Multiple Linear Regression
MATH-2050 Applied Statistics
Hunter Stevens

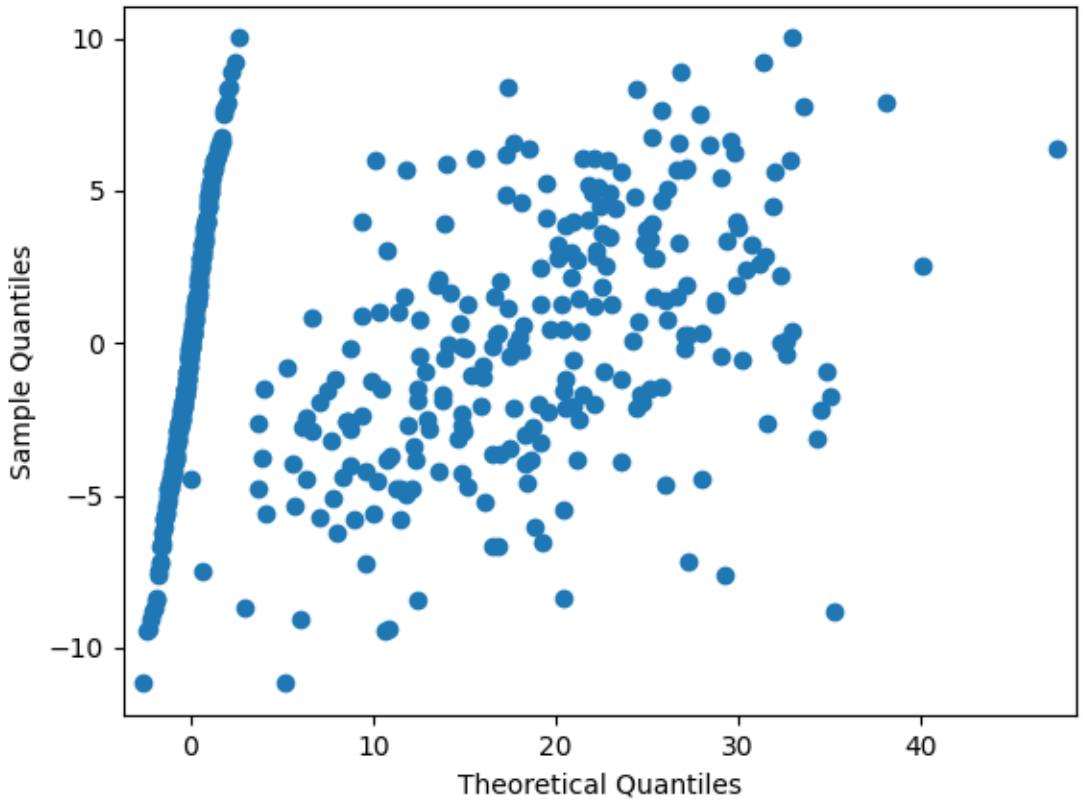
Here is the model summary for percent body fat. The R squared score on this is .749 this is for the model made with smf.OLS().fit()

OLS Regression Results

Dep. Variable:	PercentBodyFat	R-squared:	0.749
Model:	OLS	Adj. R-squared:	0.735
Method:	Least Squares	F-statistic:	54.65
Date:	Fri, 01 Nov 2024	Prob (F-statistic):	7.72e-64
Time:	21:19:14	Log-Likelihood:	-718.25
No. Observations:	252	AIC:	1465.
Df Residuals:	238	BIC:	1514.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-18.1885	17.349	-1.048	0.296	-52.365	15.988
Age	0.0621	0.032	1.919	0.056	-0.002	0.126
Weight	-0.0884	0.054	-1.652	0.100	-0.194	0.017
Height	-0.0696	0.096	-0.725	0.469	-0.259	0.120
NeckCir	-0.4706	0.232	-2.024	0.044	-0.929	-0.013
ChestCir	-0.0239	0.099	-0.241	0.810	-0.219	0.171
AbCir	0.9548	0.086	11.044	0.000	0.784	1.125
HipCir	-0.2075	0.146	-1.422	0.156	-0.495	0.080
ThighCir	0.2361	0.144	1.636	0.103	-0.048	0.520
KneeCir	0.0153	0.242	0.063	0.950	-0.461	0.492
AnkleCir	0.1740	0.221	0.786	0.433	-0.262	0.610
BicepsCir	0.1816	0.171	1.061	0.290	-0.156	0.519
ForearmCir	0.4520	0.199	2.270	0.024	0.060	0.844
WristCir	-1.6206	0.535	-3.030	0.003	-2.674	-0.567

Omnibus:	4.404	Durbin-Watson:	1.776
Prob(Omnibus):	0.111	Jarque-Bera (JB):	2.820
Skew:	-0.012	Prob(JB):	0.244
Kurtosis:	2.482	Cond. No.	1.78e+04



Here is our scatter plot of the predictions

After splitting the data into training and testing data we use the scikit learn LinearRegression() function to make a model and make predictions.

The R-squared value on the split data is:
0.679102232216163

Here is our mean absolute error value: 3.9656133034131487
and our RMSE: 4.7270146119634715

So as you can see we don't have the best model for predicting this variable, however it doesn't seem to be absolutely terrible. Our R-squared values are 70-75% accurate which leaves much to be desired. Our mean absolute error value is more accurate meaning that when we do have errors they are when we have large outliers as the mean squared error is more affected by them.

Part II

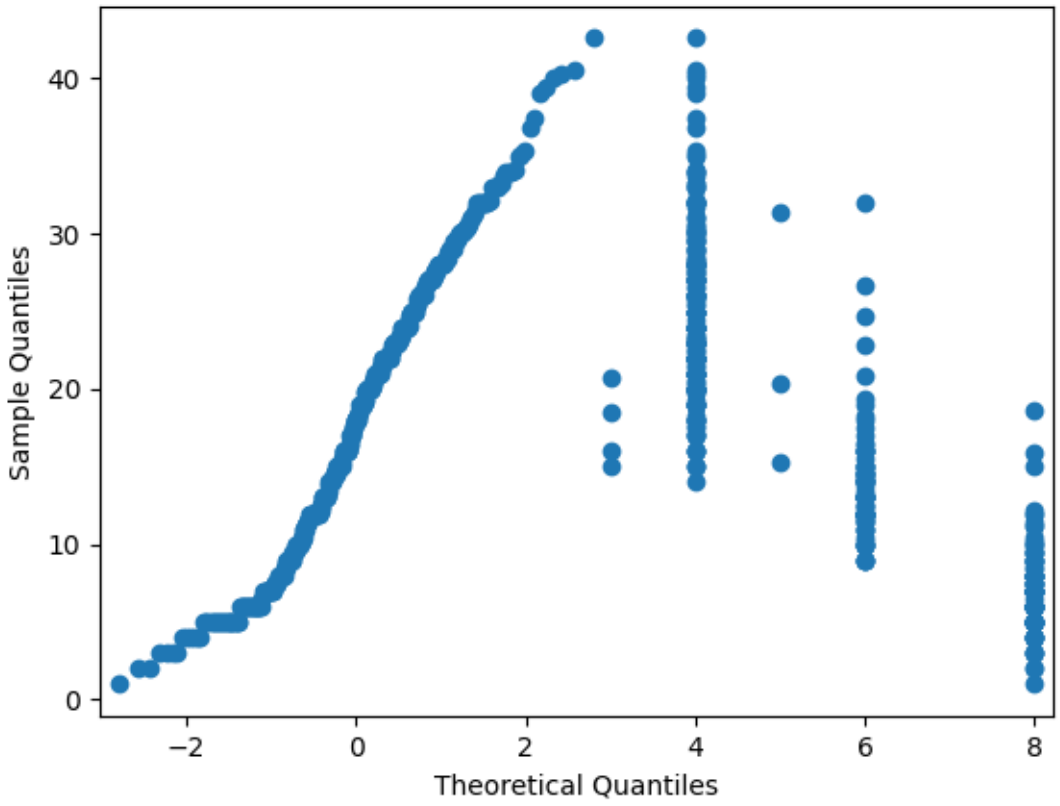
OLS Regression Results MPG

Dep. Variable:	cylinders	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	1.440e+26
Date:	Fri, 01 Nov 2024	Prob (F-statistic):	0.00
Time:	21:34:43	Log-Likelihood:	10282.
No. Observations:	392	AIC:	-2.055e+04
Df Residuals:	383	BIC:	-2.051e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.472e-13	1.05e-12	0.714	0.476	-1.31e-12	2.8e-12
cylinders	1.0000	9.65e-14	1.04e+13	0.000	1.000	1.000
displacement	-2.268e-16	2.3e-15	-0.099	0.921	-4.75e-15	4.3e-15
horsepower	-7.546e-16	4.12e-15	-0.183	0.855	-8.85e-15	7.35e-15
weight	1.884e-16	1.97e-16	0.957	0.339	-1.99e-16	5.76e-16
acceleration	9.021e-17	2.95e-14	0.003	0.998	-5.79e-14	5.81e-14
model year	3.365e-16	1.56e-14	0.022	0.983	-3.03e-14	3.09e-14
USA	2.507e-13	3.69e-13	0.680	0.497	-4.74e-13	9.76e-13
Europe	2.485e-13	3.47e-13	0.715	0.475	-4.34e-13	9.32e-13
Asia	2.436e-13	3.7e-13	0.659	0.510	-4.84e-13	9.71e-13

Omnibus:	17.977	Durbin-Watson:	0.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.409
Skew:	0.397	Prob(JB):	8.26e-06
Kurtosis:	3.895	Cond. No.	5.22e+18

R-squared seems to be 100% accurate here, meaning we have overfit this problem or I did something wrong when manipulating the data, but I don’t know enough about what I’m currently doing to know what it is that I may have messed up.



Here is our other scatter plot, a different plot seems like it would work better for this problem.

R^2 on test set mpg: 1.0

The R-squared value is still 1 after using scikit learn's LinearRegression() function. Again I think this means I set something up wrong, or we have figured out exactly how to predict miles per gallon.

MAE: 3.717929918736245e-14
MSE: 2.5290730202992693e-27
RMSE: 5.0289889841788967e-14

These values however, seem to confirm that while we do have errors they are extremely small. Our predictions may as well be perfectly precise when looking at numbers this small. Seeing the closeness of MAE and RMSE might confirm that there are not very many outliers in this dataset.