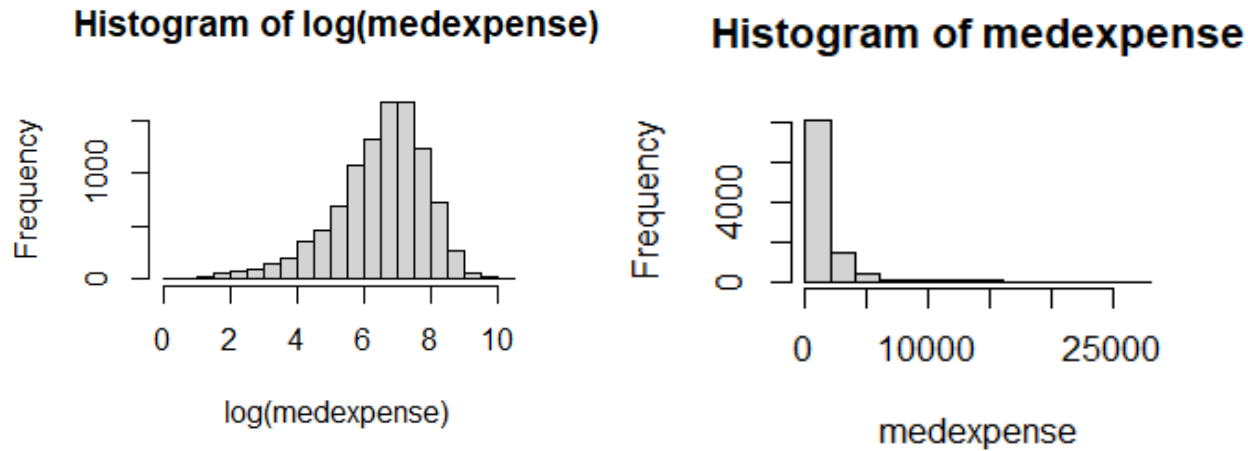


## Medical Expense Analysis

### 1. Preliminary Analysis

```
> hist(medexpense)
> hist(log(medexpense))
```



It is clearly visible that the histogram of medexpense is right skewed so, we cannot use medexpense data directly for OLS regression model as it is failing to satisfy normality assumption. Instead we can use log(medexpense) with which the data seemed to have fairly normal distribution curve.

### 2. Variable Analysis

<u>Predictor</u>	<u>Effect</u>	<u>Rationale</u>
Healthins	+	It is observed that people with health insurance gets billed more because there is a sentiment among public that nobody is paying and is a win-win situation for all. So, hospitals charge more for insured patients by adding unnecessary treatment and care.
Age	+	People with more age tend to get sick more often than average
Female	+	Females tend to have more health expenses than male as they have hospital visits more due to menstrual and pregnancy issues.
Income	-	People with more income do take care of their health properly. So, there is less probability that they will get sick.
Illnesses	+	People with more number of illnesses have higher probability of more health expenses.
Prioritylist	+	Person in priority list/elder people have more occurrences of hospital visits.
msa	+	Person living in metro areas are more accessible to health facilities than those in rural area so must have more health expense as hospital visit occurrence probability would be more.
HealthLevel		It is a feature engineered variable from 4 variables which are poor, fair, good and very good health. Person with poor health must have more health expense and person with good health must have less health expense
Age*Illness	+	People with age and higher number of illness will have higher probability of medical expenses

### 3. Model Analysis

```
> m1 = lm(log(medexpense+1)~healthins+age+female+income+illnesses+prioritylist+msa+Health
Level)
-Baseline model
> m2 = lm(log(medexpense+1)~healthins+age+female+income+illnesses+prioritylist+msa+Health
Level+age*illnesses)
-To understand the impact of adding interaction term in model
> m3 = lm(log(medexpense+1)~healthins+age+female+income+illnesses+prioritylist+msa+Health
Level+age*illnesses+*blackhisp+**private+**IncomeLevel)
-To understand the impact of adding blackhisp, private and IncomeLevel variable in the model
> stargazer(m1,m2,m3,type = "text", single.row = TRUE)
```

\*Blackhisp variable is added in model 3 to understand the impact of being racially diverse on medical expense. I personally don't feel that there is any racial bias in medical expenses so I haven't added this variable in the predictor table.

\*\*IncomeLevel and private variables are added to understand their impact on medical expenses.

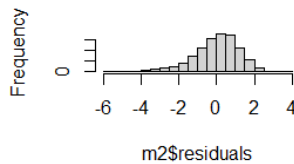
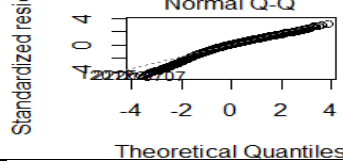
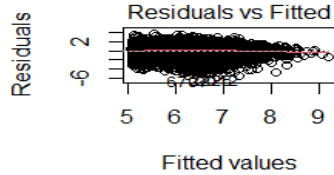
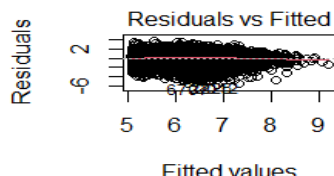
Dependent variable:			
	log(medexpense + 1)		
	(1)	(2)	(3)
healthins	0.097*** (0.026)	0.097*** (0.026)	0.086*** (0.032)
age	-0.004** (0.002)	0.004 (0.003)	0.003 (0.003)
female	0.077*** (0.025)	0.078*** (0.025)	0.078*** (0.025)
income	0.001 (0.001)	0.001* (0.001)	0.001 (0.001)
illnesses	0.354*** (0.010)	0.690*** (0.107)	0.687*** (0.107)
prioritylist	0.571*** (0.038)	0.565*** (0.038)	0.564*** (0.038)
msa	-0.048* (0.027)	-0.047* (0.027)	-0.030 (0.028)
HealthLevelGood	-0.151*** (0.035)	-0.150*** (0.035)	-0.161*** (0.035)
HealthLevelPoor	0.105* (0.055)	0.107* (0.055)	0.100* (0.055)
HealthLevelVeryGood	-0.284*** (0.035)	-0.281*** (0.035)	-0.298*** (0.035)
blackhisp			-0.173*** (0.034)
private			-0.002 (0.032)
IncomeLevelMid-Income			0.032 (0.028)
IncomeLevelPoor			-0.022 (0.037)
age:illnesses		-0.004*** (0.001)	-0.004*** (0.001)
Constant	5.749*** (0.148)	5.132*** (0.244)	5.229*** (0.246)
Observations	10,089	10,089	10,089
R2	0.203	0.204	0.206
Adjusted R2	0.202	0.203	0.205
Residual Std. Error	1.206 (df = 10078)	1.205 (df = 10077)	1.204 (df = 10073)
F Statistic	256.483*** (df = 10; 10078)	234.287*** (df = 11; 10077)	174.324*** (df = 15; 10073)

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

I am considering model 2 for further investigation.

### 4. Assumption Analysis

Assumption	Analysis
------------	----------

Normality	<div>Histogram of m2\$residuals</div>  <div>Frequency</div> <div>m2\$residuals</div>	<div>Interpretation: Data seems to follow normal distribution.</div> <div>Result: Pass</div>																																				
Linearity	<div>Standardized residuals</div> <div>Normal Q-Q</div>  <div>Theoretical Quantiles</div>	<div>Interpretation: Data is fairly linear</div> <div>Result: Pass</div>																																				
Homoskedasticity	<div>Residuals</div> <div>Residuals vs Fitted</div>  <div>Fitted values</div>	<div>Interpretation: Heteroskedastic pattern is clearly visible</div> <div>Result: Fail</div>																																				
Multicollinearity	<table><thead><tr><th></th><th>GVIF</th><th>Df</th><th>GVIF^(1/(2*Df))</th></tr></thead><tbody><tr><td>healthins</td><td>1.066095</td><td>1</td><td>1.032519</td></tr><tr><td>age</td><td>1.045588</td><td>1</td><td>1.022540</td></tr><tr><td>female</td><td>1.029049</td><td>1</td><td>1.014421</td></tr><tr><td>income</td><td>1.076788</td><td>1</td><td>1.037684</td></tr><tr><td>illnesses</td><td>1.250360</td><td>1</td><td>1.118195</td></tr><tr><td>prioritylist</td><td>1.161851</td><td>1</td><td>1.077892</td></tr><tr><td>msa</td><td>1.020446</td><td>1</td><td>1.010171</td></tr><tr><td>HealthLevel</td><td>1.130904</td><td>3</td><td>1.020715</td></tr></tbody></table>		GVIF	Df	GVIF^(1/(2*Df))	healthins	1.066095	1	1.032519	age	1.045588	1	1.022540	female	1.029049	1	1.014421	income	1.076788	1	1.037684	illnesses	1.250360	1	1.118195	prioritylist	1.161851	1	1.077892	msa	1.020446	1	1.010171	HealthLevel	1.130904	3	1.020715	<div>Interpretation: All variables have low VIF.</div> <div>Result: Pass</div>
	GVIF	Df	GVIF^(1/(2*Df))																																			
healthins	1.066095	1	1.032519																																			
age	1.045588	1	1.022540																																			
female	1.029049	1	1.014421																																			
income	1.076788	1	1.037684																																			
illnesses	1.250360	1	1.118195																																			
prioritylist	1.161851	1	1.077892																																			
msa	1.020446	1	1.010171																																			
HealthLevel	1.130904	3	1.020715																																			
Independence	<div>Residuals</div> <div>Residuals vs Fitted</div>  <div>Fitted values</div>	<div>Interpretation: There is no relation between data points.</div> <div>Result: Pass</div>																																				

## 5. Interpretations

- 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?

**Ans:** Yes. People with health insurance having expenses more by 9.7%. It is observed that people with health insurance gets billed more because there is a sentiment among public that nobody is paying and is a win-win situation for all. So, hospitals charge more for insured patients by adding unnecessary treatment and care.

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

**Ans:** People with private health insurance pay 0.1% less than those with public insurance which is not significant.

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

**Ans:** People with more illness have higher medical expenses than those with less illness. According to data, cost increases by 69% with increase in number of disease the patient having.

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

**Ans:** Females have 7.8% more medical expenses than males.

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

**Ans:** Older/people with fragile health pay 56.5% more than younger people (by variable prioritylist). But this point is debatable as data is belonged to the patients of age-group 65-91.

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

**Ans:** Minority groups pay 17.7% less medical expenses according to data.

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

**Ans:** There is no significant observation that people with higher income level are paying more or less than people with less income as all people can get all kind of diseases. Poor people pay 2.2% less than rich people but interestingly mid-income people pay 3.2% more than rich people. The reason for having high medical expenses for middle class is that they are the majorly insured as they are qualified for premium tax credits. (100% - 400% FPL) \*\*\*

\*\*\*src. - <https://www.healthcare.gov/glossary/federal-poverty-level-fpl/>