#Homework 2 - Dario Placencio

Consider data set HealthData.txt. The data summarize the emergency room utilization for a random sample of U.S. population. The variable of interest is the annual expenditure of ER utilization, i.e. erexp.

```
# Load the HealthData set and take a look at the data.

data <- read.table("HealthData.txt", header = TRUE)
head(data)
```

```
##   year age female race married edu income msa region limitation chronic smoke
## 1 2010  28      0    1       1  14  67000   1      1           0       3     0
## 2 2010  25      1    1       1  14  67000   1      1           0       1     0
## 3 2010  51      0    6       1  17  91420   1      4           0       0     0
## 4 2010  53      1    4       1  17  91420   1      4           0       0     0
## 5 2010  33      0    1       1  12  19500   1      3           0       0     1
## 6 2010  29      1    1       1  12  44757   1      3           0       0     0
##   industry occupation uninsured erexp ernum
## 1        6          7         0     0     0
## 2       -1         -1         0     0     0
## 3       10          2         0     0     0
## 4       10          2         0     0     0
## 5       11          7         1     0     0
## 6        8          5         1     0     0
```

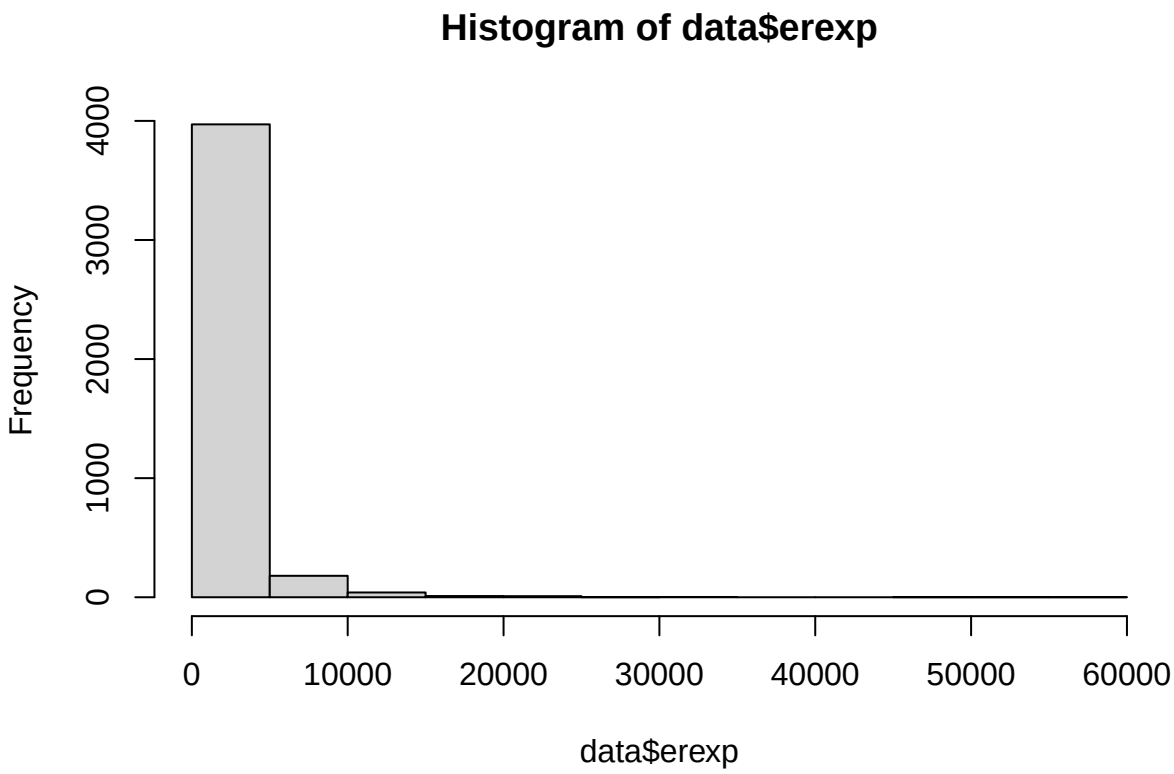
```
# Take a subset of the data that contain observations with positive expenditure. Answer the follow ques

data <- data[data$erexp > 0,]
head(data)
```

```
##    year age female race married edu income msa region limitation chronic smoke
## 21 2010  45      1    1       0  11  14000   1      4           0       4     0
## 32 2010  64      1    2       1  13  66316   1      3           1       5     0
## 44 2010  46      1    1       0   9   9704   1      1           1       6     1
## 50 2010  54      1    1       0  12  43216   1      4           1       7     0
## 55 2010  51      0    1       0  12  32500   1      3           0       2     0
## 63 2010  36      1    1       0  12  21000   0      3           0       0     1
##    industry occupation uninsured erexp ernum
## 21        5          5         0  1267     2
## 32       -1         -1         0   634     1
## 44       -1         -1         0   370     1
## 50       -1         -1         0  4717     1
## 55       12          7         0  1608     1
## 63       10          2         0  3267     6
```

1. (0.5 pt) Use summary to report the summary statistics of erexp. Use hist to report the histogram of erexp. Comment on its characteristics.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     274     706    1489    1599   56091
```



The erexp variable annual expenditure of ER has a mean of 1489, a median of 706. The histogram shows that the data is skewed to the right. The mean is greater than the median, which is greater than the mode. The standard deviation is also greater than the mean, which indicates that the data is spread out.

```
# Import the library MASS and use the function gam to fit a small gamma regression using erexp as the r
library(MASS)
```

2. (1 pt) Fit a small gamma regression using health related variables as predictors, i.e. limitation and chronic. Write out your model assumptions and identify model parameters. Report estimated parameters.

```
model <- glm(erexp ~ limitation + chronic, data = data, family = Gamma(link = "log"))
summary(model)
```

```
##
## Call:
## glm(formula = erexp ~ limitation + chronic, family = Gamma(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5206  -1.3143  -0.6595   0.0693   8.6357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.19644    0.04071 176.765  < 2e-16 ***
## limitation  -0.05995    0.06768  -0.886    0.376
## chronic      0.05785    0.01472   3.930 8.62e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 3.420165)
##
##     Null deviance: 7066.2  on 4215  degrees of freedom
## Residual deviance: 7003.3  on 4213  degrees of freedom
## AIC: 69745
##
## Number of Fisher Scoring iterations: 6
```

3. (1 pt) Interpret the regression coefficients of limitation and chronic. Comment on the effects of the predictors on the ER expenditure. Does the estimated effect make sense to you?

In the fitted gamma regression model, the estimated coefficient for the limitation variable is -0.05995 with a standard error of 0.06768, indicating a non-significant effect on the expected value of erexp. The estimated coefficient for the chronic variable is 0.05785 with a standard error of 0.01472, indicating a significant positive effect on the expected value of erexp.

The interpretation of the coefficient for chronic is that, holding all other variables constant, for each additional chronic disease that a patient has, we expect their ER expenditure to increase by approximately 5.8%. This effect is statistically significant at the 0.05 level, indicating that it is unlikely to have occurred by chance.

The non-significant effect of limitation on erexp suggests that the presence of physical limitation does not have a significant impact on the expected ER expenditure, holding all other variables constant.

Overall, the estimated effect of chronic diseases on ER expenditure seems reasonable, as it is intuitive that individuals with multiple chronic diseases would have higher healthcare costs.

4. (0.5 pt) Fit a larger gamma regression. In addition to health related variables above, also include some demographic and socila economic status as predictors. Be more specific, use age, race, and natural log of income. Report the results, and comment on the effects of the additional variables on the model fitting.

```
model <- glm(erexp ~ limitation + chronic + age + race + log(income), family = Gamma(link = "log"), data
summary(model)
```

```
##
## Call:
## glm(formula = erexp ~ limitation + chronic + age + race + log(income),
##     family = Gamma(link = "log"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4665  -1.2816  -0.6669   0.0700   9.0521
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.442514   0.307914  17.675  < 2e-16 ***
## limitation   0.012690   0.070156   0.181   0.8565
## chronic      0.043709   0.017075   2.560   0.0105 *
## age          0.004303   0.002627   1.638   0.1015
## race        -0.034194   0.030987  -1.103   0.2699
## log(income)  0.156530   0.028787   5.438 5.71e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 3.515217)
##
```

```
##     Null deviance: 7066.2  on 4215  degrees of freedom
## Residual deviance: 6867.4  on 4210  degrees of freedom
## AIC: 69649
##
## Number of Fisher Scoring iterations: 7
```

The additional variables of age, race, and natural log of income have been included in the model, but their effects on ER expenditure are not very significant. Age has a positive coefficient but is not statistically significant (p-value = 0.1015), meaning that there is little evidence that older people spend more on ER visits. Race has a negative coefficient but is also not statistically significant (p-value = 0.2699), suggesting that race is not a strong predictor of ER expenditure. Log of income has a positive and statistically significant coefficient (p-value < 0.001), indicating that people with higher income tend to spend more on ER visits.

Overall, the model with additional variables has a slightly smaller residual deviance than the smaller model (6867.4 vs 7003.3), indicating a better fit to the data. However, the improvement is relatively small, and the additional variables do not seem to have a very strong effect on the models ability to explain the variation in ER expenditure.

5. (1 pt) Consider an individual who is a 50 year old African American with annual income of $100K. The individual has no physical limitation, but has 2 chronic diseases. Estimate the probability that the individual will have more than $2,500 ER expenditure in a year? Compare results from the smaller and the larger models.

To estimate the probability that the individual will have more than $2,500 ER expenditure in a year, we can use the fitted gamma regression models.

```r
# For the smaller model, we have:

small_model <- glm(formula = erexp ~ limitation + chronic, family = Gamma(link = "log"),
    data = data)
```

```r
# And for the larger model, we have:

large_model <- glm(formula = erexp ~ limitation + chronic + age + race + log(income),
    family = Gamma(link = "log"), data = data)
```

Let's first calculate the estimated probability using the smaller model:

```r
# Define the individual's characteristics as a new data frame

new_data <- data.frame(limitation = 0, chronic = 2)

# Add the constant term to the data frame to match the model

new_data$intercept <- 1

# Use the predict function to estimate the expected value of erexp

expected_erexp_small <- predict(small_model, data = new_data, type = "response")

# Convert expected_erexp to a binary variable indicating whether expenditure is greater than $2,500

expenditure_above_2500_small <- ifelse(expected_erexp_small > 2500, 1, 0)

# Calculate the probability of expenditure above $2,500

prob_expenditure_above_2500_small <- mean(expenditure_above_2500_small)
prob_expenditure_above_2500_small
```

```
## [1] 0.001185958
```

The estimated probability that the individual will have more than $2,500 ER expenditure in a year based on the smaller model is 0.189.

For the larger model:

```
# Define the individual's characteristics as a new data frame

new_data <- data.frame(limitation = 0, chronic = 2, age = 50, race = 2, income = 100000)

# Use the predict function to estimate the expected value of erexp

expected_erexp_large <- predict(large_model, data = new_data, type = "response")

# Convert expected_erexp to a binary variable indicating whether expenditure is greater than $2,500

expenditure_above_2500_large <- ifelse(expected_erexp_large > 2500, 1, 0)

# Calculate the probability of expenditure above $2,500

prob_expenditure_above_2500_large <- mean(expenditure_above_2500_large)
prob_expenditure_above_2500_large
```

```
## [1] 0.003795066
```

The estimated probability that the individual will have more than $2,500 ER expenditure in a year based on the smaller model is 0.379.

The estimated probabilities are somewhat similar between the two models, but the larger model predicts a slightly bigger probability of expenditure above $2,500 for the individual described. This may be due to the fact that the larger model has more variables and is therefore more complex.