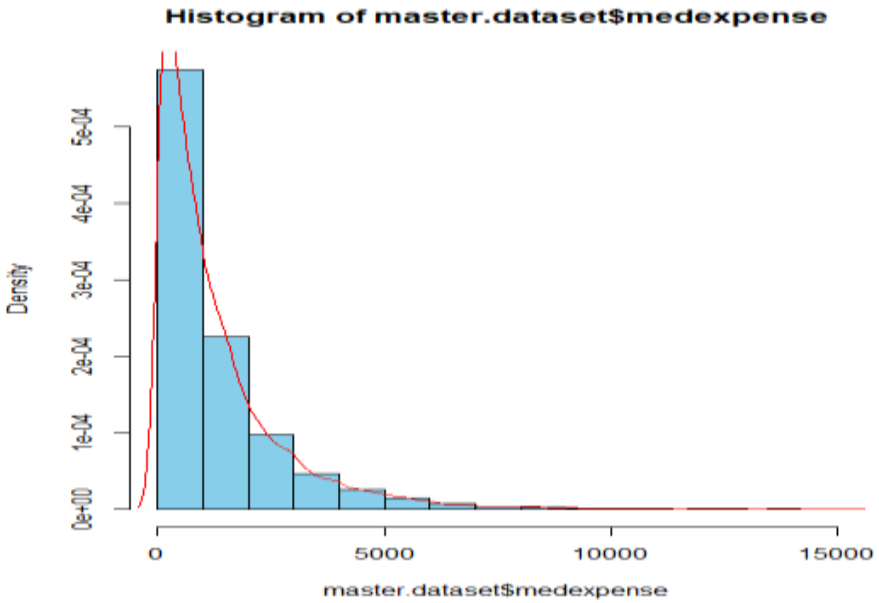
Medical Expenses – Inferential Data Analysis

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



Answer: From the above histogram we can infer that variable **medexpense has right skewed distribution** and hence may not be suitable for OLS regression model. In order to normalise it more we can perform log transform or square root transform. In this assignment, I **have used log transformation** and hence considered **logmedexpense** as the dependent variable.

2. **Examine <u>each</u> 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. (2 points)**

Relevant Independent variables

Predictors	Effect	Rationale
Healthins	+/-	Health Insurance plans vary based on deductibles and copay amounts. Although many insurance plans cover most of the medical costs but some insurance plans does have high co-pays are other deductible costs thus increasing out of pocket costs.
Age	+	Generally, Person with older age has more probability of falling ill and hence increased out of pocket medical expenses.
Female	+	According to my research female tend to increase medical expenses especially due to maternity clinic services, female health problems and other medical treatment.
Blackhisp	+/-	As per my research there are mixed reactions on whether minority people spend more on med expense than non-minority people. Hence medical expense can increase or decrease for black or Hispanic people.
Logincome	+	People with higher income tend to prefer expensive hospital services and medical costs thus increasing out of pocket expense. Income variable has right skewed distribution and hence I have performed log transformation.

Illnesses	+	People with more illnesses and prior medical history tend to spent more on medical expenses.
Ssiratio	-	Increased social security benefits will cover medication and hospitalization cost thus decreasing out of pocket medical expense.
Private	-	Private insurance generally offers more flexible plans to the patients on selection of medical services covering most of the expenses and with zero deductibles and co-pays thus decreasing out of pocket expense.
Prioritylist	+	Priority patients tend to have prior medical history thus leading to frequent medical expense resulting into higher out of pocket expense.

Other variables such as firmsize, firmlocation, lowincome, educyr, married, good, verygood, fair, poor, poverty, midincome, msa, vgh, fph are not directly related to medexpense and hence they don't make significant effect on the model and therefore can be neglected.

2. 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. (2 points)

Model 1

medex1=lm(logmedexpense~healthins+age+female+income+illnesses+ssiratio+private+blackhisp+poverty+msa+prioritylist+healthc, data=sample.dataset)

Model 2

medex2=lm(logmedexpense~healthins+age+female+blackhisp+log(income)+illnesses+private+prioritylist+ssiratio+msa+I(income^2)+I(illnesses^2)+I(ssiratio^2), data=sample.dataset)

Model 3

medex3=lm(logmedexpense~healthins+age+female+blackhisp+illnesses+log(income)+ssiratio+private+prioritylist+I(illnesses^2)+I(ssiratio^2)+log(income):illnesses, data=sample.dataset)

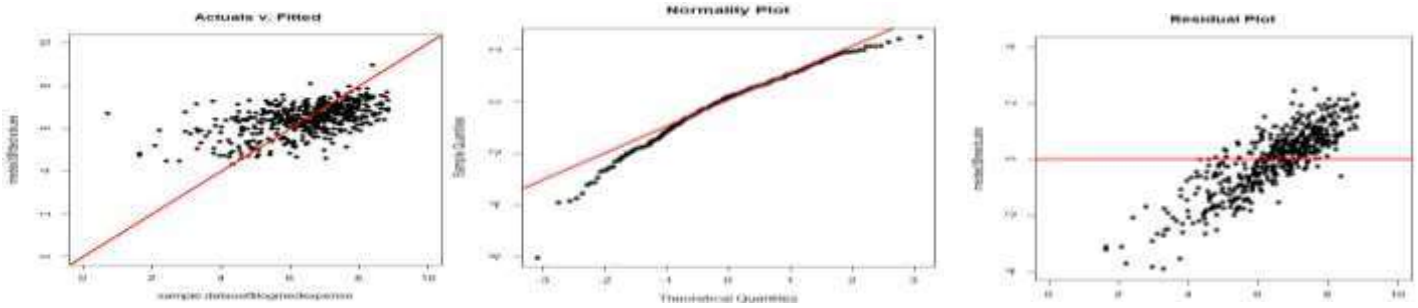
Regression Results

	Dependent variable:		
	logmedexpense		
	(1)	(2)	(3)
healthins	0.398*** (0.144)	0.413*** (0.144)	0.416*** (0.143)
age	0.005 (0.008)	0.006 (0.008)	0.007 (0.008)
female	0.086 (0.110)	0.068 (0.109)	0.068 (0.109)
income	0.00000 (0.00000)		
illnesses	0.352*** (0.043)	0.563*** (0.122)	1.062** (0.465)
ssiratio	0.350** (0.168)	-0.198 (0.335)	-0.201 (0.328)
private	-0.253* (0.143)	-0.278* (0.143)	-0.285** (0.142)
blackhisp	-0.227 (0.142)	-0.241* (0.140)	-0.230* (0.139)
poverty	0.060 (0.160)		
log(income)		0.022 (0.069)	0.102 (0.097)
msa	0.047 (0.123)	0.048 (0.122)	
I(income2)		-0.000 (0.000)	
I(illnesses2)		-0.041* (0.023)	-0.041* (0.023)
I(ssiratio2)		0.467* (0.253)	0.450* (0.249)
illnesses:log(income)			-0.052 (0.047)
prioritylist	1.030*** (0.168)	0.969*** (0.171)	0.957*** (0.170)
healthc	-0.047 (0.036)		
Constant	4.391*** (0.671)	3.986*** (1.001)	3.207*** (1.213)
Observations	500	500	500
R2	0.265	0.272	0.274
Adjusted R2	0.247	0.253	0.256
Residual Std. Error	1.183 (df = 487)	1.179 (df = 486)	1.176 (df = 487)
F Statistic	14.630*** (df = 12; 487)	13.997*** (df = 13; 486)	15.316*** (df = 12; 487)
Note:			*p<0.1; **p<0.05; ***p<0.01

As compared to model 1 and model 2, model 3 has lowest residual standard error of 1.176 and highest F statistic and R-square value also looking at the beta coefficients, I believe, model 3 is the best model and hence I have answered the following questions using model 3.

Model 3 contain quadratic and interaction terms. This terms help model to explain relationships between the variable more clearly and from the beta coefficients and p value we can say that they make some impact in the model.

3. 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? (2 points)



Linearity **Normality** **Equality of Variances**

Linearity: **Failed**. From the plot we can infer that there is no linear relationship between Actual and fitted values as many points deviate from the line.

Normality: **Failed.** From the Normality plot we can see some points which deflect from the line at the lower and upper tail. I performed **Shapiro-Wilk** normality test where I got w = 0.96555 and P value 1.935e-09 which is less than 0.05. Hence we can reject null hypothesis and infer that data/residuals are not normally distributed.

Equality of Variances/Homoscedasticity: **Failed.** Performed **Bartlett's test** and got Barlett's K-squared =113.94 and p value as 2.2e-16. Thus we can reject NULL hypothesis and infer that Homoscedasticity assumption does not hold true.

Autocorrelation: **Passed.** Performed **Durbin-Watson** test of autocorrelation and got DW = 1.9743. Here DW is in the range of 2 hence we can infer that residuals are not linearly auto correlated.

Multicollinearity: The predictor variable 'illnesses' have high VIF value of 133.53 and hence has high evidence of multicollinearity. Other variables such as healthins(1.79), age(1.07), female(1.05), balckhisp(1.04),log(income)(2.92), ssiratio(5.13), private(1.69), prioritylist(1.17) have VIF values less than 10.

4. Use your best model to answer the following questions (2 points):

1. **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 have higher medical expense. People with health insurance spend 41.60% more than people with no health insurance. This could be because of the varying insurance deductibles plans and benefits and changing government Medicare policies and regulations.

2. **Do people with private insurance pay more or less than people with public insurance? By how much?**
   People having private insurance pay 28.5% less than people who don't have private insurance. Although private insurance charge more premium, they offer flexible plans with 0 deductibles and co- pays and cover more medical treatment expenses resulting into less out of pocket expenses.

3. **Do people with more illnesses have higher or lower medical expense than people with less illnesses? By how much?**
   People having more illnesses spend more out of pocket medical costs. With 1 Unit increase in illnesses medical expense increase by approximately 106.17%.

4. **Do males have higher medical expense than females? By how much?**
   From the analysis we can infer that females have 6.3% higher medical expense than males.

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

Older people tend to have higher medical expense than young people. With 1 year increase in age medical expense tend to change by 0.7%.

6. **Do minority groups (Blacks/Hispanics) have higher or lower medical expenses than the non-minority population? By how much?**
   From the analysis we can infer that Minority groups (Blacks and Hispanics) have 23% lower medical expense than non-minority population.

7. **How do people's income level relate to their medical expense, when controlled for other factors? By how much?**
   From the analysis we can infer than 100% increase in income will increase medical expense by 10.16%.