Multi-Variate Regression on the UCI Auto MPG Data Set

Run Using Minitab

Kevin Maher
Vettejeep365@gmail.com

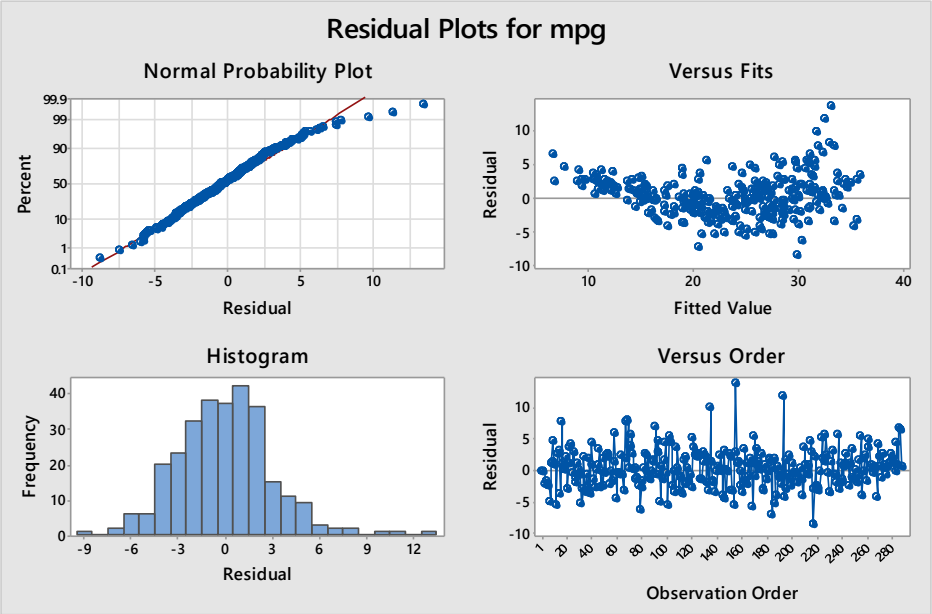
October 8, 2017

Data

- Goal: Build a predictive model of automotive fuel economy using multi-variate regression
  - Data will be split into a set consisting of 75% of the data for building the model and 25% for validating it
- Data: UCI Auto MPG Data Set
  - Missing horsepower values looked up
  - Diesel engine and Mazda rotary engine vehicles removed because there are too few of them to represent well in both training and test data sets
  - Data set from: https://archive.ics.uci.edu/ml/datasets/auto+mpg
  - Target data field: mpg
  - Continuous variables: cylinders, displacement, horsepower, weight, acceleration, year (car model year)
  - Categorical variable: origin (USA, Europe, Japan)

Initial Model

- Use all variables in simple linear regression
- Problems with residuals – low spread at low MPG, high spread at high MPG – Indicates a data transformation may be helpful
- High multi-collinearity as indicated by high VIF values – will drop displacement from the model since it has highest VIF value
- R-squared adjusted is 83.76% - will try to improve this



Model Summary

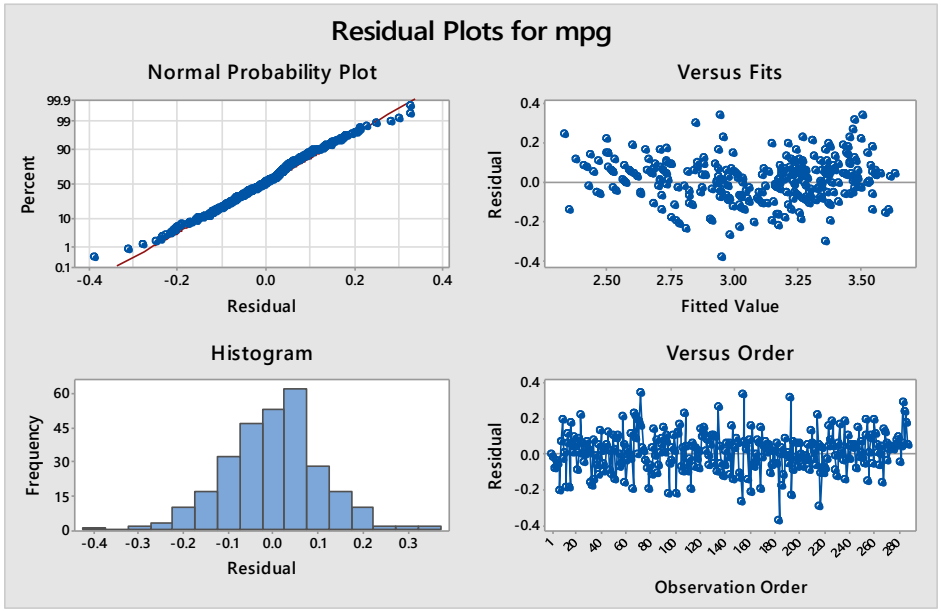
S	R-sq	R-sq(adj)	R-sq(pred)
3.09316	84.21%	83.76%	83.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-9.24	5.21	-1.77	0.077	
cylinders	-0.592	0.343	-1.73	0.085	10.54
displacement	0.01073	0.00870	1.23	0.218	24.62
horsepower	-0.0037	0.0164	-0.22	0.823	11.13
weight	-0.006055	0.000792	-7.65	0.000	14.04
acceleration	-0.080	0.110	-0.73	0.468	2.60
year	0.6912	0.0569	12.14	0.000	1.28
origin					
2	1.519	0.648	2.34	0.020	1.64
3	2.948	0.592	4.98	0.000	1.74

Second Model

- Displacement dropped due to multi-collinearity, mpg target transformed by natural log to improve dispersion of residuals
- Residuals are better distributed the upper right hand “versus fits” chart due to transformation of MPG – not perfect, but much better and the normal probability plot shows a more normal distribution
- R-Squared adjusted now up to 89.29% from 83.76%
- Variance Inflation Factors (VIF) still too high so will work to reduce this



Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.110690	89.55%	89.29%	88.85%

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.011	0.186	10.79	0.000	
cylinders	-0.02097	0.00941	-2.23	0.027	6.20
horsepower	-0.000637	0.000574	-1.11	0.268	10.67
weight	-0.000244	0.000025	-9.68	0.000	11.08
acceleration	-0.00664	0.00389	-1.71	0.089	2.52
year	0.02726	0.00202	13.49	0.000	1.26
origin					
2	0.0322	0.0218	1.48	0.141	1.44
3	0.0784	0.0202	3.89	0.000	1.58

Second Model

- Need to work to improve the model
- Multi-collinearity still too high as evidenced by the high VIF values, this can cause problems with the stability and sign of coefficients
- Weight has the highest VIF value but I am reluctant to remove it due to the physics of the mpg problem – also removing it caused a big loss of fit as expressed by R-squared adjusted – note the zero p-value for weight
- Horsepower seems a better candidate for removal – has high VIF factor and its p-value is higher and above the usual threshold of significance of 0.05

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.011	0.186	10.79	0.000	
cylinders	-0.02097	0.00941	-2.23	0.027	6.20
horsepower	-0.000637	0.000574	-1.11	0.268	10.67
weight	-0.000244	0.000025	-9.68	0.000	11.08
acceleration	-0.00664	0.00389	-1.71	0.089	2.52
year	0.02726	0.00202	13.49	0.000	1.26
origin					
2	0.0322	0.0218	1.48	0.141	1.44
3	0.0784	0.0202	3.89	0.000	1.58

Third Model

- Horsepower removed from the model
- VIF values much improved – all are below 10
  - I have seen recommendations to remove variables with VIF values anywhere from 2.5 to 10, since the model VIFs are now below 10, I will continue to improve the model from here by other means, especially since we have few predictors
- R-squared adjusted suffered only a slight loss to 89.28% from 89.29% so removing horsepower simplifies the model while not making any significant difference in the model's quality of fit

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.924	0.169	11.37	0.000	
cylinders	-0.02150	0.00940	-2.29	0.023	6.19
weight	-0.000264	0.000018	-14.86	0.000	5.50
acceleration	-0.00376	0.00290	-1.30	0.195	1.40
year	0.02777	0.00197	14.10	0.000	1.20
origin					
2	0.0314	0.0218	1.44	0.151	1.44
3	0.0747	0.0199	3.75	0.000	1.53

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.110736	89.51%	89.28%	88.89%

Fourth Model

- Add squared and interaction terms – this improves R-squared adjusted to 91.05% from 89.28%
- Some terms are not significant as evidenced by high p-values
- I removed high p-value terms one at a time, to simplify the presentation I will jump straight to the simplified model

Analysis of Variance for Transformed Response

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	16	30.0637	1.87898	183.58	0.000
cylinders	1	0.0348	0.03484	3.40	0.066
weight	1	0.0077	0.00769	0.75	0.387
acceleration	1	0.0801	0.08013	7.83	0.006
year	1	0.0751	0.07509	7.34	0.007
origin	2	0.0491	0.02456	2.40	0.093
cylinders*cylinders	1	0.0143	0.01429	1.40	0.238
weight*weight	1	0.0108	0.01078	1.05	0.306
acceleration*acceleration	1	0.0091	0.00914	0.89	0.345
year*year	1	0.0689	0.06892	6.73	0.010
cylinders*weight	1	0.0002	0.00017	0.02	0.897
cylinders*acceleration	1	0.0940	0.09397	9.18	0.003
cylinders*year	1	0.0029	0.00292	0.29	0.594
weight*acceleration	1	0.0545	0.05448	5.32	0.022
weight*year	1	0.0022	0.00221	0.22	0.643
acceleration*year	1	0.0632	0.06317	6.17	0.014
Error	271	2.7738	0.01024		
Total	287	32.8376			

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.101171	91.55%	91.05%	90.29%

Final Model

- R-squared adjusted improved to 91.12% - slight improvement over the previous model but also we now have a simpler model
- High p-value terms removed one at a time to get to here
  - cylinders*weight, weight*year and acceleration*acceleration removed
- First order variables kept where they participate in a higher order term
- Residual plots have not changed visibly during this process
- In spite of some p-values above 0.05, removing more terms hurt R-squared adjusted and caused problems with the residuals

Analysis of Variance for Transformed Response

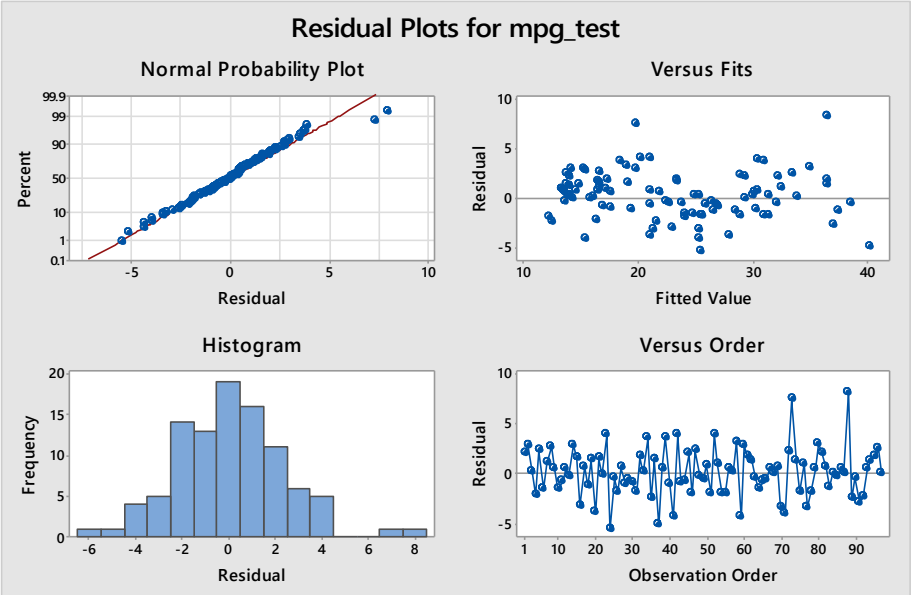
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	13	30.0528	2.31175	227.46	0.000
cylinders	1	0.1550	0.15501	15.25	0.000
weight	1	0.0147	0.01473	1.45	0.230
acceleration	1	0.0711	0.07109	6.99	0.009
year	1	0.0893	0.08926	8.78	0.003
origin	2	0.0437	0.02183	2.15	0.119
cylinders*cylinders	1	0.0485	0.04847	4.77	0.030
weight*weight	1	0.0491	0.04906	4.83	0.029
year*year	1	0.0827	0.08268	8.14	0.005
cylinders*acceleration	1	0.0893	0.08931	8.79	0.003
cylinders*year	1	0.0345	0.03447	3.39	0.067
weight*acceleration	1	0.0612	0.06123	6.02	0.015
acceleration*year	1	0.0736	0.07362	7.24	0.008
Error	274	2.7848	0.01016		
Total	287	32.8376			

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.100813	91.52%	91.12%	90.51%

Proving the Model

- Validate the model using data that was not part of the training set – validation data is 25% of overall data
- Two values with high residuals – two outliers – not bad considering the simplicity of this model for a complex response like automotive mpg and the small number of predictors
- Coefficient (slope) of fit is .09767 – very close to the value of 1 that would indicate a perfect average fit
- R-squared adjusted for the validation set is good at 90.4%



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.35684	90.50%	90.40%	90.05%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5029.3	5029.28	905.41	0.000
mpg_pred	1	5029.3	5029.28	905.41	0.000
Error	95	527.7	5.55		
Total	96	5557.0			