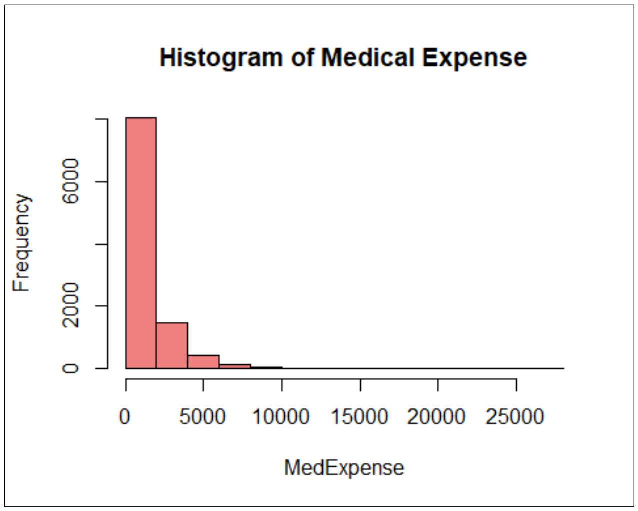
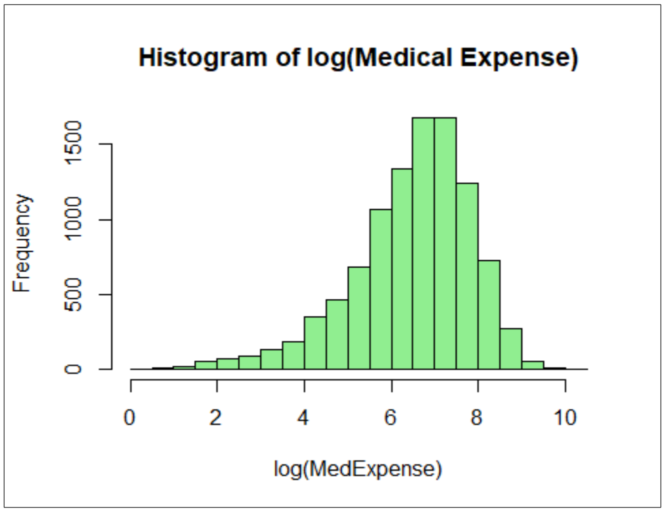
Medical Expense Prediction:

I. First draw a histogram of medexpense. Does it seem like this data is suitable as a dependent (response) variable in an ordinary least square's regression model? If not, what can you do to make it suitable for regression?

Medexpense: Medical expense in this case is an Out-of-pocket cost. These are the medical costs that are not reimbursed by the insurance. Out-of-pocket expenses include premiums, copayments, prescription drugs, deductibles, coinsurance etc.





Medical expense is highly positive (Right) skewed (Plot on the left) so it is not suitable for ordinary least squares regression. In this case we need to apply non-linear transformation which changes linear relationship between variables (i.e., changes correlation between the variables). When we try a log method on our dependent variable medexpense, it seems to follow a normal distribution (Plot on the right). So, log(medexpense) becomes our new dependent variable.

II. Examine each variable in this data set systematically on whether or not that variable should be a predictor of medexpense. Create a table with the following three columns: (1) predictor name, (2) the sign of the hypothesized effect of that variable on medexpense (hypotheses), and (3) a one-sentence rationale for that predicted effect.

Factor Effects:

Positive Effect: If there is a direct proportionality (+X then +y / -X then -y) between the predictor variable (X) and the response variable (y), we can say there is a positive effect.

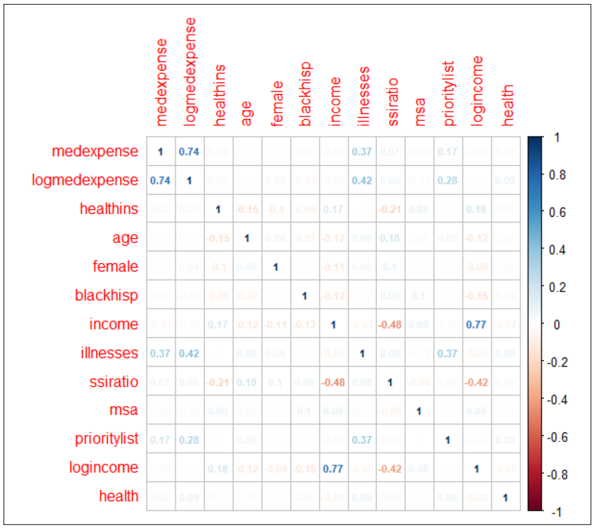
Negative Effect: If there is an inverse proportionality (+X then -y / -X then +y) between the predictor variable (X) and the response variable (y), we can say there is a negative effect.

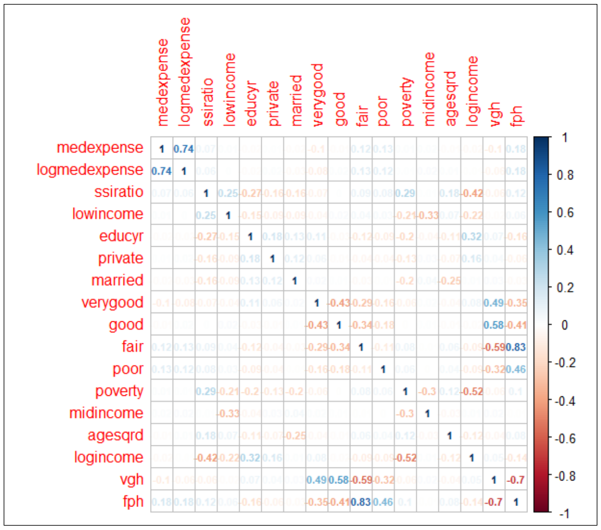
Relevant Factors for Medical Expense		
Predictor	**Effect**	**Rationale**
healthins	-	Out-of-pocket costs depend on whether patients have insurance coverage. If a patient does not have insurance coverage, then said person will pay more for medical treatments.
age	+	As a person's age increases, their health care costs increase which could include Out-of-pocket casts as well.
female	+	Considering women's reproductive health and that they live longer than men on average, they tend to have higher health care costs which includes Out-of-pocket costs as well.
blackhisp	+/-	Racial and ethnic differences could lead to a person either having better health insurance or not having any insurance at all which could have varying effects on medical expenses.

income	+/-	People with higher income could opt for private insurance plans which also have higher Out-of-pocket costs. Recent policy changes also add to the medical expenses of higher-income groups.
illnesses	+	Number of illnesses directly affects medical expenses as the patients will need to make frequent trips to their doctor.
ssiratio	-/+	Supplemental security income ratio is a function of income and is used for medicare settlement purposes or low-income payment adjustment. Although income is negatively correlated with ssiratio and explains about 50% of the data, it is worth looking at this variable as there are other factors which influence the other 50% of the data like SSI federal benefit rate, Social security benefit value etc.
verygood, good and fair	-	Patients with verygood/good/fair health does not need to pay regular visits to their doctor which could in turn reduce their medical expenses (Combined with poor as a single categorical column).
poor	+	Patients with poor health needs to pay regular visits to their doctor which adds to their medical expenses (Combined with verygood/good/fair as a single categorical column).
msa	+	Patients living in urban areas have access to better health services/clinics which could translate to higher Out-of-pocket costs.
prioritylist	+	“Elderly or in fragile health” could mean patients with genetic defects or hereditary/preexisting health conditions, which means these patients receive priority. This could directly influence Out-of-pocket costs.

Irrelevant Factors

Predictor	Effect	Rationale
lowincome	No Effect	This variable is correlated with income so can be omitted.
firmsize	No Effect	Firm size does not have any effect on predicting medical expenses.
firmlocation	No Effect	Firm location does not have any effect on predicting medical expenses.
educyr	No Effect	Level of education of a patient is not a factor for Out-of-pocket expenses.
private	No Effect	The type of health insurance does not have any effect on medical expenses.
hisp	No Effect	This variable is already included.
black	No Effect	This variable is already included.
married	No Effect	Marital status does not have any effect on medical expenses as it depends on the individual’s health.
verygood	No Effect	Heath condition correlates with good and fair (combined as single variable).
good	No Effect	Heath condition correlates with verygood and fair (combined as single variable).
fair	No Effect	Heath condition correlates with verygood and good (combined as single variable).
poverty	No Effect	Poverty is correlated with income so can be omitted.
midincome	No Effect	Since we have considered income as a variable, we can omit midincome.
agesqrd	No Effect	This variable is a function of age so can be omitted.
logincome	No Effect	This variable is highly correlated with income so can be omitted.





Correlation matrix above helps us decide which variables to consider and which to omit from our models. If there is a high correlation (>0.7) between any two variables, then we need to consider only one variable. We can see verygood, good and fair are correlated. So, these can be combined to form a single categorical column along with the poor variable. We can see some correlation between log(medexpense) and illness, prioritylist but we will consider these variables as they could contribute to medical expenses.

III. Run three "reasonably good" regression models to predict medexpense, using the variables you hypothesized in your answer to Question 2. Summarize the results of these models stargazer. Copy and paste the models and the stargazer output.

Summary:
Model m1: Consists of variables that has the highest effect on medical expenses.
Model m2: Consists of all the variables that are relevant based on our prior analysis in Question II.
Model m3: Consists of key variables along with some interaction terms to understand the relationships between the variables.

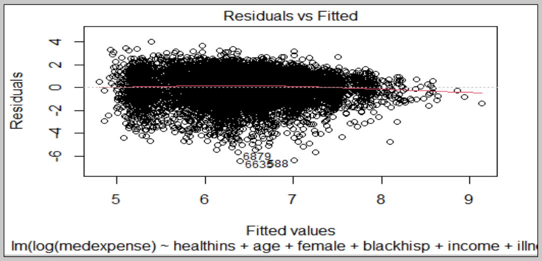
```
#Regression models
m1 <- lm(log(medexpense) ~ blackhisp + illnesses + health + prioritylist, data = hi)
m2 <- lm(log(medexpense) ~ healthins + age + female + blackhisp + income + illnesses + ssiratio + health
         + msa + prioritylist, data = hi)
m3 <- lm(log(medexpense) ~ healthins + age + female + blackhisp + illnesses*age + ssiratio*age + health
         + msa + prioritylist, data = hi)
```

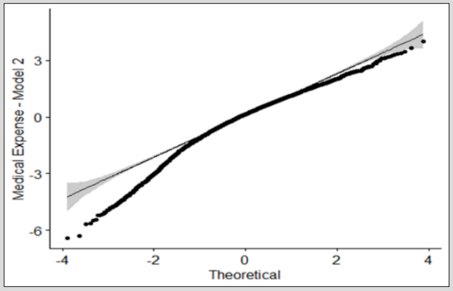
```
#Stargazer
stargazer(m1, m2, m3, type='text', single.row = TRUE)
```

```
##
## =======================================================================================================
##                                                    Dependent variable:
##                     -----------------------------------------------------------------------------------
##                                                      log(medexpense)
##                                 (1)                         (2)                         (3)
## -------------------------------------------------------------------------------------------------------
## healthins                                           0.096*** (0.026)            0.100*** (0.026)
## age                                                 -0.005*** (0.002)            0.007* (0.004)
## female                                               0.057** (0.025)             0.056** (0.025)
## blackhisp               -0.164*** (0.033)           -0.163*** (0.033)           -0.170*** (0.033)
## income                                               0.001** (0.001)
## illnesses                0.380*** (0.010)           0.377*** (0.010)            0.704*** (0.108)
## ssiratio                                            0.181*** (0.038)             0.810** (0.364)
## health                   0.088*** (0.017)           0.090*** (0.017)            0.087*** (0.017)
## msa                                                  -0.040 (0.028)              -0.035 (0.028)
## prioritylist             0.579*** (0.039)           0.581*** (0.038)            0.574*** (0.039)
## age:illnesses                                                                   -0.004*** (0.001)
## age:ssiratio                                                                     -0.009* (0.005)
## Constant                 5.156*** (0.042)           5.395*** (0.149)            4.501*** (0.304)
##
## -------------------------------------------------------------------------------------------------------
## Observations                  10,089                     10,089                      10,089
## R2                            0.196                       0.200                       0.201
## Adjusted R2                   0.196                       0.199                       0.200
## Residual Std. Error     1.221 (df = 10084)         1.219 (df = 10078)          1.218 (df = 10077)
## F Statistic         615.856*** (df = 4; 10084) 251.935*** (df = 10; 10078) 230.028*** (df = 11; 10077)
## =======================================================================================================
## Note:                                                                       *p<0.1; **p<0.05; ***p<0.01
```

IV. Select the best model for Question 3 and test if this model meets the assumptions of OLS regression. Copy and paste any appropriate graphics and/or tests. Based on your analysis, is your analysis appropriate for your data?

The best model from our analysis is m2. Model m2 is the best choice as m3 is multicollinear with various interaction terms.

Assumption	DV Model: m2 (Medical Expenses)
Linearity: Passed  	Residual vs fitted values plot does not show any patterns in the spread and the data is mostly unbiased. We can say the linearity condition is met.

Assumption	DV Model: m2 (Medical Expenses)
Normality: Failed • **Kolmogorov-Smirnov Test: Since Sample Size > 2000** **1. If p < 0.05, reject Ho ⇒ data is not normally distributed.** **2. If p > 0.05, fail to reject Ha ⇒ data may be normally distributed.**	 D = 0.059476, p-value = 7.772e-16 **Residual deviation at lower quartiles observed. This model fails Kolmogorov-Smirnov test.**

Assumption	DV Model: m2 (Medical Expenses)
Homoscedasticity: Failed • **Bartlett Test / Breusch-Pagan test** **1. If p < 0.05, reject Ho ⇒ two samples have unequal variances.** **2. If p > 0.05, fail to reject Ha ⇒ two samples may have equal variances.**	Bartlett's K-squared = 4502.3, df = 1, p-value < 2.2e-16 **This model fails Bartlett test.** BP = 204.53, df = 10, p-value < 2.2e-16 **This model fails Breusch-Pagan test.** **Note: Levene's test crashes R Studio.**

<table>
<tr><th>Assumption</th><th>DV Model: m2 (Medical Expenses)</th></tr>
<tr><td><b>Multicollinearity: Passed</b>
• <b>Variance Inflation Factor (VIF)</b>
<b>1. VIF = 1/T (Where T = 1 – R² , T < 0.1 is indicative of multicollinearity).</b>
<b>2. VIF > 5 indicates multicollinearity. VIF > 10 is strong evidence of multicollinearity.</b></td><td><pre>vif(m2)</pre><pre>##   healthins   age   female   blackhisp   income   illnesses
##   1.087503   1.068120   1.026191   1.041426   1.333087   1.174758
##   ssiratio   health   msa prioritylist
##   1.349998   1.015312   1.033319   1.160544</pre><b>This model passed the VIF test.</b></td></tr>
</table>

Assumption	DV Model: m2 (Medical Expenses)
Independence: Passed • **Durbin-Watson's Test (DW)** **1. Ho: Residuals are not linearly auto-correlated.** **2. DW ~ [0, 4]; values around 2 (i.e., 1.5 to 2.5) suggests no autocorrelation.**	DW = 1.8036, p-value < 2.2e-16 **This model passed the Durbin-Watson's test.**

V. Use your best model to answer the following questions: [Selected Model = m2]

- **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?**
  People with health insurance will have higher medical expenses by 9.6% than those without health insurance. One of the reasons for this could be that people with health insurance get preventive care, vaccines, health checkups etc. for less but in the process must deal with Out-of-pocket costs often (It is kind of like a double-edged sword where not having insurance means higher medical costs but having medical insurance means lower medical costs but more out-of-pocket costs).

- **Do people with private insurance pay more or less than people with public insurance? By how much?**
  Since we dropped this variable, we do not have any information to infer from our analysis. The reason for dropping this variable is that any health insurance (Private or Public) aims to reduce unexpected high medical costs. So, type of insurance will not have any effect on Out-of-pocket costs.

- **Do people with more illnesses have higher or lower medical expense than people with less illnesses? By how much?**
  People with more illnesses have higher medical expenses by 37.7% than those that have less illnesses.

- **Do males have higher medical expense than females? By how much?**
  Males have lower medical expenses by 5.7% than females.

- **Do older people have higher medical expense than younger people? By how much?**
  Based on our analysis, older people have lower medical expenses by 0.5% for a unit increase in age. It could be that we are seeing the combined effect of age and illnesses in the prioritylist variable at 58.1% increase in medical expenses if he/she is on the priority list.

- **Do minority groups (Blacks/Hispanics) have higher or lower medical expenses than the non-minority population? By how much?**
  Minority groups (Blacks/Hispanics) have lower medical expenses by 16.3% than the non-minority population. This could mean minority groups, even on an insurance plan are not seeking preventive care, vaccines, health checkups etc. on a frequent basis.

- **How do people's income level relate to their medical expense, when controlled for other factors? By how much?**
  People in higher Income level will have their medical expenses increased by 0.1%. It could be that we are seeing combined effect of income and other factors (SSI federal rate, Social security benefit value etc.) in ssiratio variable at 18.1% increase in medical expenses.