# Lab 7 BIS 505b

Maria Ciarleglio

4/12/2021


- Goal of Lab 7
- Analysis Data Set
- Research Questions
- Summarizing Count and Rate Data
    - Numerical Summaries
    - Graphical Summary - Barplot
- Poisson Regression of the Count
    - Simple Poisson Regression of the Count
    - Fitted Mean Count
    - Multiple Poisson Regression of the Count
- Poisson Regression of the Rate
    - Simple Poisson Regression of the Rate
    - Fitted Rate
    - Multiple Poisson Regression of the Rate
- Likelihood Ratio Test
- Checking for Overdispersion
- Negative Binomial Regression of the Count or Rate


# Goal of Lab 7

In **Lab 7**, we will analyze a **count endpoint** using **Poisson regression.** We will begin by **(1)** numerically and graphically summarizing the response variable. Next, we will model the **(2)** count and the **(3)** rate outcome using a Poisson model. We will consider binary, continuous, and categorical predictors with $> 2$ levels. We will **(4)** predict the mean count and the event rate from our fitted models, **(5)** determine if overdispersion is a problem in the data, and **(6)** fit a negative binomial regression model.

# Analysis Data Set

In this lab, we will analyze data on the number of office-based doctor visits by adults aged 25-64 ($n$ = 4408) contained in `visits.csv` . This data set is imported as the data frame `vis` in code chunk 3 above. The **Data Key** is provided below. In this lab, our endpoint of interest is number of office visits ( `visits` ). The length of follow-up in months for each patient is recorded in the `followup` variable.

| Variable Name | Definition |
|---|---|
| `visits` | Number of office-based doctor visits (**Our Response**) |
| `followup` | Follow-up period (months) |
| `age` | Age (years) |
| `educ` | Years of schooling (12 = High school) |
| `income` | Annual income (thousands of dollars) |

| Variable Name | Definition |
|---|---|
| `female` | Indicator of female sex |
| | 0 = Male |
| | 1 = Female |
| `black` | Indicator of black race |
| | 0 = Non-black |
| | 1 = Black |
| `hispanic` | Indicator of Hispanic ethnicity |
| | 0 = Non-Hispanic |
| | 1 = Hispanic |
| `location` | Location |
| | 0 = West |
| | 1 = Northeast |
| | 2 = Midwest |
| | 3 = South |
| `insured` | Health insurance status |
| | 0 = Uninsured |
| | 1 = Insured |
| `chronic` | Health status |
| | 0 = No chronic conditions |
| | 1 = Chronic conditions |

- **Creating Factor Variables:**

Next, we create factor variable versions of the **categorical variables** in this data set ( `female` , `black` , `hispanic` , `location` , `private` , and `chronic` ). The **first level** specified in the `factor()` function is the **reference level** of the variable.

```
# Creating factor variables in vis using mutate() function in "dplyr" package
vis <- mutate(vis,
              female_factor = factor(female,
                                     levels = c(0, 1),
                                     labels = c("Male", "Female")),
              black_factor = factor(black,
                                    levels = c(0, 1),
                                    labels = c("Non-black", "Black")),
              hispanic_factor = factor(hispanic,
                                       levels = c(0, 1),
                                       labels = c("Non-Hispanic", "Hispanic")),
              location_factor = factor(location,
                                       levels = c(0, 1, 2, 3),
                                       labels = c("West", "Northeast", "Midwest", "South")),
              insured_factor = factor(insured,
                                      levels = c(0, 1),
                                      labels = c("Uninsured", "Insured")),
              chronic_factor = factor(chronic,
                                      levels = c(0, 1),
                                      labels = c("No chronic conditions",
                                                 "1+ chronic conditions")))
```

# Research Questions

We are interested in identifying the characteristics that are associated with **number of doctor visits** `visits` (our response variable, $y$). We will study the effects of age ( `age` ), education ( `educ` ), sex ( `female` ), race ( `black` ), Hispanic ethnicity ( `hispanic` ), geographic location ( `location` ), health insurance status ( `insured` ), and health status ( `chronic` ) on the response. The differing lengths of follow-up for each patient should be accounted for in the analysis because a longer follow-up will likely give us the opportunity to observe a greater number of visits to the doctor. Looking at the **rate of doctor visits** will account for variable follow-up time.

# Summarizing Count and Rate Data

## Numerical Summaries

We can numerically summarize counts by reporting the estimated **mean count**, $\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n}$. When all individuals are not followed for the same length of time (i.e., there is **variable follow-up time**), we can take length of follow-up into account by estimating the **event rate**, $\hat{\lambda} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} t_i}$. Rates have units of number of events per unit of time (e.g., visits/month).

- **Summarizing the Count**

```
# Mean and SD of number of doctor visits in full sample
c(mn = mean(vis$visits, na.rm=TRUE), std = sd(vis$visits, na.rm=TRUE), rnge = range(vis$visits,
 na.rm=TRUE))
```

```
##         mn        std      rnge1      rnge2
##   3.925817   7.704503   0.000000 103.000000
```

On average, an individual in this sample visits the doctor $\hat{\mu} = 3.926$ times (SD = 7.705). The number of visits per subject has a wide range in these data– from 0 visits to 103 visits.

We can report mean number of doctor visits by levels of a factor variable such as insurance status ( `insured_factor` ), sex ( `female_factor` ), or presence of a chronic condition ( `chronic_factor` ).

- *By Insurance Status*

```
# Mean number of doctor visits by insurance status
aggregate(x = list(visits = vis$visits), by = list(group = vis$insured_factor),
          FUN = function(x) c(mn=mean(x, na.rm=TRUE), std=sd(x, na.rm=TRUE)))
```

```
##       group visits.mn visits.std
## 1 Uninsured  1.578669   5.070006
## 2   Insured  4.568044   8.164409
```

In this study, those *with insurance* visit the doctor an average of $\hat{\mu}_1 = 4.568$ times, while those *without insurance* visit the doctor an average of $\hat{\mu}_0 = 1.579$ times.

```
# Mean ratio (insurance vs. no insurance)
muhat1 <- mean(vis$visits[vis$insured_factor=="Insured"], na.rm=TRUE)
muhat0 <- mean(vis$visits[vis$insured_factor=="Uninsured"], na.rm=TRUE)
meanratio <- muhat1/muhat0
meanratio
```

```
## [1] 2.893604
```

The **ratio** of the mean number of visits in the insured group vs. uninsured group, or the estimated **mean ratio**, $\hat{\text{MR}} = \frac{\hat{\mu}_1}{\hat{\mu}_0}$, is equal to 2.894, indicating that in this study, those *with insurance* visit the doctor 2.894 times more often than those *without insurance*.

- *By Presence of Chronic Condition*

```
# Mean number of doctor visits by presence of a chronic condition
aggregate(x = list(visits = vis$visits), by = list(group = vis$chronic_factor),
          FUN = function(x) c(mn=mean(x, na.rm=TRUE), std=sd(x, na.rm=TRUE)))
```

```
##                   group visits.mn visits.std
## 1 No chronic conditions  2.222147   4.928082
## 2 1+ chronic conditions  7.448156  10.654598
```

```
# Mean ratio (chronic vs. no chronic)
muhat1 <- mean(vis$visits[vis$chronic_factor=="1+ chronic conditions"], na.rm=TRUE)
muhat0 <- mean(vis$visits[vis$chronic_factor=="No chronic conditions"], na.rm=TRUE)
meanratio <- muhat1/muhat0
meanratio
```

```
## [1] 3.351783
```

In this study, those with *at least one chronic condition* visit the doctor an average of $\hat{\mu}_1 = 7.448$ times, while those *with no chronic conditions* visit the doctor an average of $\hat{\mu}_0 = 2.222$ times. Those with *at least one chronic condition* visit the doctor 3.352 times more often than those with *no chronic conditions*.

---

**Exercise**: Report mean number of doctor visits by sex and the mean ratio of doctor visits in females vs. males.

---

▶ Answer:

- **Summarizing the Rate**

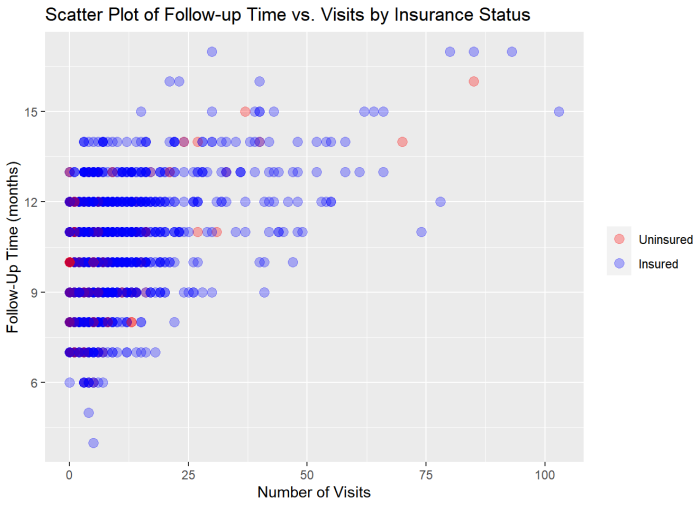
A common type of **rate** reports the number of events per unit of time, which accounts for the amount of individual follow-up time. The denominator of the rate is the total amount of time at risk for all subjects being followed. Follow-up time for individuals in this study varies from 4 to 17 months and the total follow-up time is equal to 43589 person-months.

```
# Range of follow-up time in this study (months)
range(vis$followup, na.rm = TRUE)
```

```
## [1]  4 17
```

Following individuals for a longer period of time gives you greater opportunity to observe (and count) a visit to the doctor. As we see in the scatterplot below, those with a larger number of visits tend to have a longer follow-up.

```
# Scatterplot of number of follow-up time (months) vs. number of visits by insurance status
ggplot(data = vis, aes(x = visits, y = followup,
                       color = insured_factor)) +               # color dots by insured_factor
  geom_point(size = 3, shape = 19, alpha = 0.3) +
  labs(title = "Scatter Plot of Follow-up Time vs. Visits by Insurance Status",
       x = "Number of Visits", y = "Follow-Up Time (months)") +
  theme(legend.title = element_blank()) +                       # suppress legend title
  scale_color_manual(values = c("red", "blue")) # dot color for each level of insurance status
```




We can calculate average follow-up time in the groups examined above.

```
# Mean follow-up time by insurance status
aggregate(x = list(followup = vis$followup), by = list(group = vis$insured_factor),
          FUN = function(x) c(mn=mean(x, na.rm=TRUE), rnge=range(x, na.rm=TRUE)))
```

```
##       group followup.mn followup.rnge1 followup.rnge2
## 1 Uninsured    9.661035       6.000000      16.000000
## 2   Insured    9.950881       4.000000      17.000000
```

Although those with medical insurance have a larger average number of doctor visits compared to those without medical insurance, the average length of follow-up is similar in the two groups: 9.951 months in those *with insurance* vs. 9.661 months in those *without insurance*.

To compute the **visit rate** in the overall sample, we divide the total number of events (17305 visits) by the total follow-up time (43589 person-months).

```
# Rate of doctor visits in the full sample, lambda.hat
overall.rate <- sum(vis$visits, na.rm = TRUE)/sum(vis$followup, na.rm = TRUE)
overall.rate
```

```
## [1] 0.3970038
```

The **rate** of doctor visits in the sample is $\hat{\lambda} = 0.397$ visits/month. We can similarly compute the rate within levels of a factor variable.

- *By Insurance Status*

```
# Rate of doctor visits by insurance status
bygroup.sum <- aggregate(x = list(visits = vis$visits, followup = vis$followup),
                         by = list(group = vis$insured_factor),
                         FUN = sum, na.rm = TRUE)

bygroup.rate <- cbind(bygroup.sum, rate = bygroup.sum[,2]/bygroup.sum[,3])
bygroup.rate
```

```
##       group visits followup      rate
## 1 Uninsured   1495     9149 0.1634058
## 2   Insured  15810    34440 0.4590592
```

```
# Rate ratio (insured vs. uninsured)
lamhat1 <- sum(vis$visits[vis$insured_factor=="Insured"], na.rm=TRUE)/
  sum(vis$followup[vis$insured_factor=="Insured"], na.rm=TRUE)
lamhat0 <- sum(vis$visits[vis$insured_factor=="Uninsured"], na.rm=TRUE)/
  sum(vis$followup[vis$insured_factor=="Uninsured"], na.rm=TRUE)
rateratio <- lamhat1/lamhat0
rateratio
```

```
## [1] 2.80932
```

In this study, those *with insurance* have a doctor visit rate of $\hat{\lambda}_1 = 0.459$ visits/month, while those *without insurance* have a doctor visit rate of $\hat{\lambda}_0 = 0.163$ visits/month. The **ratio** of the monthly rates in the insured vs. the uninsured group, or the **rate ratio**, $\hat{\text{RR}} = \frac{\hat{\lambda}_1}{\hat{\lambda}_0}$, is equal to 2.809, indicating that those *with insurance* have 2.809 times the rate of doctor visits compared to those *without insurance*.

# Graphical Summary - Barplot

We can graphically summarize count data using a histogram or a barplot. Recall that plotting with `ggplot()` involves adding layers of plot elements to your plot area. Since we've already covered the syntax for using `ggplot()` to create a **histogram** in **Lab 4**, today we will focus on the **R** syntax for creating **barplots** of the discrete event counts. Barplots are typically used to describe the frequency or relative frequency of observations within levels of a categorical (factor) variable. For example, a barplot is a good way to visually describe the frequency distribution of race. Barplots can be reported in the overall sample, in a subset of your sample, or by levels of another grouping variable such as exposure status. Here, we are going to create a barplot of the discrete number of visits to the doctor observed in the full sample and stratified by insurance status.

The functions `geom_bar()` and `geom_col()` can be used to create frequency and relative frequency barplots. `geom_bar()` makes the height of the bar proportional to the number of observations in each group `y=stat(count)`, while `geom_col()` makes the height of the bar represent values in the data frame. In the plot aesthetic ( `aes()` ), specify the x-variable as the count variable (i.e., `x=visits` ) and specify the height of the bar as number of observations using `y=stat(count)`. The `stat()` function tells `ggplot()` that a calculated aesthetic value produced by the statistic will be used. In the example below, the height of the bar (the y-value) will equal the number subjects with 0, 1, 2, etc. doctor visits.

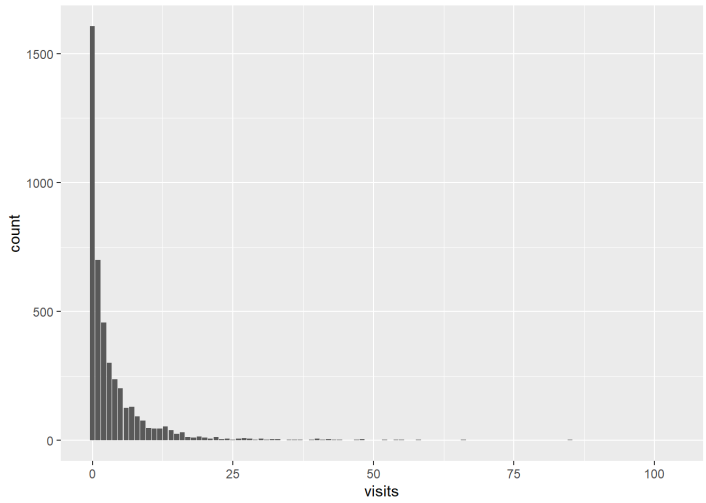
We can further customize the barplot using the following options:

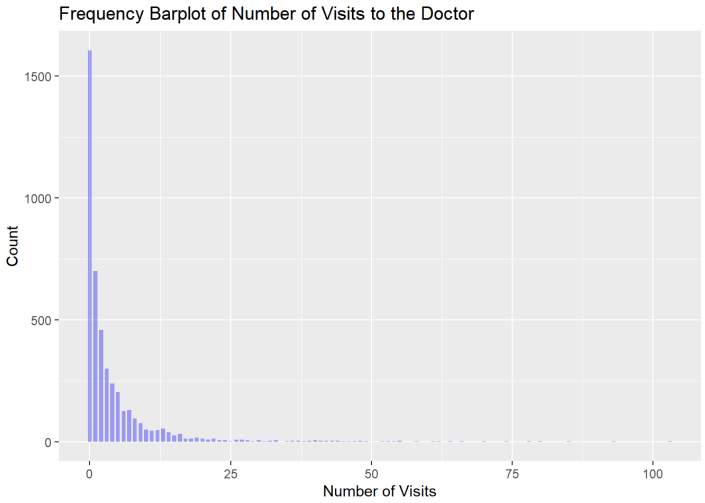
| **Option** | **Syntax in `geom_bar()`** |
|---|---|
| Bar width | `width=` |
| Bar fill color | `fill=` |
| Bar outline color | `col=` |
| Bar transparency | `alpha=` ( `0` = transparent - `1` = solid) |

| Option | Syntax in `geom_bar()` |
|---|---|
| Bar outline line type | `linetype=` (`1` = solid, `2` = dashed) |
| Grouped bar position | `position=position_dodge()` |
| Legend (if applicable) | `show.legend=` (`TRUE` or `FALSE`) |

The width, bar outline color, and transparency of the bars can be specified using the options `width=`, `col=`, and `alpha=` in the `geom_bar()` function. To specify a constant bar fill color that is different from the default color of gray, specify `fill=` in the `geom_bar()` function instead of in the `ggplot()` `aes()`. For example, `fill="blue"` will produce blue bars.

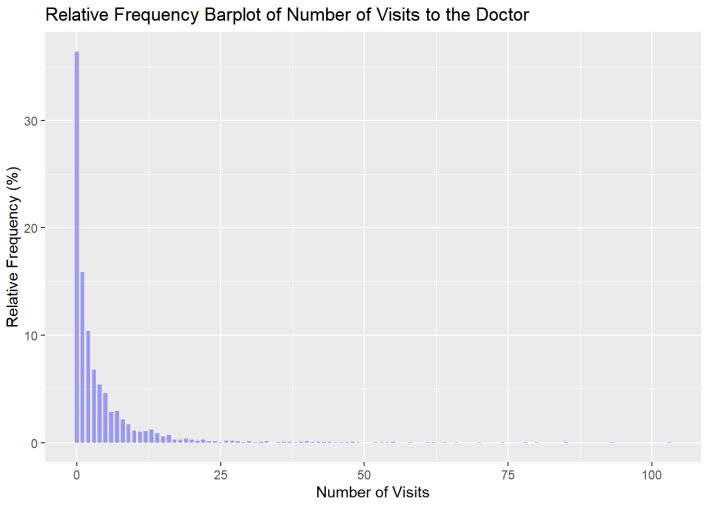
```
# Basic frequency barplot
ggplot(data = vis, aes(x = visits, y = stat(count))) +
  geom_bar()

# With customization:
ggplot(data = vis, aes(x = visits, y = stat(count))) +
  geom_bar(fill = "blue", width = 0.7, alpha = 0.35) +
  labs(title = "Frequency Barplot of Number of Visits to the Doctor",
       x = "Number of Visits", y = "Count")
```



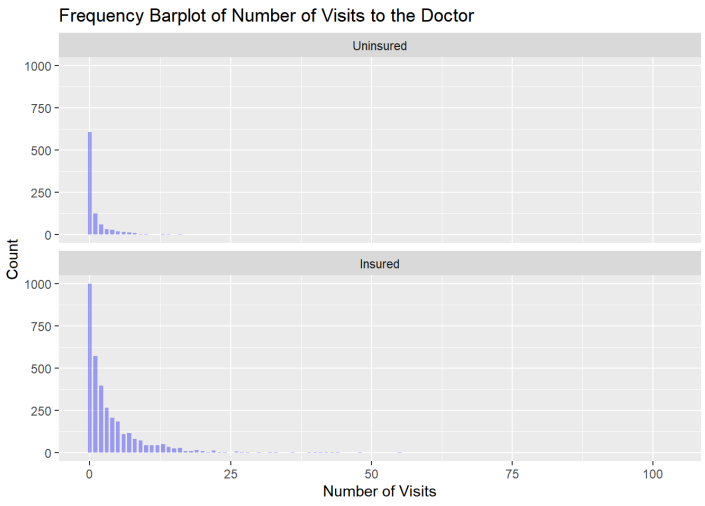

To report a relative frequency barplot, change the `y=` value in the original aesthetic to `y=100*(stat(count))/sum(stat(count))`.

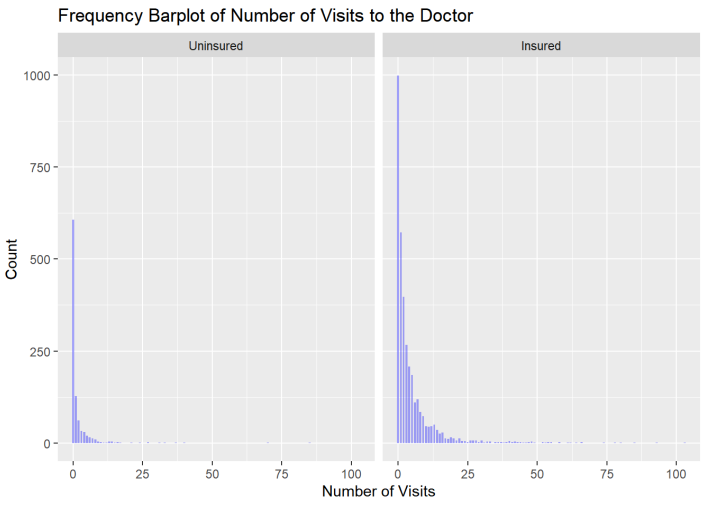
```
# Relative frequency (%) barplot
ggplot(data = vis, aes(x = visits, y = 100*(stat(count))/sum(stat(count)))) +
  geom_bar(fill = "blue", width = 0.7, alpha = 0.35) +
  labs(title = "Relative Frequency Barplot of Number of Visits to the Doctor",
       x = "Number of Visits", y = "Relative Frequency (%)")
```




An option for creating plots by group is to create **panel plots**, in which the plot for each group is presented in a different panel. To create panel plots of for each level of `insured_factor` (i.e., insured vs. uninsured) add a `facet_wrap()` layer using the syntax `+ facet_wrap(~ insured_factor)`. Panels can be horizontal (left plot) or vertical (right plot).

```
# Horizontal panel plots by group
ggplot(data = vis, aes(x = visits, y = stat(count))) +
  geom_bar(fill = "blue", width = 0.7, alpha = 0.35) +
  labs(title = "Frequency Barplot of Number of Visits to the Doctor",
       x = "Number of Visits", y = "Count") +
  facet_wrap(~ insured_factor, ncol = 1)      # plots on top of each other (one column)

# Vertical panel plots by group
ggplot(data = vis, aes(x = visits, y = stat(count))) +
  geom_bar(fill = "blue", width = 0.7, alpha = 0.35) +
  labs(title = "Frequency Barplot of Number of Visits to the Doctor",
       x = "Number of Visits", y = "Count") +
  facet_wrap(~ insured_factor, nrow = 1)      # plots side-by-side (one row)
```




As we can see, the counts are heavily **right skewed**. There are a few individuals with a large observed number of visits, giving the distribution of `visits` a long right tail.

# Poisson Regression of the Count

**Poisson regression** is a log-linear model that can be used to model a count $Y = \{0, 1, 2, \ldots\}$ response. A **Poisson distribution** is a **discrete** probability distribution describing the probability that a given number of events occur in a fixed period of time. The mean and variance of a Poisson random variable are both equal to $\mu$. When modeling a Poisson outcome $Y$, we assume the mean of $Y$ (i.e., $\mu$) is related to the linear function of the covariates $\alpha + \beta\, x$ (a.k.a., the *linear predictor*) through the **log function**, giving the following log-linear regression model:

$$\log(\mu) = \alpha + \beta_1\, x_1 + \beta_2\, x_2 + \ldots + \beta_k\, x_k$$

Thus, the *link function* in Poisson regression is the $\log$ function since the log function links $\mu$ to $\alpha + \beta\, x$. The model of the count is estimated to give $\log(\hat{\mu}) = a + b_1\, x_1 + b_2\, x_2 + \ldots + b_k\, x_k$.

- The estimated intercept $a$ is equal to the estimated **log-mean number of events** (i.e., visits to the doctor) when all values of $x = 0$.
- The estimated slope $b_j$ is equal to the estimated **log-mean ratio** associated with a 1-unit increase in $x_j$ controlling for or holding all other predictors constant. We must **exponentiate** the slope to find the estimated **mean ratio** (i.e., $\hat{\text{MR}} = e^{b_j}$).

A **hypothesis test** of the slope parameter $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ is performed using a **Wald test**, $z = \frac{b_j}{s_{b_j}}$, which is compared to a **standard Normal distribution**. Under $H_0$, $\beta_j = 0$, there is no association between $x_j$ and the outcome. When the $\log(MR) = 0$, the $MR = e^0 = 1$.

The estimated or predicted or fitted mean count $\hat{\mu}$ is equal to $\hat{\mu} = e^{a + b_1\, x_1 + b_2\, x_2 + \ldots + b_k\, x_k}$.

We use the `glm()` function in **R** to run a Poisson regression model. A Poisson regression model is specified through the function argument `family = poisson(link = "log")`.

| `glm()` **Function Arguments** | **Option Definition** |
|---|---|
| `formula=` | `analysis_variable ~ predictor_variable1 + predictor_variable2` |
| `data=` | Data frame containing sample data |
| `family=` | Error distribution and link function |
| | - Poisson regression `=poisson(link="log")` |
| | - Logistic regression `=binomial(link="logit")` |
| | - Linear regression `=gaussian(link="identity")` |

In our doctor visit data, since all subjects are not followed for the same length of time, it is more appropriate to account for follow-up time by analyzing the **rate** rather than analyzing the **count**. We will discuss analyzing the **rate** in the next section. However, we will create a Poisson model of the count as an illustration of how to analyze and interpret this type of model. A model of the count would have been appropriate if all of the individuals in this data set had been followed for a fixed period of time (e.g., number of doctor visits over a fixed 1-year period).

# Simple Poisson Regression of the Count

We begin by fitting an **unadjusted Poisson model of the count** using insurance status ( `insured_factor` ) as the only predictor. We would like to determine if there is an association between the insurance status and the average number of visits to the doctor. In other words, is there a significant difference in the average number of doctor visits in those who have insurance vs. those who do not have insurance (reference group).

The `contrasts()` function returns the dummy variable coding that **R** uses to represent a factor variable. For example, `insured_factor` is a dummy variable ($z_1$) that equals `1` for those who are insured and `0` for those who are uninsured (the reference category).

```
contrasts(vis$insured_factor)
```

```
##           Insured
## Uninsured       0
## Insured         1
```

To estimate the association between **number of visits to the doctor** ( `visits` ) and **insurance status** ( `insured_factor` ), fit the Poisson regression model, $\log(\mu) = \alpha + \beta_1 \text{Insured}$. The result of the `glm()` function is usually saved as an object ( `mod.mean1` , below) and the `summary()` function is applied to that object ( `summary(mod.mean1)` ) to output detailed results.

```
# Poisson regression of the count
mod.mean1 <- glm(visits ~ insured_factor, data = vis, family = poisson(link="log"))
summary(mod.mean1)
```

```
##
## Call:
## glm(formula = visits ~ insured_factor, family = poisson(link = "log"),
##     data = vis)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.023  -2.024  -1.354   0.199  22.601
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.45658    0.02586   17.66   <2e-16
## insured_factorInsured  1.06250    0.02706   39.27   <2e-16
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 36282  on 4407  degrees of freedom
## Residual deviance: 34215  on 4406  degrees of freedom
## AIC: 43078
##
## Number of Fisher Scoring iterations: 6
```

We can extract the **model coefficients** ($a$, $b_1$) using the `coef()` function and the **confidence intervals** of the model parameters ($\alpha$, $\beta_1$) using the `confint.default()` function. Remember that we must exponentiate $b_1$ to give an estimate of the mean ratio. Similarly, we must exponentiate the confidence interval for $\beta_1$ to give a confidence interval for the mean ratio, $e^{\beta_1}$.

```
# Slope coefficient = logMR, exponentiated slope coefficient = MR and 95% CI for MR
round(cbind(bj=coef(mod.mean1), MR=exp(coef(mod.mean1)), exp(confint.default(mod.mean1))), 5)
```

```
##                            bj      MR   2.5 %  97.5 %
## (Intercept)           0.45658 1.57867 1.50065 1.66075
## insured_factorInsured 1.06250 2.89360 2.74416 3.05119
```

- The **fitted model** is given by the equation, $\log(\hat{\mu}) = 0.457 + 1.063$ Insured
- The **estimated intercept** $a = 0.457$ is equal to the *log-mean number of doctor visits* when $z_1 = 0$ (i.e., the *log-mean number of doctor visits* in the reference category (the *uninsured*)). The exponentiated intercept $e^a = 1.579$ is equal to the *mean number of doctor visits* in those without insurance. We computed this mean when we summarized the counts by group in **Section 4** of this Lab using the `aggregate()` function.
- The **estimated slope** of `insured_factor` $b_1 = 1.063$ is equal to the *log-mean ratio* of the number of doctor visits in the *insured* vs. the *uninsured* (ref). The exponentiated slope $e^{b_1}$ gives the estimated **mean ratio**, $\hat{MR} = e^{b_1} = 2.89$ [95% CI (2.74, 3.05)]. In this study, those who are insured had 2.89 times the number of doctor visits, on average, compared to those who are uninsured. We also computed this mean ratio using the raw data in **Section 4** above.
- A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a z-statistic z = 39.27, which is compared to a standard Normal distribution. We have evidence to reject $H_0$ and conclude that the mean number of doctor visits is significantly different in those with health insurance and those without health insurance (p-value <.001).

---

**Exercise**: Construct a simple Poisson regression model of the count to determine if there is an association between number of doctor visits and sex. Report and interpret the mean ratio in females vs. males.

---

▶ Answer:

# Fitted Mean Count

The **fitted model** can be used to **estimate or predict the mean number of doctor visits** $\mu$ for given values of $x$ ( `insured_factor` ) using the `predict()` function. A data frame that contains the values of $x$ for which we would like to calculate $\hat{\mu}$ must be specified in the `newdata=` argument of the `predict()` function. The values of $x$ of interest (i.e., `insured_factor` = “Uninsured” and “Insured”) are stored in the data frame `pred.x` .

```
# Data frame that includes desired values of "x" for prediction
levels(vis$insured_factor)
```

```
## [1] "Uninsured" "Insured"
```

```
pred.x <- data.frame(insured_factor = levels(vis$insured_factor))

# Equivalently,
pred.x <- data.frame(insured_factor = c("Uninsured", "Insured"))

# Returns estimated means
muhat <- predict(mod.mean1, newdata = pred.x, type = "response")
cbind(pred.x, muhat)
```

```
##   insured_factor    muhat
## 1      Uninsured 1.578670
## 2        Insured 4.568044
```

- The **estimated mean number of doctor visits** in an **uninsured** individual is equal to $\hat{\mu}_0 = 1.579$.
- The **estimated mean number of doctor visits** in an **insured** individual is equal to $\hat{\mu}_1 = 4.568$.
- These model-estimated fitted values are equal to the raw mean number of doctor visits observed in each group:

```
# Mean number of doctor visits by insurance status
aggregate(x = list(visits = vis$visits), by = list(group = vis$insured_factor),
          FUN = mean, na.rm = TRUE)
```

```
##       group   visits
## 1 Uninsured 1.578669
## 2   Insured 4.568044
```

# Multiple Poisson Regression of the Count

Next, we extend this model to include age ( `age` ), education level ( `educ` ), sex ( `female_factor` ), race ( `black_factor` ), Hispanic ethnicity ( `hispanic_factor` ), geographic location ( `location_factor` ) and presence of chronic conditions ( `chronic_factor` ). `age` and `educ` are quantitative variables; `female_factor` (ref = “Male”), `black_factor` (ref = “Non-black”), `hispanic_factor` (ref = “Non-Hispanic”), and `chronic_factor` (ref = “No chronic conditions”) are dichotomous variables; and `location_factor` is a 4-level categorical variable represented by 3 dummy variables (ref = “West”).

```
# Poisson regression of the count
mod.mean2 <- glm(visits ~ age + educ + female_factor + black_factor + hispanic_factor +
                 location_factor + insured_factor + chronic_factor, data = vis,
                 family = poisson(link="log"))
summary(mod.mean2)
```

```
##
## Call:
## glm(formula = visits ~ age + educ + female_factor + black_factor +
##     hispanic_factor + location_factor + insured_factor + chronic_factor,
##     family = poisson(link = "log"), data = vis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9930  -1.9696  -1.1733   0.2213  23.9534
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.0881340  0.0612605 -17.762  < 2e-16
## age                                0.0084785  0.0007581  11.184  < 2e-16
## educ                               0.0676836  0.0030887  21.913  < 2e-16
## female_factorFemale                0.4840288  0.0159691  30.310  < 2e-16
## black_factorBlack                 -0.2151091  0.0368663  -5.835 5.38e-09
## hispanic_factorHispanic           -0.1871626  0.0248005  -7.547 4.46e-14
## location_factorNortheast           0.0806986  0.0240105   3.361 0.000777
## location_factorMidwest            -0.0287193  0.0228696  -1.256 0.209194
## location_factorSouth              -0.1177910  0.0208283  -5.655 1.56e-08
## insured_factorInsured              0.6376778  0.0281594  22.645  < 2e-16
## chronic_factor1+ chronic conditions  1.0324379  0.0161547  63.909  < 2e-16
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 36282  on 4407  degrees of freedom
## Residual deviance: 26743  on 4397  degrees of freedom
## AIC: 35623
##
## Number of Fisher Scoring iterations: 6
```

```
# Slope coefficient = logMR, exponentiated slope coefficient = MR and 95% CI for MR
round(cbind(bj=coef(mod.mean2), MR=exp(coef(mod.mean2)), exp(confint.default(mod.mean2))), 5)
```

```
##                                       bj      MR   2.5 %  97.5 %
## (Intercept)                       -1.08813 0.33684 0.29873 0.37982
## age                                0.00848 1.00851 1.00702 1.01001
## educ                               0.06768 1.07003 1.06357 1.07652
## female_factorFemale                0.48403 1.62260 1.57260 1.67419
## black_factorBlack                 -0.21511 0.80645 0.75024 0.86688
## hispanic_factorHispanic           -0.18716 0.82931 0.78996 0.87062
## location_factorNortheast           0.08070 1.08404 1.03421 1.13628
## location_factorMidwest            -0.02872 0.97169 0.92910 1.01623
## location_factorSouth              -0.11779 0.88888 0.85333 0.92592
## insured_factorInsured              0.63768 1.89208 1.79048 1.99944
## chronic_factor1+ chronic conditions  1.03244 2.80790 2.72039 2.89823
```

- The **fitted model** is given by the equation, $\log(\hat{\mu})$ = -1.0881 + 0.008 Age + 0.068 Education + 0.484 Female − 0.215 Black − 0.187 Hispanic + 0.081 NE − 0.029 MW − 0.118 South + 0.638 Insured + 1.032 ChronicConditions

- Wald tests of the individual slopes show that there is evidence to reject $H_0 : \beta_j = 0$ for all parameters except the indicator for Midwest geographic location (i.e., controlling for all other variables in the model, the mean number of visits to the doctor is not significantly different in those in the Midwest vs. those in the West (reference)) (p-value = 0.209).

*Controlling for all of the other variables in the model…*

- As **age** increases, the mean number of doctor visits increases. A 1-year increase in age increases the mean number of doctor visits by 1%; adjusted $\hat{MR} = e^{b_1} = 1.009$ [95% CI (1.007, 1.01)].
- As **years of education** increase, the mean number of doctor visits increases. A 1-year increase in education level increases the mean number of doctor visits by 7%, adjusted $\hat{MR} = e^{b_2} = 1.07$ [95% CI (1.064, 1.077)].
- The mean number of doctor visits in **females** is 62% higher than in *males* (ref); adjusted $\hat{MR} = e^{b_3} = 1.62$ [95% CI (1.57, 1.67)].
- The mean number of doctor visits in **blacks** is 19% lower than in *non-blacks* (ref); adjusted $\hat{MR} = e^{b_4} = 0.81$ [95% CI (0.75, 0.87)].
- The mean number of doctor visits in **Hispanics** is 17% lower than in *non-Hispanics* (ref); adjusted $\hat{MR} = e^{b_5} = 0.83$ [95% CI (0.79, 0.87)].
- The mean number of doctor visits in the **Northeast** is 8% higher than in the *West* (ref); adjusted $\hat{MR} = e^{b_6} = 1.08$ [95% CI (1.03, 1.14)].
- The mean number of doctor visits in the **Midwest** is 3% lower than in the *West* (ref); adjusted $\hat{MR} = e^{b_7} = 0.97$ [95% CI (0.93, 1.02)], (p-value = 0.209).
- The mean number of doctor visits in the **South** is 11% lower than in the *West* (ref); adjusted $\hat{MR} = e^{b_8} = 0.89$ [95% CI (0.85, 0.93)].
- The mean number of doctor visits in the **insured** is 89% higher than in the *uninsured* (ref); adjusted $\hat{MR} = e^{b_9} = 1.89$ [95% CI (1.79, 2)].
- The mean number of doctor visits in those with **1+ chronic condition** is 181% higher than in those with *no chronic conditions* (ref); adjusted $\hat{MR} = e^{b_{10}} = 2.81$ [95% CI (2.72, 2.9)].

The estimated or **predicted mean number of doctor visits** in an individual with covariate values `age = 50`, `educ = 12`, `female_factor = "Male"`, `black_factor = "Non-black"`, `hispanic_factor = "Hispanic"`, `location_factor = "Northeast"`, `insured_factor = "Insured"`, `chronic_factor = "No chronic conditions"` can be found using the `predict()` function:

```
# Data frame that includes desired values of "x" for prediction
pred.x <- data.frame(age = 50, educ = 12, female_factor = "Male", black_factor = "Non-black",
                     hispanic_factor = "Hispanic", location_factor = "Northeast",
                     insured_factor = "Insured", chronic_factor = "No chronic conditions")

# Returns estimated mean
muhat <- predict(mod.mean2, newdata = pred.x, type = "response")
muhat
```

```
##        1
## 1.972317
```

- According to our model, this individual is expected to visit the doctor 1.97 times over the course of follow-up.

# Poisson Regression of the Rate

When outcomes occur over time, it is more relevant to model the event **rate of occurrence** $\lambda$ than the raw count $\mu$. When a response count, $y$, has time at risk associated with it equal to $t$, the sample rate is equal to $\hat{\lambda} = y/t$. The expected value of $\lambda$ is equal to $\mu/t$. The **rate** of the event $\lambda$ is modeled using a log-linear model:

$$\begin{aligned} \log(\lambda) = \log\left(\frac{\mu}{t}\right) &= \alpha + \beta_1\, x_1 + \beta_2\, x_2 + \ldots + \beta_k\, x_k \\ \log(\mu) - \log(t) &= \alpha + \beta_1\, x_1 + \beta_2\, x_2 + \ldots + \beta_k\, x_k \\ \log(\mu) &= \alpha + \beta_1\, x_1 + \beta_2\, x_2 + \ldots + \beta_k\, x_k + \log(t) \end{aligned}$$

where $\log(t)$ is the **offset** term, or a term in the regression model that does not have an estimated parameter (i.e., the "slope" of $\log(t)$ is forced to equal 1). Thus, a model of the **rate** reduces to a model of the Poisson **count** with an adjustment or **offset** term $\log(t)$, that accounts for each individual's time at risk. In **R**, we specify the offset term on the right hand side of the model equation as `offset(log(followup))`, where `followup` is the time variable in the denominator of the rate.

The model of the rate is estimated to give $\log(\hat{\lambda}) = a + b_1\, x_1 + b_2\, x_2 + \ldots + b_k\, x_k$.

- The estimated intercept $a$ is equal to the estimated **log-rate of events** (i.e., visits to the doctor per month) when all values of $x = 0$.
- The estimated slope $b_j$ is equal to the estimated **log-rate ratio** associated with a 1-unit increase in $x_j$ controlling for or holding all other predictors constant. We must **exponentiate** the slope to find the estimated **rate ratio** (i.e., $\hat{\text{RR}} = e^{b_j}$).

A **hypothesis test** of the slope parameter $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ is performed using a **Wald test**, $z = \frac{b_j}{s_{b_j}}$, which is compared to a **standard Normal distribution**. Under $H_0$, $\beta_j = 0$, there is no association between $x_j$ and the rate of the event. When the $\log(RR) = 0$, the $RR = e^0 = 1$.

The estimated or predicted or fitted rate of the event $\hat{\lambda}$ is equal to $\hat{\lambda} = e^{a + b_1\, x_1 + b_2\, x_2 + \ldots + b_k\, x_k}$.

# Simple Poisson Regression of the Rate

We begin by re-fitting `mod.mean1` and running an **unadjusted Poisson model of the rate** using insurance status (`insured_factor`) as the only predictor. Here, we would like to determine if there is a significant association between insurance status and the rate of doctor visits. In other words, is there a significant difference in the rate of doctor visits (i.e., visits/month) in those who have insurance vs. those who do not have insurance (reference group). To estimate the association between **rate of doctor visits** (`visits`) and **insurance status** (`insured_factor`), fit the Poisson regression model, $\log(\lambda) = \alpha + \beta_1\, \text{Insured}$, or $\log(\mu) = \alpha + \beta_1\, \text{Insured} + \log(\text{Followup})$.

```
# Poisson regression of the rate
mod.rate1 <- glm(visits ~ insured_factor + offset(log(followup)), data = vis,
                 family = poisson(link="log"))
summary(mod.rate1)
```

```
##
## Call:
## glm(formula = visits ~ insured_factor + offset(log(followup)),
##     family = poisson(link = "log"), data = vis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4548  -2.2047  -1.5475   0.2063  20.6663
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.81152    0.02586  -70.05   <2e-16
## insured_factorInsured  1.03294    0.02706   38.18   <2e-16
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 32830  on 4407  degrees of freedom
## Residual deviance: 30892  on 4406  degrees of freedom
## AIC: 39754
##
## Number of Fisher Scoring iterations: 6
```

```
# Slope coefficient = logRR, exponentiated slope coefficient = RR and 95% CI for RR
round(cbind(bj=coef(mod.rate1), RR=exp(coef(mod.rate1)), exp(confint.default(mod.rate1))), 5)
```

```
##                            bj      RR   2.5 %  97.5 %
## (Intercept)           -1.81152 0.16341 0.15533 0.17190
## insured_factorInsured  1.03294 2.80932 2.66422 2.96232
```

Notice that we do not see the offset term in the list of estimated model coefficients. This is because the coefficient of $\log(\text{Followup})$ is not estimated and is fixed at 1.

- The **fitted model** is given by the equation, $\log(\hat{\lambda}) =$ -1.812 + 1.033 Insured
- The **estimated intercept** $a =$ -1.812 is equal to the *log-rate of doctor visits* (i.e., visits/month) in the reference category (the *uninsured*). The exponentiated intercept $e^a = 0.163$ visits/month is equal to the *monthly rate of doctor visits* in those without insurance. We computed this rate when we summarized the rates by group in **Section 4** of this Lab.
- The **estimated slope** of `insured_factor` $b_1 = 1.033$ is equal to the *log-rate ratio* of doctor visits in the *insured* vs. the *uninsured* (ref). The exponentiated slope $e^{b_1}$ gives the estimated **rate ratio**, $\hat{RR} = e^{b_1} =$ 2.81 [95% CI (2.66, 2.96)]. In this study, those who are insured had 2.81 times the rate of doctor visits of the uninsured. We also computed this rate ratio using the raw data in **Section 4** above.

- A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a z-statistic z = 38.18, which is compared to a standard Normal distribution. We have evidence to reject $H_0$ and conclude that the rate of doctor visits is significantly different in those with health insurance and those without health insurance (p-value <.001).

---

**Exercise**: Construct a simple Poisson regression model of the rate to determine if there is an association between the rate of doctor visits and sex. Report and interpret the rate ratio in females vs. males.

---

▶ Answer:

# Fitted Rate

The **fitted model** can be used to **estimate or predict the rate of doctor visits** $\lambda$ for given values of $x$ ( `insured_factor` ) and a given length of follow-up ( `followup` ).

```
# Data frame that includes desired values of "x" for prediction
pred.x <- data.frame(insured_factor = c("Uninsured", "Insured"), followup = 1)

# Returns estimated monthly rates
lambdahat <- predict(mod.rate1, newdata = pred.x, type = "response")
cbind(pred.x, lambdahat)
```

```
##   insured_factor followup lambdahat
## 1      Uninsured        1 0.1634058
## 2        Insured        1 0.4590592
```

- The **estimated visit rate/month** in an **uninsured** individual equal to $\hat{\lambda}_0 = 0.163$ visits/month. In other words, an uninsured individual is expected to visit the doctor $\hat{\lambda}_0 = 0.163$ times/month.
- The **estimated visit rate/month** in an **insured** individual equal to $\hat{\lambda}_1 = 0.459$ visits/month. In other words, an uninsured individual is expected to visit the doctor $\hat{\lambda}_1 = 0.459$ times/month.
- These numbers are equal to the raw monthly rates computed earlier.

**Note:** We can change the value of `followup` in the `pred.x` data frame to estimate the rate over a different period of time. Since follow-up time is measured on a scale of months, to estimate the **annual rate** of doctor visits, specify `followup = 12` .

---

**Exercise**: Report **annual rate** of doctor visits by in the insured and in the uninsured.

---

▶ Answer:

# Multiple Poisson Regression of the Rate

Just as we did when modeling the count, we now extend this model of the visit rate to include age ( `age` ), education level ( `educ` ), sex ( `female_factor` ), race ( `black_factor` ), Hispanic ethnicity ( `hispanic_factor` ), geographic location ( `location_factor` ) and presence of chronic conditions ( `chronic_factor` ).

```
# Poisson regression of the rate
mod.rate2 <- glm(visits ~ age + educ + female_factor + black_factor + hispanic_factor +
                 location_factor + insured_factor + chronic_factor + offset(log(followup)),
                 data = vis, family = poisson(link="log"))
summary(mod.rate2)
```

```
##
## Call:
## glm(formula = visits ~ age + educ + female_factor + black_factor +
##     hispanic_factor + location_factor + insured_factor + chronic_factor +
##     offset(log(followup)), family = poisson(link = "log"), data = vis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0880  -1.9389  -1.1584   0.2216  22.0592
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -3.279636   0.061232 -53.561  < 2e-16
## age                               0.008366   0.000759  11.021  < 2e-16
## educ                              0.062772   0.003090  20.315  < 2e-16
## female_factorFemale               0.463659   0.015974  29.025  < 2e-16
## black_factorBlack                -0.203345   0.036870  -5.515 3.48e-08
## hispanic_factorHispanic          -0.180903   0.024797  -7.295 2.98e-13
## location_factorNortheast          0.086513   0.024003   3.604 0.000313
## location_factorMidwest           -0.019072   0.022860  -0.834 0.404134
## location_factorSouth             -0.110460   0.020811  -5.308 1.11e-07
## insured_factorInsured             0.619512   0.028193  21.974  < 2e-16
## chronic_factor1+ chronic conditions  0.985802   0.016156  61.017  < 2e-16
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 32830  on 4407  degrees of freedom
## Residual deviance: 24046  on 4397  degrees of freedom
## AIC: 32927
##
## Number of Fisher Scoring iterations: 6
```

```
# Slope coefficient = logRR, exponentiated slope coefficient = RR and 95% CI for RR
round(cbind(bj=coef(mod.rate2), RR=exp(coef(mod.rate2)), exp(confint.default(mod.rate2))), 5)
```

```
##                                        bj      RR   2.5 %  97.5 %
## (Intercept)                       -3.27964 0.03764 0.03338 0.04244
## age                                0.00837 1.00840 1.00690 1.00990
## educ                               0.06277 1.06478 1.05835 1.07125
## female_factorFemale                0.46366 1.58988 1.54087 1.64045
## black_factorBlack                 -0.20335 0.81600 0.75911 0.87715
## hispanic_factorHispanic           -0.18090 0.83452 0.79493 0.87608
## location_factorNortheast           0.08651 1.09037 1.04026 1.14289
## location_factorMidwest            -0.01907 0.98111 0.93812 1.02607
## location_factorSouth              -0.11046 0.89542 0.85963 0.93270
## insured_factorInsured              0.61951 1.85802 1.75814 1.96358
## chronic_factor1+ chronic conditions  0.98580 2.67996 2.59643 2.76618
```

- The **fitted model** is given by the equation, $\log(\hat{\lambda})$ = -3.2796 + 0.008 Age + 0.063 Education + 0.464 Female − 0.203 Black − 0.181 Hispanic + 0.087 NE − 0.019 MW − 0.11 South + 0.62 Insured + 0.986 ChronicConditions
- Just as in the model of the count, the Wald tests of the individual slopes show that there is evidence to reject $H_0 : \beta_j = 0$ for all parameters except the indicator for Midwest geographic location (i.e., the rate of doctor visits is not significantly different in those in the Midwest vs. those in the West (reference)) (p-value = 0.404).

*Controlling for all of the other variables in the model…*

- As **age** increases, the rate of doctor visits increases. A 1-year increase in age increases the rate of doctor visits by 1%; adjusted $\hat{RR} = e^{b_1} =$ 1.008 [95% CI (1.007, 1.01)].
- As **years of education** increase, the rate of doctor visits increases. A 1-year increase in education level increases the rate of doctor visits by 6%, adjusted $\hat{RR} = e^{b_2} =$ 1.065 [95% CI (1.058, 1.071)].
- The rate of doctor visits in **females** is 59% higher than in *males* (ref); adjusted $\hat{RR} = e^{b_3} =$ 1.59 [95% CI (1.54, 1.64)].
- The rate of doctor visits in **blacks** is 18% lower than in *non-blacks* (ref); adjusted $\hat{RR} = e^{b_4} =$ 0.82 [95% CI (0.76, 0.88)].
- The rate of doctor visits in **Hispanics** is 17% lower than in *non-Hispanics* (ref); adjusted $\hat{RR} = e^{b_5} =$ 0.83 [95% CI (0.79, 0.88)].
- The rate of doctor visits in the **Northeast** is 9% higher than in the *West* (ref); adjusted $\hat{RR} = e^{b_6} =$ 1.09 [95% CI (1.04, 1.14)].
- The rate of doctor visits in the **Midwest** is 2% lower than in the *West* (ref); adjusted $\hat{RR} = e^{b_7} =$ 0.98 [95% CI (0.94, 1.03)], (p-value = 0.404).
- The rate of doctor visits in the **South** is 10% lower than in the *West* (ref); adjusted $\hat{RR} = e^{b_8} =$ 0.9 [95% CI (0.86, 0.93)].
- The rate of doctor visits in the **insured** is 86% higher than in the *uninsured* (ref); adjusted $\hat{RR} = e^{b_9} =$ 1.86 [95% CI (1.76, 1.96)].
- The rate of doctor visits in those with **1+ chronic condition** is 168% higher than in those with *no chronic conditions* (ref); adjusted $\hat{RR} = e^{b_{10}} =$ 2.68 [95% CI (2.6, 2.77)].

The estimated or **predicted monthly rate of doctor visits** in an individual with covariate values `age = 50`, `educ = 12`, `female_factor = "Male"`, `black_factor = "Non-black"`, `hispanic_factor = "Hispanic"`, `location_factor = "Northeast"`, `insured_factor = "Insured"`, `chronic_factor = "No chronic conditions"` can be found using the `predict()` function:

```
# Data frame that includes desired values of "x" for prediction
pred.x <- data.frame(age = 50, educ = 12, female_factor = "Male", black_factor = "Non-black",
                     hispanic_factor = "Hispanic", location_factor = "Northeast",
                     insured_factor = "Insured", chronic_factor = "No chronic conditions",
                     followup = 1)

# Returns estimated mean
lambdahat <- predict(mod.rate2, newdata = pred.x, type = "response")
lambdahat
```

```
##         1
## 0.2053641
```

- According to our model, this individual is expected to visit the doctor 0.2 times/month.

# Likelihood Ratio Test

Just as in logistic regression, a **Likelihood Ratio Test** can be used to simultaneously test the significance of a group or set of parameters when fitting a Poisson regression model of the count or of the rate. This test is commonly used to test the effect of categorical variables that are naturally made up of more than one dummy variable. For example, to test the significance of **geographic location** in the adjusted model, we would test: $H_0: \beta_6 = \beta_7 = \beta_8 = 0$ vs. $H_1: \beta_6, \beta_7, \beta_8$ not all 0.

Here, we are comparing two **nested models**,

- **Full model**: $\log(\lambda) = \alpha + \beta_1$ Age $+\beta_2$ Education $+\beta_3$ Female $+\beta_4$ Black $+\beta_5$ Hispanic $+\beta_6$ NE $+\beta_7$ MW $+\beta_8$ South $+\beta_9$ Insured $+\beta_{10}$ ChronicConditions
- **Reduced model** (i.e., model under $H_0$, without `location_factor`): $\log(\lambda) = \alpha + \beta_1$ Age $+\beta_2$ Education $+\beta_3$ Female $+\beta_4$ Black $+\beta_5$ Hispanic $+\beta_6$ Insured $+\beta_7$ ChronicConditions

The **likelihood ratio test statistic** compares the likelihood of the full and reduced models, $\mathrm{G} = -2\log\text{-likelihood}(R) - (-2\log\text{-likelihood}(F))$. The test statistic is compared to an Chi-square distribution with *degrees of freedom* equal to the number of parameters tested under $H_0$, $\chi^2_{df}$.

We can use the `Anova()` function in the `car` package to perform a likelihood ratio test for each variable included in the model without having to fit reduced models. The `Anova()` function applied to a model object (e.g., `mod.rate2`) returns individual likelihood ratio tests for each variable in the model.

```
# LRT using Anova() function in the "car" package
Anova(mod.rate2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: visits
##                 LR Chisq Df Pr(>Chisq)
## age                121.3  1  < 2.2e-16
## educ               430.3  1  < 2.2e-16
## female_factor      870.2  1  < 2.2e-16
## black_factor        32.2  1  1.367e-08
## hispanic_factor     54.9  1  1.254e-13
## location_factor     84.9  3  < 2.2e-16
## insured_factor     558.7  1  < 2.2e-16
## chronic_factor    3869.0  1  < 2.2e-16
```

- Based on the output above, the **likelihood ratio test** of `location_factor` has a test statistic G = 84.9, which is compared to an Chi-square distribution with 3 degrees of freedom. The overall effect of location is statistically significant in the presence of the other variables in this Poisson regression model (p-value <.001). We have evidence to reject $H_0$ and conclude that at least one $\beta_6$, $\beta_7$, or $\beta_8$ is not equal to 0.

# Checking for Overdispersion

When using Poisson regression, an important assumption is that the count data arose from a **Poisson distribution**. One feature of the Poisson distribution is that the **mean** of the Poisson distributed random variable equals the **variance**. If there is evidence that the variance is not similar to the mean, then the Poisson distribution, and the Poisson regression model, may not be appropriate. At the beginning of this Lab, when we summarized the raw count data, we saw that the standard deviation of the count variable ( `visits` ) was much larger than the mean. So, the variance of the number of visits is also much larger than the mean number of visits:

```
# Mean and SD of number of doctor visits in full sample
c(mn = mean(vis$visits, na.rm = TRUE), std = sd(vis$visits, na.rm = TRUE),
  var = var(vis$visits, na.rm = TRUE))
```

```
##        mn       std       var
##  3.925817  7.704503 59.359370
```

When the variance is larger than the mean, this is known as **overdispersion**. A common cause of overdispersion is subject heterogeneity. Overdispersion can significantly affect model inference because the Poisson model will *underestimate* the standard errors of the model parameters, which will affect statistical inference (i.e., confidence intervals and p-values).

If overdispersion is an issue, then one solution is to use **negative binomial regression** instead of Poisson regression. A negative binomial distribution allows the variance of $Y$ to be larger than the mean of $Y$ since it assumes $Var(Y) = \mu + \frac{\mu^2}{k}$. The value $\frac{1}{k}$ is the **dispersion parameter** that is used to adjust the variance independently of the mean. Note that in Poisson regression, $\frac{1}{k} = 0$.

To assess if overdispersion is a problem in the data, we check to see if the ratio of the **Residual Deviance** to the **Residual Degrees of Freedom** is larger than 1. Notice that these values are printed below the estimated coefficients in the model `summary()` output. As a *rule of thumb*, a ratio $>$ 1.1 suggests overdispersion in the data.

```
# See row labeled "Residual deviance:" for residual deviance and its degrees of freedom
summary(mod.rate2)
```

```
##
## Call:
## glm(formula = visits ~ age + educ + female_factor + black_factor +
##     hispanic_factor + location_factor + insured_factor + chronic_factor +
##     offset(log(followup)), family = poisson(link = "log"), data = vis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0880  -1.9389  -1.1584   0.2216  22.0592
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -3.279636   0.061232 -53.561  < 2e-16
## age                               0.008366   0.000759  11.021  < 2e-16
## educ                              0.062772   0.003090  20.315  < 2e-16
## female_factorFemale               0.463659   0.015974  29.025  < 2e-16
## black_factorBlack                -0.203345   0.036870  -5.515 3.48e-08
## hispanic_factorHispanic          -0.180903   0.024797  -7.295 2.98e-13
## location_factorNortheast          0.086513   0.024003   3.604 0.000313
## location_factorMidwest           -0.019072   0.022860  -0.834 0.404134
## location_factorSouth             -0.110460   0.020811  -5.308 1.11e-07
## insured_factorInsured             0.619512   0.028193  21.974  < 2e-16
## chronic_factor1+ chronic conditions  0.985802   0.016156  61.017  < 2e-16
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 32830  on 4407  degrees of freedom
## Residual deviance: 24046  on 4397  degrees of freedom
## AIC: 32927
##
## Number of Fisher Scoring iterations: 6
```

```
# Checking overdispersion
deviance(mod.rate2)/mod.rate2$df.residual
```

```
## [1] 5.468693
```

In our multiple Poisson regression model of the rate `mod.rate2` , the residual deviance is equal to 24046 and the residual degrees of freedom is equal to 4397. Their ratio, 5.469 is much larger than 1, indicating that overdispersion is a problem in these data.

# Negative Binomial Regression of the Count or Rate

The `glm.nb()` function in the `MASS` package can be used to fit a **negative binomial regression model**. Since this function only fits a negative binomial model, there is no `family=` argument in the `glm.nb()` function. Note that this model may be used to either fit a negative binomial model of the **count** or the **rate** and interpretation of the fitted parameters is the same as in Poisson regression. When modeling the rate, we must specify the offset term `offset(log(followup))` .

```
# Negative binomial regression of the rate using glm.nb() function in the "MASS" package
mod.NBrate <- glm.nb(visits ~ age + educ + female_factor + black_factor + hispanic_factor +
                     location_factor + insured_factor + chronic_factor + offset(log(followup)),
                     data = vis)
summary(mod.NBrate)
```

```
##
## Call:
## glm.nb(formula = visits ~ age + educ + female_factor + black_factor +
##     hispanic_factor + location_factor + insured_factor + chronic_factor +
##     offset(log(followup)), data = vis, init.theta = 0.6847357157,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0529  -1.1691  -0.5504   0.1297   7.4027
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -3.609615   0.150356 -24.007  < 2e-16
## age                              0.012936   0.002097   6.169 6.88e-10
## educ                             0.058995   0.007768   7.594 3.10e-14
## female_factorFemale              0.524418   0.041960  12.498  < 2e-16
## black_factorBlack               -0.278344   0.095197  -2.924  0.00346
## hispanic_factorHispanic         -0.170846   0.060373  -2.830  0.00466
## location_factorNortheast         0.079125   0.068594   1.154  0.24869
## location_factorMidwest           0.015953   0.063233   0.252  0.80082
## location_factorSouth            -0.149712   0.054883  -2.728  0.00638
## insured_factorInsured            0.732766   0.059664  12.281  < 2e-16
## chronic_factor1+ chronic conditions  1.017330   0.044123  23.057  < 2e-16
##
## (Dispersion parameter for Negative Binomial(0.6847) family taken to be 1)
##
##     Null deviance: 6002.9  on 4407  degrees of freedom
## Residual deviance: 4576.6  on 4397  degrees of freedom
## AIC: 19224
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.6847
##           Std. Err.:  0.0208
##
##  2 x log-likelihood:  -19199.5030
```

```
# Comparing Poisson to Negative Binomial
round(cbind(bjPoi = coef(mod.rate2), sePoi = coef(summary(mod.rate2))[,2],
            RRPoi = exp(coef(mod.rate2)),
            bjNB = coef(mod.NBrate), seNB = coef(summary(mod.NBrate))[,2],
            RRNB = exp(coef(mod.NBrate))), 3)
```

```
##                                    bjPoi sePoi RRPoi   bjNB  seNB  RRNB
## (Intercept)                       -3.280 0.061 0.038 -3.610 0.150 0.027
## age                                0.008 0.001 1.008  0.013 0.002 1.013
## educ                               0.063 0.003 1.065  0.059 0.008 1.061
## female_factorFemale                0.464 0.016 1.590  0.524 0.042 1.689
## black_factorBlack                 -0.203 0.037 0.816 -0.278 0.095 0.757
## hispanic_factorHispanic           -0.181 0.025 0.835 -0.171 0.060 0.843
## location_factorNortheast           0.087 0.024 1.090  0.079 0.069 1.082
## location_factorMidwest            -0.019 0.023 0.981  0.016 0.063 1.016
## location_factorSouth              -0.110 0.021 0.895 -0.150 0.055 0.861
## insured_factorInsured              0.620 0.028 1.858  0.733 0.060 2.081
## chronic_factor1+ chronic conditions  0.986 0.016 2.680  1.017 0.044 2.766
```

We see that the parameter estimates from the Poisson model ( `bjPoi` ) and the negative binomial model ( `bjNB` ) are very close. However, the standard errors of the parameters from the negative binomial model ( `seNB` ) are larger than in the Poisson model ( `sePoi` ). As a result, the individual Wald test p-values tend to be larger in the negative binomial model. The adjusted visit rate in the Northeast geographic location vs. West is now not significantly different in the negative binomial model (p-value = 0.249). The negative binomial model estimates the **overdispersion factor**, $1/k$ as 0.6847. In a Poisson model, this parameter equals zero.

We can perform a **likelihood ratio test** to compare the full model (the negative binomial model) to the reduced model (the Poisson model) and test if the negative binomial model is providing a significantly better fit to the data. Under $H_0$, the negative binomial model does *not* provide a significantly better fit than the Poisson model. Under $H_1$, the negative binomial model *does* provide a significantly better fit than the Poisson model. The likelihood ratio test statistic $\mathrm{G} = -2\log\text{-likelihood}(Poisson) - (-2\log\text{-likelihood}(NB))$ is compared to a chi-square distribution with 1 degree of freedom, $\chi^2_1$. Thus, the critical value of the G test statistic is equal to 3.84 for a test performed at the $\alpha$ = 0.05-level.

```
# LRT testing if NB model provides significantly better fit to data than Poisson
G <- as.numeric(-2*logLik(mod.rate2) - (-2*logLik(mod.NBrate)))
pval <- 1 - pchisq(G, df = 1)
pval
```

```
## [1] 0
```

Here, the test statistic, $\mathrm{G} =$ 32904.51 - 19199.5 = 13705, which is greater than the chi-square critical value 3.84. Thus, we have evidence to reject $H_0$ and conclude that overdispersion is present and the negative binomial model provides a significantly better fit than the Poisson model. The final adjusted rate ratios from the **negative binomial model of the rate** should be interpreted to address the research question in these data.