MTH 522: Advanced Mathematical Statistics Predicting MPG with Multiple Variables Using Multiple Regression By: Dhyey Doshi 02/19/2023

_, ..., _.._.

<u>Issues</u>

Using the predictor variables "displacement," "weight," "horsepower," and "acceleration," we design a multiple regression model in this report to predict a vehicle's miles per gallon (MPG). Multiple regression refers to the process of attempting to predict Something by combining two or more variables. We discuss examining the variable that is most useful for predicting MPG as a baseline variable to base the remainder of the model on. The key variables for predicting MPG are covered in this report. When creating a multiple regression model, it's possible that only some predictor variables will be helpful and others won't.

Findings

After receiving the results, we can draw the reasonable conclusion that the model works well, with an r-squared of 0.61, or around 61%, indicating that the model accurately describes the data. Additionally, we draw the conclusion that the residuals exhibit almost normal distribution, indicating that our fitted values are relatively correct and the predictions exhibit minimal error. The residuals do not follow a perfect normal distribution, and the variables in the model can only precisely represent roughly 61% of the miles per gallon data. This should be taken into consideration when planning to utilize the model.

Discussion

We find solutions to our problem, which enabled us to reach our conclusion, by using statsmodel in the python programming language. One of our first problems is identifying the most useful variable from which to base the prediction, and these methods help us realize that factors like weight are crucial in predicting miles per gallon for a car. Coupled with these straightforward summary statistics, we were able to properly add variables to the multiple regression model, enabling us to create a model that performs admirably and accounts for around 61% of the data related to miles per gallon.

Appendix A: Methods

The automobile data utilized in this report has five variables. Four factors displacement, weight, horsepower, and acceleration—are predictors, and one is the target value—miles per gallon, or MPG—that we want to predict. The first step is to choose a useful predictor variable, one that best aids in predicting miles per gallon. To achieve this, we compute a basic linear model for each predictor variable, i.e., displacement predicts MPG, followed by weight predicts MPG, giving us a total of four simple models. We also take model summary to look at r^2 value for each model and identify best variable for best model. After building the multiple regression model with forward selection process we must determine its accuracy, to do this check the if the residuals/errors follow a normal distribution. We have used quantile plot for the residual points on the line.

Appendix B: Results

Each Model Summary Charts

Displacement

Dep. Variable: mpg		R-sq		qua	ared:		0.623	
Model: OLS		OLS		Ad	j. R	ared:	0.622	
Method:		Leas	t Squares	s F-s	tati	stic:		656.5
Date: Sun,		19 Feb 2	023 Pr	ob (tistic):	3.69e-86		
Time:	21:2):04	Lo	g-Li	ood:	-1183.1	
No. Observatio	ons:	399		Al	:			2370.
Df Residuals:		397		BI	:			2378.
Df Model:		1						
Covariance Ty	pe:	nonro	obust					
	coe	f	std err	t	P	> t	[0.025	0.975]
const	34.9	9557	0.491	71.183	0.	000	33.990	35.921
displacement	-0.0	587	0.002	-25.623	0.	000	-0.063	-0.054
Omnibus:	2	6.466	Durbin	-Watson		2.02	24	
Prob(Omnibus): (0.000.	Jarque	e-Bera (J	B):	31.7	706	
Skew:	C	0.577 Prob(.		JB):		1.30	De-07	
Kurtosis:	3	3,760	Cond.	No.		447	6	

#Horsepower

Dep. Variable	:	mpg	3		R-s	quared	d:	0.557	
Model:		OLS	6		Adj	. R-sq	uared:	0.556	
Method:		Lea	st Square	es	F-s	tatistic	:	498.8	
Date:		Sun	, 19 Feb	2023	Pro	b (F-st	atistic):	3.74e-72	2
Time:		21:2	22:20		Log	j-Likeli	ihood:	-1215.4	
No. Observat	ions:	399			AIC	:		2435.	
Df Residuals:		397			BIC			2443.	
Df Model:		1							
Covariance T	ype:	non	robust						
	coef		std err	t		P> t	[0.025	0.975]	
const	38.98	50	0.721	54.03	4	0.000	37.567	40.403	
horsepower	-0.147	72	0.007	-22.33	84	0.000	-0.160	-0.134	
Omnibus:	1	6.637	7 Durb	in-Wat	son:	1.9	950		
Prob(Omnibu	is): 0	.000	Jarqu	ue-Bera	a (JE	3): 17	.736		
Skew:	0	.514	Prob	(JB):		0.	000141		
Kurtosis:	3	.104	Cond	I. No.		30	9.		

#Weight

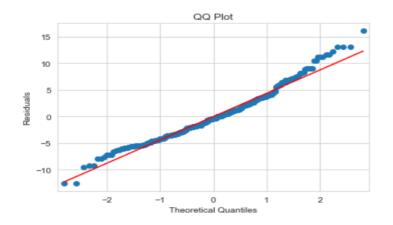
#Acceleration

odel:OLSAdj R-squared:0.667Model:OLSAdj. R-squared:ethod:Leat: SquaresF-statistic:764.1Model:Leat: SquaresF-statistic:764.1ate:Sun, 19 Feb 2023Prob (F-statistic:1.50e-94Model:Leat: SquaresProb (F-statistic:Date:Sun, 19 Feb 2023Prob (F-statistic:Date:Sun, 19 F	Dep. Variable:	mpg	R-squ	uared:	0.658	Dep. Variable		mpg			R-squa	rod		
ethod: <td cols<="" td=""><td>Model:</td><td>OLS</td><td>Adj. F</td><td>R-squared</td><td>. 0.657</td><td>•</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td>	<td>Model:</td> <td>OLS</td> <td>Adj. F</td> <td>R-squared</td> <td>. 0.657</td> <td>•</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>	Model:	OLS	Adj. F	R-squared	. 0.657	•							
ate: Sun, 19 Feb 2023 Prob (F-statistic) 1.50e-34 Date: Sun, 19 Feb 2023 Prob (F-statistic) me: 21:23:2 Log-Likelihood: -1163.7 o. Observations: 39' AIC: 2331. f Residuals: 307 BIC: 2331. f Model: 1 SIC: 2339. Df Residuals: 39' AIC: BIC:	Method:	Least Squares	s F-sta	tistic:	764.1				t Saua	res	-			
me: 21:23:2 Log Likelihoot -1163.7 o. Observations 39 I IC 2331. f Residuals: 39.7 IC 2331. f Rodel: 1 2331. f Model: 1 I I IC I	Date:	Sun, 19 Feb 2	2023 Prob	(F-statisti	c): 1.50e-94) :	
o. Observations: 399 AIC: 2331. f Residuals: 397 BIC: 2339. f Model: 1 1 Df Residuals: 397 BIC: Df Model: 397 BIC: Df Model: 397 BIC: 1 Df Model: 1 1 Df Model: 1 1 Df Model: 1 1 Df Model: 1 1 0.000 <td>Time:</td> <td>21:23:02</td> <td>Log-L</td> <td>Likelihood</td> <td>-1163.7</td> <td></td> <td></td> <td></td> <td></td> <td>. 2020</td> <td></td> <td></td> <td>·</td>	Time:	21:23:02	Log-L	Likelihood	-1163.7					. 2020			·	
Model: 1 Dr Residuals: 397 BIC: f Model: 1 Dr Model: 1 U </td <td>No. Observations:</td> <td>399</td> <td>AIC:</td> <td></td> <td>2331.</td> <td></td> <td>ons:</td> <td></td> <td></td> <td></td> <td>-</td> <td></td> <td></td>	No. Observations:	399	AIC:		2331.		ons:				-			
Dr Model: 1 covariance Type: nonrobust Covariance Type: nonrobust Covariance Type: nonrobust coef std err t P> t [0.025 0.975] onno 0.000 2.7.64 0.000 44.429 47.701 coef std err t P> t [0.025 0.371 13 roh(Omnibus): 0.000 2.7.64 0.000 2.1.38 Omnibus: 19.834 Durbin-Watson: 2.0.14 roh(Omnibus): 0.000 Jarque-Bera (JB): 2.6.981 Skew: 0.568 Prob(JB): 1.73e-05	Df Residuals:	397	BIC:		2339.	Df Residuals:		397			BIC:			
coef std err t P>[t] [0.025 0.975] onst 46.0649 0.832 55.358 0.000 44.429 47.701 eight -0.0076 0.000 -27.642 0.000 -0.008 -0.007 mnibus: 23.726 Durbin-Watson: 2.138 Omnibus: 19.834 Durbin-Watson: 2.014 rob(Omnibus): 0.000 Jarque-Bera (JB): 26.981 0.000 Jarque-Bera (JB): 21.934 kew: 0.568 Prob(JB): 1.38e-06 Skew: 0.568 Prob(JB): 1.73e-05	Df Model:	1				Df Model:		1						
const 46.0649 0.832 55.358 0.000 44.429 47.701 const 9.3004 1.999 4.653 0.000 5.371 13 acceleration 0.9286 0.125 7.433 0.000 6.663 1.000 1.000	Covariance Type:	nonrobust				Covariance T	/pe:	nonro	obust					
Const 9.3004 1.999 4.653 0.000 5.371 13 eight -0.0076 0.000 -27.642 0.000 -0.008 -0.007 mnibus: 23.726 Durbin-Watson: 2.138 Omnibus: 19.834 Durbin-Watson: 2.014 rob(Omnibus): 0.000 Jarque-Bera (JB): 2.6981 Prob(Omnibus): 0.000 Jarque-Bera (JB): 2.1934 kew: 0.568 Prob(JB): 1.73e-05 Skew: 0.568 Prob(JB): 1.73e-05	coef	std err t	P> t	[0.025 0	0.975]		coef	st	d err	t	P> t	[0.025	0.	
Commibus: 23.726 Durbin-Watson: 2.138 Omnibus: 19.834 Durbin-Watson: 2.014 rob(Omnibus): 0.000 Jarque-Bera (JB): 26.981 Prob(Omnibus): 0.000 Jarque-Bera (JB): 21.934 kew: 0.562 Prob(JB): 1.38e-06 Skew: 0.568 Prob(JB): 1.73e-05	const 46.0649	0.832 55.35	0.000 86	44.429 4	\$7.701	const	9.30	04 1.	999	4.653	0.000	5.371	13	
Prob(Omnibus): 0.000 Jarque-Bera (JB): 26.981 Prob(Omnibus): 0.000 Jarque-Bera (JB): 21.934 kew: 0.562 Prob(JB): 1.38e-06 Skew: 0.568 Prob(JB): 1.73e-05	weight -0.0076	0.000 -27.64	42 0.000	-0.008 -	0.007	acceleration	0.92	86 0.	125	7.433	0.000	0.683	1.	
Kew: 0.562 Prob(JB): 1.38e-06 Skew: 0.568 Prob(JB): 1.73e-05	Omnibus: 2	23.726 Durbi r	n-Watson:	2.138		Omnibus:		19.834	Dur	bin-Wat	son:	2.014		
Skew. 0.008 FID(05). 1.736-00	Prob(Omnibus):	0.000 Jarqu	e-Bera (JB):	26.981		Prob(Omnibu	s): (0.000	Jaro	que-Ber	a (JB):	21.934		
urtosis: 3.599 Cond. No. 1.13e+04 Kurtosis: 2.834 Cond. No. 89.3	Skew:	0.562 Prob(JB):	1.38e-0	6	Skew:	C	0.568	Pro	b(JB):		1.73e-05		
	Kurtosis:	3.599 Cond.	No.	1.13e+0	94	Kurtosis:	2	2.834	Cor	d. No.		89.3		

#Multiple Regression Summary

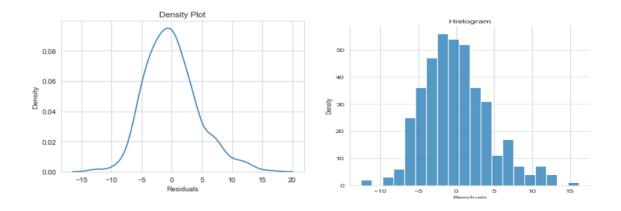
Dep. Variable:	m	pg		R-square	d:	0.671		
Model:	0	LS		Adj. R-sq	uared:	0.670		
Method:	Le	east Squar	es	F-statisti	c:	404.2		
Date:	S	un, 19 Feb	Feb 2023 Pro		tatistic):	2.27e-96		
Time:	21	1:42:17		Log-Like	lihood:	-1155.9		
No. Observatio	ons: 39	99		AIC:		2318.		
Df Residuals:	39	96		BIC:		2330.		
Df Model:	2							
Covariance Ty	pe: no	onrobust						
	coef	std er	r t	P> t	[0.025	0.975]		
const	coef 42.899		r t 37.59		•	0.975] 45.143		
const displacement		2 1.141		92 0.000	40.656			
	42.899	2 1.141 3 0.005	37.5	92 0.000 74 0.000	40.656	45.143		
displacement	42.899	2 1.141 3 0.005 1 0.001	37.59 -3.97	92 0.000 74 0.000 94 0.000	40.656	45.143 -0.011		
displacement weight	42.899 -0.0213 -0.005 ⁷ 26.6	2 1.141 3 0.005 1 0.001 39 Durk	37.59 -3.97 -7.60	2 0.000 4 0.000 04 0.000	40.656 -0.032 -0.006	45.143 -0.011		
displacement weight Omnibus:	42.899 -0.0213 -0.005 ⁷ 26.6	2 1.141 3 0.005 1 0.001 339 Durb 0 Jarq	37.59 -3.97 -7.60	2 0.000 4 0.000 04 0.000 con: 2 (JB): 3	40.656 -0.032 -0.006	45.143 -0.011		
displacement weight Omnibus: Prob(Omnibus	42.899 -0.0213 -0.005 26.6): 0.00	2 1.141 3 0.005 1 0.001 39 Durb 0 Jarq 0 Prob	37.59 -3.97 -7.60 Din-Wats ue-Bera	2 0.000 4 0.000 4 0.000 5001: 2 (JB): 3 1	40.656 -0.032 -0.006 107 1.398	45.143 -0.011		

Below a quantile plot can be found of the multiple regression models residuals, this is testing whether the residuals follow a normal distribution.

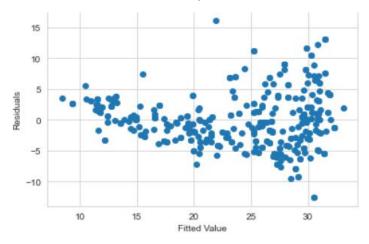


#Density Plot

#Histogram



#Test of heteroscedasticity



Appendix C: Code

import numpy as np import pandas as pd import scipy.stats as sts import statsmodels.api as sm from statsmodels.formula.api import ols import matplotlib.pyplot as plt import seaborn as sns

df = pd.read_excel('E:\\Study\\MTH\\auto_data_doshi_dhyey.xlsx')

df.describe()

```
displacement = df['displacement']
horsepower = df['horsepower']
weight = df['weight']
acceleration = df['acceleration']
mpg = df['mpg']
```

#Displacement

```
displacement = sm.add_constant(displacement)
displacementModel = sm.OLS(mpg, displacement).fit()
displacementModel.summary()
```

#Horsepower

```
horsepower = sm.add_constant(horsepower)
horsepowerModel = sm.OLS(mpg, horsepower).fit()
horsepowerModel.summary()
```

#Weight

```
weight = sm.add_constant(weight)
weightModel = sm.OLS(mpg, weight).fit()
weightModel.summary()
```

#Acceleration

```
acceleration = sm.add_constant(acceleration)
accelerationModel = sm.OLS(mpg, acceleration).fit()
accelerationModel.summary()
```

```
print('Displacement:', np.sum(np.square(displacementModel.resid)))
print('Horsepower:', np.sum(np.square(horsepowerModel.resid)))
print('Weight:', np.sum(np.square(weightModel.resid)))
print('Acceleration:', np.sum(np.square(accelerationModel.resid)))
```

```
X = df[['displacement', 'weight']]
y = df[['mpg']]
X = sm.add_constant(X)
multiRegModel = sm.OLS(y, X).fit()
multiRegResiduals = multiRegModel.resid
multiRegModel.summary()
```

```
sm.qqplot(multiRegResiduals, line='s')
plt.title('QQ Plot')
plt.ylabel('Residuals')
plt.show()
```

```
sns.kdeplot(multiRegResiduals)
plt.ylabel('Density')
plt.xlabel('Residuals')
plt.title('Density Plot')
plt.show()
```

```
sns.displot(multiRegResiduals)
plt.ylabel('Density')
plt.xlabel('Residuals')
plt.title('Histogram')
plt.show()
```

```
plt.scatter(multiRegModel.fittedvalues, multiRegResiduals)
plt.xlabel("Fitted Value")
plt.ylabel("Residuals")
plt.show()
```