=\left(e^{\widehat{\beta}_1}-1\right)\cdot 100\%.$$

GAMMA REGRESSION MODEL: THEORY (CONT.)

- If x_1 is a 0 - 1 variable, $e^{\widehat{\beta}_1} \cdot 100\%$ represents percent ratio of the estimated mean responses for x_1=1 and x_1=0, controlling for the other predictors. Indeed,

$$\frac{\hat{E}y|_{x_1=1}}{\hat{E}y|_{x_1=0}} \cdot 100\% = \frac{\exp(\widehat{\beta}_0+\widehat{\beta}_1 \cdot 1+\cdots+\widehat{\beta}_k x_k)}{\exp(\widehat{\beta}_0+\widehat{\beta}_1 \cdot 0+\cdots+\widehat{\beta}_k x_k)} \cdot 100\% = e^{\widehat{\beta}_1} \cdot 100\%.$$

❑ Prediction: $y^0 = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0)$.

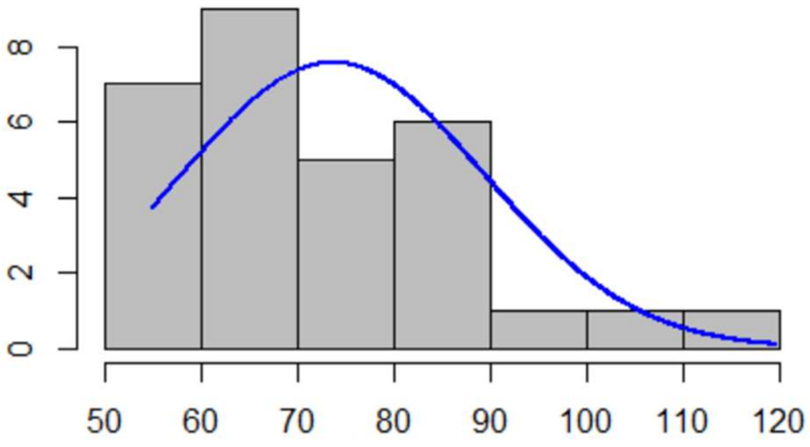
GAMMA REGRESSION MODEL: EXAMPLE

❑ A real estate specialist is interested in modeling house prices in a certain U.S. region. He suspects that house prices depend on such characteristics as the number of bedrooms, number of bathrooms, square footage of the house, type of heating (central/electrical/none), presence of air conditioner (A/C) (yes/no), and lot size. He obtains the data on 30 houses currently on the market.

GAMMA REGRESSION MODEL: EXAMPLE (CONT.)

❑ Construct histogram for house prices.

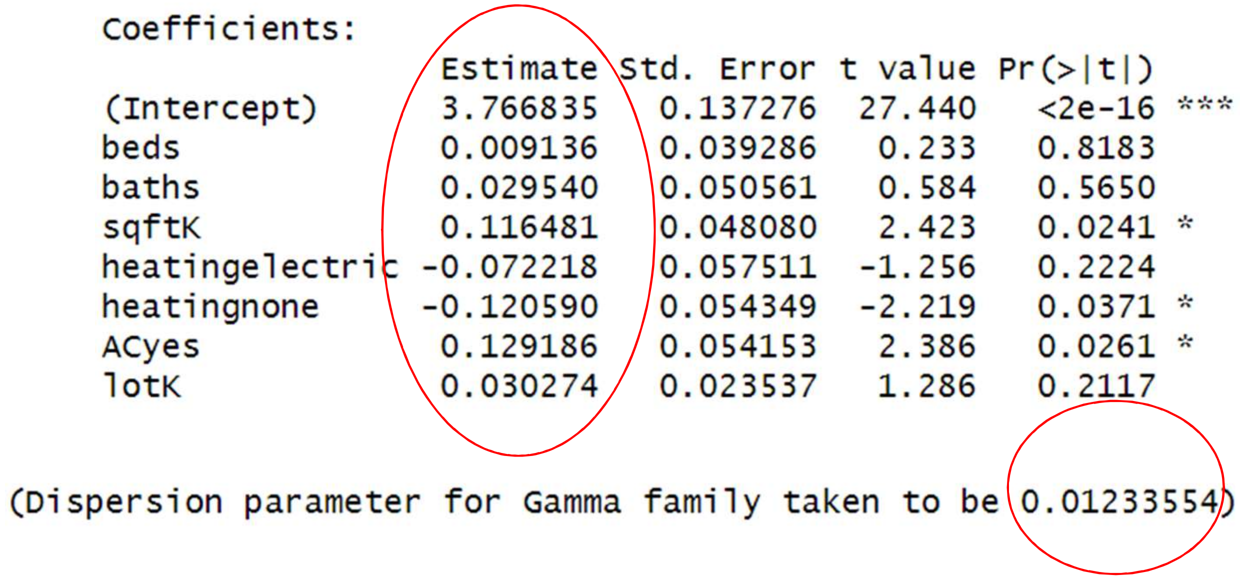
```
real.estate.data<- read.csv(file="./GammaExampleData.csv",  header=TRUE,sep=",")

#rescaling variables
price10K<- real.estate.data$price/10000
sqftK<-real.estate.data$sqft/1000
lotK<-real.estate.data$lot/1000

#plotting histogram with fitted normal density
install.packages("rcompanion")
library(rcompanion)
plotNormalHistogram(price10K)
```



GAMMA REGRESSION MODEL: EXAMPLE (CONT.)

❑ Fit a gamma regression model.

```
summary(fitted.model<- glm(price10K ~ beds + baths + sqftK + heating.rel
+ AC.rel + lotK, data=real.estate.data, family=Gamma(link=log)))
```



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.766835   0.137276  27.440   <2e-16 ***
beds             0.009136   0.039286   0.233   0.8183    
baths            0.029540   0.050561   0.584   0.5650    
sqftK            0.116481   0.048080   2.423   0.0241 *  
heatingelectric -0.072218   0.057511  -1.256   0.2224    
heatingnone     -0.120590   0.054349  -2.219   0.0371 *  
ACyes            0.129186   0.054153   2.386   0.0261 *  
lotK             0.030274   0.023537   1.286   0.2117    

(Dispersion parameter for Gamma family taken to be 0.01233554)
```

GAMMA REGRESSION MODEL: EXAMPLE (CONT.)

- Fitted gamma regression model is

$\hat{E}(price10K) = \exp\{3.7668 + 0.0091 \cdot beds + 0.0295 \cdot baths + 0.1165 \cdot sqftK - 0.0722 \cdot electric\_heater - 0.1206 \cdot no\_heater + 0.1292 \cdot A/C + 0.0303 \cdot lotK\}$, and $\hat{\alpha} = 0.0123$.

- Interpretation of the estimated coefficients. For example,

  - As the number of bedrooms increases by one, the estimated mean house price increases by $(\exp(0.0091) - 1) \cdot 100\% = 0.91\%$.

  - The estimated average price for air-conditioned houses is $\exp(0.1292) \cdot 100\% = 113.79\%$ of that for non-air-conditioned ones.

GAMMA REGRESSION MODEL: EXAMPLE (CONT.)

$$\hat{E}(price10K) = \exp\{3.7668 + 0.0091 \cdot beds + 0.0295 \cdot baths + 0.1165 \cdot sqftK - 0.0722 \cdot \ electric\_heater \ - 0.1206 \cdot no\_heater + 0.1292 \cdot A/C + 0.0303 \cdot lotK\}$$

- Predict the price of a house that has four bedrooms, two bathrooms, area of 1,680 squared feet, central heater, no A/C, and lot size of 5,000 squared feet.

$$price^0 = \$10{,}000 \cdot \exp(3.7668 + 0.0091 \cdot 4 + 0.0295 \cdot 2 + 0.1165 \cdot 1.68 + 0.0303 \cdot 5) = \$673{,}174.84.$$

```
print(10000*predict(fitted.model, type="response", data.frame(beds=4, baths=2,
sqftK=1.68, heating="central", AC="no", lotK=5)))
```

```
673237.9
```

GAMMA REGRESSION MODEL: EXERCISE

❑ Investigators at a large medical center conducted a quality improvement (QI) study which consisted of a six-month-long series of seminars and practical instructional tools on how to improve quality assurance for future projects at this center. Data were collected on participants' designation (nurse/doctor/staff), years of work at the center, whether had a prior experience with QI projects, and the score on the knowledge and attitude test taken at the end of the study. The score was constructed as the sum of 20 questions on a 5-point Likert scale, thus potentially ranging between 20 and 100. The large value indicates better knowledge about QI and more confidence and desire to use it in upcoming projects. The data on 45 study participants are available.

GAMMA REGRESSION MODEL: EXERCISE (CONT.)

(1) Check that the distribution of the response variable is right-skewed.

(2) Fit a gamma regression model. Write down the fitted model.

(3) Give interpretation of the estimated regression coefficients.

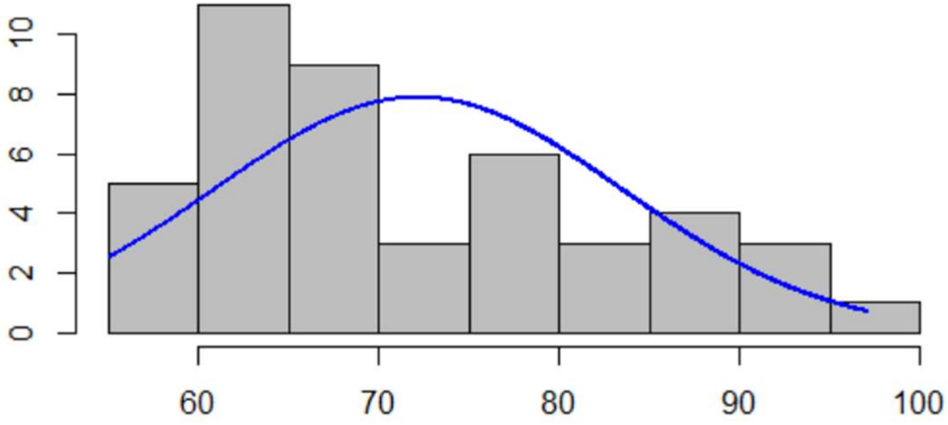
(4) Predict the score for a nurse who has worked at the center for seven years and who had previously been a co-PI on a grant that involved quality assurance component.

GAMMA REGRESSION MODEL: EXERCISE SOLUTION

```
QIscore.data<- read.csv(file="./GammaExerciseData.csv", header=TRUE, sep=",")
```

❑ Construct a histogram.

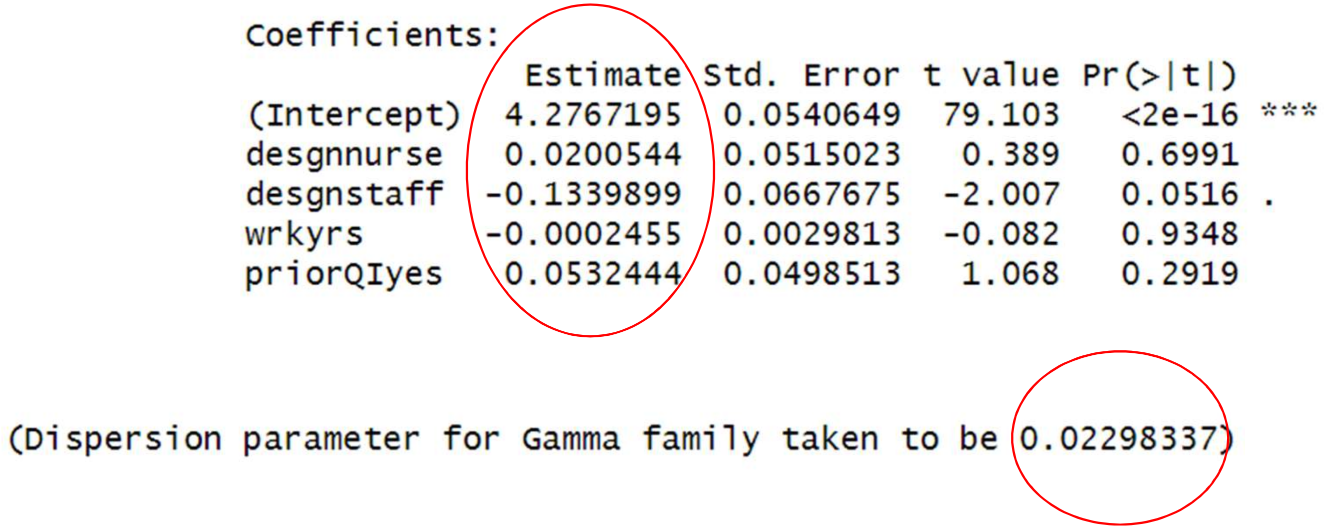
```
install.packages("rcompanion")
library(rcompanion)
plotNormalHistogram(QIscore.data$score)
```



GAMMA REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

- ❑ Fit a gamma regression model.

```
summary(fitted.model<- glm(score ~ desgn + wrkyrs + priorQI, data=QIscore.data,
family=Gamma(link=log)))
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.2767195  0.0540649  79.103   <2e-16 ***
desgnnurse   0.0200544  0.0515023   0.389   0.6991    
desgnstaff  -0.1339899  0.0667675  -2.007   0.0516 .  
wrkyrs      -0.0002455  0.0029813  -0.082   0.9348    
priorQIyes   0.0532444  0.0498513   1.068   0.2919    


(Dispersion parameter for Gamma family taken to be 0.02298337)
```

GAMMA REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

- The fitted gamma regression model is

$$\hat{E}(score) = \exp(4.2767 + 0.0201 \cdot nurse \; - 0.1340 \cdot staff - 0.0002 \cdot wrkyrs + 0.0532 \cdot QIyes),$$

and $\hat{\alpha} = 0.0230$.

- Interpretation of estimated coefficients. For example,

- The estimated mean score for nurses is $\exp(0.0201) \cdot 100\% = 102.03\%$ of that for doctors.
- As the number of years of work at the center increases by one, the estimated mean score changes by $(\exp(-0.0002) - 1) \cdot 100\% = -0.02\%$, that is, decreases by 0.02%.

GAMMA REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

$$\hat{E}(score) = \exp(4.2767 + 0.0201 \cdot nurse\ - 0.1340 \cdot staff - 0.0002 \cdot wrkyrs + 0.0532 \cdot QIyes)$$

- Predict the score for a nurse who has worked at the center for seven years and who had previously been a co-PI on a grant that involved quality assurance component.

$$score^0 = \exp(4.2767 + 0.0201 - 0.0002 \cdot 7 + 0.0532) = 77.37.$$

```
print(pred.score<- predict(fitted.model, type="response", data.frame(desgn="nurse",
wrkyrs=7, priorQI="yes")))
```

```
77.34687
```

BINARY LOGISTIC REGRESSION MODEL: THEORY

- ❑ Suppose $y = 1$ with probability $\pi = P(y = 1)$, and 0, otherwise. Then y has a *Bernoulli* (or *binary*) distribution with the mean $Ey = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi = P(y = 1)$.

This mean lies between 0 and 1, so we can relate it to the linear regression via the *logistic* function $\frac{\exp(x)}{1+\exp(x)}$:

$$\pi = P(y = 1) = \frac{Exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}{1+Ex\quad(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}.$$

Binary logistic regression model is the generalized linear model with the *logit* link function $g(x) = ln\frac{x}{1-x}$:

$$ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

BINARY LOGISTIC REGRESSION MODEL: THEORY (CONT.)

- ❑ Fitted model is $\hat{\pi} = \hat{P}(y = 1) = \frac{Exp(\widehat{\beta}_0+\widehat{\beta}_1 x_1+\cdots+\widehat{\beta}_k x_k)}{1+Ex\ \ (\widehat{\beta}_0+\widehat{\beta}_1 x_1+\cdots+\widehat{\beta}_k x_k)}$. Equivalently, the fitted *odds in favor of* $y = 1$ can be written as

$$\frac{\hat{\pi}}{1-\hat{\pi}} = Exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k).$$

- ❑ Interpretation:
  - If x_1 is continuous, as x_1 increases by one unit, the estimated odds change by $\frac{\widehat{odds}_{x_1+1}-\widehat{odds}_{x_1}}{\widehat{odds}_{x_1}} \cdot 100\% = \left(Exp\left(\hat{\beta}_1\right) - 1\right) \cdot 100\%$.
  - If x_1 is a 0 - 1 variable, the percent ratio of estimated odds for $x_1 = 1$ and x_1 =0 is $\frac{\widehat{odds}_{x_1=1}}{\widehat{odds}_{x_1=0}} \cdot 100\% = Exp\left(\hat{\beta}_1\right) \cdot 100\%$.

BINARY LOGISTIC REGRESSION MODEL: THEORY (CONT.)

- Prediction:

$$\pi^0 = \frac{\exp\{\widehat{\beta}_0+\widehat{\beta}_1 x_1^0+\cdots+\widehat{\beta}_k x_k^0\}}{1+\exp\{\widehat{\beta}_0+\widehat{\beta}_1 x_1^0+\cdots+\widehat{\beta}_k x_k^0\}}.$$

BINARY LOGISTIC REGRESSION MODEL: EXAMPLE

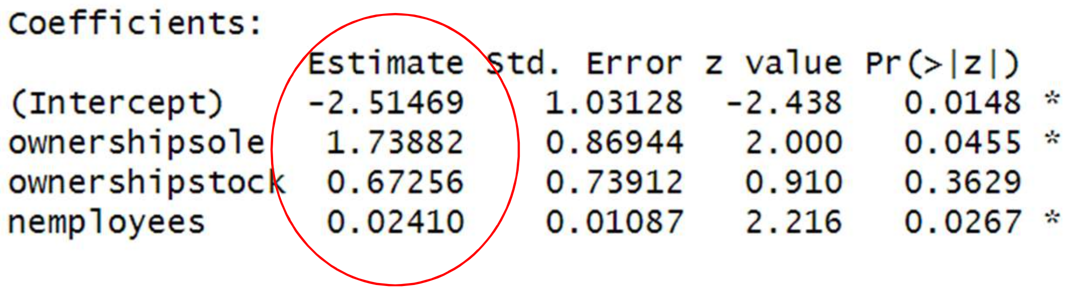
- ❑ A professor of organization and management is interested in studying the factors that influence the approach that company managers promote among their employees, competition or collaboration. The data on 50 companies are collected. The variables are the type of company ownership (sole ownership, partnership, or stock company), the number of employees, and the promoted approach (competition or collaboration). We model the probability of collaboration via the binary logistic regression.

BINARY LOGISTIC REGRESSION MODEL: EXAMPLE (CONT.)

```
companies.data<- read.csv(file="./LogisticExampleData.csv", header=TRUE, sep=",")
```

- Fitting a binary logistic regression model.

```
#specifying reference category
approach.rel<- relevel(companies.data$approach, ref="comp")

#fitting logistic model
summary(fitted.model<- glm(approach.rel ~ ownership + nemployees,
data=companies.data, family=binomial(link=logit)))
```



```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.51469    1.03128  -2.438   0.0148 *
ownershipsole   1.73882    0.86944   2.000   0.0455 *
ownershipstock  0.67256    0.73912   0.910   0.3629
nemployees      0.02410    0.01087   2.216   0.0267 *
```

BINARY LOGISTIC REGRESSION MODEL: EXAMPLE (CONT.)

❑ The fitted model is

$$\hat{P}(collaboration) = \frac{Exp(-2.5147+\ \ .7388\cdot sole+0.6726\cdot stock+0.0241\cdot nemployee\ \ )}{1+Exp(-2.5147+1.7388\cdot sole+\ \ .6726\cdot stock+0.0241\cdot nemployees)}.$$

❑ Interpretation of estimated regression coefficients. For example,

- For sole owned companies, the estimated odds in favor of collaboration are $\exp(1.7388)\cdot 100\% = 569.05\%$ of those for partnership companies.
- As the number of employees increases by one, the estimated odds in favor of collaboration increase by $(\exp(0.0241)-1)\cdot 100\% = 2.44\%$.

BINARY LOGISTIC REGRESSION MODEL: EXAMPLE (CONT.)

$$\hat{P}(collaboration) = \frac{exp(-2.5147+1.7388 \cdot sole+0.6726 \cdot stock+0.0241 \cdot nemployee\ \ )}{1+exp(-2.5147\quad .7388 \cdot sole+0.6726 \cdot stock+0.0241 \cdot nemployees)}$$

- ❑ Suppose the professor would like predict the probability of the collaborative approach in a solely owned company with 40 employees.

$$P^0(collaboration) = \frac{ex\ \ (-2.5147+\ \ .7388+\ \ .0241 \cdot\ \ \ )}{1+exp(-2.5147+1.7388+0.0241 \cdot 40)} = 0.5469.$$

```
print(predict(fitted.model, type="response", data.frame(ownership="sole",
nemployees=40)))
```

0.5468756

BINARY LOGISTIC REGRESSION MODEL: EXERCISE

- ❑ A bank needs to estimate the default rate of customers' home equity loans. The selected variables are loan-to-value (LTV) ratio defined as the ratio of a loan to the value of an asset purchased (in percent), age (in years), income (high/low), and response (yes=default, no=payoff). The data for 35 customers are available.

(1) Fit a binary logistic regression to model default.

(2) Give interpretation of the estimated regression coefficients.

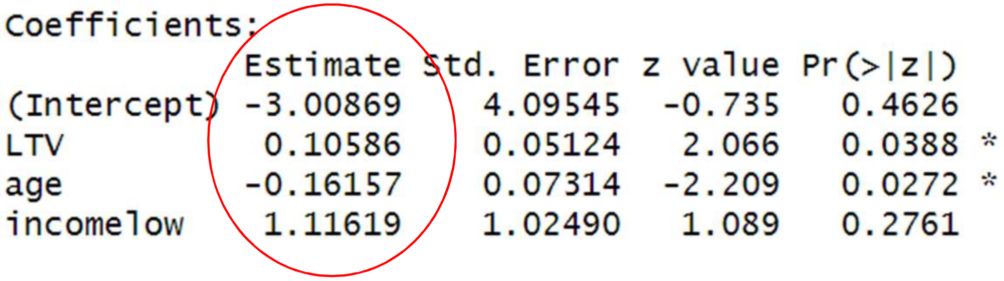
(3) Find predicted probability of loan default if the LTV ratio is 50%, and the borrower is a 50-year old man with a high income.

BINARY LOGISTIC REGRESSION MODEL: EXERCISE SOLUTION

```
rate.data<- read.csv(file="./LogisticExerciseData.csv", header=TRUE, sep=",")
```

❑ Fitting a binary logistic model.

```
#specifying reference category

default.rel<- relevel(rate.data$default, ref="No")

summary(fitted.model.logit<- glm(default.rel ~ LTV + age + income, data=rate.data,
family=binomial(link=logit)))
```



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.00869    4.09545  -0.735   0.4626
LTV          0.10586    0.05124   2.066   0.0388 *
age         -0.16157    0.07314  -2.209   0.0272 *
incomelow    1.11619    1.02490   1.089   0.2761
```

BINARY LOGISTIC REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

- ❑ The fitted model is

$$\hat{P}(default) = \frac{exp(-3.0087+0.1059\cdot LTV-0.1616\cdot age+1.1162\cdot low\_income)}{1+exp(-3.0087\quad .1059\cdot LTV-0.1616\cdot age+1.1162\cdot low\_income)}.$$

- ❑ Interpretation of estimated regression coefficients. For example,
  - As the LTV ratio increases by one percent, the estimated odds in favor of default increase by $(\exp(0.1059)-1)\cdot 100\% = 11.17\%$.
  - For people with low income, the estimated odds in favor of default are $\exp(1.1162)\cdot 100\% = 305.32\%$ of those for people with high income.

BINARY LOGISTIC REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

$$\hat{P}(default) = \frac{exp(-3.0087+\ .1059\cdot LTV-0.1616\cdot age+\ .1162\cdot low\_income)}{1+exp(-3.0087+\ .1059\cdot LTV-0.1616\cdot age+1.1162\cdot low\_income)}$$

- ❑ Find predicted probability of loan default if LTV ratio is 50%, and the borrower is a 50-year old men with a high income.

$$P^0(default) = \frac{ex\ \ (-3.0087+0.1059\cdot 50-0.1616\cdot\ \ \ )}{1+exp(-3.0087\ \ \ \ .1059\cdot 50-0.1616\cdot 50)} = 0.0030.$$

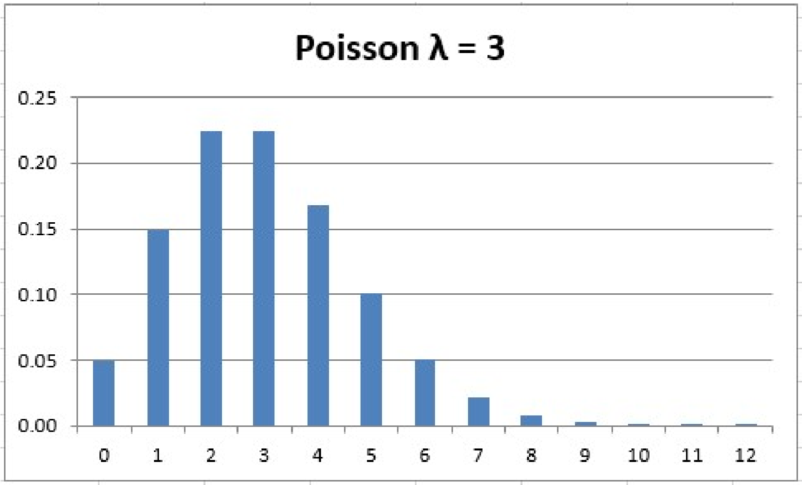
```
print(predict(fitted.model.logit, type="response", data.frame(LTV=50, age=50, 
income="high")))
```

```
0.00303576
```

POISSON REGRESSION MODEL: THEORY

- ❑ Suppose the response y assumes values 0, 1, 2, etc. The measurements like these are called *count data*.
- ❑ We can model y as having a Poisson distribution with mean λ and probability mass function

$$P(Y = y) = \frac{\lambda^y}{y!}e^{-\lambda}, y = 0, 1, 2, \ldots .$$



POISSON REGRESSION MODEL: THEORY

- We know that λ must be positive, thus we can model

$$\lambda = Ey = Exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

- A *Poisson regression* models y as having Poisson distribution, and the mean relating to the linear regression term through the *log* link function

$$ln(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

POISSON REGRESSION MODEL: THEORY (CONT.)

- The fitted model is $\hat{\lambda} = \hat{E}y = Exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\right)$.

- Prediction: $y^0 = Exp\ (\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0)$.

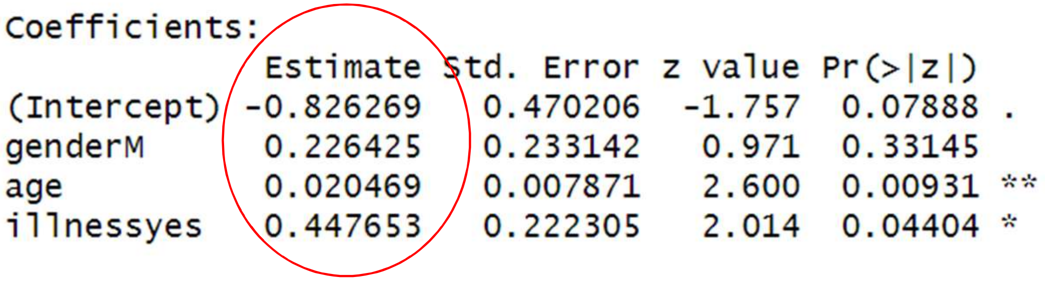
- Interpretation of estimated regression coefficients:
  - If x_1 is continuous, as x_1 increases by one unit, the estimated mean response changes by $\frac{\widehat{\lambda}_{x_1+1}-\widehat{\lambda}_{x_1}}{\widehat{\lambda}_{x_1}} \cdot 100\% = \left(Exp\left(\hat{\beta}_1\right) - 1\right) \cdot 100\%$.
  - If x_1 is a 0 -1 variable, the ratio of estimated mean responses for $x_1 = 1$ and x_1 =0 is $\frac{\widehat{\lambda}_{x_1=1}}{\widehat{\lambda}_{x_1=0}} \cdot 100\% = Exp\left(\hat{\beta}_1\right) \cdot 100\%$.

POISSON REGRESSION MODEL: EXAMPLE

❑ Number of days of hospital stay was recorded for 45 patients along with their gender, age, and history of chronical cardiac illness. The data are here.

❑ We fit the Poisson regression model.

```
hospitalstay.data<- read.csv(file="./PoissonExampleData.csv", header=TRUE, sep=",")

summary(fitted.model<- glm(days ~ gender + age + illness, data=hospitalstay.data,
family=poisson(link=log)))
```



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.826269   0.470206  -1.757  0.07888 .
genderM      0.226425   0.233142   0.971  0.33145
age          0.020469   0.007871   2.600  0.00931 **
illnessyes   0.447653   0.222305   2.014  0.04404 *
```

POISSON REGRESSION MODEL: EXAMPLE (CONT.)

- We write the fitted model as

$$\hat{\lambda} = \exp(-0.8263 + 0.2264 \cdot male + 0.0205 \cdot age + 0.4477 \cdot illness).$$

- We interpret the estimated regression coefficients. For example,

  - The estimated average length of hospital stay for males is $\exp(0.2264) \cdot 100\% = 125.41\%$ of that for females.

  - For a one-year increase in patient's age, the estimated average number of days of hospital stay increases by $(\exp(0.0205) - 1) \cdot 100\% = 2.07\%$.

POISSON REGRESSION MODEL: EXAMPLE (CONT.)

$$\hat{\lambda} = \exp(-0.8263 + 0.2264 \cdot male + 0.0205 \cdot age + 0.4477 \cdot illness).$$

- The predicted length of stay for a 55-year old male with no chronic cardiac illness is computed as

$$y^0 = exp(-0.8263+0.2264+0.0205 \cdot 55) = 1.6949.$$

```
print(predict(fitted.model, data.frame(gender="M", age=55, illness="no"),
type="response"))
```

```
1.692066
```

POISSON REGRESSION MODEL: EXERCISE

❑ A large automobile insurance company is studying the relation between the total number of auto accidents (including minor) that a policyholder had caused, and the policyholder's gender, age, and total number of miles driven (in thousands). The data for 45 randomly chosen policyholders are given here.

(1) Write down the fitted Poisson regression model.

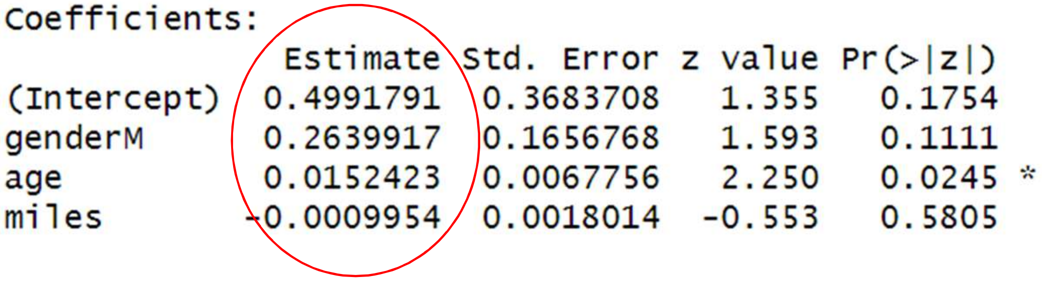
(2) Interpret estimated regression coefficients.

(3) Give a predicted value of the total number of auto accidents caused by a 35-year-old woman who has driven a total of one hundred thousand miles.

POISSON REGRESSION MODEL: EXERCISE SOLUTION

- ❑ We fit the Poisson regression model.

```
insurance.data<- read.csv(file="./PoissonExerciseData.csv", header=TRUE, sep=",")

summary(fitted.model<- glm(accidents ~ gender + age + miles, data=insurance.data,
family=poisson(link=log)))
```



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.4991791  0.3683708   1.355   0.1754
genderM      0.2639917  0.1656768   1.593   0.1111
age          0.0152423  0.0067756   2.250   0.0245 *
miles       -0.0009954  0.0018014  -0.553   0.5805
```

POISSON REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

- The fitted model is

$$\hat{\lambda} = exp(0.4992+0.2640 \cdot male + 0.0152 \cdot age - 0.00099 \cdot \ miles).$$

- Interpret the estimated regression coefficients. For example,

  - The estimated average number of auto accidents for males is $exp(0.2640) \cdot \ 100\% = 130.21\%$ of that for females.

  - For a one-year increase in policyholder's age, the estimated average number of auto accidents increases by $(exp(0.0152) - 1) \cdot 100\% \ = \ 1.53\%$.

POISSON REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

$$\hat{\lambda} = exp(0.4992{+}0.2640 \cdot male{+}\ 0.0152 \cdot age - 0.00099 \cdot\ miles)$$

- ❑ To give a predicted value of the total number of auto accidents caused by a 35-year-old woman who has driven a total of one hundred thousand miles, we compute:

$$y^0 = exp(0.4992{+}\ 0.0152 \cdot 35 - 0.00099 \cdot 100){=}2.5401.$$

```
print(predict(fitted.model, data.frame(gender="F", age=35, miles=100),
 type="response"))
```

```
2.542427
```

thank
you