Time Series Analysis and Models Lab 1 Report

Edison Murairi

October 13, 2023

Problem 1

We display the size of the train and test set.

- The size of the train set is 160
- The size of the test set is 41

Problem 2

Figure 1 shows the correlation map.

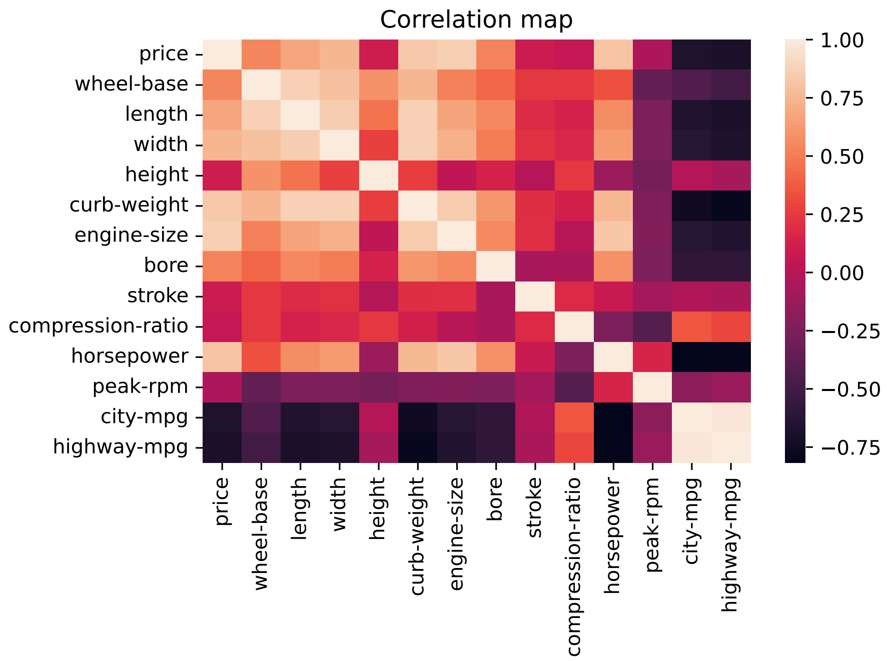

Figure 1: Correlation map

Some observations we make from this map is that some variables are strongly positively correlated with the price while other are strongly negatively correlated to the price. The variables that are strongly positively correlated to the price are the 'length', 'width', 'curb-weight', 'engine-size' and 'horsepower'. The ones that are the most strongly negatively correlated to 'price' are 'highway-mpg' and 'city-mpg'. So, we expect that these variables will be important in our model. Moreover, there are some parameters that are strongly correlated. For example, 'length' and 'city-mpg' are strongly negatively correlated. Therefore, we will try to find these variables in our feature selection.

Problem 3

(a): SVD

The eigenvalues obtained from SVD decomposition are:

$$\{7.28 \times 10^4, 6.97 \times 10^3, 4.13 \times 10^2, 2.25 \times 10^2, 1.23 \times 10^2, 6.93 \times 10^1, 3.67 \times 10^1, 2.94 \times 10^1 \\ 2.15 \times 10^1, 1.76 \times 10^1, 1.16 \times 10^1, 3.76, 2.42\} \tag{1}$$

Our observation is that we do not expect a big co-linearity issue because the smallest eigenvalue we see is $\lambda_{\mathrm{min}} = 2.42$ which is different from zero. However, since $\lambda_{\mathrm{min}} = 2.42$ is small compared to the largest $\lambda_{\mathrm{max}} = 7.28 \times 10^4$, there must be some a strong degree of collinearity, which we will confirm from the condition number.

(b): Condition Number

The condition number is

$$k = \frac{\lambda_{max}}{\lambda_{min}} = 30040.45 \tag{2}$$

With such a large $k > 100$, we conclude that we have a severe degree of collinearity in our dataset.

(c): Number of variables to remove

If we want the collinearity to be moderate at worst, we want $k < 1000$. Therefore, we want to remove all the variables with $\lambda < \frac{\lambda_{max}}{1000} = 7.28 \times 10^1$. From Eq. 1, there are 7 variables with $\lambda < 7.28 \times 10^1$. Therefore, we will remove 7 variables to make sure that correlation is moderate at worst.

Problem 4

The dataset has been standardized.

Problem 5

The regression coefficeints are

$$\begin{aligned} &\{13453.49\,, 1043.56\,, -1504.92\,, 1395.17\,, 353.68\,, 1274.03\,, 4434.77 \\ &-100.05\,, -904.36\,, 1222.88\,, 1531.53\,, 1404.04\,, -1868.49\,, 1516.00\} \end{aligned} \tag{3}$$

Problem 6

Table 1 shows the results of the linear regression model when we consider all the predictors. We see that the regression coefficients agree with what we found in the previous problem.

Table 1: Linear Regression Model with all the predictors

Dep. Variable:	y	**R-squared:**	0.841
Model:	OLS	**Adj. R-squared:**	0.827
Method:	Least Squares	**F-statistic:**	59.53
Date:	Thu, 12 Oct 2023	**Prob (F-statistic):**	1.35e-51
Time:	16:46:17	**Log-Likelihood:**	-1520.0
No. Observations:	160	**AIC:**	3068.
Df Residuals:	146	**BIC:**	3111.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	267.558	50.282	0.000	1.29e+04	1.4e+04
x1	1043.5589	722.888	1.444	0.151	-385.117	2472.235
x2	-1504.9222	797.762	-1.886	0.061	-3081.575	71.731
x3	1395.1717	659.123	2.117	0.036	92.516	2697.827
x4	353.6814	396.861	0.891	0.374	-430.653	1138.016
x5	1274.0268	1026.135	1.242	0.216	-753.970	3302.024
x6	4434.7706	723.877	6.126	0.000	3004.139	5865.402
x7	-100.0455	395.357	-0.253	0.801	-881.408	681.317
x8	-904.3627	296.608	-3.049	0.003	-1490.562	-318.163
x9	1222.8808	383.235	3.191	0.002	465.476	1980.286
x10	1531.5339	811.564	1.887	0.061	-72.397	3135.465
x11	1404.0386	398.006	3.528	0.001	617.442	2190.636
x12	-1868.4934	1463.885	-1.276	0.204	-4761.637	1024.650
x13	1516.0009	1428.070	1.062	0.290	-1306.360	4338.362

Omnibus:	17.759	**Durbin-Watson:**	2.208
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	54.136
Skew:	0.299	**Prob(JB):**	1.76e-12
Kurtosis:	5.786	**Cond. No.**	19.3

Problem 7

Problem 7 (a): Backward regression with adjusted r^2

Table 2 shows the regression final regression when we use the adjusted r^2 has the metric, and we iteratively remove the predictor with the largest P-value. The final predictors are: wheel-base, length, width, curb-weight, engine-size, stroke, compression-ratio, horsepower, peak-rpm, city-mpg and highway-mpg.

Table 2: Final regression model after iteratively removing the predictor with the highest P-value and using the adjusted r^2 as the metric.

Dep. Variable:	price	**R-squared:**	0.840
Model:	OLS	**Adj. R-squared:**	0.828
Method:	Least Squares	**F-statistic:**	70.82
Date:	Thu, 12 Oct 2023	**Prob (F-statistic):**	2.90e-53
Time:	23:13:47	**Log-Likelihood:**	-1520.5
No. Observations:	160	**AIC:**	3065.
Df Residuals:	148	**BIC:**	3102.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	266.522	50.478	0.000	1.29e+04	1.4e+04
x1	1262.7468	669.661	1.886	0.061	-60.585	2586.079
x2	-1355.4501	771.794	-1.756	0.081	-2880.609	169.708
x3	1270.4253	628.348	2.022	0.045	28.733	2512.118
x4	1353.1621	1017.583	1.330	0.186	-657.706	3364.030
x5	4373.6048	714.823	6.118	0.000	2961.028	5786.182
x6	-938.3654	285.607	-3.286	0.001	-1502.759	-373.972
x7	1210.1871	381.357	3.173	0.002	456.579	1963.796
x8	1431.6507	781.387	1.832	0.069	-112.466	2975.767
x9	1447.1458	367.673	3.936	0.000	720.580	2173.712
x10	-1733.6651	1440.553	-1.203	0.231	-4580.374	1113.044
x11	1498.1224	1419.636	1.055	0.293	-1307.252	4303.497

Omnibus:	18.158	**Durbin-Watson:**	2.200
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	54.555
Skew:	0.323	**Prob(JB):**	1.42e-12
Kurtosis:	5.787	**Cond. No.**	18.4

Problem (b): Backward regression with AIC

Table 3 shows the final regression model using the AIC as the metric. The selected predictors are wheel-base, length, width, engine-size, strok, compression-ratio, horsepower, peak-rpm and curb-weight.

Table 3: Final linear regression model after iteratively removing predictors with the highest P-Value and using AIC as the metric

Dep. Variable:	price	**R-squared:**	0.839
Model:	OLS	**Adj. R-squared:**	0.829
Method:	Least Squares	**F-statistic:**	86.72
Date:	Thu, 12 Oct 2023	**Prob (F-statistic):**	6.80e-55
Time:	23:13:47	**Log-Likelihood:**	-1521.3
No. Observations:	160	**AIC:**	3063.
Df Residuals:	150	**BIC:**	3093.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	266.032	50.571	0.000	1.29e+04	1.4e+04
x1	1076.8714	650.001	1.657	0.100	-207.468	2361.211
x2	-1106.2411	738.560	-1.498	0.136	-2565.565	353.083
x3	1277.4807	625.616	2.042	0.043	41.323	2513.639
x4	1324.4560	895.441	1.479	0.141	-444.851	3093.763
x5	4224.8997	672.820	6.279	0.000	2895.471	5554.328
x6	-911.0149	283.480	-3.214	0.002	-1471.144	-350.886
x7	1120.8269	323.648	3.463	0.001	481.329	1760.325
x8	1679.5837	720.419	2.331	0.021	256.103	3103.064
x9	1463.5389	363.094	4.031	0.000	746.100	2180.978

Omnibus:	18.403	**Durbin-Watson:**	2.206
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	56.487
Skew:	0.321	**Prob(JB):**	5.42e-13
Kurtosis:	5.839	**Cond. No.**	8.41

Problem 7 (c): Backward regression with BIC

The selected predictors are: width, engine-size, stroke, compression-ratio, peak-rpm and horsepower.

Table 4: Final linear regression model after iteratively removing predictors with the highest P-Value and using BIC as the metric

Dep. Variable:	price	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	126.3
Date:	Thu, 12 Oct 2023	**Prob (F-statistic):**	1.12e-56
Time:	23:13:47	**Log-Likelihood:**	-1524.5
No. Observations:	160	**AIC:**	3063.
Df Residuals:	153	**BIC:**	3085.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	268.846	50.042	0.000	1.29e+04	1.4e+04
x1	2013.7996	416.708	4.833	0.000	1190.555	2837.044
x2	4570.6618	647.835	7.055	0.000	3290.804	5850.519
x3	-845.6888	284.062	-2.977	0.003	-1406.879	-284.499
x4	1198.4789	319.675	3.749	0.000	566.932	1830.026
x5	1701.1253	643.013	2.646	0.009	430.795	2971.456
x6	1324.7155	361.861	3.661	0.000	609.826	2039.605

Omnibus:	14.730	**Durbin-Watson:**	2.132
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	40.738
Skew:	0.222	**Prob(JB):**	1.43e-09
Kurtosis:	5.432	**Cond. No.**	5.28

Problem 8

Problem 8 (a): VIF Method with Adjusted r^2

The predictors selected are: engine-size, stroke, compression-ratio, peak-rpm, horsepower and width.

Table 5: Final regression using the VIF method with the adjusted r^2 as the metric

Dep. Variable:	price	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	126.3
Date:	Fri, 13 Oct 2023	**Prob (F-statistic):**	1.12e-56
Time:	13:54:47	**Log-Likelihood:**	-1524.5
No. Observations:	160	**AIC:**	3063.
Df Residuals:	153	**BIC:**	3085.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	268.846	50.042	0.000	1.29e+04	1.4e+04
x1	2013.7996	416.708	4.833	0.000	1190.555	2837.044
x2	4570.6618	647.835	7.055	0.000	3290.804	5850.519
x3	-845.6888	284.062	-2.977	0.003	-1406.879	-284.499
x4	1198.4789	319.675	3.749	0.000	566.932	1830.026
x5	1324.7155	361.861	3.661	0.000	609.826	2039.605
x6	1701.1253	643.013	2.646	0.009	430.795	2971.456

Omnibus:	14.730	**Durbin-Watson:**	2.132
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	40.738
Skew:	0.222	**Prob(JB):**	1.43e-09
Kurtosis:	5.432	**Cond. No.**	5.28

Problem 8 (b): VIF method with AIC

The predictors selected are: engine-size, stroke, compression-ratio, peak-rpm, horsepower and width.

Table 6: Final regression using the VIF method with AIC as the metric

Dep. Variable:	price	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	126.3
Date:	Fri, 13 Oct 2023	**Prob (F-statistic):**	1.12e-56
Time:	13:54:47	**Log-Likelihood:**	-1524.5
No. Observations:	160	**AIC:**	3063.
Df Residuals:	153	**BIC:**	3085.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	268.846	50.042	0.000	1.29e+04	1.4e+04
x1	4570.6618	647.835	7.055	0.000	3290.804	5850.519
x2	-845.6888	284.062	-2.977	0.003	-1406.879	-284.499
x3	1198.4789	319.675	3.749	0.000	566.932	1830.026
x4	1324.7155	361.861	3.661	0.000	609.826	2039.605
x5	1701.1253	643.013	2.646	0.009	430.795	2971.456
x6	2013.7996	416.708	4.833	0.000	1190.555	2837.044

Omnibus:	14.730	**Durbin-Watson:**	2.132
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	40.738
Skew:	0.222	**Prob(JB):**	1.43e-09
Kurtosis:	5.432	**Cond. No.**	5.28

Problem 8(c): VIF methid with BIC

The predictors selected are: engine-size, stroke, compression-ratio, peak-rpm, horsepower and width.

Table 7: Final regression with VIF method using the BIC as the metric.

Dep. Variable:	price	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	126.3
Date:	Fri, 13 Oct 2023	**Prob (F-statistic):**	1.12e-56
Time:	13:54:47	**Log-Likelihood:**	-1524.5
No. Observations:	160	**AIC:**	3063.
Df Residuals:	153	**BIC:**	3085.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	268.846	50.042	0.000	1.29e+04	1.4e+04
x1	4570.6618	647.835	7.055	0.000	3290.804	5850.519
x2	-845.6888	284.062	-2.977	0.003	-1406.879	-284.499
x3	1198.4789	319.675	3.749	0.000	566.932	1830.026
x4	1324.7155	361.861	3.661	0.000	609.826	2039.605
x5	1701.1253	643.013	2.646	0.009	430.795	2971.456
x6	2013.7996	416.708	4.833	0.000	1190.555	2837.044

Omnibus:	14.730	**Durbin-Watson:**	2.132
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	40.738
Skew:	0.222	**Prob(JB):**	1.43e-09
Kurtosis:	5.432	**Cond. No.**	5.28

Problem 9

- In Problem 7, the model we select is the one that uses BIC. It has only 6 parameters while its adjusted r^2 is not drastically lower compared to the one of the AIC which uses 9 parameters.
  - The 6 parameters are: width, engine-size, stroke, compression-ratio, peak-rpm and horsepower

- In problem 8, all the models are similar. The final model has 6 parameters.
  - The 6 parameters are: width, engine-size, stroke, compression-ratio, peak-rpm and horsepower
- The parameters in the best models of Problem 7 and Problem 8 are all identical.

Problem 10

The best model in step (7) and (8) are identical, and we pick them as our final model.

Table 8: Final regression with VIF method using the BIC as the metric.

Dep. Variable:	price	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	126.3
Date:	Fri, 13 Oct 2023	**Prob (F-statistic):**	1.12e-56
Time:	13:54:47	**Log-Likelihood:**	-1524.5
No. Observations:	160	**AIC:**	3063.
Df Residuals:	153	**BIC:**	3085.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1.345e+04	268.846	50.042	0.000	1.29e+04	1.4e+04
x1	4570.6618	647.835	7.055	0.000	3290.804	5850.519
x2	-845.6888	284.062	-2.977	0.003	-1406.879	-284.499
x3	1198.4789	319.675	3.749	0.000	566.932	1830.026
x4	1324.7155	361.861	3.661	0.000	609.826	2039.605
x5	1701.1253	643.013	2.646	0.009	430.795	2971.456
x6	2013.7996	416.708	4.833	0.000	1190.555	2837.044

Omnibus:	14.730	**Durbin-Watson:**	2.132
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	40.738
Skew:	0.222	**Prob(JB):**	1.43e-09
Kurtosis:	5.432	**Cond. No.**	5.28

Problem 11

Figure 2 shows the plot.

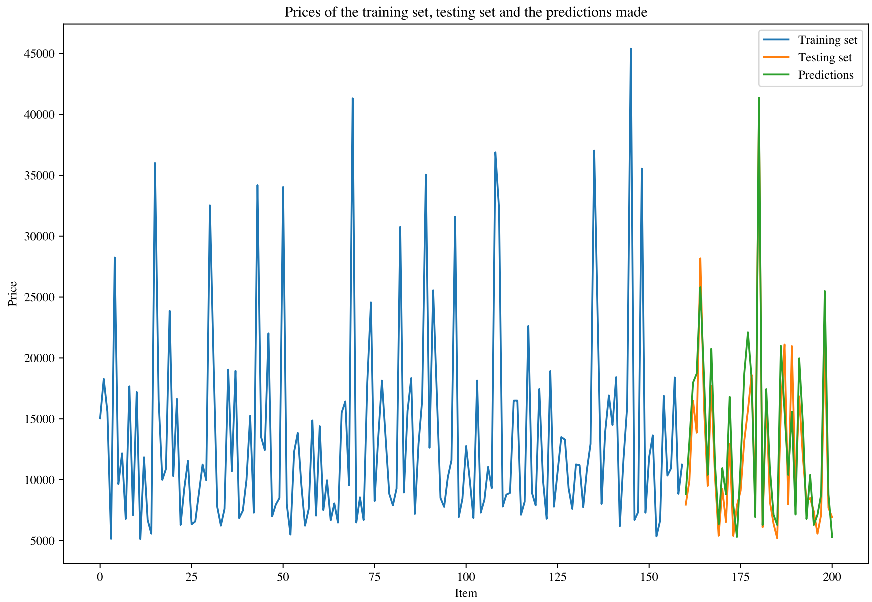

Figure 2: Plot of prices for the training dataset, the test set and the predictions overlayed with the testing set.

Problem 12

Figure 3 shows the ACF plot of the prediction errors. This ACF plot is consistent with a white noise. We see that for $\tau \neq 0$, the ACF drops close to 0. Therefore, we expect that our model captures well the dynamics in the dataset.

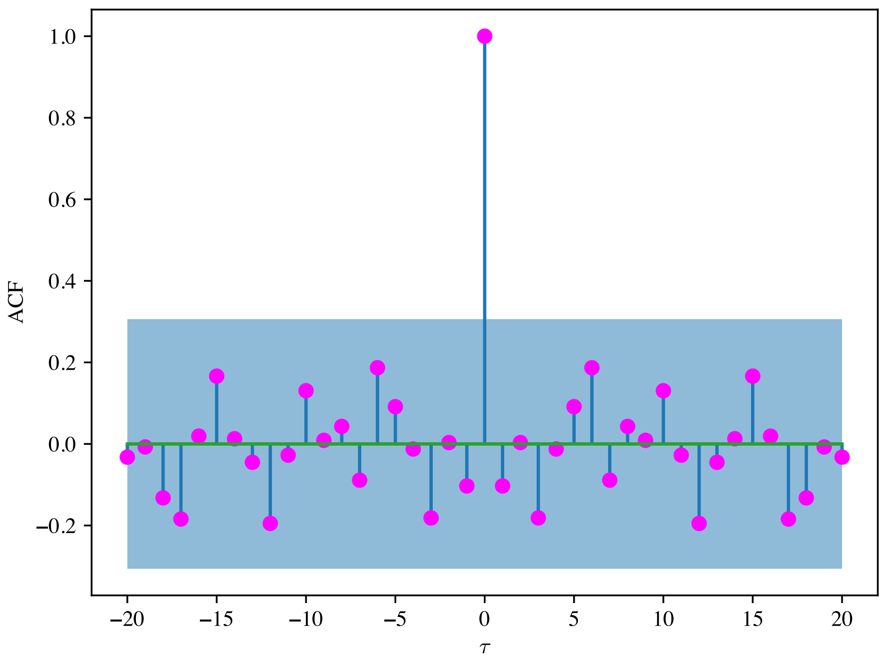

Figure 3: ACF plot of the prediction error

Problem 13

T Test

Null Hypothesis: There is no difference between the mean of the test set price and the mean of the price predictions. For the T-Test between the test set and the predictions, we obtain

- T-Test = -0.7525
- P-Value = 0.45395

We use the significance level $\alpha = 0.05$. With P-value > 0.05, we fail to reject the null hypothesis. Therefore, we conclude that there is no difference, suggesting that the model performed well.

F-Test

Null Hypothesis: There is no difference between our model and the intercept-only model. For the F-Test between the test set and the predictions, we obtain

- F-Test = 466.037
- P-Value = 9.65×10^{-100}

With P-value < 0.05, we conclude that there is a difference between the intercept-only model and our model. Therefore, the predictors we included are useful in explaining the variations in the prices. Therefore, we conclude again that our model performs well.