Prices of Automobiles regressed STatistically (PAST model)

Gian Alix
gian.alix@gmail.com

Celina Landolfi
clandolfi17@gmail.com

Department of Mathematics and Statistics
York University

GitHub Repository: https://github.com/techGIAN/PAST_AutoPrice_Regressor/tree/master
Paper: https://github.com/techGIAN/PAST_AutoPrice_Regressor/blob/master/PAST_Project_Paper.pdf

Preliminary Scan

Variables to Keep

- x_5 (door number)
- x_7 (drive wheel)
- x_9 (wheel base)
- x_{10} (car length)
- x_{11} (car width)
- x_{12} (car height)
- x_{13} (curb weight)
- x_{16} (engine size)
- x_{20} (compression ratio)
- x_{21} (horse power)
- x_{23} (city mpg)
- x_{24} (highway mpg)

Variables to Drop

- x_1 (symboling)
- x_2 (make)
- x_3 (fuel type)
- x_4 (aspiration)
- x_6 (car body)
- x_8 (engine location)
- x_{14} (engine type)
- x_{15} (cylinder number)
- x_{17} (fuel system)
- x_{18} (bore ratio)
- x_{19} (stroke ratio)
- x_{22} (peak rpm)

Multicollinearity

Pearson Correlation Coefficients, N = 164
Prob > |r| under H0: Rho=0

x16	x18	x19	x20	x21	x22	x23	x24
0.00576 0.9416	0.25415 0.0010	-0.08752 0.2651	0.15735 0.0442	0.02529 0.7478	-0.22123 0.0044	0.02587 0.7423	0.02704 0.7311
0.55544 <.0001	0.44102 <.0001	0.19782 0.0111	0.28581 0.0002	0.30342 <.0001	-0.40858 <.0001	-0.42650 <.0001	-0.51836 <.0001
0.67863 <.0001	0.62606 <.0001	0.18129 0.0202	0.18482 0.0178	0.52039 <.0001	-0.33008 <.0001	-0.65905 <.0001	-0.69918 <.0001
0.71080 <.0001	0.55512 <.0001	0.21414 0.0059	0.22157 0.0044	0.60532 <.0001	-0.26411 0.0006	-0.60826 <.0001	-0.65287 <.0001
0.14941 0.0562	0.18099 0.0204	-0.00978 0.9011	0.25989 0.0008	-0.06047 0.4418	-0.31917 <.0001	-0.08228 0.2949	-0.15284 0.0507
0.84307 <.0001	0.68614 <.0001	0.18697 0.0165	0.19049 0.0146	0.73041 <.0001	-0.30122 <.0001	-0.75340 <.0001	-0.80135 <.0001
1.00000	0.63909 <.0001	0.21479 0.0057	0.04496 0.5675	0.78328 <.0001	-0.29090 0.0002	-0.64499 <.0001	-0.67367 <.0001
0.63909 <.0001	1.00000	-0.07268 0.3550	0.03203 0.6839	0.65147 <.0001	-0.27146 0.0004	-0.62299 <.0001	-0.60884 <.0001
0.21479 0.0057	-0.07268 0.3550	1.00000	0.15437 0.0484	0.06582 0.4024	-0.10251 0.1915	-0.08684 0.2689	-0.09579 0.2224
0.04496 0.5675	0.03203 0.6839	0.15437 0.0484	1.00000	-0.17960 0.0214	-0.40269 <.0001	0.26477 0.0006	0.19419 0.0127
0.78328 <.0001	0.65147 <.0001	0.06582 0.4024	-0.17960 0.0214	1.00000	0.11183 0.1540	-0.79623 <.0001	-0.75748 <.0001
-0.29090 0.0002	-0.27146 0.0004	-0.10251 0.1915	-0.40269 <.0001	0.11183 0.1540	1.00000	-0.08704 0.2677	-0.01575 0.8413
-0.64499 <.0001	-0.62299 <.0001	-0.08684 0.2689	0.26477 0.0006	-0.79623 <.0001	-0.08704 0.2677	1.00000	0.96871 <.0001
-0.67367 <.0001	-0.60884 <.0001	-0.09579 0.2224	0.19419 0.0127	-0.75748 <.0001	-0.01575 0.8413	0.96871 <.0001	1.00000

Pearson Coefficient Correlation Matrix

x7fwd =	Intercept - x74wd - x7rwd

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	-40761	17642	-2.31	0.0223	0
x9	1	-11.88128	121.37798	-0.10	0.9222	8.51081
x10	1	17.97974	67.92063	0.26	0.7916	10.83747
x11	1	620.22894	273.62426	2.27	0.0249	5.66657
x12	1	139.11947	159.72615	0.87	0.3852	2.61839
x13	1	-1.00051	2.07218	-0.48	0.6299	18.49816
x16	1	137.46546	15.43544	8.91	<.0001	6.47347
x18	1	-4140.24348	1649.05116	-2.51	0.0131	2.98968
x19	1	-4209.52131	1065.15964	-3.95	0.0001	1.50962
x20	1	358.22988	92.88234	3.86	0.0002	2.05963
x21	1	30.83365	17.48133	1.76	0.0798	7.71545
x22	1	2.29921	0.75954	3.03	0.0029	2.21585
x23	1	-238.95990	207.09985	-1.15	0.2504	28.76805
x24	1	105.43637	187.36754	0.56	0.5745	26.00287
x5	1	-64.05359	331.24655	-0.19	0.8469	1.84462
x74wd	B	1757.54683	1628.79147	1.08	0.2823	1.59574
x7rwd	B	2160.83681	899.82541	2.40	0.0176	3.15845
x7fwd	0	0	.	.	.	.

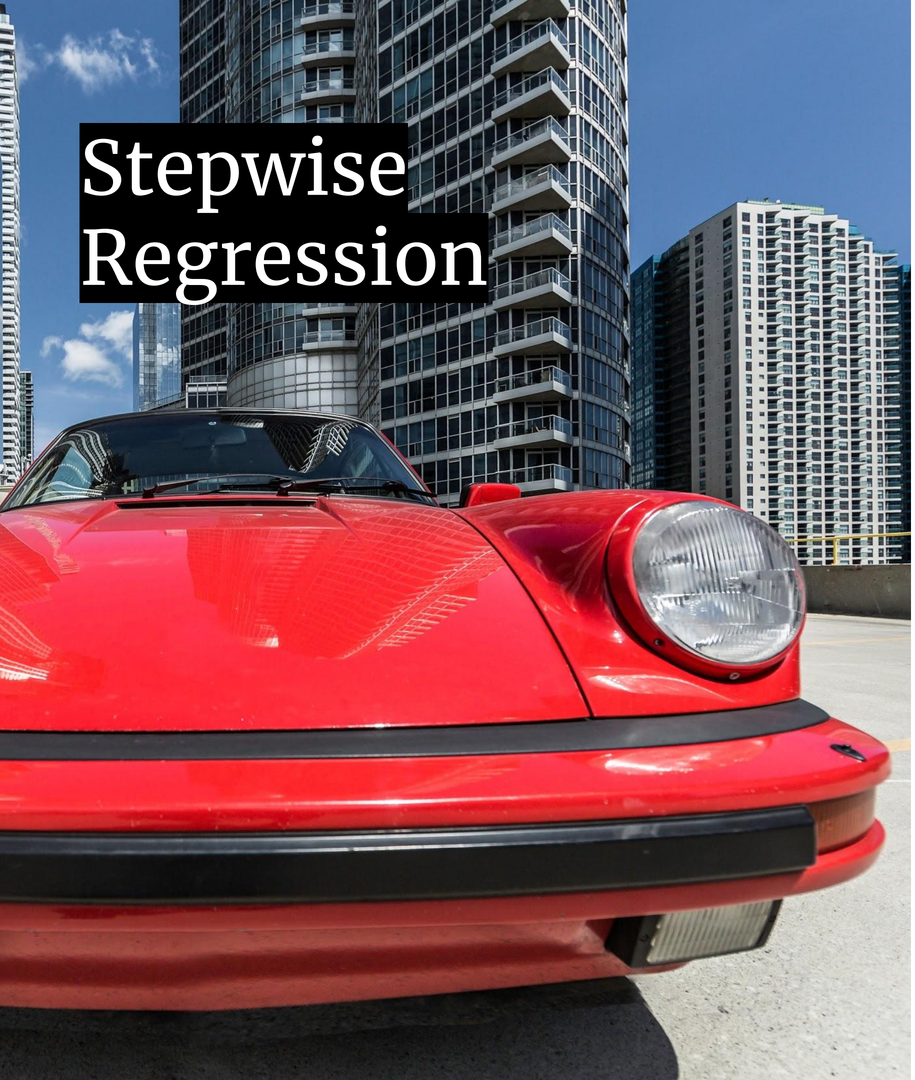
VIF table before dropping independent variables

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-44357	15608	-2.84	0.0051	0
x5	1	181.88488	353.86046	0.51	0.6080	1.74502
x74wd	1	1169.93248	1506.33419	0.78	0.4386	1.13138
x7rwd	1	2561.30389	804.75111	3.18	0.0018	2.09418
x9	1	-71.30813	124.41696	-0.57	0.5674	7.41283
x10	1	-61.09567	68.11441	-0.90	0.3712	9.03517
x11	1	667.70451	285.55369	2.34	0.0207	5.11588
x12	1	233.10343	170.89063	1.36	0.1746	2.48458
x16	1	99.75264	13.26619	7.52	<.0001	3.96392
x20	1	215.16945	91.73459	2.35	0.0203	1.66542
x21	1	48.96166	16.47623	2.97	0.0034	5.68149
x23	1	-105.26576	96.00303	-1.10	0.2746	5.12453

VIF table after dropping independent variables

Stepwise Regression



Stepwise Selection: Step 5

Variable x11 Entered: R-Square = 0.8265 and C(p) = 4.7783

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	8670813472	1734162694	150.55	<.0001
Error	158	1819947483	11518655		
Corrected Total	163	10490760955			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-42772	11490	159632613	13.86	0.0003
x11	536.19671	190.82757	90942693	7.90	0.0056
x16	94.03199	12.35653	667052080	57.91	<.0001
x20	186.26463	81.91760	59553612	5.17	0.0243
x21	58.69996	12.71267	245585988	21.32	<.0001
x7rwd	2225.73164	705.08890	114778253	9.96	0.0019

Bounds on condition number: 3.4612, 60.597

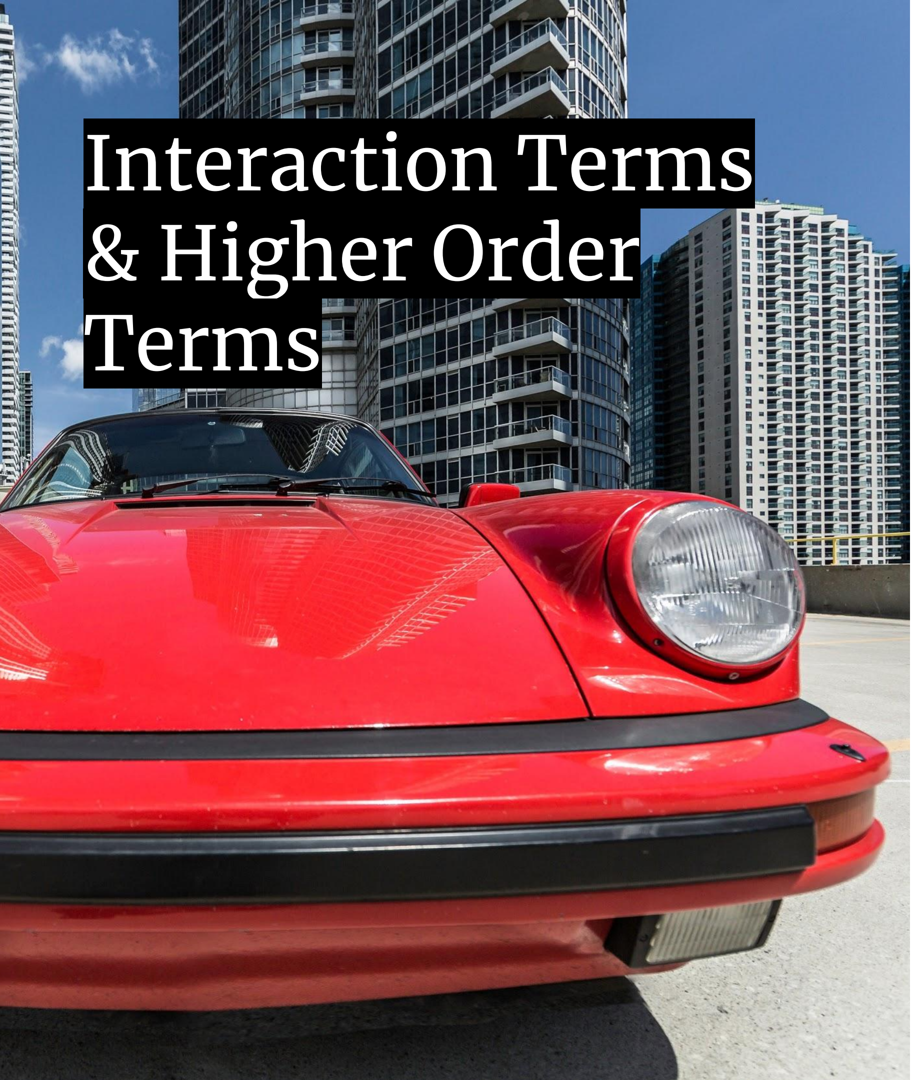
All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	x16		1	0.7531	0.7531	63.1233	494.15	<.0001
2	x21		2	0.0339	0.7870	34.5175	25.60	<.0001
3	x20		3	0.0194	0.8063	19.0288	15.99	<.0001
4	x7rwd		4	0.0115	0.8179	10.6125	10.06	0.0018
5	x11		5	0.0087	0.8265	4.7783	7.90	0.0056

The results of the first pass of Stepwise Regression

Interaction Terms & Higher Order Terms



Stepwise Selection: Step 7

Variable x16x11 Entered: R-Square = 0.8671 and C(p) = 25.3799

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	9096642717	1299520388	145.41	<.0001
Error	156	1394118238	8936655		
Corrected Total	163	10490760955			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-109201	35057	86712911	9.70	0.0022
x11	1942.08386	582.63190	99293873	11.11	0.0011
x16	-802.57998	251.31723	91139603	10.20	0.0017
x21	1320.31950	255.06352	239461879	26.80	<.0001
x11x21	-21.30604	3.94945	260080392	29.10	<.0001
x20x7rwd	262.37438	50.83685	238045819	26.64	<.0001
x16x11	10.33606	3.82595	65223765	7.30	0.0077
x16x21	1.25063	0.24838	226565251	25.35	<.0001

Bounds on condition number: 2290.2, 57246

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x16x21		1	0.7742	0.7742	134.636	555.36	<.0001
2	x20x7rwd		2	0.0468	0.8210	75.5293	42.13	<.0001
3	x11		3	0.0171	0.8381	55.2446	16.88	<.0001
4	x11x21		4	0.0050	0.8431	50.6648	5.11	0.0251
5	x21		5	0.0064	0.8495	44.3168	6.72	0.0104
6	x16		6	0.0114	0.8609	31.4915	12.82	0.0005
7	x16x11		7	0.0062	0.8671	25.3799	7.30	0.0077

The results of the second pass of Stepwise Regression

Model Comparison

Number	Model
1	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21}$
2	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{11} x_{21}$
3	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd}$
4	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd} + \beta_7 x_{11} x_{21}$
5	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{21} + \beta_5 x_{21} x_{7,rwd}$
6	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd} + \beta_7 x_{11} x_{21} + \beta_8 x_{16}$
7	$\hat{y} = \beta_0 + \beta_1 x_{7,4wd} + \beta_2 x_{7,rwd} + \beta_3 x_{11} + \beta_4 x_{20} + \beta_5 x_{21} + \beta_6 x_{21} x_{7,rwd} + \beta_7 x_{11} x_{21} + \beta_8 x_{16} + \beta_9 x_{16} x_{21}$

The seven “best” models used for comparison.

Model	k	C_k	R^2	$\bar{R}^2$	s	$PRESS$
1	6	6	0.7637	0.7563	3960.65	2,787,102,686
2	7	7	0.7646	0.7556	3966.35	3,841,614,615
3	7	7	0.7795	0.7710	3838.75	2,782,421,119
4	7	7	0.7820	0.7722	3829.10	3,547,983,422
5	6	6	0.7728	0.7657	3883.58	2,817,907,210
6	9	9	0.8373	0.8289	3318.50	2,271,013,952
7	10	10	0.8622	0.8542	3063.75	1,828,314,522

Thus, it is evident that **model 7** is the “best” model

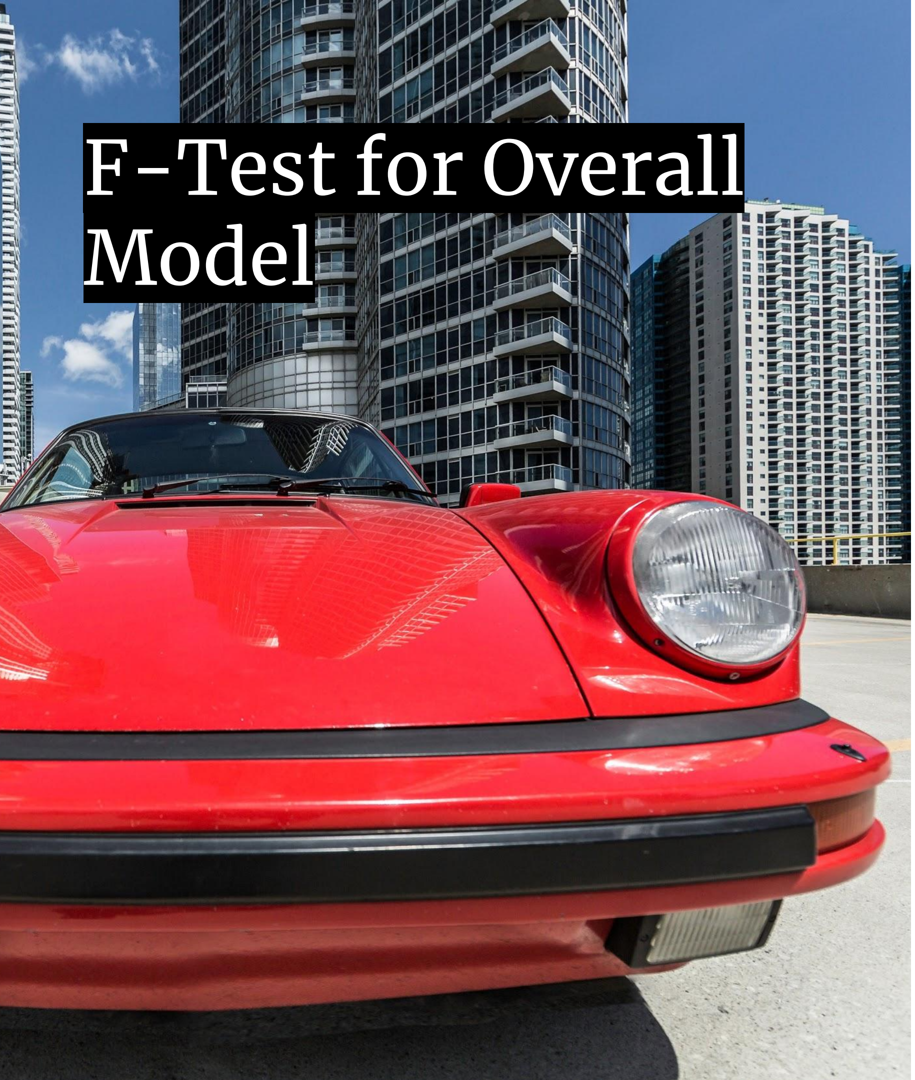
Outliers & Influential Points

Test Statistic	Description	Threshold	Applicable Observations (i)
Leverage Point (h_{ii})	Outlier with respect to x test	$h_{ii} > 0.06097$	i=8, 60, 62, 85, 87, 92, 109
Studentized Residual ($\frac{d_i}{s_{di}}$)	Outlier with respect to y test	$\lvert\frac{d_i}{s_{di}}\rvert>1.97559$	i=14, 16, 60, 62, 85, 87, 89, 92, 109
Cook's Distance (D_i)	Influential point test	$D_i>0.938263$	i=109
Difference of Betas ($g_j^{(i)}/s_{g_j}^{(i)}$)	a test for whether or not removing observation i will substantially change the parameter estimates	$\lvert\frac{f_i}{s_{d_i}}\rvert>2$	i=109 for x_{21} and $x_{11}x_{21}$
Difference in Fits Statistic (f_i/s_{d_i})	difference between the point predictions of y_i made with and without using the ith observation	$\lvert\frac{f_i}{s_{d_i}}\rvert>2$	i=109
(a) Covariance Ratio (CVR_i)	removing obs i enhances model precision	$CVR_i<0.817$	i=8, 14, 16, 89
(b) Covariance Ratio (CVR_i)	removing obs i damages model precision	$CVR_i > 1.1829$	i=109

Outlying and influential observations

Thus, observation 109 was **kept** in the training data &
Observations 14, 16 and 89 were **dropped**

C_k	k	R^2	$\bar{R}^2$	s
10	10	0.8603	0.8520	3057.83921



F-Test for Overall Model

- H_0: $\beta_1 = \beta_2 = ... = \beta_9 = 0$
  (no relation between y and the independent variables, i.e. no significant independent variables in the model)
- H_a: At least one in $\{\beta_1, \beta_2, ..., \beta_9\}$ is non-zero
  (at least one independent variable has significant relation with y)

Complete F-Test

The REG Procedure
Model: MODEL1
Dependent Variable: y

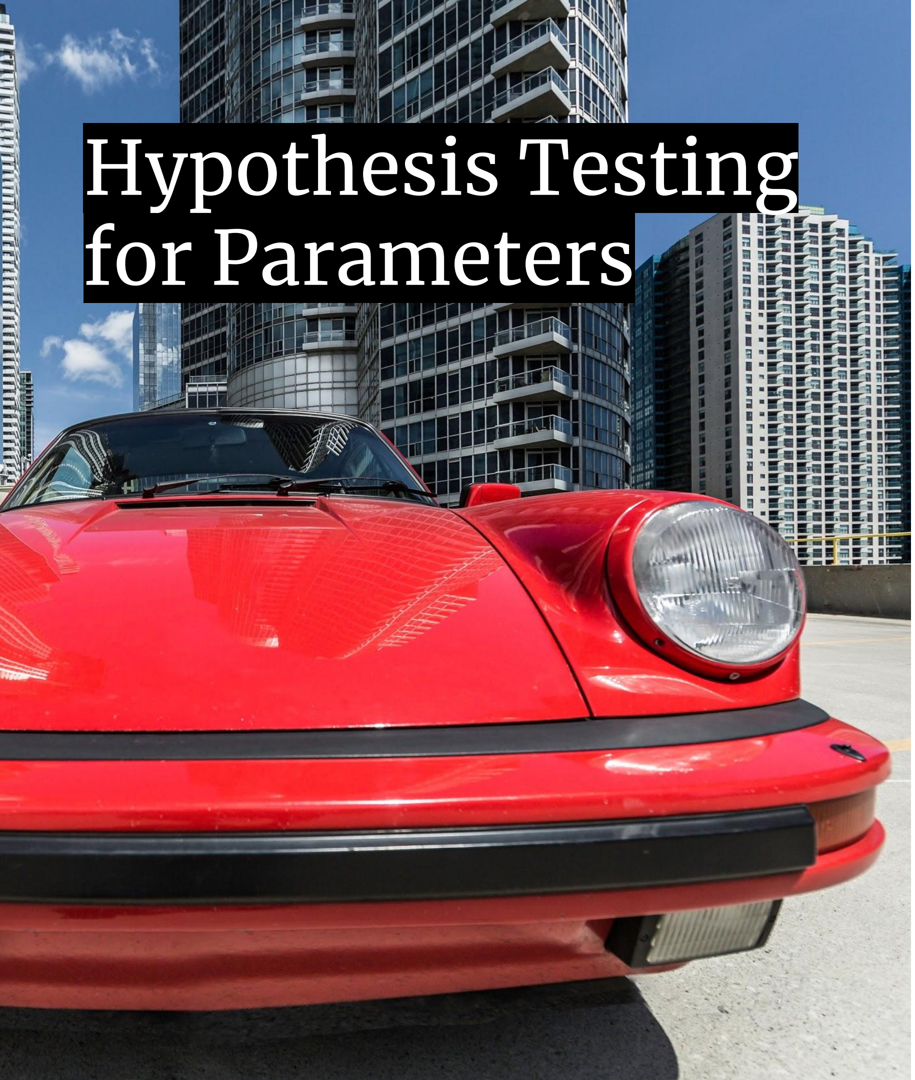
Number of Observations Read	161
Number of Observations Used	161

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	8697637432	966404159	103.35	<.0001
Error	151	1411907476	9350381		
Corrected Total	160	10109544907			

We **reject** H_0 since p-value < alpha

Hypothesis Testing for Parameters



Hypothesis Testing for b_j

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	161
Number of Observations Used	161

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	8697637432	966404159	103.35	<.0001
Error	151	1411907476	9350381		
Corrected Total	160	10109544907			

Root MSE	3057.83921	R-Square	0.8603
Dependent Mean	13252	Adj R-Sq	0.8520
Coeff Var	23.07542		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-169318	30000	-5.64	<.0001
x11	1	2876.88967	506.16133	5.68	<.0001
x20	1	181.76961	75.69650	2.40	0.0176
x21	1	913.89498	217.09589	4.21	<.0001
x74wd	1	2655.83332	1307.74064	2.03	0.0440
x7rwd	1	3231.48238	2587.91389	1.25	0.2137
x21x7rwd	1	-5.37010	22.50410	-0.24	0.8117
x11x21	1	-15.92851	3.59548	-4.43	<.0001
x16	1	-131.38162	41.71614	-3.15	0.0020
x16x21	1	1.50720	0.27730	5.44	<.0001

- H_0: $\beta_i = 0$, for $i = 1, 2, ..., 9$
- H_a: $\beta_i \neq 0$, for $i = 1, 2, ..., 9$

p-value of $x_{7,rwd}$ and $x_{21}x_{7,rwd}$ are > alpha; p-value of $x_{7,4wd}$ < alpha

$x_{7,rwd}$ is insignificant. $x_{7,4wd}$ is significant. Hence x_7 is important so **keep** both $x_{7,rwd}$ and $x_{7,4wd}$

$x_{21}x_{7,rwd}$ is insignificant. So **drop** $x_{21}x_{7,rwd}$

Do not reject H_0 in β_i=0, for when i=4.
Reject H_0 when i=1,2,3,5,6,7,8,9

Is wheel drive (x_7) significant?

- H_0: $\beta_1 = \beta_2 = 0$
  (the independent variables to be dropped, $x_{7,4wd}$ and $x_{7,rwd}$ are not significant to y)
- H_a: At least one of β_1, β_2 is non-zero
  (at least one of the independent variables to be dropped, $x_{7,4wd}$ and $x_{7,rwd}$ are significant to y)

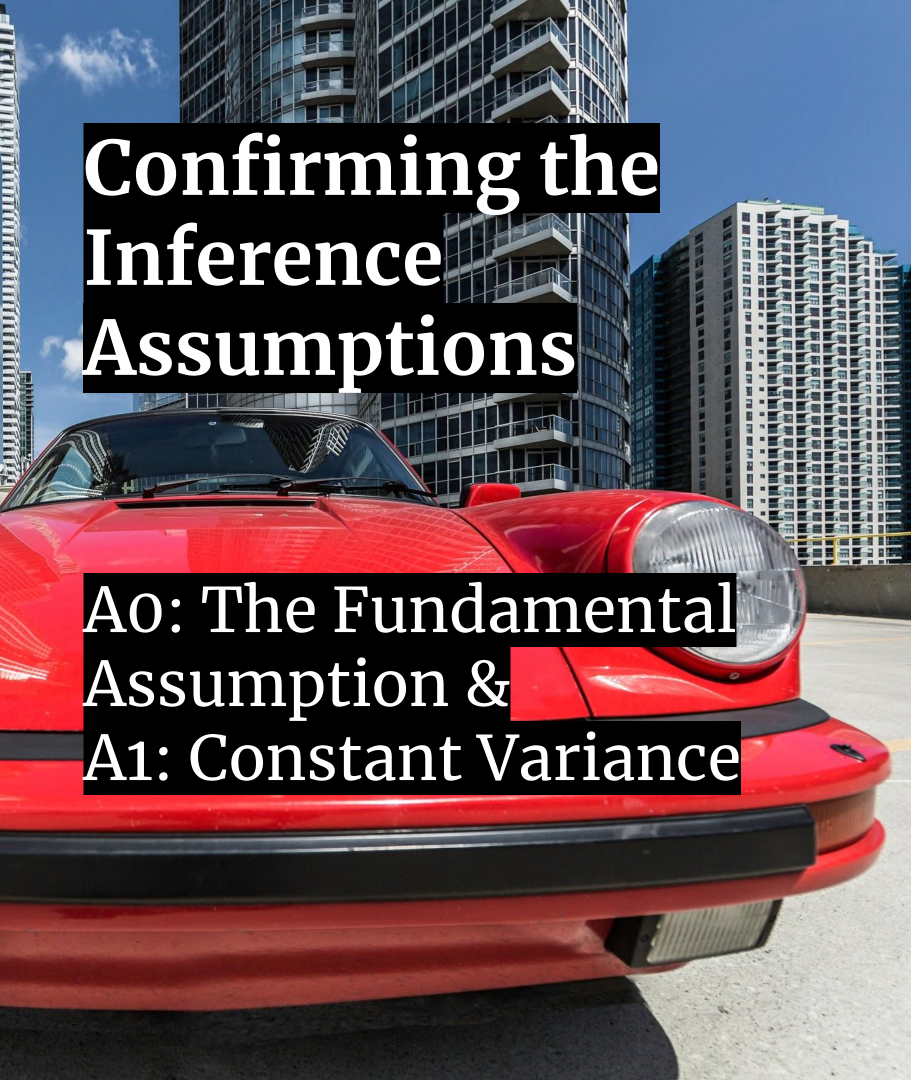
Partial F-Testing - drop x74wd and x7rwd?

The REG Procedure
Model: MODEL1

Test pft Results for Dependent Variable y

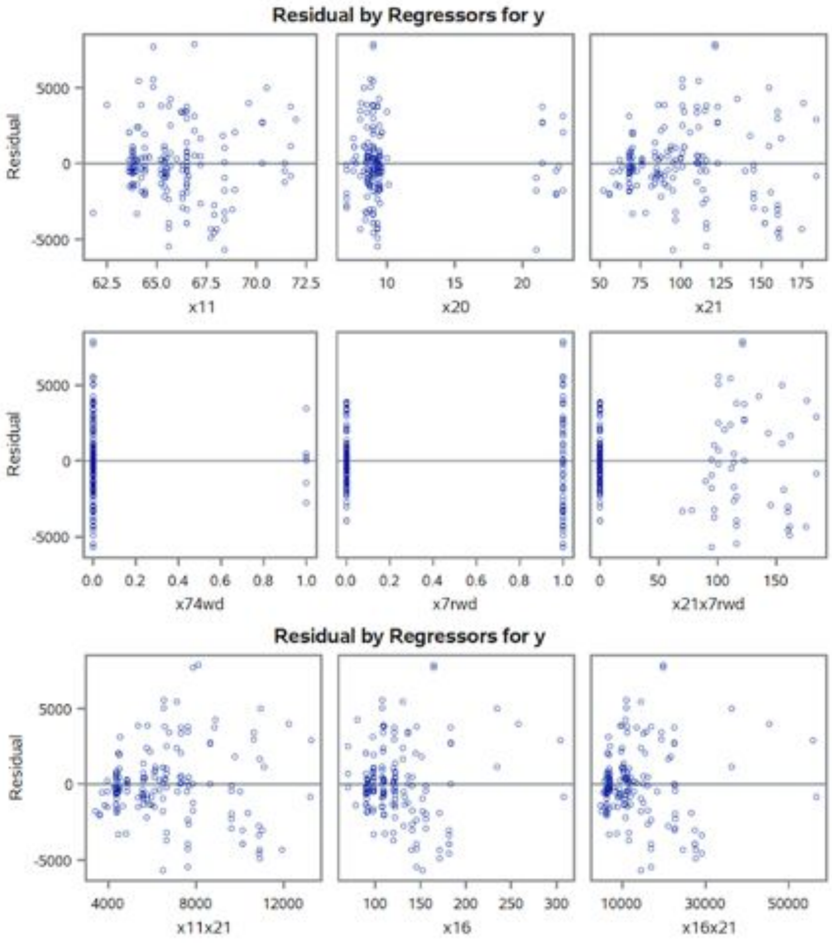
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	85562895	9.21	$\boxed{0.0002}$
Denominator	152	9292368		

We **reject** H_0 since p-value < alpha.
Wheel drive is significant to the model.



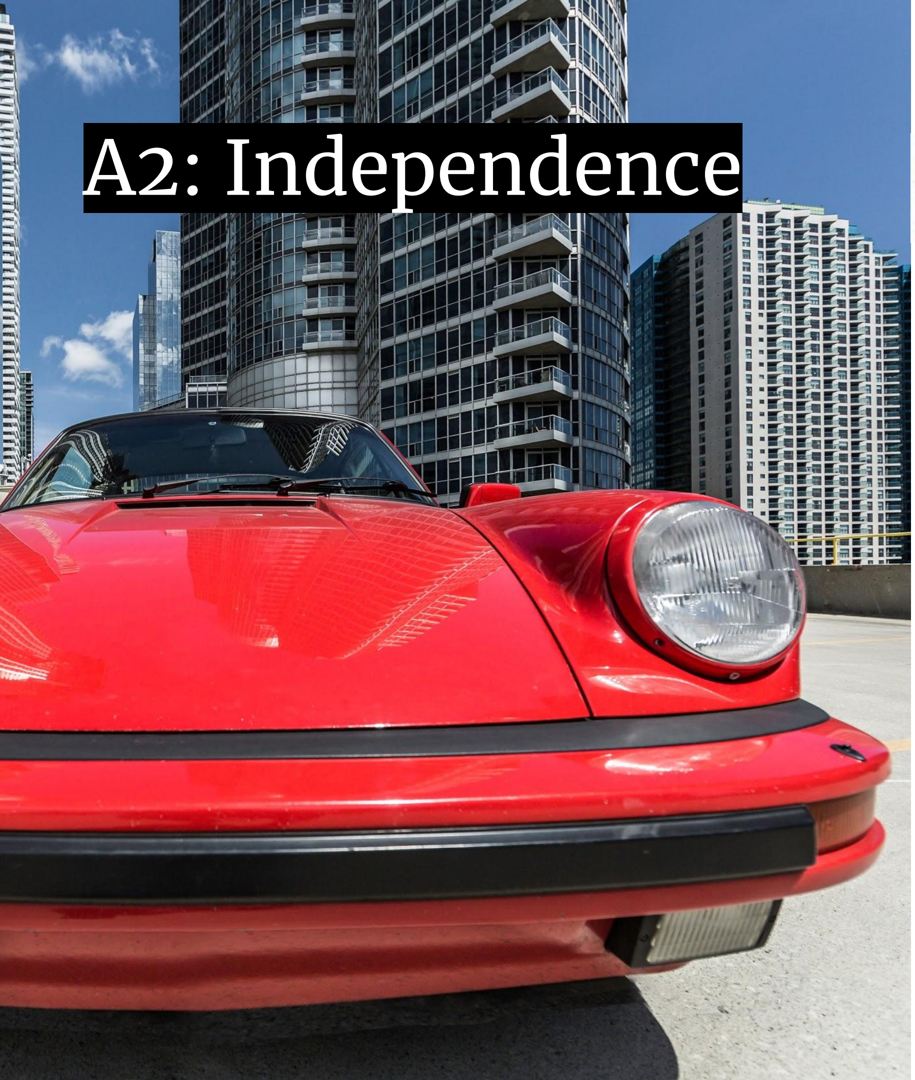
Confirming the Inference Assumptions

A0: The Fundamental Assumption & A1: Constant Variance



A0 and A1 **hold**

A2: Independence



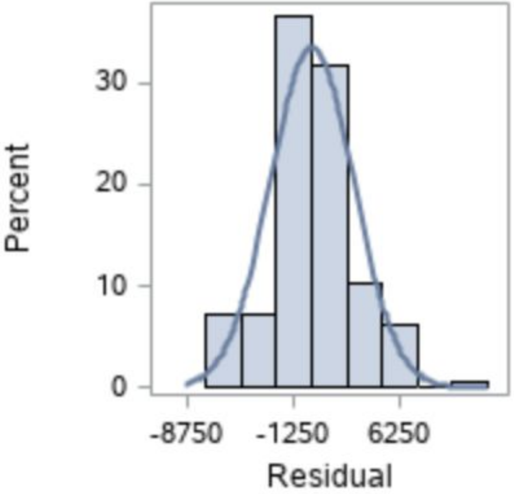
H_0: Error terms are not autocorrelated.
H_a: Error terms are positively or negatively autocorrelated.

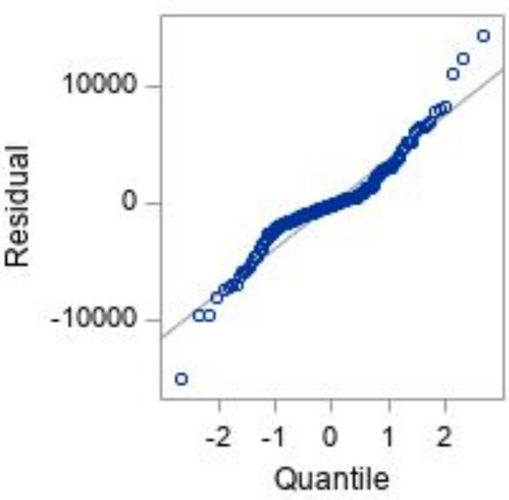
Durbin-Watson D	0.900
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	164
1st Order Autocorrelation	0.548

We **reject** H_0 since p-value < alpha and there seems to be autocorrelation.

This conclusion can be explained.

A3: Normality





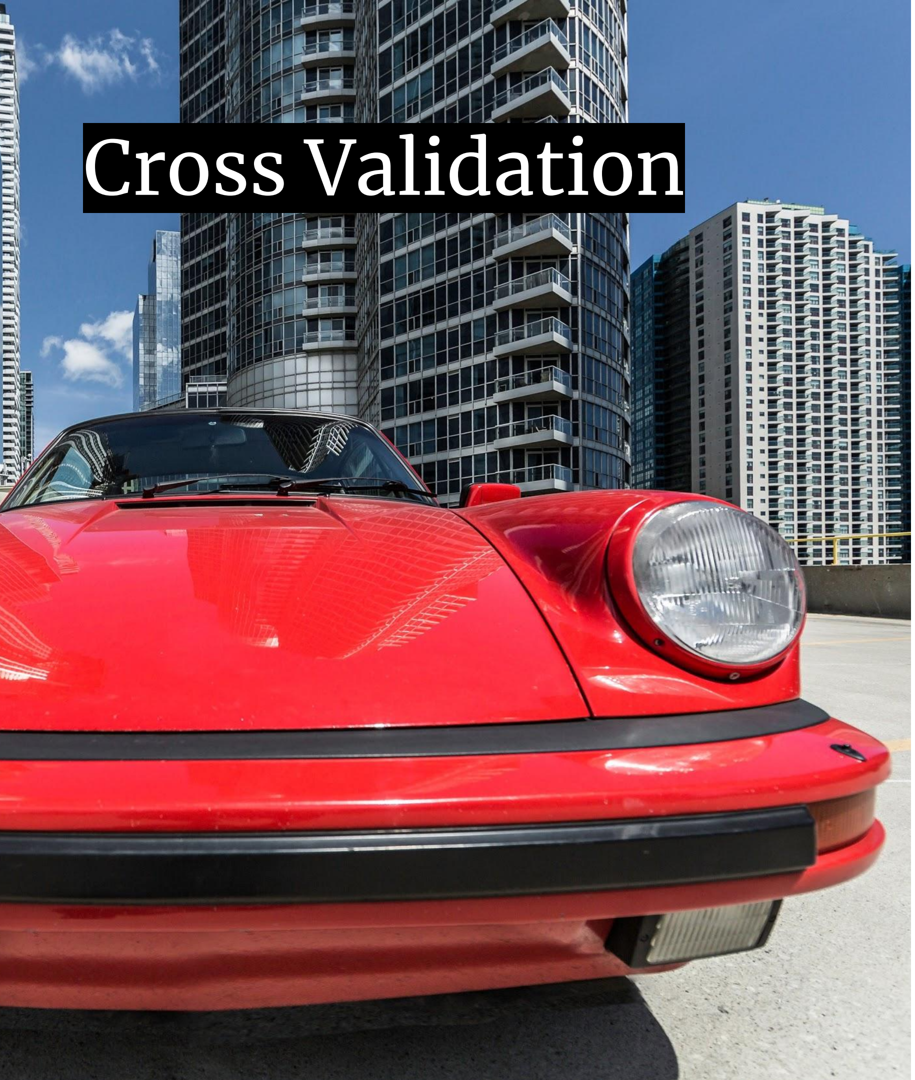
A3
holds.

Final Model & Interpretation

$$
\begin{aligned}
\hat{y} = -170238 &+ 2680.67x_{7,4wd} + 2634.04x_{7,rwd} \\
&+ 2886.72x_{11} + 177.98x_{20} + 918.89x_{21} \\
&- 15.98x_{11}x_{21} - 126.73x_{16} + 1.47x_{16}x_{21}
\end{aligned}
$$

C_k	k	R^2	$\bar{R}^2$	s
9	9	0.8603	0.8529	3048.34

Cross Validation



Prediction Intervals

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	18538	711.3975	17133	19944	12335	24741	.
2	.	29025	858.5085	27328	30721	22749	35300	.
3	.	47.7111	1219	-2361	2457	-6457	6552	.
4	.	7361	461.7648	6448	8273	1250	13471	.
5	.	6906	380.4661	6155	7658	818.1008	12995	.
6	.	8978	330.7515	8325	9632	2901	15055	.
7	.	34942	1020	32927	36958	28573	41311	.
8	.	68415	4557	59412	77419	57573	79258	.
9	.	6993	430.0375	6143	7842	891.4605	13094	.
10	.	15136	536.5173	14076	16196	9002	21270	.
11	.	14618	1135	12376	16860	8174	21062	.
12	.	7395	427.0560	6551	8239	1295	13495	.
13	.	7361	461.7648	6448	8273	1250	13471	.
14	.	7586	1005	5599	9572	1226	13945	.
15	.	6218	418.4994	5391	7045	120.3015	12316	.
16	.	6218	418.4994	5391	7045	120.3015	12316	.
17	.	6218	418.4994	5391	7045	120.3015	12316	.
18	.	19110	854.9251	17421	20800	12837	25384	.
19	.	19110	854.9251	17421	20800	12837	25384	.
20	.	16442	851.2178	14761	18124	10171	22714	.
21	.	17278	725.1280	15845	18710	11068	23487	.
22	.	11861	424.2840	11023	12699	5761	17960	.
23	.	11861	424.2840	11023	12699	5761	17960	.
24	.	5434	474.0620	4498	6371	-679.3889	11548	.
25	.	5689	536.3952	4629	6749	-445.1897	11823	.
26	.	11799	1298	9234	14364	5235	18363	.
27	.	9143	353.6704	8444	9842	3061	15225	.
28	.	11799	1298	9234	14364	5235	18363	.
29	.	5650	484.0919	4694	6607	-466.4950	11767	.
30	.	5650	484.0919	4694	6607	-466.4950	11767	.
31	.	8306	1356	5627	10985	1697	14915	.
32	.	9372	1123	7152	11591	2935	15808	.
33	.	10990	679.8378	9647	12333	4801	17179	.
34	.	14852	585.5228	13695	16008	8700	21003	.
35	.	20089	628.4541	18847	21331	13921	26257	.
36	.	11427	1013	9425	13430	5063	17792	.
37	.	11696	1008	9704	13688	5334	18057	.
38	.	16216	501.5951	15225	17207	10094	22338	.
39	.	17043	878.7472	15307	18779	10757	23329	.
40	.	20623	627.5936	19383	21863	14455	26790	.
41	.	18020	628.1006	16779	19261	11852	24188	.

The prediction intervals obtained for each observation y_i.

39/41 ~ 95.12% of the observations have an actual value y that falls within their respective P.I.

An indication of the model's good predictive power!