MATH 4044 – Statistics for Data Science

Practical Week 6 Solutions

The data was collected from Kelly Blue Book http://www.kbb.com for several hundred 2005 used General Motors (GM) cars. The goal is to develop a multivariate regression model to determine car values based on a variety of characteristics such as mileage, make, model, engine size, interior style, and cruise control. All cars in this data set were less than one year old when priced and considered to be in excellent condition.

(a) Fit a simple linear regression model with *Price* as the dependent variable and *Mileage* as the independent variable. Discuss the resulting model in terms of goodness of fit.

Intuitively, we would expect miles travelled to be an important factor in determining the retail price of a car. We would also expect cars with lower mileage to be worth more. The scatterplot in Figure 1 does show a negative association between *Price* and *Mileage*, however it is very weak. Interestingly, there is a group of 10 cars with much higher prices for which there is a very strong negative association. Examining the data file we find that these are all cars of the same make and model, Cadillac XLR-V8. Hence powerful sports cars appear to have higher prices. This suggests that there are most likely other variables besides mileage that explain the price of a used car and that make may be one of them.

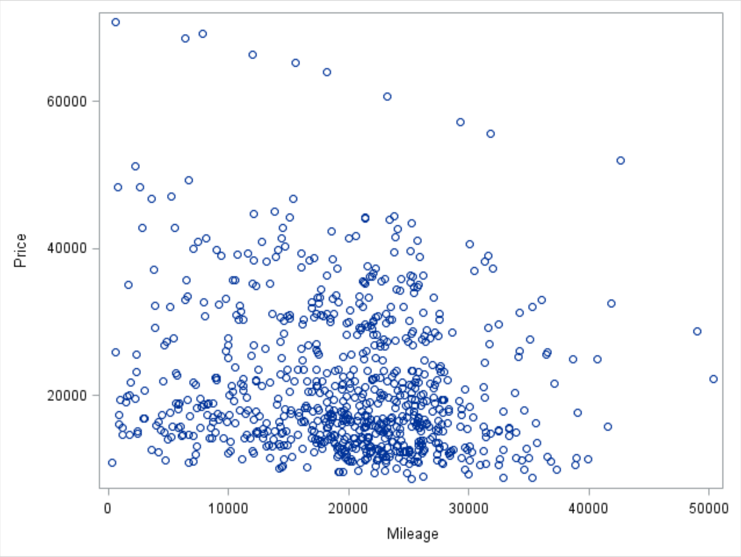


Figure 1. Scatterplot of *Price* vs *Mileage*

Although we do not expect the model based on mileage to have a lot of explanatory power, we use it as a starting point for this model building exercise.

The results of fitting a simple linear regression model to Price with Mileage as the only explanatory variable are reported in Tables 1 to 3, and the fitted line plot is shown in Figure 2. The plot confirms that a model based on mileage only is not going to have much predictive power as there is a lot of scatter and many data points actually lie outside the 95% prediction limits.

Using parameter estimates in Table 3, the fitted regression equation is:

Price = 24,765 – 0.1725 *Mileage*.

Our model suggests that on average, the price of a car decreases by 17 cents per additional travelled. From Table 2, the coefficient of determination R^2 is 0.0205, indicating that mileage explains only 2% of overall variability in price. The model is however statistically significant. The P-value shown in the Table 1 is less than 0.0001 and the corresponding F-statistic is F = 16.75 with 1 and 802 degrees of freedom. From Table 3, the t-statistic for the slope is t = -4.09 (802 df) and the P-value is again less than 0.0001. Therefore, *Mileage* is an important factor in determining *Price*, but it is clearly not the only one.

Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	1	1605590375	1605590375	16.75	<.0001
Error	802	76855792486	95830165		
Corrected Total	803	78461382861			

Table 1. Analysis of Variance table for the regression model of *Price* vs *Mileage*

Root MSE	9789.28829	R-Square	0.0205
Dependent Mean	21343	Adj R-Sq	0.0192
Coeff Var	45.86620		

Table 2. Goodness of fit statistics for *Price* vs *Mileage*

Parameter Estimates					
		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	24765	904.36328	27.38	<.0001
Mileage	1	-0.17252	0.04215	-4.09	<.0001

Table 3. Parameter estimates for *Price* vs *Mileage*

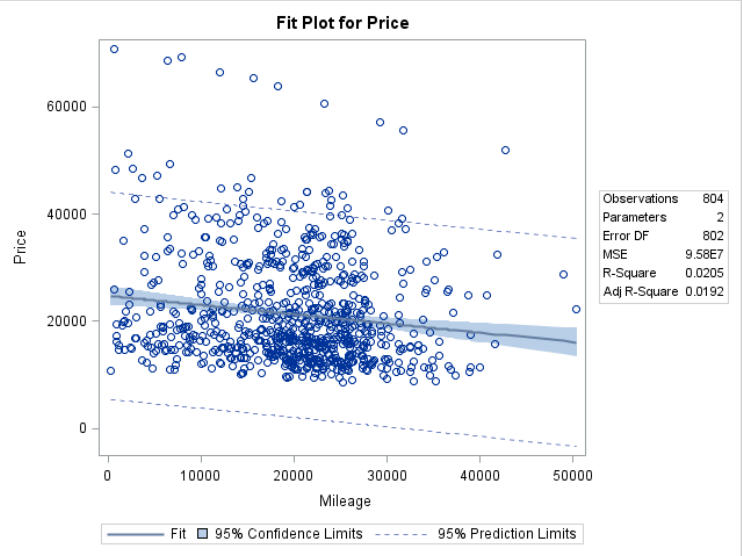


Figure 2. Fitted line plot for *Price* vs *Mileage*

(b) Use model selection techniques available in SAS to identify a better model with multiple predictors. A good model should have a high R-squared and adjusted R-squared, and a C_p value that is close to the number of predictors contained in the model.

Table 4 below shows the results of the R-squared selection method. Only the best three models for each choice of the number of explanatory variables are listed.

Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	0.3239	0.3230	171.9587	Cylinder
1	0.3115	0.3107	189.6865	Liter
1	0.1856	0.1846	370.6599	Cruise
2	0.3839	0.3824	87.5787	Cylinder Cruise
2	0.3680	0.3665	110.4398	Liter Cruise
2	0.3435	0.3418	145.7814	Cylinder Doors
3	0.4038	0.4016	61.0390	Cylinder Cruise Leather
3	0.4024	0.4001	63.0919	Mileage Cylinder Cruise
3	0.4002	0.3979	66.2195	Cylinder Doors Cruise
4	0.4225	0.4196	36.1508	Mileage Cylinder Cruise Leather
4	0.4191	0.4162	40.9776	Mileage Cylinder Doors Cruise
4	0.4178	0.4148	42.9749	Cylinder Doors Cruise Leather
5	0.4369	0.4334	17.4035	Mileage Cylinder Doors Cruise Leather
5	0.4300	0.4264	27.3540	Mileage Cylinder Cruise Sound Leather
5	0.4258	0.4222	33.4635	Cylinder Doors Cruise Sound Leather
6	0.4457	0.4415	6.8243	Mileage Cylinder Doors Cruise Sound Leather
6	0.4378	0.4336	18.1593	Mileage Cylinder Liter Doors Cruise Leather
6	0.4301	0.4259	29.1768	Mileage Cylinder Liter Cruise Sound Leather
7	0.4463	0.4414	8.0000	Mileage Cylinder Liter Doors Cruise Sound Leather

Table 4. Best three models for each choice of the number of explanatory variables

Based on the suggested criteria of a high R^2, high adjusted R^2 and a C_p value close to the number of predictors, there are two candidate models, one with six and the other with seven explanatory variables. These models are highlighted in Table 4 and will be analysed in part (c).

(c) Fit the model identified in part (b) and discuss goodness-of-fit. Also examine and discuss residual patterns. Are there any issues with collinearity?

Consider first the model with seven explanatory variables. The results of fitting this model to our data are shown in Tables 5 to 7. The model is highly statistically significant. The P-value shown in the Table 5 is less than 0.0001 and the corresponding F-statistic is F = 91.64 with 7 and 796 degrees of freedom.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	35014518311	5002074044	91.64	<.0001
Error	796	43446864550	54581488		
Corrected Total	803	78461382861			

Table 5. Analysis of Variance table for the regression model for *Price* with all seven explanatory variables

From Table 6, the coefficient of determination R^2 is 0.4463, indicating that the chosen variables together explain 44.63% of overall variability in price. Using parameter estimates listed in Table 7, the fitted regression equation is:

$$\text{Price} = 6758.76 - 0.1698\ \text{Mileage} + 3792.38\ \text{Cylinder} - 787.22\ \text{Liter} - 1542.75\ \text{Doors} + 6289.00\ \text{Cruise} - 1993.80\ \text{Sound} + 3349.36\ \text{Leather}$$

All coefficients are highly statistically significant except for *Liter*. The P-value for that coefficient estimate is 0.3642, which means that the hypothesis that beta for *Liter* is zero cannot be rejected.

Variance inflation column in Table 7 shows two values above 10, 13.22 for *Cylinder* and 13.52 for *Liter*, indicating that these two variables are highly correlated and we have multicollinearity in our model. If the purpose of the model is to explain price, only one of these variables should be included. Since the coefficient for *Liter* is not statistically significant, it is *Liter* that gets omitted. Note that the resulting model is the six explanatory variable model identified as good in part (b). As there is multicollinearity present, we do not interpret coefficients.

Remark. Both *Cylinder* and *Liter* are measures of engine size, and there is a strong relationship between them. This can be verified using correlation analysis. From the current results it may appear that *Liter* is not a useful predictor, however further analysis of the relationship between *Price*, *Mileage*, *Liter* and *Cylinder* would show that both *Cylinder* and *Liter* are useful predictors, but may not always register as statistically significant, particularly if included together in a particular model, because of multicollinearity.

Root MSE	7387.92854	R-Square	0.4463
Dependent Mean	21343	Adj R-Sq	0.4414
Coeff Var	34.61500		

Table 6. Goodness of fit statistics

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6758.75514	1876.96724	3.60	0.0003	0
Mileage	1	-0.16975	0.03187	-5.33	<.0001	1.00413
Cylinder	1	3792.37893	683.17989	5.55	<.0001	13.21983
Liter	1	-787.22073	867.06176	-0.91	0.3642	13.51875
Doors	1	-1542.74585	320.45564	-4.81	<.0001	1.09198
Cruise	1	6288.99715	657.99212	9.56	<.0001	1.18781
Sound	1	-1993.79528	571.77573	-3.49	0.0005	1.04945
Leather	1	3349.36162	597.68128	5.60	<.0001	1.05175

Table 7. Parameter estimates and variance inflation factors for the regression model for *Price* with all seven explanatory variables

We now fit the model with six explanatory variables, *Mileage*, *Cylinder*, *Doors*, *Cruise*, *Sound* and *Leather*. The results of fitting this model are shown in Tables 8 to 10. The model is highly statistically significant. The P-value shown in the Table 8 is less than 0.0001 and the corresponding F-statistic is F = 106.80 with 6 and 797 degrees of freedom.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	34969525995	5828254332	106.80	<.0001
Error	797	43491856866	54569457		
Corrected Total	803	78461382861			

Table 8. Analysis of Variance table for the regression model for *Price* with six explanatory variables

From Table 9, the coefficient of determination R^2 is 0.4457, indicating that the chosen variables together explain 44.57% of overall variability in price. Note that this is only slightly less than the R^2 value for the model that also included *Liter*. Using parameter estimates listed in Table 10, the fitted regression equation is:

Price = 7323.16 – 0.1705 Mileage + 3200.12 Cylinder – 1463.40 Doors + 6205.51 Cruise – 2024.40 Sound + 3327.14 Leather

All coefficients are highly statistically significant with all P-values less than 0.0001. Interpreting the estimated coefficients we find that on average, the price decreases by 17 cents for each extra mile travelled, and increases by $3,200.17 for an extra cylinder. All else kept fixed, cruise control adds on average $6,205.51 to the price, while an upgraded sound system decreases the price by $2,024.40. Leather seats increase the price by $3,327.14 on average. Finally, the price decreases on average by $1,463.40 per each additional door. The intercept does not have a useful interpretation in this case.

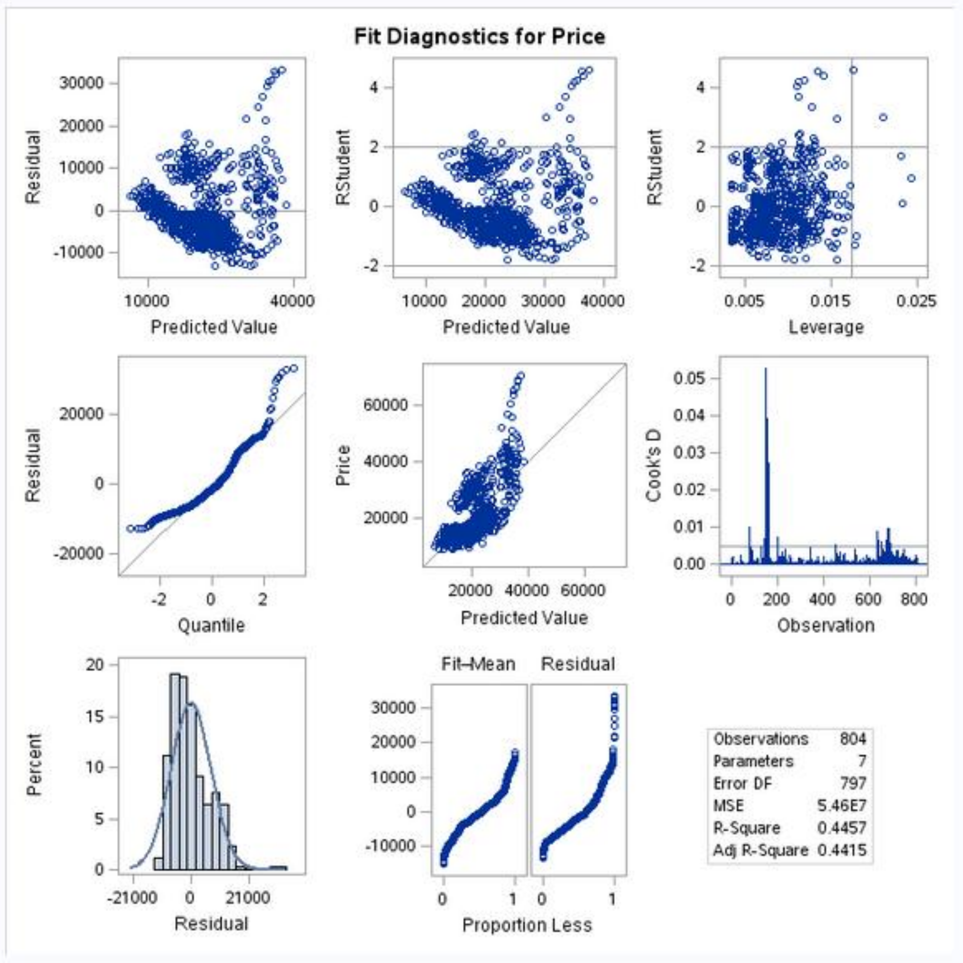
Root MSE	7387.11422	R-Square	0.4457
Dependent Mean	21343	Adj R-Sq	0.4415
Coeff Var	34.61118		

Table 9. Goodness of fit statistics

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7323.16431	1770.83680	4.14	<.0001
Mileage	1	-0.17052	0.03186	-5.35	<.0001
Cylinder	1	3200.12460	202.98326	15.77	<.0001
Doors	1	-1463.39906	308.27441	-4.75	<.0001
Cruise	1	6205.51127	651.46348	9.53	<.0001
Sound	1	-2024.40071	570.71826	-3.55	0.0004
Leather	1	3327.14331	597.11426	5.57	<.0001

Table 10. Parameter estimates and variance inflation factors for the regression model for *Price* with six explanatory variables

Figures 3 to 5 show regression diagnostics. Residual plots in Figure 3 indicate that there are quite a few outliers and influential observations. There is a cluster of observations with studentised residuals greater than 2; these are the Cadillacs identified in part (a). The Rstudent vs Leverage plot indicates seven points of high leverage. This is confirmed by the Cook's D plot which shows a handful of observations with Cook's D values of up to 10 times the cut-off value of 4/n = 4/804, two of which also have high studentised residuals. The residual vs predicted value plot shows some evidence of heteroskedasticity (unequal variance). From the histogram and the Q-Q plot, residuals are strongly skewed to the right hence non-Normal.



Observations	804
Parameters	7
Error DF	797
MSE	5.46E7
R-Square	0.4457
Adj R-Square	0.4415

Figure 3. Diagnostic plots for the regression model with six explanatory variables

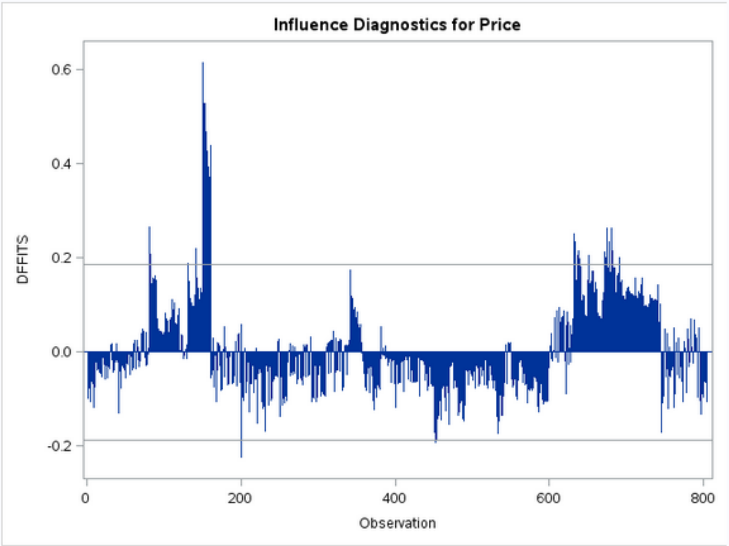


Figure 4. DFFITS vs observation order for the model with six explanatory variables

Examining the DFFITS plot in Figure 4 we find that there are a number of observations with dfits values up to three times higher than the cut-off value of 2 x sqrt(7/804). For these observations, adjusted predicted values are much higher than the predicted values generated by the model that includes these observations. These observations may therefore be highly influential.

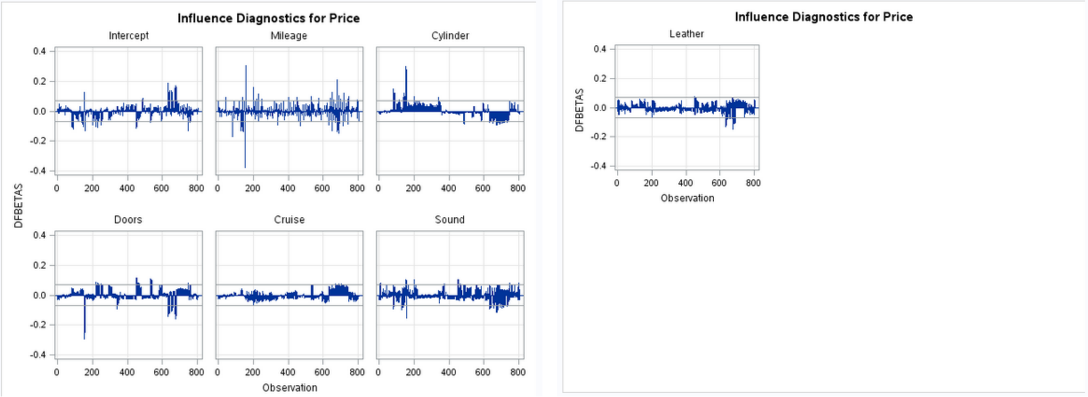


Figure 5. DFBETAS vs observation order for the regression model with six explanatory variables

More specific measures of influence are given by DFBETAS shown in Figure 5. For every explanatory variable, there are a number of observations with either higher or lower dfbeta values than the cut-off point of 2/sqrt(804). This means that when these observations are included in the model, slope coefficient estimates are either much higher or much lower than without them.

If desired, all outliers and influential observations can be identified by examining diagnostics output in more detail. Please refer to Appendix 1 for an example of SAS code that allows for this.

As many of the regression assumptions are violated, care must be taken when attempting to generalise the results to the population of all used cars in good condition that are less than one year old.

(d) Create dummy variables based on the makes of cars in this data set (Buick, Cadillac, Chevrolet, Pontiac, SAAB, and Saturn). Also create a new variable *LPrice* = log(*Price*).

Please refer to Appendix 1 for the SAS code that was used to create the new variables.

Note that the dummy variables for *Make* were defined in relation to 'Saturn'. In other words, the model for Saturn is obtained by setting all the other dummy variables to value zero.

(e) Fit a multiple regression model with *LPrice* as the dependent variable and *Mileage*, *Cylinder* and *Make* dummy variables as explanatory variables. Examine residuals and comment on the goodness of fit. Try other models including additional explanatory variables and comment on the results.

We first consider the multiple regression model for log price that uses *Mileage*, *Cylinder* and *Make* dummy variables as explanatory variables. The results of fitting this model are shown in Tables 11 to 13.

Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	7	119.41280	17.05897	867.94	<.0001
Error	796	15.64494	0.01965		
Corrected Total	803	135.05773			

Table 11. Analysis of Variance table for the regression model for *log Price* with *Mileage*, *Cylinder* as well as *Make* dummy variables

Root MSE	0.14019	R-Square	0.8842
Dependent Mean	9.87905	Adj R-Sq	0.8831
Coeff Var	1.41911		

Table 12. Goodness of fit statistics

Parameter Estimates					
		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	8.86432	0.03032	292.37	<.0001
Mileage	1	-0.00000789	6.053592E-7	-13.03	<.0001
Cylinder	1	0.19226	0.00487	39.49	<.0001
Buick	1	0.07680	0.02528	3.04	0.0025
Cadillac	1	0.43708	0.02849	15.34	<.0001
Chev	1	-0.02673	0.01999	-1.34	0.1817
Pontiac	1	0.00991	0.02230	0.44	0.6568
SAAB	1	0.81681	0.02242	36.43	<.0001

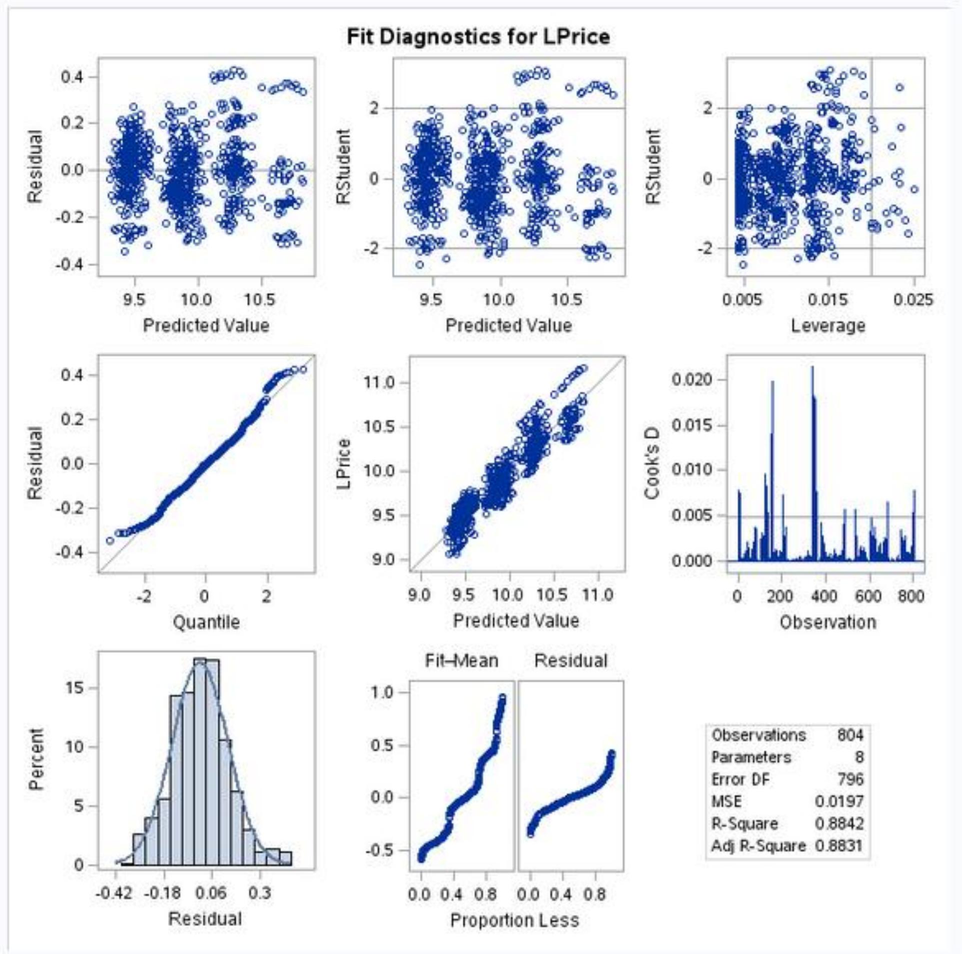
Table 13. Parameter estimates for the regression model for *log Price* with *Mileage*, *Cylinder* as well as *Make* dummy variables

The model is highly statistically significant. The P-value shown in the Table 11 is less than 0.0001 and the corresponding F-statistic is F = 867.94 with 7 and 796 degrees of freedom.

From Table 12, the coefficient of determination R^2 is 0.8842, indicating that the chosen variables together explain 88.42% of overall variability in log price. Note that the adjusted R^2 is only slightly lower at 0.8831, showing very little loss of predictive power if we consider the population rather than a sample of cars. There is also a clear improvement compared to the models considered in part (c). Using parameter estimates listed in Table 13, the fitted regression equation is:

$$\text{LPrice} = 8.86 - 0.000008\ \text{Mileage} + 0.1923\ \text{Cylinder} + 0.0768\ \text{Buick} + 0.4371\ \text{Cadillac} \\ - 0.0267\ \text{Chev} + 0.0099\ \text{Pontiac} + 0.8168\ \text{SAAB}$$

All coefficients are highly statistically significant except for *Chev* and *Pontiac*. These two variables could therefore be eliminated from the model. Interpreting the estimated coefficients we find that on average, log price decreases by 0.008 for every thousand miles travelled, and increases by 0.1923 for an extra cylinder. As the dummy variables were defined using Saturn as the baseline, the preceding sentence describes the model for a Saturn. All else kept fixed, log price increases by 0.0768 on average if the car is a Buick, by 0.4371 if it a Cadillac and by 0.8168 if it is a SAAB.



Observations	804
Parameters	8
Error DF	796
MSE	0.0197
R-Square	0.8842
Adj R-Square	0.8831

Figure 6. Diagnostic plots for the regression model for *log Price* vs *Mileage*, *Cylinder* as well as *Make* dummy variables

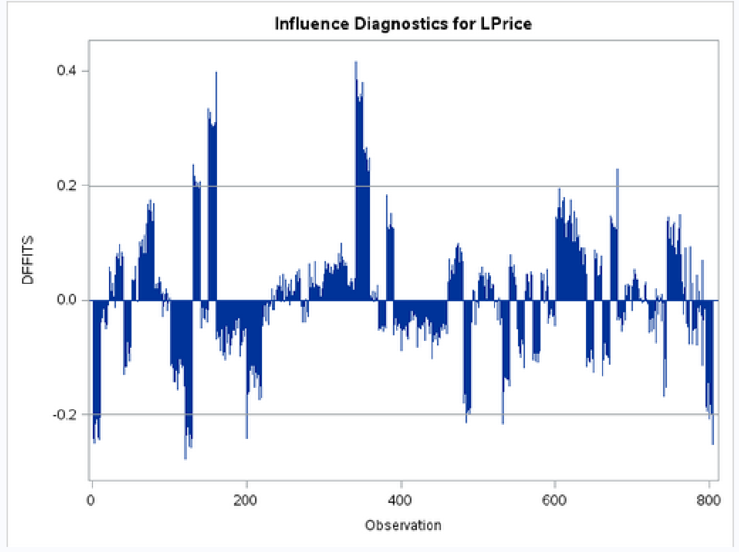


Figure 7. DFFITS vs observation order for the model with *Make* dummy variables

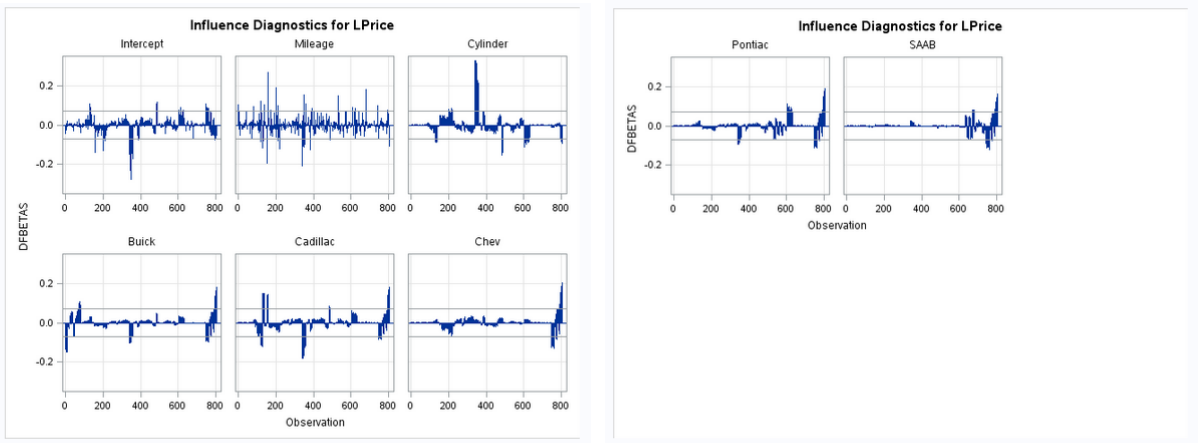


Figure 8. DFBETAS vs observation order for the regression model with *Make* dummy variables

Diagnostic plots in Figures 6 to 8 show improvement in the model fit, but there is some clustering of residuals which still appear to be non-Normal. Formal Normality tests could be applied to confirm the visual impression from the Q-Q plot and the histogram of the residuals. There are still outliers and influential observations present, but they are somewhat less extreme.

Additional dummy variables have been created based on categorical variable *Type* which indicates whether a car is a convertible, hatchback, sedan, coup or wagon. We now consider a model that uses also includes *Type* dummy variables with wagon as the baseline. Results for this model are shown in Tables 14 to 16.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	124.70987	12.47099	955.70	<.0001
Error	793	10.34786	0.01305		
Corrected Total	803	135.05773			

Table 14. Analysis of Variance table for the regression model for log price with *Mileage*, *Cylinder* as well as *Make* and *Type* dummy variables

Root MSE	0.11423	R-Square	0.9234
Dependent Mean	9.87905	Adj R-Sq	0.9224
Coeff Var	1.15631		

Table 15. Goodness of fit statistics

The model is highly statistically significant. The P-value shown in the Table 14 is less than 0.0001 and the corresponding F-statistic is F = 955.70 with 10 and 793 degrees of freedom. From Table 15, the coefficient of determination R^2 is 0.9234, indicating that the chosen variables together now explain 92.34% of overall variability in log price. Note that the adjusted R^2 is again only slightly lower. Using parameter estimates listed in Table 16, the fitted regression equation is:

$$\begin{gathered}\text{LPrice} = 8.97 - 0.000008\ \text{Mileage} + 0.1826\ \text{Cylinder} + 0.1225\ \text{Buick} + 0.4578\ \text{Cadillac} \\ - 0.0059\ \text{Chev} + 0.0160\ \text{Pontiac} + 0.7345\ \text{SAAB} + 0.2243\ \text{Convertible} \\ - 0.1939\ \text{Hatchback} - 0.0886\ \text{Sedan}\end{gathered}$$

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.97066	0.02559	350.53	<.0001
Mileage	1	-0.00000814	4.934706E-7	-16.50	<.0001
Cylinder	1	0.18256	0.00405	45.07	<.0001
Buick	1	0.12252	0.02084	5.88	<.0001
Cadillac	1	0.45784	0.02331	19.64	<.0001
Chev	1	-0.00586	0.01658	-0.35	0.7240
Pontiac	1	0.01602	0.01822	0.88	0.3795
SAAB	1	0.73450	0.01896	38.73	<.0001
Convertible	1	0.22430	0.01968	11.40	<.0001
Hatchback	1	-0.19391	0.01762	-11.00	<.0001
Sedan	1	-0.08860	0.01011	-8.76	<.0001

Table 16. Parameter estimates for the regression model for log price with *Mileage*, *Cylinder* as well as *Make* and *Type* dummy variables

All coefficients are again highly statistically significant except for *Chev* and *Pontiac*. These two variables could therefore be eliminated from the model. Interpreting the estimated coefficients we find that on average, log price decreases by 0.008 for every thousand miles travelled, and increases by 0.1826 for an extra cylinder. As the *Make* dummy variables were defined using Saturn as the baseline, and the *Type* dummy variables using wagon as the baseline, the preceding sentence describes the model for a Saturn and a wagon. All else kept fixed, log price increases by 0.1225 on average if the car is a Buick, by 0.4578 if it a Cadillac and by 0.7345 if it is a SAAB. All else equal, log price increases by 0.2243 on average if the car is a convertible, decreases by 0.1939 if it a hatchback and decreases by 0.0886 if it is a sedan. Note that the dummy variable *Coup* was eliminated from the analysis because it is highly correlated with other explanatory variables.

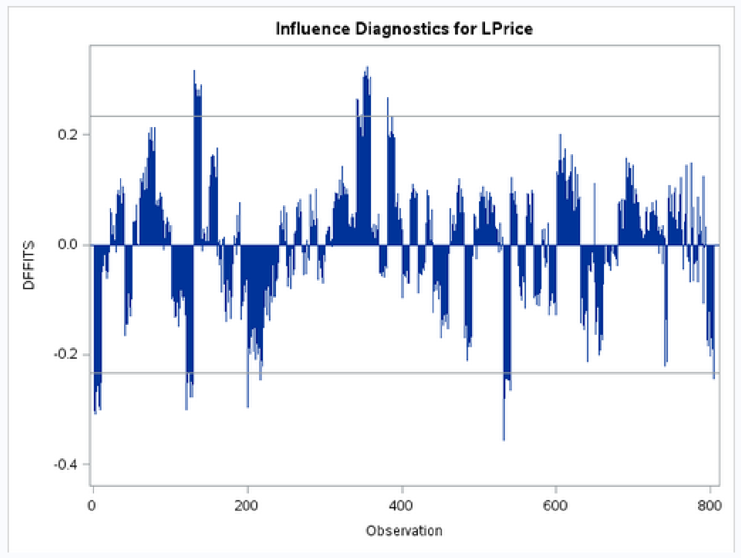
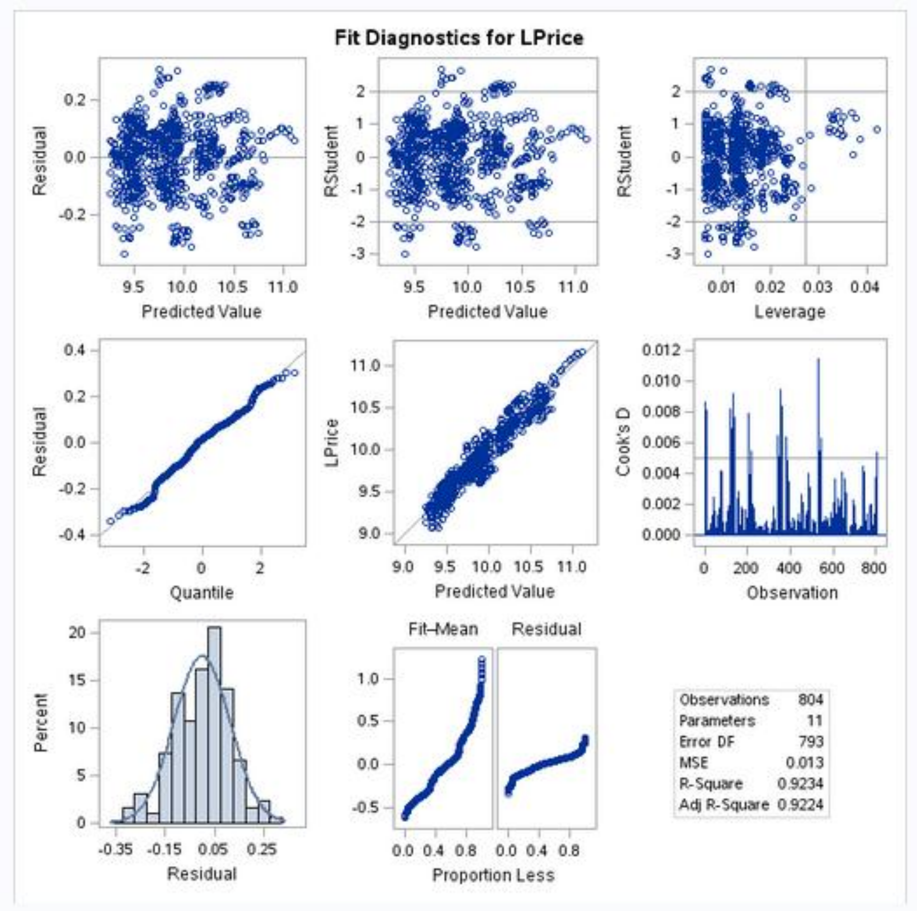


Figure 9. DFFITS vs observation order for the model with *Make* and *Type* dummy variables



Observations	804
Parameters	11
Error DF	793
MSE	0.013
R-Square	0.9234
Adj R-Square	0.9224

Figure 10. Diagnostic plots for the regression model for *log Price* vs *Mileage*, *Cylinder* as well as *Make* and *Type* dummy variables

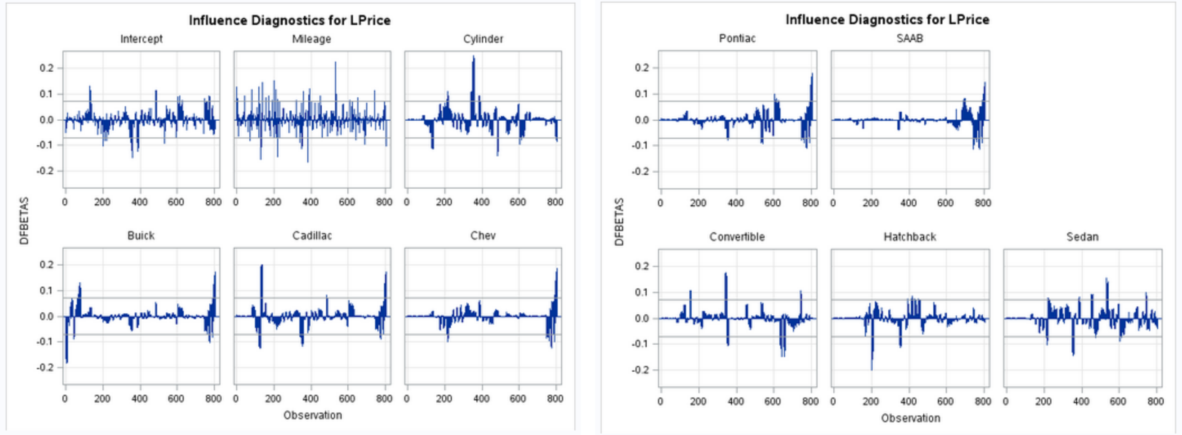


Figure 11. DFBETAS vs observation order for the regression model with *Make* and *Type* dummy variables

Diagnostic plots in Figures 9 to 11 show further improvement to the model fit. There is less clustering of residuals and they appear closer to being Normally distributed. There are still some outliers and potentially influential observations present, but the diagnostic measures for these observations are now much closer to the cut-off values.

Finally, we consider a model that uses *Liter* instead of *Cylinder.* Results for this model are shown in Tables 17 to 19.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	127.79519	12.77952	1395.40	<.0001
Error	793	7.26254	0.00916		
Corrected Total	803	135.05773			

Table 17. Analysis of Variance table for the regression model for log price with *Mileage*, *Liter* as well as *Make* and *Type* dummy variables

Root MSE	0.09570	R-Square	0.9462
Dependent Mean	9.87905	Adj R-Sq	0.9455
Coeff Var	0.96871		

Table 18. Goodness of fit statistics

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.23900	0.01826	505.94	<.0001
Mileage	1	-0.00000818	4.134128E-7	-19.78	<.0001
Liter	1	0.21536	0.00379	56.84	<.0001
Buick	1	0.13008	0.01733	7.50	<.0001
Cadillac	1	0.58330	0.01824	31.97	<.0001
Chev	1	-0.01484	0.01390	-1.07	0.2860
Pontiac	1	0.04115	0.01509	2.73	0.0065
SAAB	1	0.71275	0.01584	45.01	<.0001
Convertible	1	0.25376	0.01636	15.51	<.0001
Hatchback	1	-0.08899	0.01482	-6.00	<.0001
Sedan	1	-0.05720	0.00845	-6.77	<.0001

Table 19. Parameter estimates for the regression model for log price with *Mileage*, *Liter* as well as *Make* and *Type* dummy variables

The model is again highly statistically significant. The P-value shown in the Table 17 is less than 0.0001 and the corresponding F-statistic is F = 1395.40 with 10 and 793 degrees of freedom.

From Table 18, the coefficient of determination R^2 is 0.9462, indicating that the chosen variables together now explain 94.62% of overall variability in log price. There is therefore a slight improvement in R^2 value compared to the model with *Cylinder*. Using parameter estimates listed in Table 19, the fitted regression equation is:

$$\begin{gathered}\text{LPrice} = 9.24 - 0.000008\ \text{Mileage} + 0.2154\ \text{Liter} + 0.1301\ \text{Buick} + 0.5833\ \text{Cadillac} \\ - 0.0148\ \text{Chev} + 0.0412\ \text{Pontiac} + 0.7128\ \text{SAAB} + 0.2538\ \text{Convertible} \\ - 0.0890\ \text{Hatchback} - 0.0572\ \text{Sedan}\end{gathered}$$

All coefficients are again highly statistically significant except for *Chev*. There is now only one variable that could be eliminated from the model. Interpreting the estimated coefficients we find that on average, log price decreases by 0.008 for every thousand miles travelled, and increases by 0.2154 for an extra unit increase in engine size as measured by *Liter*. As the *Make* dummy variables were defined using Saturn as the baseline, and the *Type* dummy variables using wagon as the baseline, the preceding sentence describes the model for a Saturn and a wagon. All else equal, log price increases by 0.1301 on average if the car is a Buick, by 0.5833 if it a Cadillac, by 0.0412 if it is a Pontiac and by 0.7128 if it is a SAAB. All else equal, log price increases by 0.2538 on average if the car is a convertible, decreases by 0.0890 if it a hatchback and decreases by 0.0572 if it is a sedan. Note that the dummy variable *Coup* was again eliminated from the analysis because it is highly correlated with other explanatory variables.

Diagnostic plots in Figures 12 to 14 show further improvement to the model fit. There is much less clustering of residuals and they appear even closer to being Normally distributed. There are still some outliers and influential observations present, but the diagnostic measures for these observations are not too far off the cut-off values.

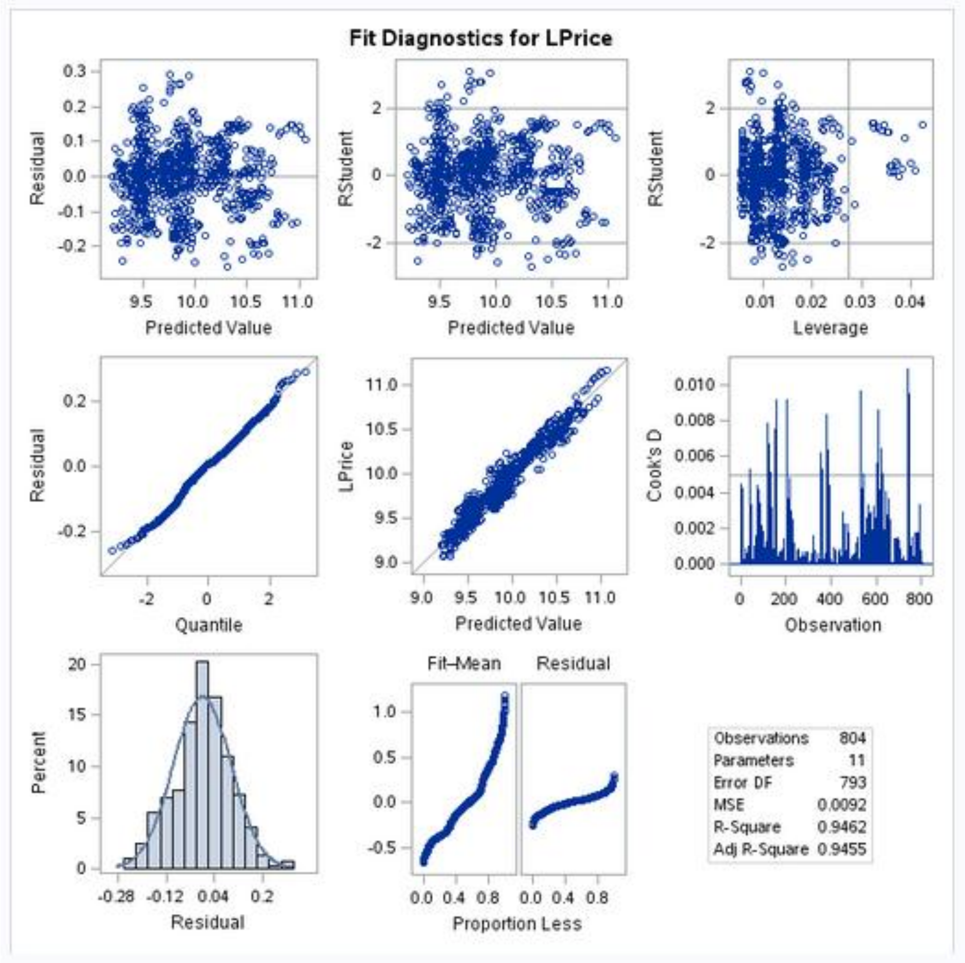


Figure 12. Diagnostic plots for the regression model for *log Price* vs *Mileage*, *Liter* as well as *Make* and *Type* dummy variables

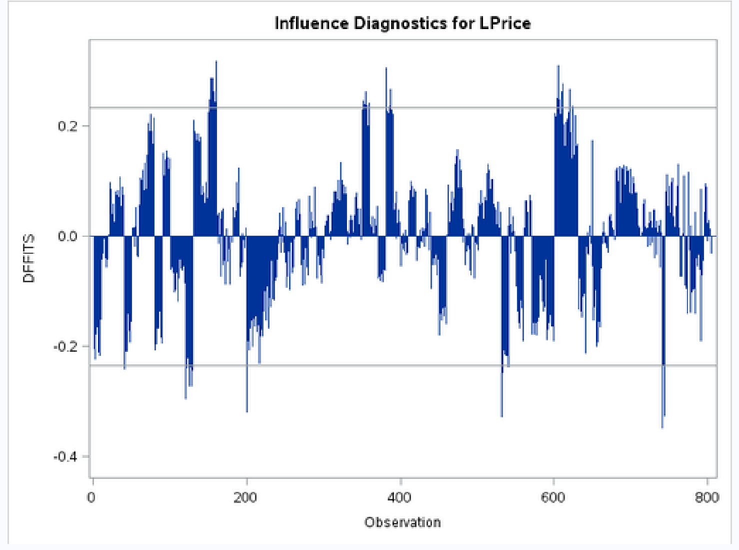


Figure 13. DFFITS vs observation order for the model with *Liter* as well as *Make* and *Type* dummy variables

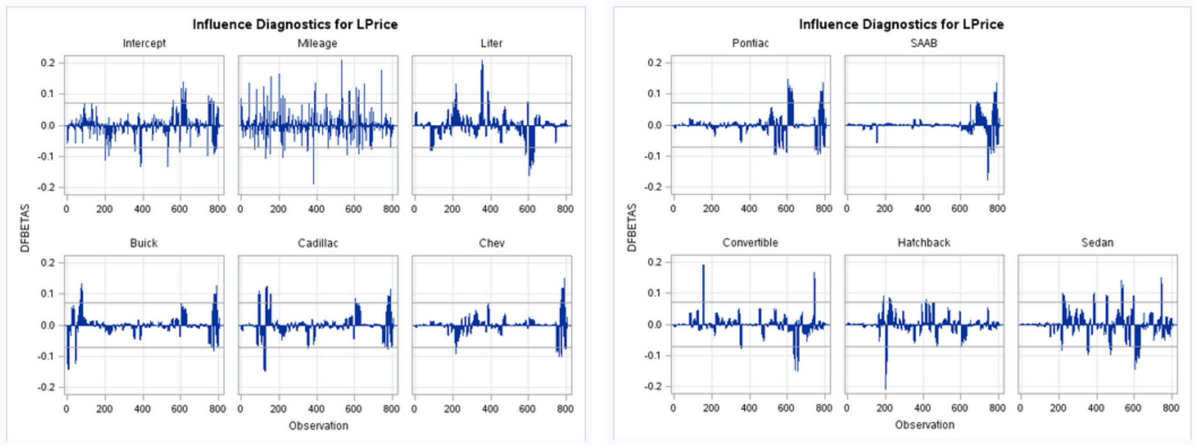


Figure 14. DFBETAS vs observation order for the regression model with *Liter* as well as *Make* and *Type* dummy variables

Conclusion

Despite starting with a simple regression model with a very low coefficient of determination, we were able to propose a number of much stronger models to predict or explain the price of a used car based on a variety of characteristics. This example illustrates that there is no single 'best' regression model, and that the same explanatory variable can be both significant and non-significant, depending on what other explanatory are included in the model.

The goals of the study will often determine which variables are included in a multivariate regression model. If the coefficients are not being interpreted, highly correlated explanatory variables could all be kept in the model. However, redundant variables should be eliminated from the model if the goal is to perform inferences.

Appendix 1. SAS code used to produce the output required for this practical

```
ods graphics on;

proc sgscatter data=mydata.kbb;
    plot price*mileage;
run;

title "Regression model for Price with Mileage only";

proc reg data=mydata.kbb PLOTS=ALL;
    model Price=Mileage;
    run;
quit;

/* Identifying best three models for each choice of the number of explanatory
variables */

title "Model selection using R-square criterion";

proc reg data=mydata.kbb PLOTS=NONE;
/* PLOTS=NONE suppresses diagnostic plots */
    model Price=Mileage Cylinder Liter Doors Cruise Sound Leather /
        selection=rsquare adjrsq cp best=3;
    run;

title "Regression models for Price with six and seven explanatory variables";

proc reg data=mydata.kbb PLOTS=ALL;
    model Price=Mileage Cylinder Doors Cruise Sound Leather;
    model Price=Mileage Cylinder Liter Doors Cruise Sound Leather / VIF;
    run;

/* Creating a temporary data file called ‘kbb_output’ in which the regression
diagnostics will be stored */

proc reg data=mydata.kbb PLOTS=NONE NOPRINT;
/* Commands Plots=NONE and NOPRINT suppress all output */
    model Price=Mileage Cylinder Doors Cruise Sound Leather / influence r;
    output out=kbb_output residual=price_residual rstudent=price_rstudent
        h=price_leverage dffits=price_dffits cookd=price_cookd
        ~~Price Make Model~~;
    run;

/* Listing observations identified as outliers or influential using rules of thumb
and the diagnostic output file */

title "Observations with large residuals";

proc print data=work.kbb_output;
    var Price Make Model;
    where abs(price_rstudent)>2;
run;

title "Observations with high leverage";

proc print data=work.kbb_output;
    var Price Make Model;
    where price_leverage > 14/804;
run;

title "Observations with high Cook's D values";

proc print data=work.kbb_output;
    var Price Make Model;
    where price_cookd > 4/804;
run;
```

```
title "Observations with high dffits values";

proc print data=work.kbb_output;
    var Price Make Model;
    where price_dffits > 2*sqrt(6/804);
run;

/* Creating log(Price) and dummy variables for Make and Type */

data work.kbb_transf;
    set mydata.kbb;

    LPrice=log(Price);

    if Make='Buick' then
        Buick=1;
    else Buick=0;

    if Make='Cadillac' then
        Cadillac=1;
    else Cadillac=0;

    if Make='Chevrolet' then
        Chev=1;
    else Chev=0;

    if Make='Pontiac' then
        Pontiac=1;
    else Pontiac=0;

    if Make='SAAB' then
        SAAB=1;
    else SAAB=0;

    if Type='Convertible' then
        Convertible=1;
    else Convertible=0;

    if Type='Hatchback' then
        Hatchback=1;
    else Hatchback=0;

    If Type='Sedan' then
        Sedan=1;
    else Sedan=0;

    if Type='Coup' then
        Coup=1;
    else Coup=0;
run;

title "Regression models for log price with Make and Type dummy variables";

proc reg data=work.kbb_transf PLOTS=ALL;
    model LPrice=Mileage Cylinder Buick Cadillac Chev Pontiac SAAB;
    model LPrice=Mileage Cylinder Buick Cadillac Chev Pontiac SAAB Convertible
        Hatchback Sedan;
    model LPrice=Mileage Liter Buick Cadillac Chev Pontiac SAAB Convertible
        Hatchback Sedan;
   run;
quit;

ods graphics off;
```