

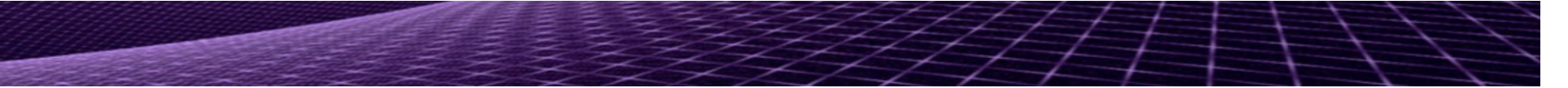


# **ADVANCED REGRESSION MODELS WITH R APPLICATIONS**

by

**Olga Korosteleva, Ph.D.**  
CSULB

Southern California R User Group  
October 1, 2022



## GENERAL LINEAR REGRESSION: OVERVIEW

---

- *General Linear Regression* model is

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$  where  $\varepsilon$  is a  $N(0, \sigma^2)$  *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  $\sigma^2$ .

- Parameters are  $\beta_0, \beta_1, \dots, \beta_k$ , and  $\sigma^2$ .

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

## 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,

$$\begin{aligned}\hat{E}y|_{x_1+1} - \hat{E}y|_{x_1} &= \hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \hat{\beta}_2x_2 \dots + \hat{\beta}_kx_k \\ &\quad - (\hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 \dots + \hat{\beta}_kx_k) = \hat{\beta}_1.\end{aligned}$$

- 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,

$$\begin{aligned}\hat{E}y|_{x_1=1} - \hat{E}y|_{x_1=0} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2x_2 \dots + \hat{\beta}_kx_k \\ &\quad - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2x_2 \dots + \hat{\beta}_kx_k) = \hat{\beta}_1.\end{aligned}$$

## 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:

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

---

## GENERAL LINEAR REGRESSION: EXAMPLE

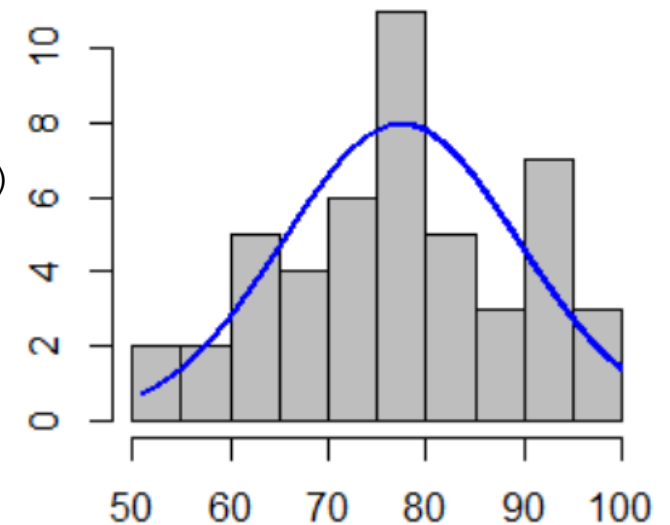
---

- 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
```

$$\hat{E}(\text{score}) = 88.0983 + 7.4876 * \text{male} - 0.3330 * \text{age} + 3.7229 * \text{doctoral} - 3.8742 * \text{masters}$$

## GENERAL LINEAR REGRESSION: EXAMPLE (CONT.)

$$\hat{E}(\text{score}) = 88.0983 + 7.4876 * \text{male} - 0.3330 * \text{age} + 3.7229 * \text{doctoral} - 3.8742 * \text{masters}$$

### □ 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.)

$$\hat{E}(\text{score}) = 88.0983 + 7.4876 * \text{male} - 0.3330 * \text{age} + 3.7229 * \text{doctoral} - 3.8742 * \text{masters}$$

- 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.

$$\text{predicted score} = 88.0983 - 0.3330 * 40 = 74.7783$$

```
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.)

---

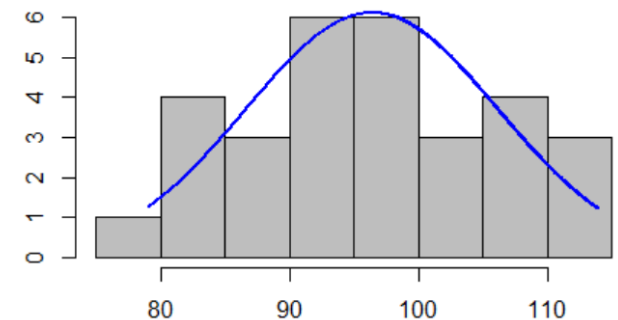
- (1) Check normality of the heart rate measurements.
  - (2) Fit the general linear regression model. Write down the fitted model.
  - (3) Give interpretation of the estimated regression coefficients.
  - (4) Compute the predicted heart rate of a 50-year-old Hispanic male who has a BMI of 20, is not 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:  
(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
```

## GENERAL LINEAR REGRESSION: EXERCISE SOLUTION (CONT.)

---

- Write down the fitted model.

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

- 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.)

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

- 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 =  $97.2961 + 0.1073 * 50 - 3.1295 - 9.0546 - 0.3230 * 20 + 10.5783 = 94.5953$

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

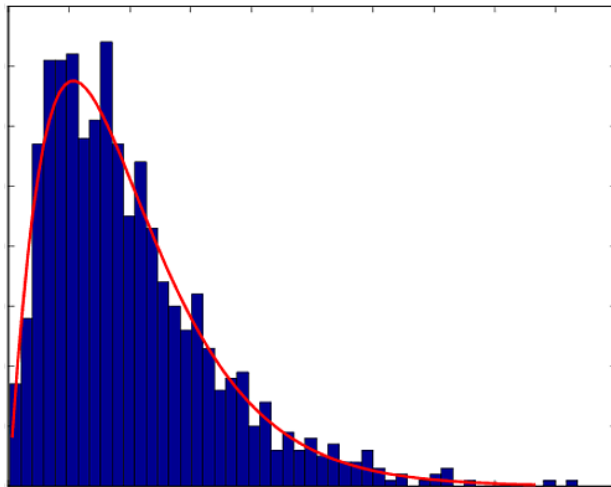
94.59647

## GENERALIZED LINEAR REGRESSION MODELS: THEORY

- Model response  $y$  as having a certain distribution defined by the setting.
- Model mean  $Ey$  as a certain function of linear regression term  $g(Ey) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ , where  $g(\cdot)$  is called a *link function*.
- Fitted model looks like:  $g(\hat{E}y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ .
- 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), gamma regression is appropriate.



- $f(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-y/\beta}, y > 0, \alpha, \beta > 0.$
- $Ey = \alpha\beta = \exp\{\beta_0 + \beta_1x_1 + \cdots + \beta_kx_k\}.$
- Generalized linear model with log link function:  
$$\ln(Ey) = \beta_0 + \beta_1x_1 + \cdots + \beta_kx_k.$$
- Parameters of the model are  $\alpha, \beta_0, \beta_1, \dots, \beta_k$ , where  $\alpha$  is called a *scale* or *dispersion* parameter.



## GAMMA REGRESSION MODEL: THEORY (CONT.)

### □ Interpretation of the estimated regression coefficients:

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

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

## GAMMA REGRESSION MODEL: THEORY (CONT.)

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

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

□ Prediction:  $y^0 = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \dots + \hat{\beta}_k x_k^0\}.$

## 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
```

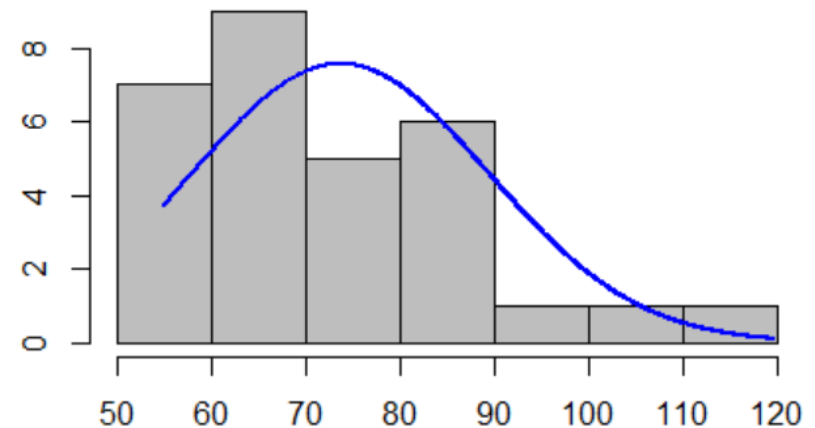
```
lotK<-real.estate.data$lot/1000
```

```
#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

$\hat{E}(\text{price10K}) = \exp\{3.7668 + 0.0091 * \text{beds} + 0.0295 * \text{baths} + 0.1165 * \text{sqftK} - 0.0722 * \text{elctrheater} - 0.1206 * \text{noheater} + 0.1292 * A/C + 0.0303 * \text{lotK}\}$ , and  $\hat{\alpha} = 0.0123$ .

□ Interpretation of 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  $e^{0.1292} \cdot 100\% = 113.79\%$  of that for non air-conditioned ones.

## GAMMA REGRESSION MODEL: EXAMPLE (CONT.)

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

$$\begin{aligned} price^0 &= \$10,000 * \exp\{3.7668 + 0.0091 * 4 + 0.0295 * 2 + 0.1165 * 1.68 \\ &\quad + 0.0303 * 5\} = \$673,174.84. \end{aligned}$$

```
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.)

---

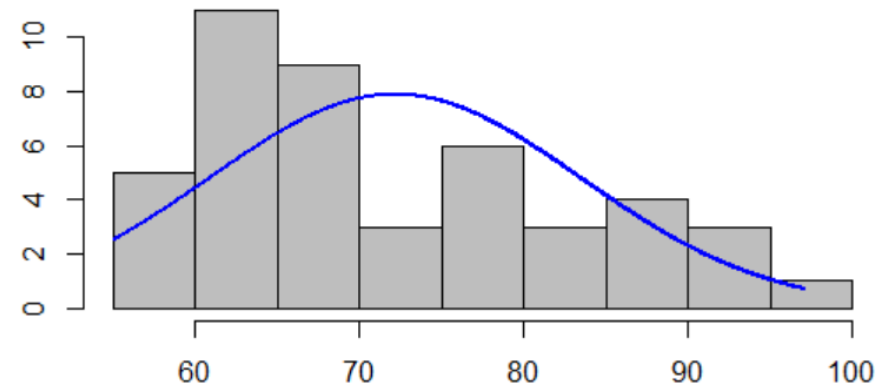
- (1) Check that the distribution of the response variable is right-skewed.
  - (2) Fit a gamma regression model. Write down the fitted model.
  - (3) Give interpretation of the estimated regression coefficients.
  - (4) Predict the score for a nurse who has worked at the center for seven years and who 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 ***
desgnurse	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

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

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  $(\exp\{-0.0002\} - 1) * 100\% = -0.02\%$ , that is, decreases by 0.02%.

## GAMMA REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

- 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.

$$score^0 = \exp\{4.2767 + 0.0201 - 0.0002 * 7 + 0.0532\} = 77.37.$$

```
print(pred.score<- predict(fitted.model, type="response",  
data.frame(design="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  $\frac{\exp(x)}{1+\exp(x)}$ :

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

*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  $\hat{\pi} = \hat{P}(y = 1) = \frac{\text{Exp}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k)}{1 + \text{Exp}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k)}$ . Equivalently, the fitted *odds in favor of*  $y = 1$  can be written as

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

- Interpretation:

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

$$x_1 = 1 \text{ and } x_1 = 0 \text{ is } \frac{\widehat{odds}_{x_1=1}}{\widehat{odds}_{x_1=0}} \cdot 100\% = \text{Exp}(\hat{\beta}_1) \cdot 100\%.$$

## BINARY LOGISTIC REGRESSION MODEL: THEORY (CONT.)

□ Prediction:

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



## 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

$$\hat{P}(\text{collaboration}) = \frac{\text{Exp}(-2.5147 + 1.7388 \cdot \text{sole} + 0.6726 \cdot \text{stock} + 0.0241 \cdot \text{employees})}{1 + \text{Exp}(-2.5147 + 1.7388 \cdot \text{sole} + 0.6726 \cdot \text{stock} + 0.0241 \cdot \text{employees})}.$$

□ 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 increase 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.)

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

$$P^0(collaboration) = \frac{\text{Exp}(-2.5147+1.7388+0.0241*40)}{1+\text{Exp}(-2.5147+1.7388+0.0241*40)}=0.5469.$$

```
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 LTV ratio is 50%, and the borrower is a 50-year old man with 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	*
income_low	1.11619	1.02490	1.089	0.2761	

## BINARY LOGISTIC REGRESSION MODEL: EXERCISE SOLUTION (CONT.)

□ The fitted model is

$$\hat{P}(\text{default}) = \frac{\text{Exp}(-3.0087 + 0.1059 \cdot \text{LTV} - 0.1616 \cdot \text{age} + 1.1162 \cdot \text{low\_income})}{1 + \text{Exp}(-3.0087 + 0.1059 \cdot \text{LTV} - 0.1616 \cdot \text{age} + 1.1162 \cdot \text{low\_income})}$$

□ Interpretation of estimated regression coefficients. For example,

- As 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.)

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

$$P^0(\text{default}) = \frac{\text{Exp}(-3.0087 + 0.1059 * 50 - 0.1616 * 50)}{1 + \text{Exp}(-3.0087 + 0.1059 * 50 - 0.1616 * 50)} = 0.0030.$$

```
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*.
- Suppose 0 is quite a common value and so is 1; 2 is more rare; 3, 4, 5 are even less frequent; 6, 7, 8 are very infrequent. Overall, we can model  $y$  as having a Poisson distribution with mean  $\lambda$  and probability mass function  $P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}$ ,  $y = 0, 1, 2, \dots$ .
- We know that  $\lambda$  must be positive, thus we can model
$$\lambda = Ey = \text{Exp}(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k).$$
- *Poisson regression* models  $y$  as having Poisson distribution, and the mean relating to the linear regression term through the *log* link function
$$\ln(\lambda) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

## POISSON REGRESSION MODEL: THEORY (CONT.)

- ❑ The fitted model is  $\hat{\lambda} = \hat{E}y = \text{Exp}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$ .
- ❑ Prediction:  $y^0 = \text{Exp}(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0)$ .
- ❑ Interpretation of estimated regression coefficients:
  - If  $x_1$  is continuous, as  $x_1$  increases by one unit, the estimated mean response changes by  $\frac{\hat{\lambda}_{x_1+1} - \hat{\lambda}_{x_1}}{\hat{\lambda}_{x_1}} \cdot 100\% = (\text{Exp}(\hat{\beta}_1) - 1) \cdot 100\%$ .
  - If  $x_1$  is a 0 -1 variable, the ratio of estimated mean responses for  $x_1 = 1$  and  $x_1 = 0$  is  $\frac{\hat{\lambda}_{x_1=1}}{\hat{\lambda}_{x_1=0}} \cdot 100\% = \text{Exp}(\hat{\beta}_1) \cdot 100\%$ .

## POISSON REGRESSION MODEL: EXAMPLE

- Number of days of hospital stay was recorded for 45 patients along with their gender, age, and history of chronic 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

$$\hat{\lambda} = \text{Exp}(-0.8263 + 0.2264 * \text{male} + 0.0205 * \text{age} + 0.4477 * \text{illness}).$$

□ 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.)

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

$$y^0 = \text{Exp}(-0.8263 + 0.2264 + 0.0205 * 55) = 1.6949.$$

```
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

$$\hat{\lambda} = \text{Exp}(0.4992 + 0.2640 * \text{male} + 0.0152 * \text{age} - 0.00099 * \text{miles}).$$

□ 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.)

---

- ❑ 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:

$$y^0 = \text{Exp}(0.4992 + 0.0152 * 35 - 0.00099 * 100) = 2.5401.$$

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

2.542427

---

## ZERO-INFLATED POISSON REGRESSION: THEORY

---

- Suppose  $y$  follows a Poisson distribution but too many zeros are observed. For example, suppose that one of the variables recorded during a health survey is the number of cigarettes the respondent smoked yesterday. Some respondents may have reported zero number of cigarettes smoked because they either do not smoke at all (*structural zero*), or they happened not to smoke a single cigarette that day (*chance zero*).
-

## ZERO-INFLATED POISSON REGRESSION: THEORY (CONT.)

- Then  $y$  can be modeled via a *zero-inflated Poisson (ZIP) regression* where the distribution of  $y$  is

$$\mathbb{P}(Y = y) = \begin{cases} \pi + (1 - \pi) \exp\{-\lambda\}, & \text{if } y = 0, \\ (1 - \pi) \frac{\lambda^y \exp\{-\lambda\}}{y!}, & \text{if } y = 1, 2, \dots, \end{cases}$$

where

$$\pi = \frac{\exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m\}},$$

and

$$\lambda = \exp\{\gamma_0 + \gamma_1 x_{m+1} + \dots + \gamma_{k-m} x_k\}.$$

- Note that the sets of predictors in the expressions for  $\pi$  and  $\lambda$  are chosen to be non-overlapping. This allows for interpretation of estimated regression coefficients. If we want only prediction, then the two sets can overlap.

## ZERO-INFLATED POISSON REGRESSION: THEORY (CONT.)

- The fitted ZIP model is

$$\hat{\pi} = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m\}},$$

$$\hat{\lambda} = \exp\{\hat{\gamma}_0 + \hat{\gamma}_1 x_{m+1} + \cdots + \hat{\gamma}_{k-m} x_k\}.$$

- The fitted mean is

$$\hat{E}y = (1 - \hat{\pi}) \cdot \hat{\lambda} = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_{m+1} + \cdots + \hat{\gamma}_{k-m} x_k)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m)}.$$

## ZERO-INFLATED POISSON REGRESSION: THEORY (CONT.)

### □ Interpretation of estimated regression coefficients:

- Probability of the structural zero  $\pi$  is modeled as in the binary logistic regression, thus, estimated beta coefficients are interpreted in terms of estimated odds.
- The mean of  $y$  is  $Ey = (1 - \pi) \cdot \lambda$ , and since we assume  $x$  variables are non-overlapping in  $\pi$  and  $\lambda$ , interpretation of gamma coefficients in  $\lambda$  is the same as in the Poisson regression model.

### □ Prediction:

$$y^0 = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_{m+1}^0 + \hat{\gamma}_{k-m} x_k^0)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_m x_m^0)}.$$

## ZERO-INFLATED POISSON REGRESSION: EXAMPLE

---

- ❑ A health survey has been administered to a random sample of 40 people aged between 25 and 50. Their gender, self-reported health condition (excellent or good), age, and the number of cigarettes they smoked yesterday were recorded. The data set is [here](#).
  - ❑ Since those respondents who don't smoke were included in the survey, it is expected that the number of cigarettes smoked would have a Poisson distribution with an inflated number of zeros.
-

## ZERO-INFLATED POISSON REGRESSION: EXAMPLE (CONT.)

---

- We fit a ZIP model where, for example, health condition is used as the predictor of structural zero, while gender and age are the count model predictors.

```
smoking.data<- read.csv(file="./ZIPExampleData.csv", header=TRUE,  
sep=",")  
install.packages("pscl")  
library(pscl)
```

```
summary(fitted.model<- zeroinfl(cigarettes ~ gender + age | health,  
data=smoking.data))
```

---

## ZERO-INFLATED POISSON REGRESSION: EXAMPLE (CONT.)

□ The output is

```
Count model coefficients (poisson with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.13809    0.83796  -0.165   0.8691
genderM      0.72684    0.28473   2.553   0.0107 *
age          0.01863    0.01996   0.933   0.3507
```

```
Zero-inflation model coefficients (binomial with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.1245     0.6806   1.652   0.0985 .
healthgood   -4.9195     2.3213  -2.119   0.0341 *
```

□ The fitted model is

$$\hat{\pi} = \frac{\exp\{1.1245 - 4.9195 * \text{good\_health}\}}{1 + \exp\{1.1245 - 4.9195 * \text{good\_health}\}}$$

$$\text{and } \hat{\lambda} = \exp\{-0.1381 + 0.7268 * \text{male} + 0.0186 * \text{age}\}$$



## ZERO-INFLATED POISSON REGRESSION: EXAMPLE (CONT.)

---

### □ Interpretation of estimated regression coefficients:

- The estimated odds of not smoking for people in excellent health is  $\exp\{4.9195\} \cdot 100\% = 13,694.26\%$  of those for people in good health.
  - As age increases by one year, the estimated average number of cigarettes smoked in a day increases by  $(\exp\{0.0186\}-1) \cdot 100\% = 1.88\%$ .
  - The estimated average number of cigarettes smoked in a day by men is  $\exp\{0.7268\} \cdot 100\% = 206.85\%$  of that by women.
-

## ZERO-INFLATED POISSON REGRESSION: EXAMPLE (CONT.)

- The predicted number of cigarettes smoked per day by a 50-year old male who is in good health is found as

$$y^0 = \frac{\exp(-0.1381 + 0.7268 + 0.0186 * 50)}{1 + \exp(1.1245 - 4.9195)} = 4.4659.$$

```
print(predict(fitted.model, data.frame(gender="M", health="good",  
age=50)))
```

4.473327

## ZERO-INFLATED POISSON REGRESSION: EXERCISE

---

- Thirty five patients in a large hospital were randomly chosen for a survey. The variables recorded were patient's BMI, age, gender, indicator of current smoking, and the number of mild to severe asthma attacks in the past three months. The data are [here](#).
- (1) Argue that a ZIP model is appropriate to model the number of asthma attacks.
  - (2) Give the fitted model.
  - (3) Interpret estimated regression coefficients.
  - (4) Calculate the predicted value for the number of asthma attacks for a male patient, aged 60, whose BMI is 21.2, and who is currently a smoker.
-

## ZERO-INFLATED POISSON REGRESSION: EXERCISE SOLUTION

---

- When no asthma attacks are observed in a patient, it could be a structural zero (the person never had an asthma attack) or a chance zero (the person happened not to have a single attack in the past three months).
- Fit the ZIP model.

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

```
install.packages("pscl")
```

```
library(pscl)
```

```
summary(fitted.model<- zeroinfl(attacks ~ BMI + age + gender | smoking,  
data=health.data))
```

---

## ZERO-INFLATED POISSON REGRESSION: EXERCISE SOLUTION (CONT.)

□ The output is

```
Count model coefficients (poisson with log link):  
      Estimate Std. Error z value Pr(>|z|)  
(Intercept)  2.22876    1.23213   1.809   0.0705 .  
BMI           -0.09537    0.04177  -2.283   0.0224 *  
age           0.01181    0.01012   1.168   0.2428  
genderM       0.71609    0.34639   2.067   0.0387 *
```

```
Zero-inflation model coefficients (binomial with logit link):  
      Estimate Std. Error z value Pr(>|z|)  
(Intercept)  -0.8700     0.7231  -1.203   0.229  
smokingyes   -9.3323    66.3837  -0.141   0.888
```

□ The fitted model looks like this:  $\hat{\pi} = \frac{\exp\{-0.8700 - 9.3323 * smoking\}}{1 + \exp\{-0.8700 - 9.3323 * smoking\}}$

and  $\hat{\lambda} = \exp\{2.2288 - 0.0954 * BMI + 0.0118 * age + 0.7161 * male\}$ .

## ZERO-INFLATED POISSON REGRESSION: EXERCISE SOLUTION (CONT.)

- Estimated regression coefficients yield the following interpretation (for example):

$$\hat{\pi} = \frac{\exp\{-0.8700 - 9.3323 * smoking\}}{1 + \exp\{-0.8700 - 9.3323 * smoking\}}$$

$$\hat{\lambda} = \exp\{2.2288 - 0.0954 * BMI + 0.0118 * age + 0.7161 * male\}.$$

- The estimated odds of not having asthma attacks for people in who smoke is  $\exp\{-9.3323\} \cdot 100\% = 0.009\%$  of those for people who don't smoke.
- As BMI increases by one point, the estimated average number of asthma attacks changes by  $(\exp\{-0.0954\}-1) \cdot 100\% = -9.10\%$ , that is, decreases by 9.1%.
- The estimated average number of asthma attacks for men is  $\exp\{0.7161\} \cdot 100\% = 204.64\%$  of that for women.

## ZERO-INFLATED POISSON REGRESSION: EXERCISE SOLUTION (CONT.)

- To predict the number of asthma attacks for a male patient, aged 60, whose BMI is 21.2, and who is currently a smoker, compute

$$y^0 = \frac{\exp\{2.2288 - 0.0954 * 21.2 + 0.0118 * 60 + 0.7161\}}{1 + \exp\{-0.8700 - 9.3323\}} = 5.1058.$$

```
print(predict(fitted.model, data.frame(BMI=21.2, age=60, gender="M",  
smoking="yes")))
```

5.112566

## BETA REGRESSION: THEORY

- When the response variable  $y$  assumes real values between 0 and 1, a beta regression is appropriate. The distribution of  $y$  is assumed to have density

$$f(y) = \frac{y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}}{B(\mu\phi, (1-\mu)\phi)}, \quad 0 < y < 1,$$

where the *location parameter*

$$\mu = \frac{\exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}},$$

and  $\phi$  is the *dispersion* parameter. The mean of  $y$  is  $\mu$  and variance  $\frac{\mu(1-\mu)}{\phi}$ .

- The fitted model is  $\hat{\mu} = \hat{E}(y) = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\}}$  and  $\hat{\phi}$ .



## BETA REGRESSION: THEORY (CONT.)

### □ Interpretation of estimated regression coefficients.

Note that  $\frac{\hat{\mu}}{1-\hat{\mu}} = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\}$ . Therefore, estimated regression coefficients can be interpreted as follows:

- For a one-unit increase in a numerical predictor  $x_1$ , the percent change in the estimated ratio  $\frac{\hat{\mu}}{1-\hat{\mu}} = \frac{\hat{E}(y)}{1-\hat{E}(y)}$  is  $(\exp\{\hat{\beta}_1\} - 1) \cdot 100\%$ , controlling for the other predictors.
- If  $x_1$  is a 0-1 variable, then  $\exp\{\hat{\beta}_1\} \cdot 100\%$  represents the percent ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  for  $x_1=1$  and  $x_1=0$ , keeping the other predictors fixed.

## BETA REGRESSION: THEORY (CONT.)

□ Prediction:

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

## BETA REGRESSION: EXAMPLE

---

- A professor of Library and Information Science has collected a random sample of 28 libraries and recorded total number of books each library has (in thousands), number of card holders (in thousands), library location (urban or rural), number of books checked out during a one-month period, and number of books that were returned on-time. The professor has calculated the proportion of books returned on-time as the ratio between the number of books returned on-time and number of books checked out, and is interested in studying associations between this proportion and the three predictor variables. The data are [here](#).
-

## BETA REGRESSION: EXAMPLE (CONT.)

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

□ We fit a beta regression model.

```
library(betareg)
```

```
summary(fitted.model<- betareg(propontime ~ nbooks + ncardholders +  
location, data=libraries.data, link="logit"))
```

```
Coefficients (mean model with logit link):  
      Estimate Std. Error z value Pr(>|z|)  
(Intercept) -0.20579    0.23095  -0.891   0.3729  
nbooks       0.02624    0.01307   2.008   0.0447 *  
ncardholders 0.04494    0.04300   1.045   0.2960  
locationurban 0.52357    0.23185   2.258   0.0239 *  
  
      Estimate  
(phi)    20.648
```

## BETA REGRESSION: EXAMPLE (CONT.)

□ The fitted model has

$$\hat{E}(\text{prop ontime}) = \frac{\exp\{-0.2058 + 0.0262 * \text{nbooks} + 0.0449 * \text{ncardholders} + 0.5236 * \text{urban}\}}{1 + \exp\{-0.2058 + 0.0262 * \text{nbooks} + 0.0449 * \text{ncardholders} + 0.5236 * \text{urban}\}},$$

and  $\hat{\phi} = 20.648$ .

□ Interpretation of the estimated regression coefficients. For example,

- As the number of books increases by one thousand, the estimated ratio of the mean proportion of books returned on-time and the mean proportion of books not returned on-time increases by  $(\exp\{0.0262\} - 1) \cdot 100\% = 2.64\%$ .
- This ratio for urban libraries is  $\exp\{0.5236\} \cdot 100\% = 168.81\%$  of that for rural libraries.

## BETA REGRESSION: EXAMPLE (CONT.)

- Suppose we would like to predict the proportion of books that are returned on time for a library in a rural area with 15,000 books and 2,500 card holders. We calculate

$$\text{prop on time}^0 = \frac{\exp\{-0.2058 + 0.0262 \cdot 15 + 0.0449 \cdot 2.5\}}{1 + \exp\{-0.2058 + 0.0262 \cdot 15 + 0.0449 \cdot 2.5\}} = 0.57448.$$

```
print(predict(fitted.model, data.frame(nbooks=15, ncardholders=2.5,
location="rural")))
```

0.5744721

## BETA REGRESSION: EXERCISE

- Ornithologists have collected data on migration of birds. They ringed 19 flocks of migratory birds prior to migration, and recorded for each flock the number of ringed birds, average mass (in kg), and average wingspan (in cm). The ringed flocks were observed later at the wintering grounds and the number of successfully migrated birds were counted. Proportion of successfully migrated birds was computed. The distances traveled (in Kkm) were also observed. The data are presented [here](#).
- (1) Fit a beta regression to model the proportion of successfully migrated birds (compute this variable first). Write down the fitted model.
  - (2) Interpret estimated regression coefficients.
  - (3) Predict the number of birds that successfully reach the winter grounds for a flock of birds with average mass of 600 g, average wingspan of 65 cm, that travel a distance of 1650 km, and typically travel in flocks of about 70 birds.

## BETA REGRESSION: EXERCISE SOLUTION

□ We fit the beta regression.

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

```
library(betareg)
```

```
summary(fitted.model<- betareg(propmigrated ~ mass + wingspan + distance,  
data = birds.data, link="logit"))
```

```
Coefficients (mean model with logit link):  
      Estimate Std. Error z value Pr(>|z|)  
(Intercept)  2.7830365  0.7820109   3.559 0.000373 ***  
mass         -0.0015016  0.0410881  -0.037 0.970848  
wingspan      0.0004487  0.0093935   0.048 0.961901  
distance     -1.3185658  0.4223698  -3.122 0.001797 **  
  
      Estimate  
(phi)      4.173
```



## BETA REGRESSION: EXERCISE SOLUTION (CONT.)

□ The fitted model is

$$\hat{E}(\text{prop migrated}) = \frac{\exp\{2.7830 - 0.0015 \cdot \text{mass} + 0.0004 \cdot \text{wingspan} - 1.3186 \cdot \text{distance}\}}{1 + \exp\{2.7830 - 0.0015 \cdot \text{mass} + 0.0004 \cdot \text{wingspan} - 1.3186 \cdot \text{distance}\}}$$

and  $\hat{\phi} = 4.173$ .

□ Interpretation of the estimated regression coefficients. For example,

- As the wing span increases by 1 cm, the estimated ratio of the mean proportion of successfully migrated birds and those who didn't succeed increases by  $(\exp\{0.0004\} - 1) \cdot 100\% = 0.04\%$ .
- If the distance increases by one thousand km, his ratio changes by  $(\exp\{-1.3186\} - 1) \cdot 100\% = -73.25\%$ , that is, decreases by 73.25%.

## BETA REGRESSION: EXERCISE SOLUTION (CONT.)

- (3) Predict the number of birds that successfully reach the winter grounds for a flock of birds with average mass of 600 g, average wingspan of 65 cm, that travel a distance of 1650 km, and typically travel in flocks of about 70 birds. We compute

$$\text{prop migrated}^0 \cdot 70 = \frac{\exp\{2.7830 - 0.0015 \cdot 0.6 + 0.0004 \cdot 65 - 1.3186 \cdot 1.65\}}{1 + \exp\{2.7830 - 0.0015 \cdot 0.6 + 0.0004 \cdot 65 - 1.3186 \cdot 1.65\}} \cdot 70$$
$$= (0.6530)(70) = 45.71.$$

```
prop.pred<- predict(fitted.model, data.frame(mass=0.6, wingspan=65,  
distance=1.65))
```

```
print(num.pred<- prop.pred*70)
```

45.76414

## LONGITUDINAL NORMAL REGRESSION: THEORY

- Suppose data are collected longitudinally at times  $t_1, \dots, t_p$ . For each individual  $i, i = 1, \dots, n$ , at time  $t_j, j = 1, \dots, p$ , the observations are  $x_{1ij}, \dots, x_{kij}$ , and  $y_{ij}$ .

The *random slope and intercept model* for a normally distributed response variable is defined as

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}$$

where  $u_{1i}$ 's are independent  $\mathcal{N}(0, \sigma_{u_1}^2)$  *random intercepts*,  $u_{2i}$ 's are independent  $\mathcal{N}(0, \sigma_{u_2}^2)$  *random slopes*, and  $\varepsilon_{ij}$ 's are independent  $\mathcal{N}(0, \sigma^2)$  *errors* that are also independent of  $u_{1i}$ 's and  $u_{2i}$ 's. It is assumed that  $\text{Cov}(u_{1i}, u_{2i}) = \sigma_{u_1 u_2}$ , and  $\text{Cov}(u_{1i}, u_{2i'}) = 0$  for  $i \neq i'$ .

## LONGITUDINAL NORMAL REGRESSION: THEORY (CONT.)

□ It can be shown that  $\text{Cov}(y_{ij}, y_{i'j'}) = 0$ ,  $i \neq i'$ , meaning that the responses for different individuals are uncorrelated for any time points. It can also be shown that responses between different time points for the same individual may be correlated, since  $\text{Cov}(y_{ij}, y_{ij'}) = \sigma_{u_1}^2 + \sigma_{u_1 u_2} (t_j + t_{j'}) + \sigma_{u_2}^2 t_j t_{j'}$ , for  $j \neq j'$ . In addition, it can be verified that the response variable  $y_{ij}$  is normally distributed with the mean

$$\mathbb{E}(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} t$$

and variance  $\text{Var}(y_{ij}) = \sigma_{u_1}^2 + 2\sigma_{u_1 u_2} t_j + \sigma_{u_2}^2 t_j^2 + \sigma^2$ .

## LONGITUDINAL NORMAL REGRESSION: THEORY (CONT.)

---

- The fitted model is

$$\hat{E}y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t.$$

- Estimated regression coefficients are interpreted the same as in a general linear regression model (for normal response).
- Likewise, predicted response is computed the same way as in a general linear regression model, that is,

$$y^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0 + \hat{\beta}_{k+1} t^0.$$

## LONGITUDINAL NORMAL REGRESSION: EXAMPLE

---

- ❑ In a clinic, doctors are testing a certain cholesterol lowering medication. Patients' gender and age at the beginning of the study are recorded for 27 patients. The low-density lipoprotein (LDL) cholesterol levels are measured in all the patients at the baseline, and then at 6-, 9-, and 24-month follow-up visits. The data are [here](#).
- ❑ We read in the data set.

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

## LONGITUDINAL NORMAL REGRESSION: EXAMPLE (CONT.)

---

- We create a long-form data set and a numeric variable for time.

```
library(reshape2)
```

```
longform.data<- melt(cholesterol.data, id.vars=c("id", "gender", "age"),  
variable.name = "LDLmonth", value.name="LDL")
```

```
#creating numeric variable for time
```

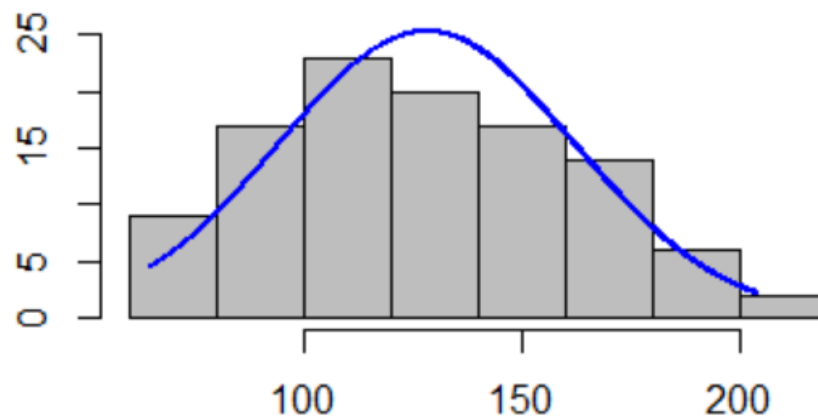
```
month<- ifelse(longform.data$LDLmonth=="LDL0", 0,  
ifelse(longform.data$LDLmonth=="LDL6", 6,  
ifelse(longform.data$LDLmonth=="LDL9", 9, 24)))
```

The output is [here](#).

## LONGITUDINAL NORMAL REGRESSION: EXAMPLE (CONT.)

- We plot a histogram to convince ourselves that LDL is a normally distributed variable.

```
library(rcompanion)
plotNormalHistogram(longform.data$LDL)
```





## LONGITUDINAL NORMAL REGRESSION: EXAMPLE (CONT.)

- We fit a random slope and intercept model.

```
library(nlme)
```

```
summary(fitted.model<- lme(LDL ~ gender+age+month,  
random =~ 1+month|id, data=longform.data))
```

	StdDev	Corr
(Intercept)	22.8072879	(Intr)
month	0.8857844	-0.812
Residual	8.3579458	

Fixed effects: LDL ~ gender + age + month					
	Value	Std.Error	DF	t-value	p-value
(Intercept)	94.82670	23.378841	80	4.056091	0.0001
genderM	-29.81056	6.971807	24	-4.275873	0.0003
age	0.92033	0.336820	24	2.732411	0.0116
month	-1.09566	0.193216	80	-5.670641	0.0000

## LONGITUDINAL NORMAL REGRESSION: EXAMPLE (CONT.)

---

□ The fitted model is

$$\hat{E}(LDL) = 94.8267 - 29.8106 * male + 0.9203 * age - 1.0957 * month.$$

□ Interpretation of the estimated regression coefficients. For example,

- As age increases by one year, the estimated mean LDL increases by 0.9203 points.
- The estimated average LDL for males is 29.8106 points lower than that for females.

## LONGITUDINAL NORMAL REGRESSION: EXAMPLE (CONT.)

---

- To predict the LDL level at 3 months for a 48-year-old female patient, we compute  $LDL^0 = 94.8267 + 0.9203 * 48 - 1.0957 * 3 = 135.7141$ .

```
print(predict(fitted.model, data.frame(gender="F", age=48, month=3),  
level=0))
```

135.7156

---

## LONGITUDINAL NORMAL REGRESSION: EXERCISE

---

- ☐ Measurements were taken on 20 people involved in a physical fitness course. The [data](#) contain participants' gender, age, oxygen intake (in ml per kg body weight per minute), run time (time to run 1 mile, in minutes), and pulse (average heart rate while running). The running was done under three different conditions: the first one on a treadmill, the second one on an indoor running track, and the third one on an outdoor running track.
  - ☐ Note that the data were collected not over time but under three different conditions. We call this type of data not longitudinal, but *repeated measures*. Random slope and intercept model still applies.
-

## LONGITUDINAL NORMAL REGRESSION: EXERCISE (CONT.)

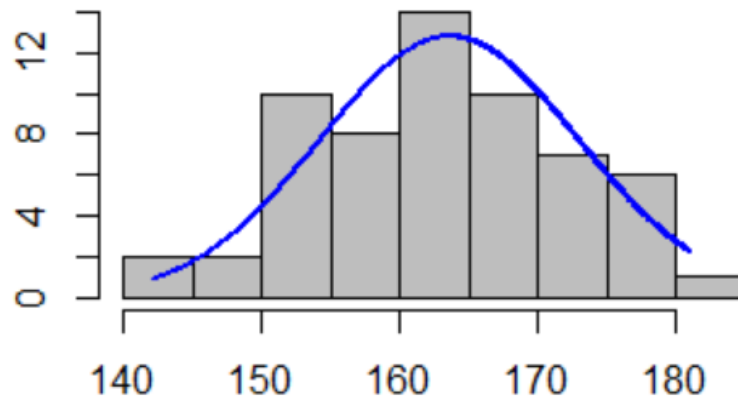
---

- (1) Study the code that creates a long-form data set and defines a numeric variable with condition (1=treadmill, 2=indoor track, or 3=outdoor track).
- (2) Plot a histogram for pulse to see that the underlying distribution is normal.
- (3) Fit a random slope and intercept model, regressing pulse of the rest of the variables. State the fitted model.
- (4) Interpret estimated regression coefficients.
- (5) Predict an average pulse for a 36-year-old woman who is running on a treadmill, if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile.

## LONGITUDINAL NORMAL REGRESSION: EXERCISE SOLUTION

□ We plot the histogram for pulse.

```
library(rcompanion)  
plotNormalHistogram(longform.data$pulse)
```



## LONGITUDINAL NORMAL REGRESSION: EXERCISE SOLUTION (CONT.)

### □ We fit the model.

```
library(nlme)
summary(fitted.model<- lme(pulse ~ gender + age + oxygen + runtime +
condition, random =~ 1 + condition|id, control=lmeControl(opt="optim"),
data=longform.data))
```

	StdDev	Corr
(Intercept)	9.0095266	(Intr)
condition	0.9085088	-0.941
Residual	5.6180510	

Fixed effects: pulse ~ gender + age + oxygen + runtime + condition					
	Value	Std.Error	DF	t-value	p-value
(Intercept)	167.18704	15.533284	37	10.763148	0.0000
genderM	-0.37349	3.641045	17	-0.102577	0.9195
age	-0.69932	0.326905	17	-2.139227	0.0472
oxygen	0.03058	0.226694	37	0.134881	0.8934
runtime	1.95905	0.968942	37	2.021844	0.0505
condition	0.75782	1.034507	37	0.732544	0.4685

## LONGITUDINAL NORMAL REGRESSION: EXERCISE SOLUTION (CONT.)

---

□ The fitted model is

$$\hat{E}(\text{pulse}) = 167.1870 - 0.3735 * \text{male} - 0.6993 * \text{age} + 0.0306 * \text{oxygen} \\ + 1.9591 * \text{runtime} + 0.7578 * \text{condition}$$

□ Interpretation of the estimated regression coefficients. For example,

- As age increases by one year, the estimated average pulse decreases by 0.6993 points.
- The estimated average pulse for males is 0.3735 points less than that for females.



## LONGITUDINAL NORMAL REGRESSION: EXERCISE SOLUTION (CONT.)

---

□ We predict an average pulse for a 36-year-old woman who is running on a treadmill, if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile. We calculate

$$\begin{aligned} pulse^0 &= 167.1870 - 0.6993 * 36 + 0.0306 * 40.2 \\ &\quad + 1.9591 * 10.3 + 0.7578 * 1 = 164.17885. \end{aligned}$$

```
print(predict(fitted.model, data.frame(id=21, gender="F", age=36,  
condition=1, oxygen=40.2, runtime=10.3), level=0))
```

164.1766

---

## LONGITUDINAL LOGISTIC REGRESSION: THEORY

---

It is no extra work to run a random slope and intercept model for a generalized linear regression (for example, gamma, logistic, Poisson, or beta). The fitted model would look like

$$g(\hat{\mathbb{E}}(y)) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t$$

where  $g(\cdot)$  is the link function that corresponds to the underlying distribution.

---

## LONGITUDINAL LOGISTIC REGRESSION: EXAMPLE

---

- A pharmaceutical company conducted a dosage trial for a painkiller medication. Two dosages (A and B) were identified for a long-run safety check. An experiment was set up with 14 subjects in each of the two groups, taking dosages A and B, respectively. The goal of the experiment was to identify which dosage is less likely to cause side effects. The [data](#) set contains participants' IDs, dosage (A or B), gender, and presence or absence of side effects (1=present, or 0=absent) at 1, 3, 7, and 16 weeks.
-

## LONGITUDINAL LOGISTIC REGRESSION: EXAMPLE (CONT.)

---

- We create a long-form [data](#) set and a time variable

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

#creating longform dataset and time variable
library(reshape2)
longform<- melt(dosages.data, id.vars=c("patid","dosage", "gender"),
variable.name="weekn", value.name="effects")
week<- ifelse(longform$weekn=="week1",1,ifelse(longform$weekn=="week3",3,
ifelse(longform$weekn=="week7",7,16)))
```

## LONGITUDINAL LOGISTIC REGRESSION: EXAMPLE (CONT.)

- We run logistic regression model on the long-form data set.

```
library(lme4)
summary(fitted.model<- glmer(effects~dosage+gender+week+(1+(week|patid)),
data=longform,family=binomial(link='logit')))
```

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-9.1020	5.6114	-1.622	0.1048
dosageB	4.6269	3.1336	1.477	0.1398
genderM	0.1295	1.5427	0.084	0.9331
week	1.0423	0.6122	1.703	0.0886

## LONGITUDINAL LOGISTIC REGRESSION: EXAMPLE (CONT.)

- The fitted model is

$$\begin{aligned} & \hat{P}(\text{side effects}) \\ &= \frac{\exp\{-9.1020 + 4.6269 * \text{dosage } B + 0.1295 * \text{male} + 1.0423 * \text{week}\}}{1 + \exp\{-9.1020 + 4.6269 * \text{dosage } B + 0.1295 * \text{male} + 1.0423 * \text{week}\}} \end{aligned}$$

- Interpretation of the estimated regression coefficients is like in a logistic regression (in terms of odds in favor of side effects). Important observation: The estimated odds in favor of side effects for dosage B are  $\exp\{4.6269\} \cdot 100\% = 10219.68\%$  of those for dosage A. The study conclusion is that dosage A outperforms dosage B.

## LONGITUDINAL LOGISTIC REGRESSION: EXAMPLE (CONT.)

---

- Suppose we want to predict the probability of a side effect occurring at week 7 for a woman taking dosage A. We compute

$$P^0(\text{side effect}) = \frac{\exp\{-9.1020 + 1.0423 \cdot 7\}}{1 + \exp\{-9.1020 + 1.0423 \cdot 7\}} = 0.1411.$$

```
print(predict(fitted.model, data.frame(patid=29, dosage="A", gender="F",  
week=7), re.form=NA, type="response"))
```

0.1411383

---

## LONGITUDINAL POISSON REGRESSION: EXERCISE

---

- A dermatologist tests a new ointment treatment for psoriasis. He administers the ointment to five patients, and keeps five patients as control. The control patients take a medication that is commonly prescribed against the disease. The doctor sees the patients next day, then after one week, two weeks, five weeks, and finally, after three months. He records the number of psoriatic patches that are visible on patients' bodies. The data are [here](#).
-



## LONGITUDINAL POISSON REGRESSION: EXERCISE (CONT.)

---

- (1) Study the code that creates a long-form data set and defines a numeric variable for time (in weeks).
  - (2) Fit a random slope and intercept model Poisson regression to model the number of psoriatic patches. State the fitted model.
  - (3) Interpret estimated regression coefficients.
  - (4) Predict the number of psoriatic patches for a patient in the treatment group at five weeks.
-

## LONGITUDINAL POISSON REGRESSION: EXERCISE SOLUTION

□ We fit the model.

```
library(lme4)

summary(fitted.model<- glmer(npatches ~ group + weeks + (1 + weeks|patid),
data=longform.data, family=poisson(link="log")))
```

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	2.45862	0.21107	11.649	< 2e-16	***
groupTx	0.19144	0.33729	0.568	0.57031	
weeks	-0.15682	0.05667	-2.767	0.00565	**

## LONGITUDINAL POISSON REGRESSION: EXERCISE SOLUTION (CONT.)

---

□ The fitted model is

$$\hat{E}(npatches) = \exp\{2.4586 + 0.1914 * Tx - 0.1568 * weeks\}.$$

□ Interpretation of the estimated coefficients:

- The estimated average number of psoriatic patches for patients in the treatment group is  $\exp\{0.1914\} \cdot 100\% = 121.09\%$  of that for patients in the control group.
- As the time increases by one week, the estimated average number of psoriatic patches changes by  $(\exp\{-0.1568\} - 1) \cdot 100\% = -14.51\%$ , that is, decreases by 14.51%.

## LONGITUDINAL POISSON REGRESSION: EXERCISE SOLUTION (CONT.)

---

- To predict the number of psoriatic patches for a patient in the treatment group at five weeks, we compute

$$npatches^0 = \exp\{2.4586 + 0.1914 - 0.1568 * 5\} = 6.4624.$$

```
print(predict(fitted.model, data.frame(patid=11, group="Tx", weeks=5),  
re.form=NA, type="response"))
```

6.462278

---

## HIERARCHICAL REGRESSION FOR NORMAL RESPONSE: THEORY

---

- ❑ IDEA: Suppose data are collected for school children. We cannot model all responses as completely independent from each other. We expect children within the same classroom (same teacher) to have correlated responses. Also, children within the same school will have correlated responses. So, we need our regression to be able model these dependences (at level 1 = school, level 2= classrooms within the same school, and level 3= children within the same classroom). This can be done with hierarchical regression model with three levels of hierarchy.
- ❑ We will look at the case when we have several clusters, several individuals within each cluster, and longitudinally collected data for each individual.

## HIERARCHICAL REGRESSION FOR NORMAL RESPONSE: THEORY (CONT.)

---

- We model a normal response for individual  $i$ , in cluster  $m$ , at time  $t_j$  as

$$y_{ijm} = \beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}$$

where betas are *fixed effects*, and the other are *random effects*.

- Fitted model is  $\hat{E}y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t$ .
  - Interpretation and prediction is done like in the general linear model.
-

## HIERARCHICAL REGRESSION FOR NORMAL RESPONSE: EXAMPLE

---

- ❑ Mothers and daughters from 24 families with signs of depression were invited for a study of efficacy of a new method of intensive psychotherapy. At the baseline, one- and three-month visits, the quality of life (QOL) questionnaire was filled out by each of the participant, and a QOL score was computed. Higher values of this score indicate better quality of life. Whether signs of depression were present was also recorded (1=present, or 0=absent). The study was done on mother-daughter dyads. This type of study is called *familial* or *dyadic*. The scores are logically expected to be correlated over time for each individual, and also members of the same family might have correlated responses. We fit a three-stage hierarchical model for these [data](#).
-

## HIERARCHICAL REGRESSION FOR NORMAL RESPONSE: EXAMPLE (CONT.)

- We create a long-form data set and run the regression.

```
summary(fitted.model<- lmer(gol ~ relation + depression + visit  
+ (1 + visit|family)+ (1 + visit|family:individual),  
control=lmerControl(calc.derivs = FALSE), data=longform.data))
```

### Random effects:

Groups	Name	Variance	Std.Dev.	Corr
family:individual	(Intercept)	0.53806	0.7335	
	visit	0.03392	0.1842	-1.00
family	(Intercept)	0.28573	0.5345	
	visit	0.05603	0.2367	-0.90
Residual		0.36030	0.6003	

Number of obs: 153, groups: family:individual, 51; family, 24

### Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	2.704732	0.284817	9.496
relationM	0.603270	0.136033	4.435
depression	0.001397	0.146239	0.010
visit	0.271363	0.096004	2.827



## HIERARCHICAL REGRESSION FOR NORMAL RESPONSE: EXERCISE

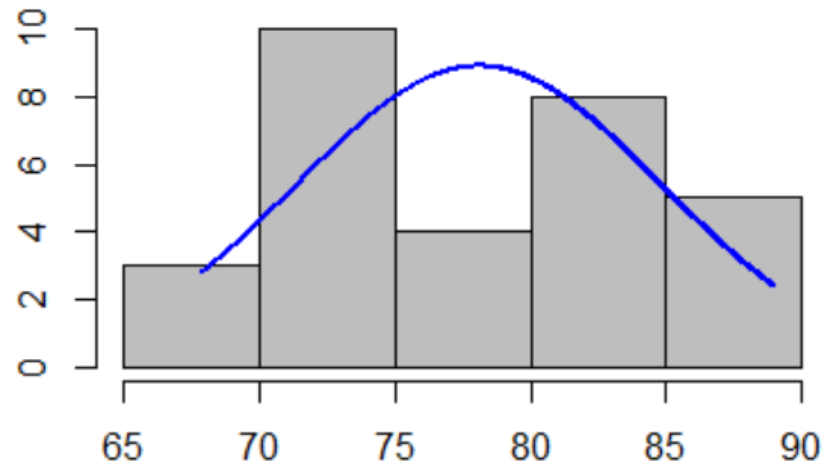
---

- A team of school inspectors is studying scores on tests in English Language Arts (ELA), Mathematics, and Science and their relation to schools' Academic Performance Index (API) and classroom size. [Data](#) on average classroom scores for five consecutive years at two schools are available. Note that the data are already in the long-form format.
- (1) Identify the three levels of hierarchy in this model.
  - (2) Plot a histogram for API to confirm that it has normal distribution.
  - (3) Fit a proper hierarchical regression to model scores.
-

## HIERARCHICAL NORMAL REGRESSION: EXERCISE SOLUTION

- The levels of hierarchy are: schools (level 1), subjects within school (level 2), and different years within each subject (level 3).
- Plot a histogram for API to confirm that it has normal distribution.

```
library(rcompanion)  
plotNormalHistogram(school$score)
```



## HIERARCHICAL NORMAL REGRESSION: EXERCISE SOLUTION (CONT.)

- Fit a proper hierarchical regression to model API.

```
library(lme4)
summary(fitted.model<- lmer(score ~ API + classsize + year
+ (1 + year|school) + (1 + year|school:subject), data=school))
```

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-6.93233	20.00680	-0.346
API	0.08582	0.02100	4.087
classsize	-0.06865	0.18869	-0.364
year	1.04652	0.35906	2.915



*THANK YOU!*

