TECHNICAL UNIVERSITY OF MUNICH

CHAIR OF TRANSPORTATION SYSTEMS ENGINEERING

STATISTICAL LEARNING AND DATA ANALYTICS FOR TRANSPORTATION SYSTEMS

PROBLEM SET 1

Author

NATALIA ALLMI

Matriculation Number

03755024

Problem 1

Suppose you have fit a multiple linear regression model and the $(X'X)^{-1}$ matrix is:

$$(X'X)^{-1} = \begin{bmatrix} 0.893758 & -0.0282448 & -0.0175641 \\ -0.0282448 & 0.0013329 & 0.0001547 \\ -0.0175641 & 0.0001547 & 0.0009108 \end{bmatrix}$$

1.1 : How many regressor variables are in this model?

In the model presented, there are 2 regressor variables. Regressor variables refer to the independent variables. In order to achieve a 3x3 matrix, X must have 3 columns and k rows.

$$X'.X = X'X$$
$$(3xk).(kx3) = (3x3)$$
$$\rightarrow (X'X)^{-1}$$
$$(3x3)$$

The first column of X is filled with 1's, because they correspond to $\beta 0$. The second and third columns have the values x_i of the independent variables used to predict the outcome. Therefore, the model has two regressors and 3 parameters.

1.2: If the error sum of squares is 307 and there are 15 observations, what is the estimate of σ^2?

$$\sigma^2 = \frac{SSE}{n-p}$$

Where:

σ^2 = unbiased estimator

SSE = Square sum of errors

n = number of observations

p = number of parameters

In our case:

SSE = 307 (from problem description)

n = 15 (from problem description)

p = 3 (from 1.1)

$$\sigma^2 = \frac{307}{15-3}$$
$$\sigma^2 = 25.583$$

1.3: What is the standard error of the regression coefficient $\hat{\beta}_1$?

$$Se(\hat{\beta}i) = \sqrt{\sigma^2 * Cii}$$

Where:

$Se(\hat{\beta}i)$ = standard error of the regression coefficient

σ^2 = unbiased estimator

C_{ii} = matrix coefficient corresponding to $\hat{\beta}i$

In our case:

C_{11} = 0.0013329 (from problem description)

σ^2 = 25.583 (from 1.2)

$$Se(\hat{\beta}1) = \sqrt{\sigma^2 * C11}$$
$$Se(\hat{\beta}1) = \sqrt{25.583 * 0.0013329}$$
$$Se(\hat{\beta}1) = \sqrt{0.03409958}$$
$$Se(\hat{\beta}1) = 0.18467$$

Problem 2

Consider the linear regression model: $y_i = \beta_0 + \beta_1(x_{i1} - \bar{x_1}) + \beta_2(x_{i2} - \bar{x_2}) + \varepsilon_i$ where $\hat{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{i1}$ and $\hat{x_2} = \frac{1}{n}\sum_{i=1}^{n} x_{i2}$ and n is the number of observations.

2.1: Write out the least squares normal equations for this model.

The model is:

$$y_i = \beta_0 + \beta_1 * (x_{i1} - \bar{x_1}) + \beta_2 * (x_{i2} - \bar{x_2}) + \varepsilon i$$

To obtain the normal equations we need to minimize the function L. Therefore we calculate the derivatives respect to each β and equate it to 0.

$$L = \sum_{i=1}^{n} {\varepsilon_i}^2$$

$\frac{\theta L}{\theta \beta_0} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n}(x_{i1} - \bar{x_1}) + \hat{\beta}_2 \sum_{i=1}^{n}(x_{i2} - \bar{x_2})$

$\frac{\theta L}{\theta \beta_1} = 0 \quad \Rightarrow$

$\sum_{i=1}^{n} y_i(x_{i1} - \bar{x_1}) = \hat{\beta}_0 \sum_{i=1}^{n}(x_{i1} - \bar{x_1}) + \hat{\beta}_1 \sum_{i=1}^{n}(x_{i1} - \bar{x_1})^2 + \hat{\beta}_2 \sum_{i=1}^{n}(x_{i2} - \bar{x_2})(x_{i1} - \bar{x_1})$

$\frac{\theta L}{\theta \beta_2} = 0 \quad \Rightarrow$

$\sum_{i=1}^{n} y_i(x_{i2} - \bar{x_2}) = \hat{\beta}_0 \sum_{i=1}^{n}(x_{i2} - \bar{x_2}) + \hat{\beta}_1 \sum_{i=1}^{n}(x_{i1} - \bar{x_1})(x_{i2} - \bar{x_2}) + \hat{\beta}_2 \sum_{i=1}^{n}(x_{i2} - \bar{x_2})^2$

2.2: Verify that the least squares estimate of the intercept in this model is $\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

$$\text{Model} : y_i = \beta_0 + \beta_1 * (x_{i1} - \bar{x_1}) + \beta_2 * (x_{i2} - \bar{x_2}) + \varepsilon i$$

To obtain the least square normal equation for β_0 we have to minimize the square errors. Therefore, we want to minimize:

$$L = \sum_{i=1}^{n} \varepsilon_i^2$$

For the β_0 least square normal equation:

$$\frac{\theta L}{\theta \beta_0} = 0$$
$$\frac{\theta L}{\theta \beta_0} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1(x_{i1} - \bar{x_1}) - \hat{\beta}_2(x_{i2} - \bar{x_2}))^2 = 0 \qquad Eq.(1)$$

Using the chain rule:

$$\sum_{i=1}^{n} 2 * (-1) * (y_i - \hat{\beta}_0 - \hat{\beta}_1(x_{i1} - \bar{x_1}) - \hat{\beta}_2(x_{i2} - \bar{x_2})) = 0$$
$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 n\frac{1}{n}\sum_{i=1}^{n} x_{i1} - \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_2 n\frac{1}{n}\sum_{i=1}^{n} x_{i2} = 0$$

The terms with $\hat{\beta}_1$ and $\hat{\beta}_2$ cancel each other out. Therefore:

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 = 0$$
$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

2.3: Suppose that we use $y_i - \hat{y}$ as the response variable in this model. What effect will this have on the least squares estimate of the intercept?

If we replace y_i in Eq. (1) with $y_i - \bar{y}$:

$$\frac{\theta L}{\theta \hat{\beta}_0} = \sum_{i=1}^{n}((y_i - \bar{y}) - \hat{\beta}_0 - \hat{\beta}_1(x_{i1} - \bar{x_1}) - \hat{\beta}_2(x_{i2} - \bar{x_2}))^2 = 0$$

And therefore, we would get:

$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} - n\hat{\beta}_0 = 0$

$\sum_{i=1}^{n} y_i - n\frac{1}{n}\sum_{i=1}^{n} y_i - n\hat{\beta}_0 = 0$

Resulting in:

$\hat{\beta}_0 = 0 \quad \Rightarrow \quad$ The intercept term becomes 0.

2.4: Assuming x_1 and x_2 are positively correlated to each other, please discuss the influence of multicollinearity on the model.

If x1 and x2 are positively correlated, it means that a positive change in x1 would signify a positive change in x2. The more correlated these variables are the harder it becomes to estimate the relationship between each independent variable and the dependent variable, because they both change at the same time. Therefore, this generates problems to interpret the model and it's coefficients.

Additionally, in the most extreme case, if there is complete correlation between x_1 and x_2 then there is lineal dependency, and the matrix X is not invertible. This means we can't estimate the parameters. In other un-extreme cases, numerically, it may still result hard to estimate parameters computationally. Also, it may be hard to evaluate the coefficients and determine their significance.

Problem 3

The data which can be found in a separate spreadsheet provides the highway gasoline mileage test results for 2005 model year vehicles from DaimlerChrysler. The variable description is written in the spreadsheet.

3.1: Fit a multiple linear regression model to these data to estimate gasoline mileage that uses the following regressions: cid, rhp, etw, cmp, axle, n/v

Using the data provided, I fit a linear regression using 'cid','rhp','etw','cmp','axle','n/v' as regressors, and 'mpg' as the target variable. Because the variables have very different ranges, I standardize the data before fitting, to measure variable importance by comparing the values of the betas.

variable	Value
intercept	29.547619
cid	-1.112567
rhp	-0.121529
etw	-2.388151
cmp	0.111749
axle	-1.892741
n/v	1.079257

3.2: Estimate σ^2 and the standard errors of the regression coefficients.

$$\hat{\sigma}^2 = \frac{SS_E}{n-p}$$

$$se(\hat{\beta}_i) = \sqrt{(\hat{\sigma}^2 * C_{ii})}$$

$$C = (X'X)^{-1}$$

Result: $\sigma^2 = 4.965$

Standard errors of regression coefficients:

variable	β_hat	se
intercept	29.547619	0.486254
cid	-1.112567	2.489655
rhp	-0.121529	1.645582
etw	-2.388151	0.697992
cmp	0.111749	0.674378
axle	-1.892741	0.652282
n/v	1.079257	1.552950

(See notebook for calculations.)

3.3: Test for significance of regression using significance level $\alpha = 0.05$. What conclusions can you draw?

The hypothesis to test the significance of the regression as a whole are:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$H_1 : \beta_i \neq 0$ for at least one i

We reject H_0 when $F_0 \geq f_{\alpha,k,n-p}$

Wih $\alpha = 0.05$, k = 6, and n-p =14 From the table (http://socr.ucla.edu/Applets.dir/F_Table.html) we find that $F = 2.8477$

$F_0 = \frac{SS_R/k}{SS_E/n-p}$

$F_0 = 19.532$

Because $F_0 \geq F$, this means that we can reject the null hypothesis that all of the $\beta = 0$, and we can conclude that our model is significant.

(See notebook for calculations.)

3.4: Find the t-test statistic for each regressor. Using $\alpha = 0.05$, what conclusions can you draw? Does each regressor contribute to the model?

We perform hypothesis tests on the estimated coefficients by constructing our hypotheses:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

The test statistic is:

$T_0 = \frac{\hat{\beta}_i - 0}{se(\hat{\beta}_i)}$

$T = 2.1448$

T-statistics:

variable	T statistic
intercept	60.765831
cid	-0.446876
rhp	-0.073852
etw	-3.421457
cmp	0.165706
axle	-2.901724
n/v	0.694972

If we compare the value obtained for t with the absolute values of each T_0, we can see that for the regressors etw and axle, we can reject the null hypothesis, and therefore they are significant. The rest of the regressors (cid, rhp,cmp, and n/v) are not significant and therefore do not contribute to the model.

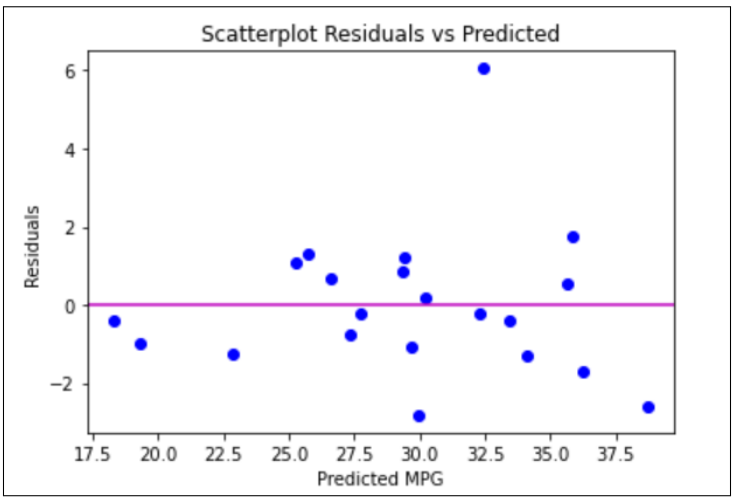
3.5: Find 99% confidence intervals on the regression coefficients.

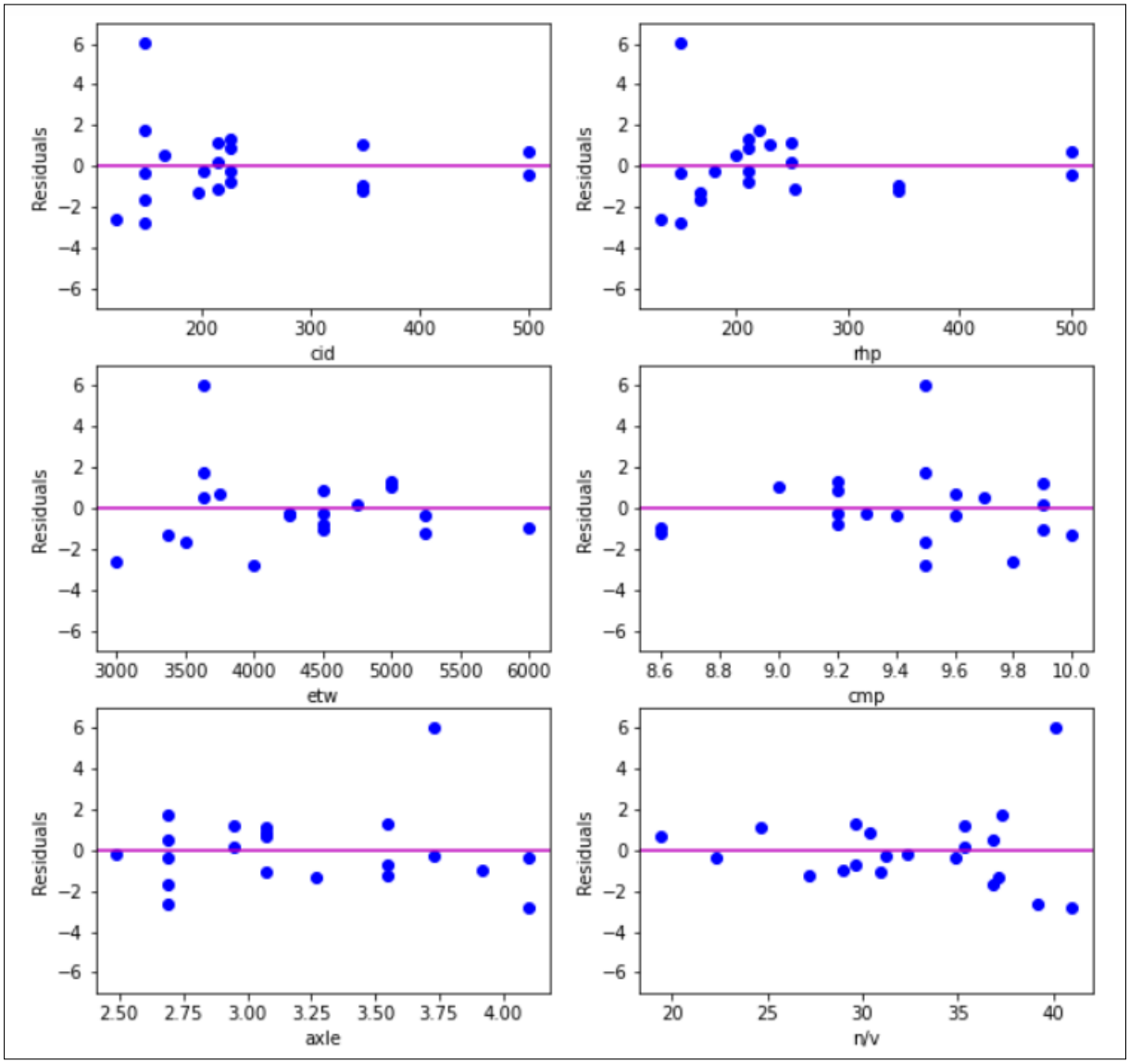
Formulas for confidence intervals on individual coefficients: $\hat{\beta}_i \pm t_{\alpha/2,n-p} * \sqrt{(\hat{\sigma}^2 C_{ii})}$

coefficient	β_hat	Lower_CI	Upper_CI
β0	29.547619	28.100118	30.995120
β1	-1.112567	-8.523880	6.298745
β2	-0.121529	-5.020170	4.777111
β3	-2.388151	-4.465964	-0.310337
β4	0.111749	-1.895769	2.119266
β5	-1.892741	-3.834480	0.048999
β6	1.079257	-3.543631	5.702146

3.6: Plot residuals versus $\hat{y}$ and versus each regressor. Discuss these residual plots.

I create the plots with the original data in order to facilitate the interpretation according to the regressor variable.





The linear regression model assumes that errors are independent and normally distributed, therefore, I analyze the residual plots to see if they satisfy this assumption. I do not see any obvious patterns for the value of the residuals. Most points seem to be concentrated close to the

origin, and the plot seems to be symmetric respect to the origin. We can say that the assumption is valid.

The data only has 21 samples, which is why it is hard to observe from the residual plots if there could be over- or under-fitting. However, keeping in mind that $e = \hat{y} - y$, for the plot residuals vs rhp, we could say that for $rhp < 200$ the model tends to underpredicts mpg. Also, for $etw > 4000$ the and for $axle > 3.2$, the model underpredicts mpg.

Additionally, there seems to be one particular data point for which the residual is much higher than the others. It may be worth to look into when building the model.

3.7: Using $\alpha = 0.05$, plot the confidence limits and prediction limits for the mean response.

I calculate confidence and prediction limits for the mean response at each point x_i. Because this is a multiple linear regression, it is not possible to plot these values, given the number regressors.

Formula for confidence limits of the mean response: $\hat{\mu}_{Y|x_0} \pm t_{\alpha/2,n-p} * \sqrt{(\hat{\sigma}^2 x_0'(X'X)^{-1}x_0)}$

Formula for prediction limits of the mean response: $\hat{\mu}_{Y|x_0} \pm t_{\alpha/2,n-p} * \sqrt{(\hat{\sigma}^2(1 + x_0'(X'X)^{-1}x_0))}$

Xi	y_hat	Lower_CL	Upper_CL	Lower_PL	Upper_PL
0	29.710411	26.718230	32.702593	24.071788	35.349034
1	32.270833	29.930287	34.611378	26.949266	37.592400
2	34.086592	31.249815	36.923369	28.528875	39.644310
3	27.345376	24.843659	29.847092	21.950980	32.739771
4	25.727175	23.213599	28.240751	20.327269	31.127081
5	22.850158	19.967737	25.732580	17.269005	28.431312
6	29.347718	27.705963	30.989473	24.294375	34.401062
7	25.291019	21.405799	29.176239	19.131810	31.450228
8	29.980694	26.924333	33.037055	24.307752	35.653635
9	27.764045	25.429521	30.098570	22.445124	33.082967
10	38.692720	36.407772	40.977667	33.395371	43.990069
11	30.203493	27.624988	32.781998	24.773059	35.633927
12	29.394393	26.431417	32.357369	23.771213	35.017572
13	35.844187	33.037350	38.651024	30.301692	41.386682
14	18.317895	14.927666	21.708123	12.458324	24.177465
15	19.337849	15.979368	22.696331	13.496589	25.179109
16	35.654296	33.811518	37.497075	30.532114	40.776479
17	36.217707	34.273420	38.161994	31.058136	41.377278
18	33.421188	31.607137	35.235239	28.309270	38.533105
19	26.592161	22.624325	30.559996	20.380507	32.803814
20	32.450091	29.914281	34.985900	27.039800	37.860382

Problem 4

The data which can be found in a separate spreadsheet provides the traffic data of a two-lane segment of the SR241-N freeway in California. The data were collected every five minutes. Each observation consists of the flow, average occupancy, and average speed on the segment of interest, the flow ratio between two lanes, and the speed information in the downstream of the segment. The variable description is written in the spreadsheet. This problem asks to predict the average speed of the segment (i.e., variable speed) by using the rest variables and their transformation (e.g., x^2).

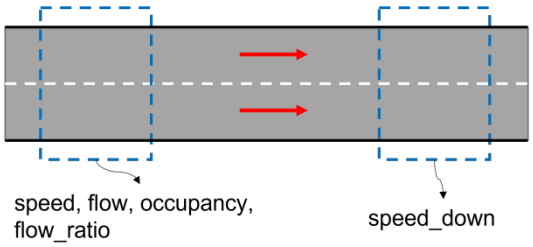


Figure 1: Conceptual illustration of the source of different variables.

a: Visualize the relationship between every two variables.

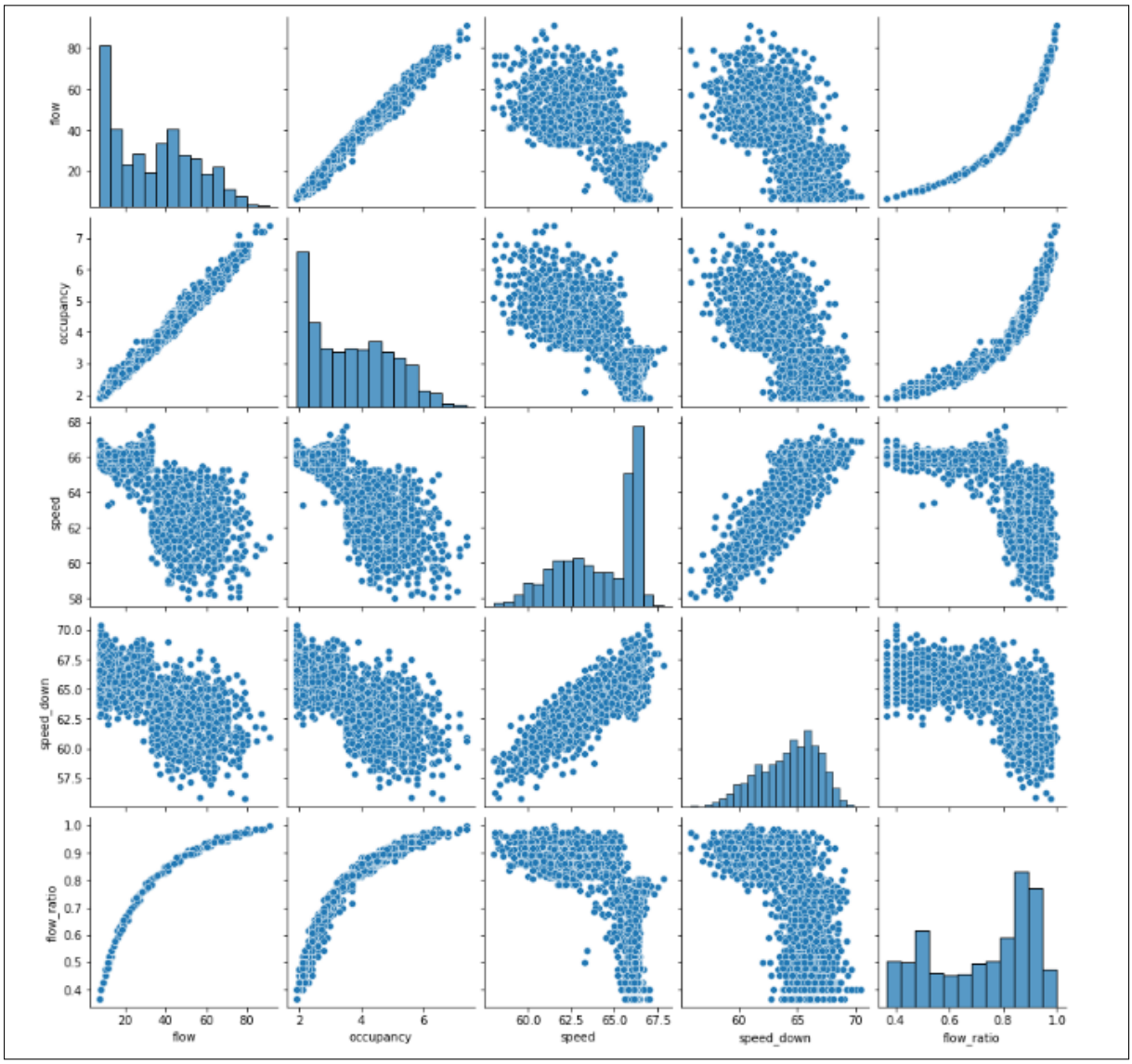


Figure 1: Pair Plots

The pair plots show high correlation between flow and occupancy. They also show an approximate linear relationship between flow, occupancy, speed_down and the target variable speed. There also seems to be a quadratic relation between flow and flow_ratio as well as occupancy and flow_ratio.

b: Fit a multiple linear regression model to the original dataset. Does the model suffer a multicollinearity problem? If yes, which variables cause this issue?

Given the results (see Figure 2), the model evidently suffers from collinearity.

Variable	VIF	Rj
flow	104.237266	0.990407
occupancy	94.881279	0.989461
speed_down	2.096501	0.523015
flow_ratio	9.116577	0.890310

Figure 2: Multicollinearity Evaluation

Because Rj is very high for variables flow and occupancy, this means that these variables are very well explained by the other variables in the model. This is why VIF for both variables is high, indicating multicollinearity (I consider 10 to be the limit value for the VIFs). This makes sense because we can see from problem part a) (see Figure 1) that flow and occupancy are highly correlated, which is why it would be prudent to eliminate one of the variables from the model.

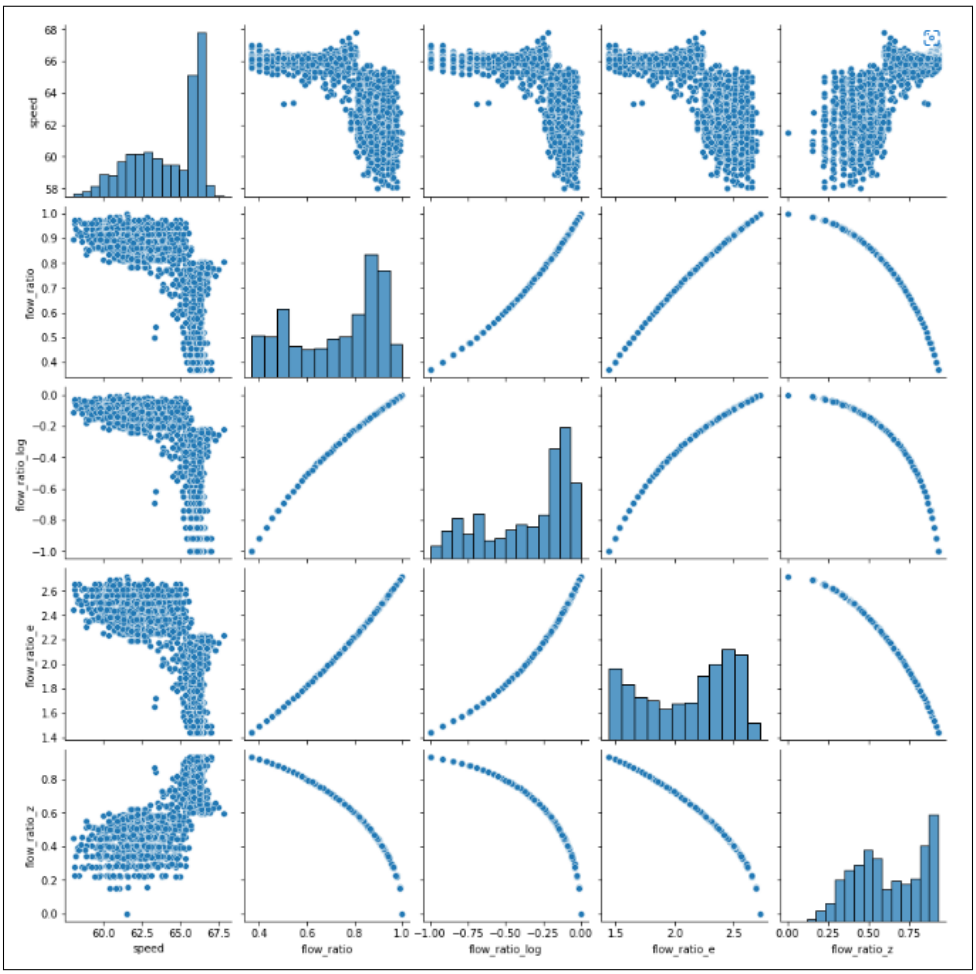
c: Select suitable regressors for the regression model. Note, you may need to do some transformations on the original variables. Explain why you do these transformations, and why you choose these variables?

For this problem, I stardardize all the data because the variables have different ranges. The standardization is done using StandardScaler, and the fit is done only with the training data because in real life we wouldn't have the test data. Additionally, I use train test split to to eliminate the temporal dependancy. We want to predict y independently of 'time'.

The first thing is to drop either occupancy or flow because we have seen that they are highly correlated, and this can introduce problems to the model. In order to decide which one to keep, I run a quick regression using statsmodel OLS and check the R^2 for each. Because the R^2 for flow is 0.611 and for occupancy is 0.651 I decide to keep occupancy.

From the pair plots, we saw that speed_down, occupancy, and flow have an approximate linear relationship with speed. However, flow_ratio does not have a linear relationship with speed, so we try some transformations on this variable, so as to fullfil the assumption that the

independent variables have a linear relation with the dependent variable. Creating scatterplots for each new variable allows us to see its relationship with speed.



$\sqrt{1-y^2}$ (flow_ratio_z) turns out to be the best transformation. I run a quick regression for flow_ratio and flow_ratio_z to compare the R^2 and I verify that flow_ratio_z has a higher R^2, which is why I replace the original with the transformation.

Once this is done I create three new variables, each variable squared, to see if the addition of these variables helps the regression. Using the Extra Sum of Squares Method we determine if at least one of the new variables contributes significantly to the model.

The null and alternative hypotheses are:

$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$

$H_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$ or $\beta_6 \neq 0$

T0 test this we need the extra Sum of Squares due to β_4, β_5, and β_6.

$$SS_R(\beta_6, \beta_5, \beta_4 | \beta_3, \beta_2, \beta_1, \beta_0) = SS_R(\beta_6, \beta_5, \beta_4, \beta_3, \beta_2, \beta_1, \beta 0) - SS_R(\beta_3, \beta_2, \beta_1, \beta 0)$$

$$F_0 = \frac{SS_R(\beta_6, \beta_5, \beta_4 | \beta_3, \beta_2, \beta_1, \beta_0)/2}{MS_E}$$

β_1 is the coefficient for speed_down,β_2 is the coefficient for occupancy,β_3 is the coefficient for flow_ratio_z,β_4 is the coefficient for speed_down squared,β_5 is the coefficient for occupancy squared, β_6 is the coefficient for flow_ratio_z squared.

We compare F_0 to $f_{\alpha,r,n-p}$

$\alpha = 0.05$

$n - p = 1452 - 7 = 1445$

$r = 3$

From the table:

$f_{0.05,3,120} = 2.6802$

$f_{0.05,3,\infty} = 2.6049$

And from the calculations, $F_0 = 131.95$

Which is why we conclude that at least one of the new variables contributes significantly to the model, and we reject the null hypothesis.

In order to explore the model further to determine which of the variables helps the model, I use stepwise regression. For the first iteration the result is:

	predictor	r-squared
0	speed_down_std	0.748251
5	speed_down_2_std	0.743198
1	occupancy_std	0.651375
2	flow_ratio_z_std	0.614845
4	occupancy_2_std	0.608199

So the first regressor we select is speed_down_std.

For the second iteration the result is:

	predictor	r-squared
0	occupancy_std	0.829823
1	flow_ratio_z_std	0.822298
3	occupancy_2_std	0.820396
2	flow_ratio_z_2_std	0.820373
4	speed_down_2_std	0.761106

So the second regressor we select is occupancy_std.

For the third iteration the result is:

	predictor	r-squared
3	speed_down_2_std	0.845274
2	occupancy_2_std	0.832217
0	flow_ratio_z_std	0.830480
1	flow_ratio_z_2_std	0.829965

So the third regressor we select is speed_down_2_std.

For the fourth iteration the result is:

	predictor	r-squared
2	occupancy_2_std	0.850752
1	flow_ratio_z_2_std	0.846461
0	flow_ratio_z_std	0.845276

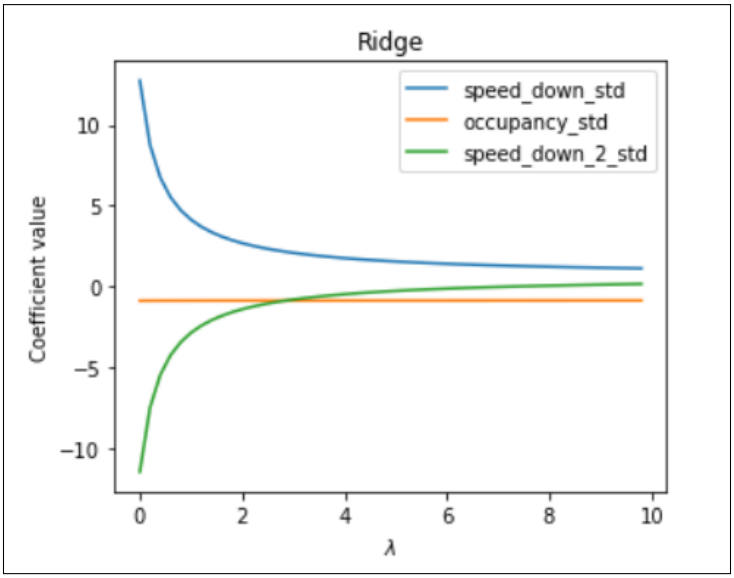
We can see in this case that the improvement in R^2 is minor, which is why I decide to keep the first three selected regressors. I calculate the model again to check that the P-value for each regressor is less than 0.05 (this checks out).

	coef	std err	t	P>|t|	[0.025	0.975]
const	64.1410	0.023	2843.253	0.000	64.097	64.185
speed_down_std	12.7404	0.953	13.365	0.000	10.870	14.610
occupancy_std	-0.8823	0.031	-28.066	0.000	-0.944	-0.821
speed_down_2_std	-11.4698	0.954	-12.025	0.000	-13.341	-9.599

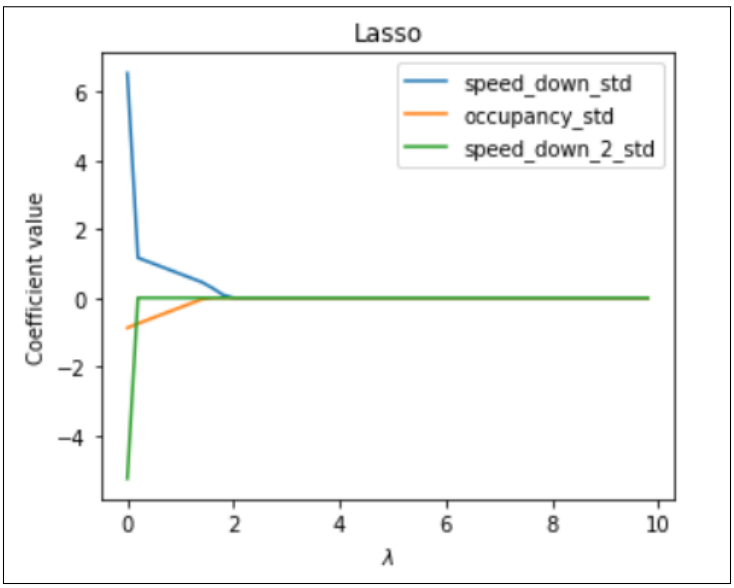
d: Apply linear regression, ridge regression, and lasso regression to estimate the relationship between the average speed of the segment and the regressors you selected. Compare the results of different models and verify the variable you selected

With the selected regressors, I run a simple linear regression, a Lasso regression. and a Ridge regression. For all of the models, $R^2 = 0.84$ is the highest they reach.

From the Ridge plot (regularization vs. coefficient value) we can see that in the beginning, speed_down_std and speed_down_2_std have the highest coefficient values, while occupancy_std remains very small. However, as the regularization parameter increases, speed_down_2_std becomes 0, and occupancy_std remains small, but larger.



We can see from the Lasso plot (regularization vs. coefficient value) that occupancy_std and speed_down_std make the most contribution to the model, since the value of speed_down_2_std becomes 0 very fast as lambda increases.



This plots verify that speed_down_2_std makes a contribution to the model, but that contribution is small (as we can see from the stepwise regression). Because the R^2 value doesn't change significantly when we add this regressor, it may be best to have a model with just occupancy_std and speed_down_std because it reduces the model complexity, while not diminishing the performance greatly.