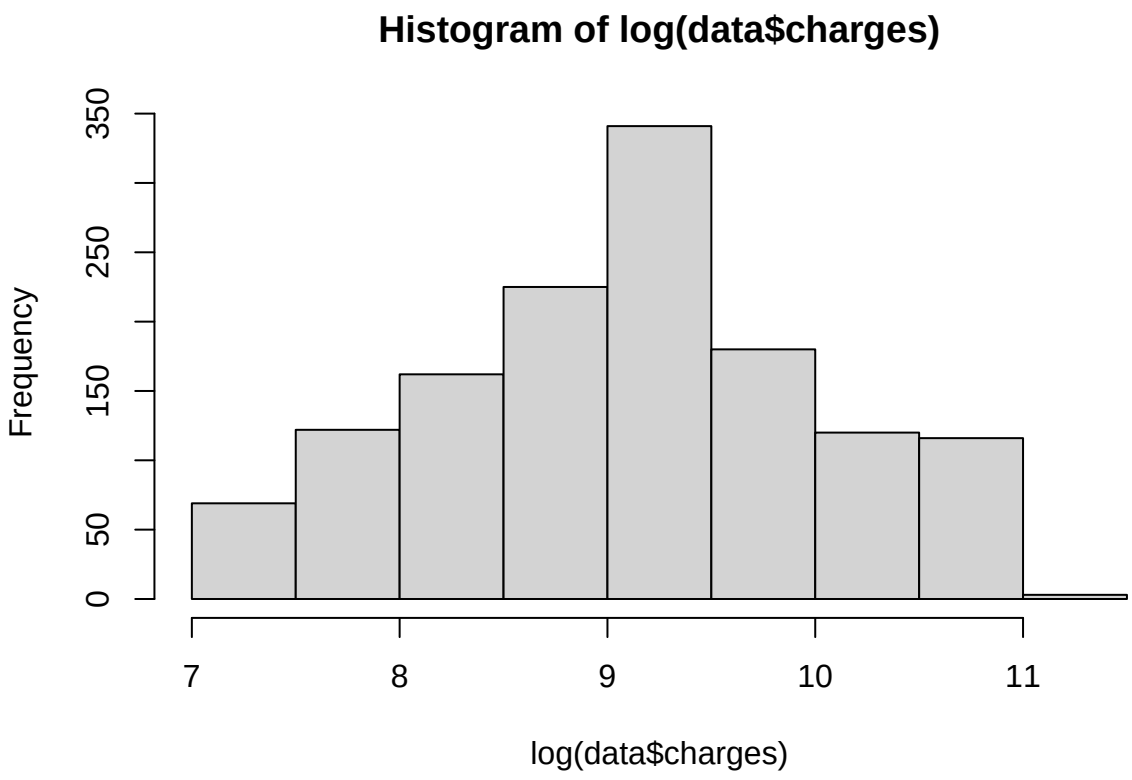
Untitled

Nikhil Gopal

11/8/2020

Question 1

```
setwd("/Users/nikhilgopal/Google Drive/Notability/Applied Linear Regression Analysis/psets")

data <- read.csv("insurance.csv")

#Question 1

lin_mod <- lm(charges~age+sex+bmi+children+smoker+region, data = data)

lin_mod_log <-lm(log(charges)~age+sex+bmi+children+smoker+region, data = data)

hist(x=log(data$charges))
```



```r
#log is the best transformation
```

Question 2

```r
#Age
mean(data$age)
```

```
## [1] 39.20703
```

```r
sd(data$age)
```

```
## [1] 14.04996
```

```r
min(data$age)
```

```
## [1] 18
```

```r
max(data$age)
```

```
## [1] 64
```

```r
#Sex
prop.table(table(data$sex))
```

```
##
##    female      male
## 0.4947683 0.5052317
```

```r
#Smoker
prop.table(table(data$smoker))
```

```
##
##        no       yes
## 0.7952167 0.2047833
```

```r
#Children
mean(data$children)
```

```
## [1] 1.094918
```

```r
sd(data$children)
```

```
## [1] 1.205493
```

```r
min(data$children)
```

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

```r
max(data$children)
```

```
## [1] 5
```

```r
#BMI
mean(data$bmi)
```

```
## [1] 30.6634
```

```r
sd(data$bmi)
```

```
## [1] 6.098187
```

```r
min(data$bmi)
```

```
## [1] 15.96
```

```r
max(data$bmi)
```

```
## [1] 53.13
```

```r
#Region
prop.table(table(data$region))
```

```
##
## northeast northwest southeast southwest
## 0.2421525 0.2428999 0.2720478 0.2428999
```

Question 3

```r
#Question 3
lm_age <- lm(charges~age, data = data)
summary(lm_age)
```

```
##
## Call:
## lm(formula = charges ~ age, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8059  -6671  -5939   5440  47829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3165.9      937.1   3.378 0.000751 ***
## age            257.7       22.5  11.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 1336 degrees of freedom
## Multiple R-squared:  0.08941,    Adjusted R-squared:  0.08872
## F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

```r
#f=131.2 on 1336 DF

lm_sex <- lm(charges~sex, data=data)
summary(lm_sex)
```

```
##
## Call:
## lm(formula = charges ~ sex, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12835  -8435  -3980   3476  51201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12569.6      470.1  26.740   <2e-16 ***
## sexmale       1387.2      661.3   2.098   0.0361 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12090 on 1336 degrees of freedom
## Multiple R-squared:  0.003282,   Adjusted R-squared:  0.002536
## F-statistic:   4.4 on 1 and 1336 DF,  p-value: 0.03613
```

```r
anova(lm_sex)
```

```
## Analysis of Variance Table
##
## Response: charges
##             Df     Sum Sq   Mean Sq F value  Pr(>F)
## sex          1 6.4359e+08 643590180  4.3997 0.03613 *
## Residuals 1336 1.9543e+11 146280413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#p = 0.03615

lm_bmi <- lm(charges~bmi, data=data)
summary(lm_bmi)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20956  -8118  -3757   4722  49442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1192.94    1664.80   0.717    0.474
## bmi           393.87      53.25   7.397 2.46e-13 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

```
#F = 54.71 on 1336 DF

lm_children <- lm(charges~children, data = data)
summary(lm_children)
```

```
##
## Call:
## lm(formula = charges ~ children, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11585  -8759  -4071   3468  51248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12522.5      446.5  28.049   <2e-16 ***
## children       683.1      274.2   2.491   0.0129 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12090 on 1336 degrees of freedom
## Multiple R-squared:  0.004624,   Adjusted R-squared:  0.003879
## F-statistic: 6.206 on 1 and 1336 DF,  p-value: 0.01285
```

```
#F = 6.206 on 1336 DF

lm_smoker <- lm(charges~smoker, data = data)
anova(lm_smoker)
```

```
## Analysis of Variance Table
##
## Response: charges
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## smoker       1 1.2152e+11 1.2152e+11  2177.6 < 2.2e-16 ***
## Residuals 1336 7.4554e+10 5.5804e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#p is basically zero 2.2e^-16

lm_region <- lm(charges~region, data=data)
anova(lm_region)
```

```
## Analysis of Variance Table
##
```

```
## Response: charges
##              Df     Sum Sq   Mean Sq F value  Pr(>F)
## region        3 1.3008e+09 433586560  2.9696 0.03089 *
## Residuals  1334 1.9477e+11 146007093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#P = 0.03089
```

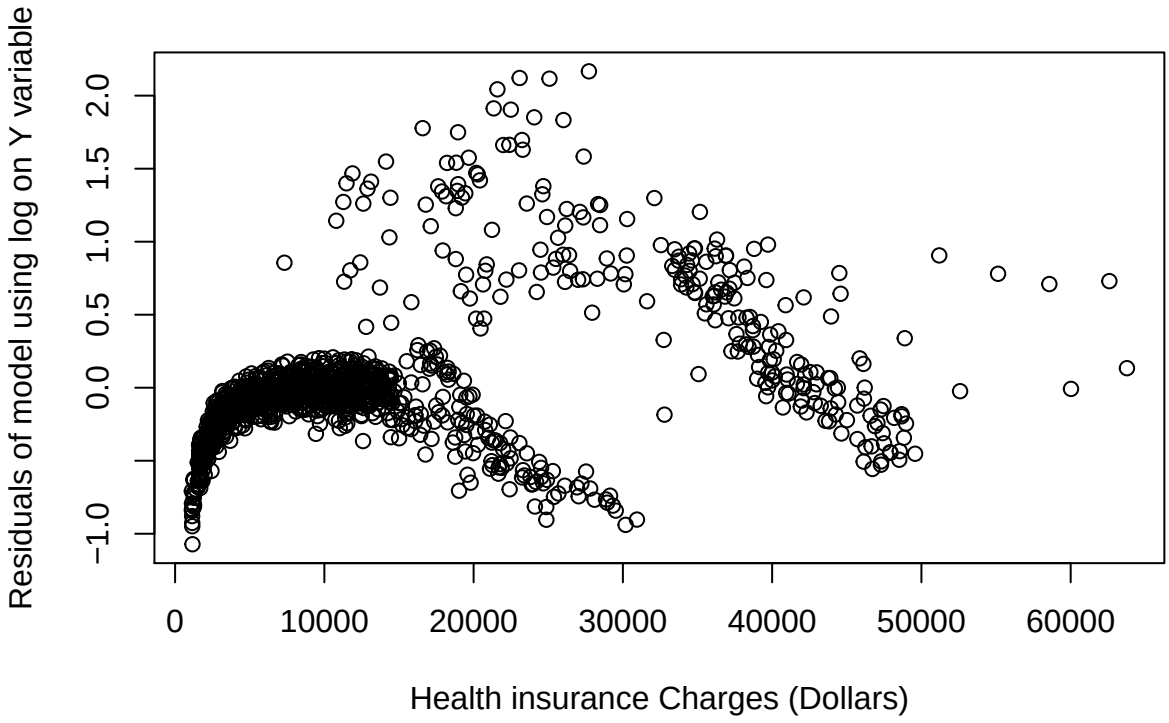
Question 4

```
#Question 4
summary(lin_mod_log)
```

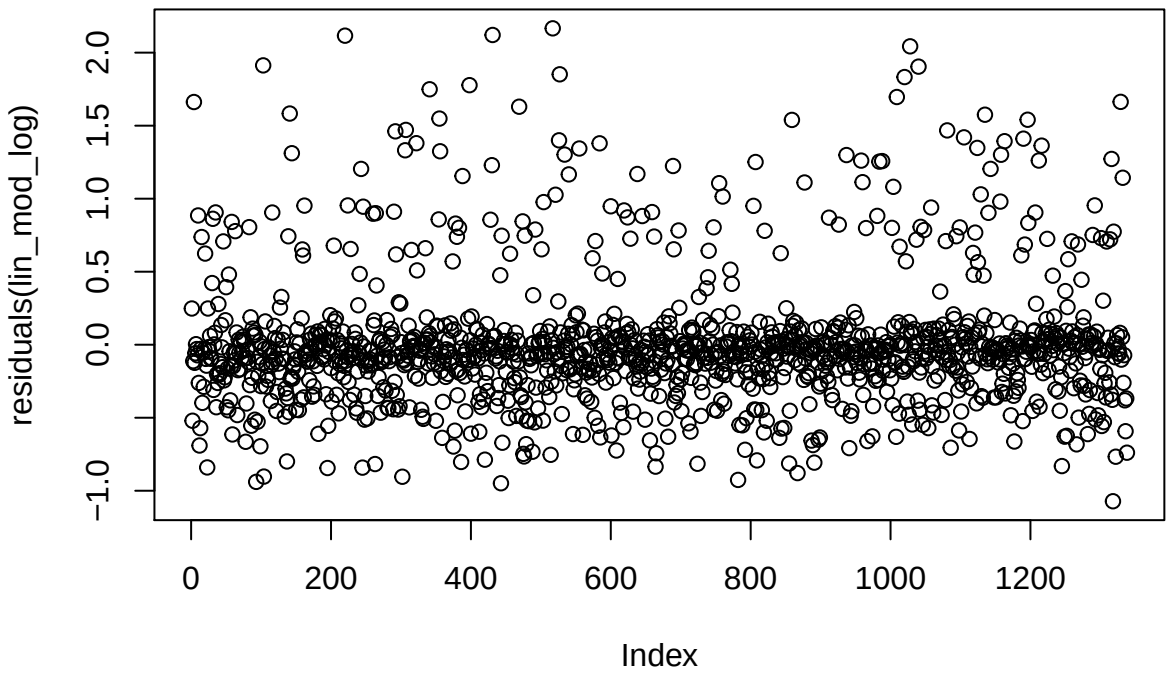
```
##
## Call:
## lm(formula = log(charges) ~ age + sex + bmi + children + smoker +
##     region, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07186 -0.19835 -0.04917  0.06598  2.16636
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
## age              0.0345816  0.0008721  39.655  < 2e-16 ***
## sexmale         -0.0754164  0.0244012  -3.091 0.002038 **
## bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
## children         0.1018568  0.0100995  10.085  < 2e-16 ***
## smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
## regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .
## regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
## regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4443 on 1329 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7666
## F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Question 4.5

```
#Question 4.5
plot(x=data$charges,y=residuals(lin_mod_log), xlab = "Health insurance Charges (Dollars)", ylab = "Resi
```




```
plot(residuals(lin_mod_log))
```



Question 5

Question 6