Medical Expenses Analysis

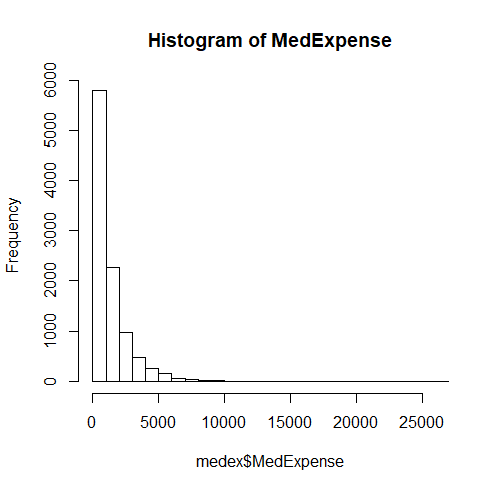
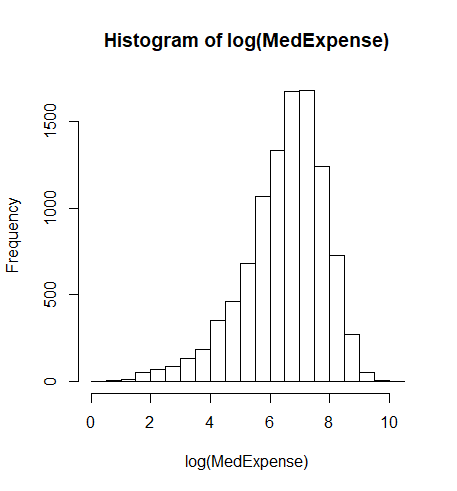
1. **First draw a histogram of MedExpense. Does it seem like this data is suitable as a dependent (response) variable in a linear regression model? If not, what can you do to make it suitable for linear regression?**

#histogram of MedExpense

hist(MedExpense, breaks=20, main="Histogram of MedExpense")

#histogram of a log transformed MedExpense

hist(log(MedExpense), breaks=20)

****

This data may not be suitable as the dependent variable in a linear regression model because it displays an exponential trend downward in frequency of data points as MedExpense increases, whereas a linear regression is used in scenarios where there is a linear relationship between X and Y. A possible solution to this is applying a logarithmic function to MedExpense to try and achieve a more normal distribution.

1. **Which of the above variables do you think will influence MedExpense? How? Write down the corresponding alternative hypotheses, with a one-sentence rationale for each hypothesis. Be sure to include the right signs (positive or negative) for each hypothesis.**

* Age will have a positive effect on MedExpense
  + Older people are more likely to visit the hospital for more complicated issues
  + Alternative - age has a negative or no significant effect on MedExpense
* Number of illnesses will have a positive effect on MedExpense
  + Higher number of illnesses means a longer stay and more complicated treatment
  + Alternative - Illnesses has a negative or no significant effect on MedExpense
* Hispanic will have a positive effect on MedExpense
  + Hispanic people typically have worse health outcomes due to demographic factors
  + Alternative - Hispanic will have a negative or no significant effect on MedExpense
* Black will have a positive effect on MedExpense
  + Black people typically have worse health outcomes due to demographic factors
  + Alternative - Black will have a negative or no significant effect on MedExpense
* Married will have a negative effect on MedExpense
  + Married people have spouses who will make them go to the hospital before conditions worsen
  + Alternative - married has no significant effect on MedExpense
* Priority has a positive effect on MedExpense
  + Higher priority patients probably have more extreme illnesses
  + Alternative - priority has no significant effect on MedExpense

1. **Run a linear regression equation to test the hypotheses listed in your answer in question 2. Copy and paste the output in your answer. Is this a reasonable model? Why or why not?**

#convert all factor columns to factors

columns <- c("publicins","privateins","female","hisp","black","married","urban","priority")

medex[columns]<- lapply(medex[columns],as.factor)

str(medex)

#create a model for log(MedExpense) using variables from previous question

m1 = lm(log(MedExpense) ~ age + illnesses + hisp + black + married + priority, data=medex)

summary(m1)

plot(m1)

Summary output:

Call:

lm(formula = log(MedExpense) ~ age + illnesses + hisp + black +

married + priority, data = medex)

Residuals:

Min 1Q Median 3Q Max

-6.3468 -0.6662 0.1469 0.8337 4.1374

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.683060 0.149501 38.013 < 2e-16 \*\*\*

age -0.005267 0.001894 -2.781 0.005436 \*\*

illnesses 0.384437 0.010162 37.831 < 2e-16 \*\*\*

hisp1 -0.178248 0.048222 -3.696 0.000220 \*\*\*

black1 -0.147927 0.041515 -3.563 0.000368 \*\*\*

married1 -0.013827 0.025582 -0.541 0.588862

priority1 0.591365 0.038562 15.335 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.223 on 10082 degrees of freedom

Multiple R-squared: 0.1947, Adjusted R-squared: 0.1943

F-statistic: 406.4 on 6 and 10082 DF, p-value: < 2.2e-16

This is a reasonable model because it shows that four of the six variables I selected as predictors have lower than a .001 p-value, with one additional being lower than .01, showing that these predictors have a significant impact on MedExpense. It shows that my hypotheses regarding the age, hispanic, black, and married variables were wrong, while my hypotheses regarding the illnesses and priority variables were correct.

1. **Test the assumptions of this model to see if it fits the requirements of a linear regression equation. Copy and paste any appropriate graphics and/or tests. Based on your analysis, is there multivariate normality? Homoskedasticity? Linearity? Multicollinearity? Independence?**

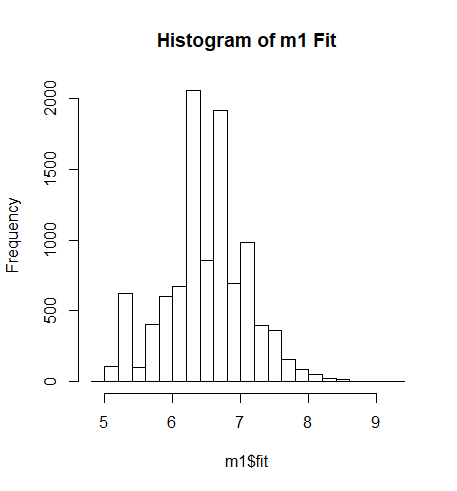
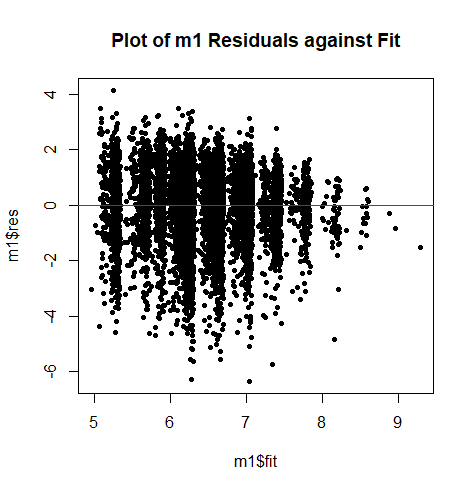
# Residual plot and histogram

plot(m1$res ~ m1$fit, main = "Plot of m1 Residuals against Fit", pch =20)

#draw a flat line at 0 to compare residual dispersion

abline(c(0,0),col ="red")

hist(m1$fit, breaks=20, main = "Histogram of m1 Fit")

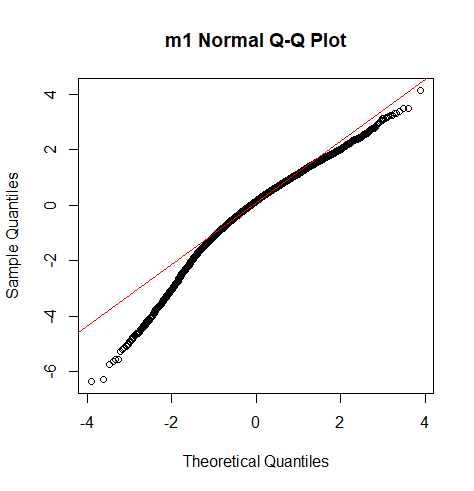


The plot of residuals shows that they are generally evenly distributed about the (0,0) line across the plot, showing that the data is generally unbiased. However, the fanning shape with a bulge in the middle of the data points suggests that the set is heteroscedastic and lacks linearity. The histogram shows a somewhat normal distribution with a couple exceptions, namely a very short bar in the middle of the distribution and an outlying bar near the left end of the chart.

# Q-Q plot

qqnorm(m1$res, main = "m1 Normal Q-Q Plot")

qqline(m1$res, col="red")



The Normal Q-Q plot shows that the dataset adheres mostly to a theoretical normal distribution within the first two quantiles on each side, diverting more significantly from the line on the left side. This generally supports the requirements of a linear regression.

# Shapiro-Wilk test of multivariate normality

resSample <- sample(m1$res, 5000, replace = FALSE)

shapiro.test(resSample)

Shapiro-Wilk normality test

data: resSample

W = 0.96532, p-value < 2.2e-16

To use the Shapiro-Wilk test I first took a sample of 5000 residuals since that is the max input allowed. A p-value close to zero indicates that the null hypothesis is rejected and that the sample data is not normally distributed.

# Kolmogorov-Smirnov test

#generate random normal dist of 10,000 points

norm <- rnorm(10000)

ks.test(norm, m1$res)

Two-sample Kolmogorov-Smirnov test

data: norm and m1$res

D = 0.073019, p-value < 2.2e-16

alternative hypothesis: two-sided

To use the KS test first a random normal distribution of 10,000 values is generated for comparison of goodness of fit. The KS tests states a p-value of near zero which means that the null hypothesis is rejected, indicating that the residuals are not normally distributed.

# Bartlett's test of homoskedasticity

bartlett.test(list(m1$res, m1$fit))

Bartlett test of homogeneity of variances

data: list(m1$res, m1$fit)

Bartlett's K-squared = 4704.3, df = 1, p-value < 2.2e-16

The Bartlett test’s null hypothesis is that the variances of the two compared sets are equal. A p-value close to zero says that we reject the null hypothesis, indicating that variance for the two sets is different, meaning they are heteroskedastic.

# Levene's test of homoskedasticity

leveneTest(m1$res, m1$fit, center=mean)

Levene's Test for Homogeneity of Variance (center = mean)

Df F value Pr(>F)

group 1129 1.5982 < 2.2e-16 \*\*\*

8959

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Levene’s test is less sensitive to deviations from a normal distribution, which might be suitable for this data. Like Bartlett, Levene’s null hypothesis is that the variances between the two sets are equal. Again, we reject the null hypothesis given the p-value, meaning that there is a difference between the sets; they are heteroskedastic.

# Variance inflation factor - test of multicollinearity

vif(m1)

age illnesses hisp black married priority

1.081278 1.164897 1.008028 1.019250 1.086459 1.158204

In the VIF test, a 1 means that there is no correlation among a given predictor and the remaining predictor variables. Each of the inflation factors is close to 1, indicating that each variable is not at all significantly correlated with any other variable; multicollinearity is very low.

# Durbin-Watson test of autocorrelation

dwtest(m1)

Durbin-Watson test

data: m1

DW = 0.38047, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

The Durbin-Watson statistic null hypothesis is that there is no autocorrelation. Since p is close to zero we reject the null hypothesis and say that there is autocorrelation. The DW statistic of .38 is between 0 and 2, indicating positive autocorrelation.

1. **What can you do to further improve your model? Run additional analysis if necessary and test your new analysis.**

To further improve the model we can attempt different combinations or exclusions of variables using a stepwise approach, as well as integrating compound variables such as “hasInsurance” or “isMinority”. We can also try different transformations of the variables, such as log or exponential transforms.

The first step is to add compound variables to the dataset, as mentioned above, the two best candidates are to combine the insurance variables and the minority variables. The code for this involves if/else statements and is shown in the relevant parts of the next question.

The second step is to conduct a stepwise regression; this is done in both directions to find the best combination of available variables to maximize the R-squared value.

#Stepwise regression

#Create a full "kitchen-sink" model

library(MASS)

mFull <- lm(log(MedExpense) ~.-medexpense, data=medex)

#stepwise model in both directions

stepwise <- stepAIC(mFull, direction = "both", trace = FALSE)

summary(stepwise)

Call:

lm(formula = log(MedExpense) ~ publicins + age + female + income +

illnesses + ssiratio + educyr + urban + priority + isMinority,

data = medex)

Residuals:

Min 1Q Median 3Q Max

-6.3667 -0.6602 0.1434 0.8312 4.0136

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.4304656 0.1579700 34.377 < 2e-16 \*\*\*

publicins1 0.0907243 0.0261921 3.464 0.000535 \*\*\*

age -0.0051038 0.0018842 -2.709 0.006766 \*\*

female1 0.0582993 0.0249136 2.340 0.019300 \*

income 0.0010030 0.0006588 1.522 0.127925

illnesses 0.3810972 0.0101816 37.430 < 2e-16 \*\*\*

ssiratio 0.1859119 0.0385805 4.819 1.47e-06 \*\*\*

educyr 0.0066460 0.0041688 1.594 0.110916

urban1 -0.0464212 0.0280813 -1.653 0.098342 .

priority1 0.5927249 0.0384859 15.401 < 2e-16 \*\*\*

isMinority1 -0.1449067 0.0343549 -4.218 2.49e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.22 on 10078 degrees of freedom

Multiple R-squared: 0.1979, Adjusted R-squared: 0.1971

F-statistic: 248.7 on 10 and 10078 DF, p-value: < 2.2e-16

The stepwise regression chose most of the available variables, with the exception of PrivateIns, Hisp, Black, and Married. Instead of Hisp or Black, it is using the compound variable that combines these two, isMinority.

1. **Based on your analysis, answer the following questions:**
   * **Do people with health insurance have higher or lower medical expense than people without health insurance, when other variables are controlled? By how much? Why do you think this happens?**

#Create new "hasInsurance" variable by combining publicins and privateins using if/else

medex$hasInsurance <- ifelse(medex$publicins==1 | medex$privateins==1,1,0)

#Change to factor

medex$hasInsurance <- as.factor(medex$hasInsurance)

#Shows that hasInsurance is now a factor with 2 levels, no insurance and has insurance

str(medex)

m6A = lm(MedExpense ~ hasInsurance, data=medex)

summary(m6A)

Call:

lm(formula = MedExpense ~ hasInsurance, data = medex)

Residuals:

Min 1Q Median 3Q Max

-1309.4 -972.4 -489.4 403.1 25064.6

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1216.85 30.82 39.480 < 2e-16 \*\*\*

hasInsurance1 93.56 35.45 2.639 0.00833 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1530 on 10087 degrees of freedom

Multiple R-squared: 0.0006899, Adjusted R-squared: 0.0005908

F-statistic: 6.964 on 1 and 10087 DF, p-value: 0.008331

People with insurance have higher medical expenses, about $93.56 more, than people without insurance. This probably happens because hospitals or other providers are more willing to bill for a higher number of, or more expensive procedures when they know that an insurance company will pick up the cost. People without insurance are expected to not be able to pay, and are therefore charged less.

* + **Do people with more illnesses have higher or lower medical expense than people with less illnesses? By how much?**

m6B = lm(MedExpense ~ illnesses, data=medex)

summary(m6B)

Call:

lm(formula = MedExpense ~ illnesses, data = medex)

Residuals:

Min 1Q Median 3Q Max

-3069.1 -804.7 -352.5 369.6 24589.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 473.98 24.82 19.10 <2e-16 \*\*\*

illnesses 437.19 10.95 39.91 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1422 on 10087 degrees of freedom

Multiple R-squared: 0.1364, Adjusted R-squared: 0.1363

F-statistic: 1593 on 1 and 10087 DF, p-value: < 2.2e-16

People with more illnesses have a higher medical expense than people with less illnesses. For each additional illness, a person is expected to pay about $437.19 more.

* + **Do males have higher medical expense than females? By how much?**

m6C = lm(MedExpense ~ female, data=medex)

summary(m6C)

Call:

lm(formula = MedExpense ~ female, data = medex)

Residuals:

Min 1Q Median 3Q Max

-1306.3 -974.3 -491.3 395.7 25066.7

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1259.33 23.43 53.756 <2e-16 \*\*\*

female1 48.92 30.84 1.586 0.113

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1530 on 10087 degrees of freedom

Multiple R-squared: 0.0002494, Adjusted R-squared: 0.0001503

F-statistic: 2.517 on 1 and 10087 DF, p-value: 0.1127

Females have higher medical expenses than males by about $48.92.

* + **Do older people have higher medical expense than younger people? By how much?**

m6D = lm(MedExpense ~ age, data=medex)

summary(m6D)

Call:

lm(formula = MedExpense ~ age, data = medex)

Residuals:

Min 1Q Median 3Q Max

-1302.3 -974.0 -490.8 395.6 25079.9

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1426.494 171.821 8.302 <2e-16 \*\*\*

age -1.851 2.280 -0.812 0.417

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1530 on 10087 degrees of freedom

Multiple R-squared: 6.533e-05, Adjusted R-squared: -3.381e-05

F-statistic: 0.659 on 1 and 10087 DF, p-value: 0.4169

Older people have lower medical expenses than younger people by about $1.85 per year of age.

* + **Do minority groups (Blacks/Hispanics) have higher or lower medical expenses than the non-minority population? By how much?**

#Create new "hasInsurance" variable by combining publicins and privateins using ifelse

medex$isMinority <- ifelse(medex$hisp==1 | medex$black==1,1,0)

#Change to factor

medex$isMinority <- as.factor(medex$isMinority)

#Shows that hasInsurance is now a factor with 2 levels, no insurance and hasinsurance

str(medex)

m6E1 = lm(MedExpense ~ isMinority, data=medex)

summary(m6E1)

Call:

lm(formula = MedExpense ~ isMinority, data = medex)

Residuals:

Min 1Q Median 3Q Max

-1308.3 -972.3 -489.3 397.7 25065.7

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1309.29 16.65 78.627 < 2e-16 \*\*\*

isMinority1 -132.83 41.18 -3.226 0.00126 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1530 on 10087 degrees of freedom

Multiple R-squared: 0.001031, Adjusted R-squared: 0.0009316

F-statistic: 10.41 on 1 and 10087 DF, p-value: 0.00126

m6E2 = lm(MedExpense ~ hisp + black, data=medex)

summary(m6E2)

Call:

lm(formula = MedExpense ~ hisp + black, data = medex)

Residuals:

Min 1Q Median 3Q Max

-1308.2 -972.2 -489.2 398.8 25065.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1309.17 16.63 78.735 <2e-16 \*\*\*

hisp1 -198.04 60.18 -3.291 0.001 \*\*

black1 -81.87 51.53 -1.589 0.112

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1530 on 10086 degrees of freedom

Multiple R-squared: 0.001268, Adjusted R-squared: 0.00107

F-statistic: 6.403 on 2 and 10086 DF, p-value: 0.001664

Minority groups have lower medical expenses than the non-minority population. Looking at the two minority groups together, the minority population has lower expenses of about $132.83. Separately, Hispanics have about $198 lower expenses while Blacks have about $82 lower expenses.

* + **Do people with private insurance pay more or less than people with public insurance? By how much?**

m6F = lm(MedExpense ~ publicins + privateins, data=medex)

summary(m6F)

Call:

lm(formula = MedExpense ~ publicins + privateins, data = medex)

Residuals:

Min 1Q Median 3Q Max

-1347.9 -972.9 -484.0 400.1 25026.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1216.85 30.82 39.488 < 2e-16 \*\*\*

publicins1 132.09 39.45 3.348 0.000817 \*\*\*

privateins1 54.14 39.63 1.366 0.171940

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1530 on 10086 degrees of freedom

Multiple R-squared: 0.00118, Adjusted R-squared: 0.000982

F-statistic: 5.958 on 2 and 10086 DF, p-value: 0.002594

People with private insurance pay less than those with public insurance. They pay about $78 less. However, there is not a strong significance of the effect of private insurance.

* + **How do people’s income level relate to their medical expense, when controlled for other factors? By how much?**

m6G = lm(MedExpense ~ income, data=medex)

summary(m6G)

Call:

lm(formula = MedExpense ~ income, data = medex)

Residuals:

Min 1Q Median 3Q Max

-1329.9 -974.5 -487.9 396.1 25036.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1339.6429 21.7268 61.658 < 2e-16 \*\*\*

income -2.3542 0.7005 -3.361 0.00078 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1530 on 10087 degrees of freedom

Multiple R-squared: 0.001118, Adjusted R-squared: 0.001019

F-statistic: 11.29 on 1 and 10087 DF, p-value: 0.0007802

Income is negatively correlated with medical expense; those with higher incomes typically pay less. For every dollar of income, medical expenses are expected to be about $2.35 lower.

## Full Code

rm(list=ls())

library(ggplot2)

library("car")

library(lmtest)

setwd("C:/Users/ajohnson/Downloads")

medex = read.csv("medicalexpenses.csv")

attach(medex)

names(medex) <- tolower(names(medex))

#Question 1 - Histogram

hist(MedExpense, breaks=20, main="Histogram of MedExpense")

hist(log(MedExpense), breaks=20)

#Question 3 - Linear Regression Test

#change relevant variables into factors

columns <- c("publicins","privateins","female","hisp","black","married","urban","priority")

medex[columns]<- lapply(medex[columns],as.factor)

str(medex)

m1 = lm(log(MedExpense) ~ age + illnesses + hisp + black + married + priority, data=medex)

summary(m1)

plot(m1)

#Question 4 - Test Assumptions

# Linearity

plot(MedExpense ~m1$fit)

abline(0,1,lwd=3,col="red")

# Residual plot

plot(m1$res ~ m1$fit, main = "Plot of m1 Residuals against Fit", pch =20)

abline(c(0,0),col ="red")

hist(m1$fit, breaks=20, main = "Histogram of m1 Fit")

# Q-Q plot

qqnorm(m1$res, main = "m1 Normal Q-Q Plot")

qqline(m1$res, col="red")

# Shapiro-Wilk's test of multivariate normality

resSample <- sample(m1$res, 2000, replace = FALSE)

shapiro.test(resSample)

# Kolmogorov-Smirnov test

norm <- rnorm(10000)

ks.test(norm, m1$res)

# Bartlett's test of homoskedasticity

bartlett.test(list(m1$res, m1$fit))

# Levene's test of homoskedasticity

leveneTest(m1$res, m1$fit, center=mean)

# Test of multicollinearity - Variance inflation factor

vif(m1)

# Durbin-Watson test of autocorrelation

dwtest(m1)

#Question 5

library(MASS)

library(leaps)

#Stepwise regression

#Create a full "kitchen-sink" mode

mFull <- lm(log(MedExpense) ~.-medexpense, data=medex)

stepwise <- stepAIC(mFull, direction = "both", trace = FALSE)

summary(stepwise)

#Question 6

#Create new "hasInsurance" variable by combining publicins and privateins using ifelse

medex$hasInsurance <- ifelse(medex$publicins==1 | medex$privateins==1,1,0)

#Change to factor

medex$hasInsurance <- as.factor(medex$hasInsurance)

#Shows that hasInsurance is now a factor with 2 levels, no insurance and hasinsurance

str(medex)

m6A = lm(MedExpense ~ hasInsurance, data=medex)

summary(m6A)

plot(m6A)

m6B = lm(MedExpense ~ illnesses, data=medex)

summary(m6B)

plot(m6B)

m6C = lm(MedExpense ~ female, data=medex)

summary(m6C)

plot(m6C)

m6D = lm(MedExpense ~ age, data=medex)

summary(m6D)

plot(m6D)

#Create new "hasInsurance" variable by combining publicins and privateins using ifelse

medex$isMinority <- ifelse(medex$hisp==1 | medex$black==1,1,0)

#Change to factor

medex$isMinority <- as.factor(medex$isMinority)

#Shows that hasInsurance is now a factor with 2 levels, no insurance and hasinsurance

str(medex)

m6E1 = lm(MedExpense ~ isMinority, data=medex)

summary(m6E1)

m6E2 = lm(MedExpense ~ hisp + black, data=medex)

summary(m6E2)

plot(m6E)

m6F = lm(MedExpense ~ publicins + privateins, data=medex)

summary(m6F)

plot(m6F)

m6G = lm(MedExpense ~ income, data=medex)

summary(m6G)

plot(m6G)