Report

Submitted By: Ketan Walia

Topic: Simple Linear Models

Data Description:

The data file "HW2Pb1Data.txt" has some sample data of 90 instances with two categorical variables, Discipline and Method. The dependent variable to predict in a model is called Result. There is also a variable called Result_with_noise where random noise has been added to Result by generating a random number between -0.1 and +0.1.

STEPS:

a) We will create a table of the actual result, the result with noise, and the fitted result with the lower and upper confidence interval for all 90 samples:

R Output:

S.NO	Result	Result_with_noise	fit	lwr	upr
1	22	22.04416841	22.00910266	21.99178546	22.02641985
2	10	10.02157246	10.0174139	10.00009671	10.03473109
3	10	9.956093715	9.990902303	9.973585109	10.0082195
4	21	21.03480913	20.99040808	20.97309088	21.00772527
5	11	11.02888051	11.00874683	10.99142964	11.02606403
6	10	10.01787953	10.01986895	10.00255175	10.03718614
7	16	15.97686327	15.99851609	15.9811989	16.01583329
8	10	10.02340154	10.0050061	9.987688904	10.02232329
9	7	7.024088694	7.001248156	6.983930961	7.01856535
10	22	22.01553019	22.00910266	21.99178546	22.02641985
11	10	10.02732564	10.0174139	10.00009671	10.03473109
12	10	9.969931737	9.990902303	9.973585109	10.0082195
13	21	21.03240134	20.99040808	20.97309088	21.00772527
14	11	11.02914484	11.00874683	10.99142964	11.02606403
15	10	10.01852897	10.01986895	10.00255175	10.03718614
16	16	15.97581785	15.99851609	15.9811989	16.01583329
17	10	10.00398572	10.0050061	9.987688904	10.02232329
18	7	7.044812247	7.001248156	6.983930961	7.01856535
19	22	22.00133338	22.00910266	21.99178546	22.02641985
20	10	10.04554169	10.0174139	10.00009671	10.03473109
21	10	9.964840277	9.990902303	9.973585109	10.0082195
22	21	20.9692505	20.99040808	20.97309088	21.00772527

23	11	11.02917373	11.00874683	10.99142964	11.02606403
24	10	10.03103316	10.01986895	10.00255175	10.03718614
25	16	15.96461028	15.99851609	15.9811989	16.01583329
26	10	9.987445031	10.0050061	9.987688904	10.02232329
27	7	6.952889901	7.001248156	6.983930961	7.01856535
28	22	21.95050601	22.00910266	21.99178546	22.02641985
29	10	10.00916992	10.0174139	10.00009671	10.03473109
30	10	10.01008757	9.990902303	9.973585109	10.0082195
31	21	20.98239513	20.99040808	20.97309088	21.00772527
32	11	11.00135723	11.00874683	10.99142964	11.02606403
33	10	10.04375687	10.01986895	10.00255175	10.03718614
34	16	16.00979311	15.99851609	15.9811989	16.01583329
35	10	10.03462346	10.0050061	9.987688904	10.02232329
36	7	6.996586225	7.001248156	6.983930961	7.01856535
37	22	22.03460752	22.00910266	21.99178546	22.02641985
38	10	9.981633999	10.0174139	10.00009671	10.03473109
39	10	9.97731486	9.990902303	9.973585109	10.0082195
40	21	20.95005846	20.99040808	20.97309088	21.00772527
41	11	10.98307491	11.00874683	10.99142964	11.02606403
42	10	9.999199226	10.01986895	10.00255175	10.03718614
43	16	16.01307059	15.99851609	15.9811989	16.01583329
44	10	10.03983966	10.0050061	9.987688904	10.02232329
45	7	6.997639494	7.001248156	6.983930961	7.01856535
46	22	21.95388647	22.00910266	21.99178546	22.02641985
47	10	9.994905597	10.0174139	10.00009671	10.03473109
48	10	9.993308789	9.990902303	9.973585109	10.0082195
49	21	20.96167042	20.99040808	20.97309088	21.00772527
50	11	11.03018306	11.00874683	10.99142964	11.02606403
51	10	10.02527435	10.01986895	10.00255175	10.03718614
52	16	16.00344517	15.99851609	15.9811989	16.01583329
53	10	9.997096101	10.0050061	9.987688904	10.02232329
54	7	7.001595732	7.001248156	6.983930961	7.01856535
55	22	22.00852616	22.00910266	21.99178546	22.02641985
56	10	10.01067358	10.0174139	10.00009671	10.03473109
57	10	9.986824897	9.990902303	9.973585109	10.0082195
58	21	20.9805464	20.99040808	20.97309088	21.00772527
59	11	10.9590577	11.00874683	10.99142964	11.02606403
60	10	9.994660232	10.01986895	10.00255175	10.03718614
61	16	15.9882893	15.99851609	15.9811989	16.01583329

62	10	9.977891967	10.0050061	9.987688904	10.02232329
63	7	7.00434842	7.001248156	6.983930961	7.01856535
64	22	22.00180782	22.00910266	21.99178546	22.02641985
65	10	10.01692451	10.0174139	10.00009671	10.03473109
66	10	9.982141692	9.990902303	9.973585109	10.0082195
67	21	20.98174372	20.99040808	20.97309088	21.00772527
68	11	10.97232799	11.00874683	10.99142964	11.02606403
69	10	9.998933037	10.01986895	10.00255175	10.03718614
70	16	16.0153859	15.99851609	15.9811989	16.01583329
71	10	10.04388197	10.0050061	9.987688904	10.02232329
72	7	6.985396086	7.001248156	6.983930961	7.01856535
73	22	22.04616642	22.00910266	21.99178546	22.02641985
74	10	10.03426609	10.0174139	10.00009671	10.03473109
75	10	10.01905209	9.990902303	9.973585109	10.0082195
76	21	21.04165498	20.99040808	20.97309088	21.00772527
77	11	11.02708952	11.00874683	10.99142964	11.02606403
78	10	10.03139592	10.01986895	10.00255175	10.03718614
79	16	16.01997899	15.99851609	15.9811989	16.01583329
80	10	9.950211334	10.0050061	9.987688904	10.02232329
81	7	7.046122961	7.001248156	6.983930961	7.01856535
82	22	22.0344942	22.00910266	21.99178546	22.02641985
83	10	10.03212552	10.0174139	10.00009671	10.03473109
84	10	10.0494274	9.990902303	9.973585109	10.0082195
85	21	20.9695507	20.99040808	20.97309088	21.00772527
86	11	11.02717883	11.00874683	10.99142964	11.02606403
87	10	10.03802816	10.01986895	10.00255175	10.03718614
88	16	16.01790647	15.99851609	15.9811989	16.01583329
89	10	9.991684198	10.0050061	9.987688904	10.02232329
90	7	6.959001796	7.001248156	6.983930961	7.01856535

b) We will find out how many of the 90 actual results are within the 95% confidence interval

There are 70 actual results that are within the 95% confidence interval.

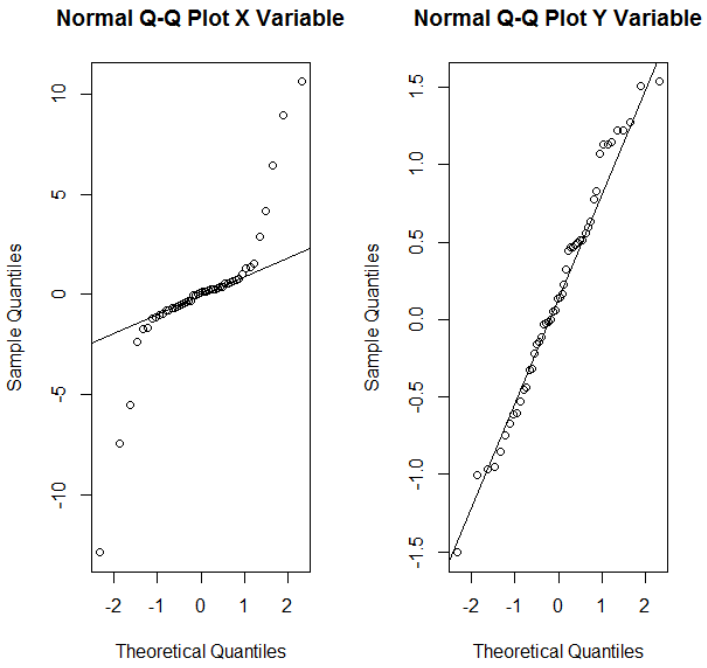
Topic 2: SIMPLE STATISTICS

Data: This file has two variables, X and Y.

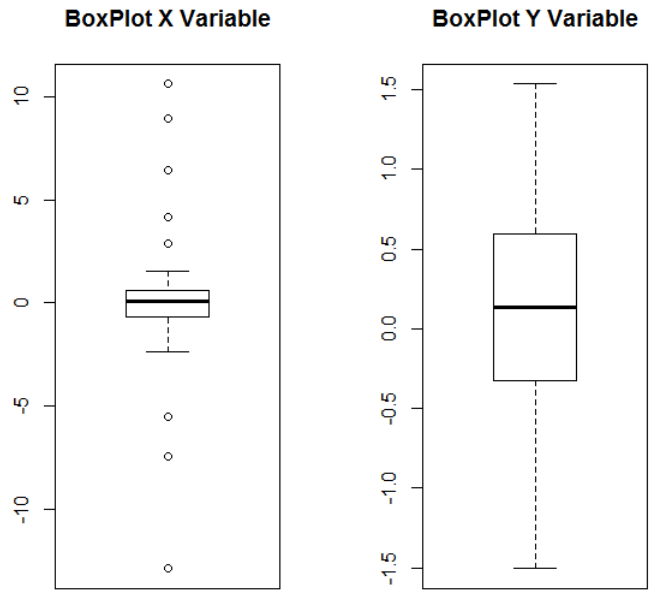
2.1 For both X and Y, we will plot the QQ-plot using the functions qqnorm and qqline and then Interpret our results.

We can see the QQ-plot for both the variables below. As per the QQ plot for “X” variable, it seems to deviate from normal distribution. It is rather having heavy tailed distribution. We can see sharp upward and downward curve at both the extremes of the distribution indicating that the tails of this distribution are too heavy for it to be considered normal.

As per the QQ plot for “Y” variable, it seems to be pretty much normally distributed. We can see the fitted line very close to the data points. The distribution of variable Y very closely mimicks the theoretical normal distribution.



2.2 For both X and Y, we will draw the boxplot using the function boxplot and interpret our results.



We can see the boxplot for both the variables above. As per the boxplot for "X" variable, we can see quite a bit of data points far away from the mean. It corroborates with the QQ plot we saw in the previous question. The box plot also shows the deviation of variable X from normal distribution. The distribution seems to have potential outliers as we can see some points beyond the whiskers on both sides of the boxplot.

As per the boxplot for "Y" variable, we cannot data points far away from the mean. Infact, there no data points beyond the whiskers. It corroborates with the QQ plot we saw in the previous question. The box plot shows that variable Y is following a normal distribution.

2.3 We will calculate the mean, standard deviation, skewness and kurtosis of X and Y and Interpret.

Please find below the table depicting the mentioned statistics:

	mu	Std	Skew	kurt
X	0.04067866	3.3291654	-0.30269818	9.011688
Y	0.16857373	0.7236973	-0.03793123	2.365179

For variable X from the above table we can observe that the skewness is negative indicating that the distribution is a bit tailed to the left. Also, we can see that kurtosis is 9 indicating that the data is highly peaked. The spread of the data i.e variability of X is also comparatively higher than for variable Y. For variable, Y the skewness is close to zero indicating that the variable is following normal distribution. Also, the kurtosis is 2.36 which is near the expected value of 3 (for a standard normal distribution) indicating

that the distribution for variable Y is peaked close enough that of a standard normal distribution. The spread of the data i.e variability of Y is pretty small indicating high density of points near the mean.

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

2.4 We will use the bootstrap sampling-with-replacement technique to create datasets of 1000 samples from both X and Y. We will use the bootstrap samples to calculate the estimated variance of the both the median and mean of X and Y.

The estimated variance of the both the median and mean of X and Y:

	med_x	mean_x	med_y	mean_y
1	0.03324475	0.2060669	0.02283347	0.01014802

For both X and Y, we will analyze how does the bootstrap-estimated variance of the mean compare to that of the median and if there is a difference we will discuss what could be causing that

For variable X, the variance of median is moderately higher that the variance of its mean. One of the reason is that the distribution of variable X is not normal. The distribution has potential outliers resulting in skewness. We saw in the QQ plot for variable X showing that the data is skewed, attributing to the gap between mean and median value for samples collected. Consequently, the rate of change for the two statistics across the samples is different resulting in difference of variance. For variable Y, the variance for both the statistics is pretty close. Since Y is nearly normally distributed the difference between the mean and the median of a given sample does not vary much and the rate of change for two statistics across the samples is pretty much the same.

2.5 We will perform paired and un-paired t-tests to determine if X and Y have the same mean and interpret our results.

Un-Paired t test:

Null hypothesis: true difference in means is equal to 0

alternative hypothesis: true difference in means is not equal to 0

p-value = 0.7917

Paired t test:

Null hypothesis: true difference in means is equal to 0

alternative hypothesis: true difference in means is not equal to 0

p-value = 0.7747

For both the tests above we can see that the p value is higher than significance level of 0.05 as such we do not reject the Null Hypothesis for both the tests. The results are trustworthy as they corroborate with our earlier findings. We saw from the boxplot of the two samples that their means are very close to each other.

2.6 We will perform the Wilcoxon paired and unpaired rank tests to determine whether X and Y have t he same mean and interpret your results. We will also discuss if these results more or less trust-worth y than the t-test results

Wilcox Unpaired Test:

```
Null hypothesis: true location shift is equal to 0

alternative hypothesis: true location shift is not equal to 0

p-value = 0.372
```

Wilcox paired test:

```
Null hypothesis: true location shift is equal to 0

alternative hypothesis: true location shift is not equal to 0

p-value = 0.009412
```

For both the tests above we can see that the p value is higher than significance level of 0.05 as such we do not reject the Null Hypothesis for both the tests. The results are consistent with our findings, however for Wilcox paired test we see that the p value is very close to 0.05 which is a reason for concern. When we compare these results to what we got from t tests in the previous question where the p value was significantly higher for both the tests we can say that the Wilcox tests are not as trustworthy as t tests from previous section.

Topic 3: More Modelling

Data: This file has three variables, independent variables x1 and x2, and a dependent variable y.

3.1 We will build a linear least squares model (model1) to predict y using a linear combination of x1 and x2. We will show the model summary and interpret the results.

Model Summary:

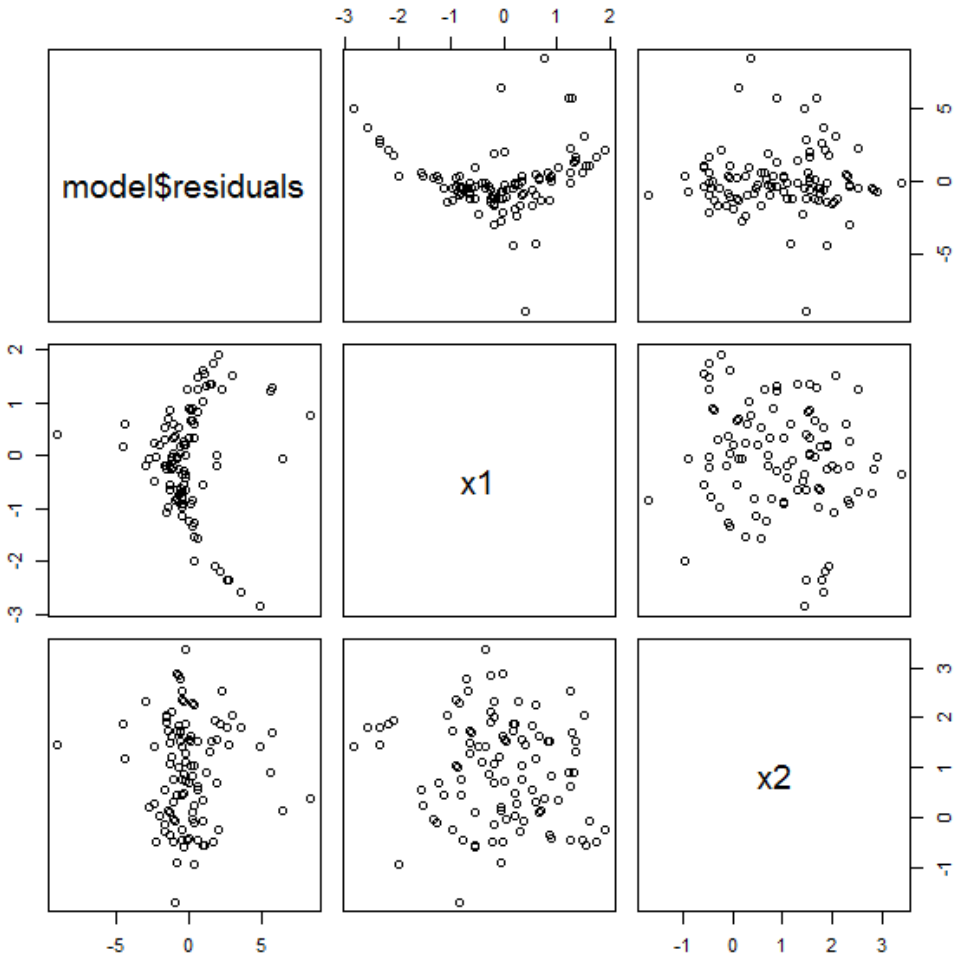
```
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9954 -1.0318 -0.3327  0.6067  8.4260

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6731     0.3017  12.173  < 2e-16 ***
x1            1.4741     0.2229   6.612 2.07e-09 ***
x2           -2.1320     0.2188  -9.745 4.76e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.223 on 97 degrees of freedom
Multiple R-squared:  0.6236,    Adjusted R-squared:  0.6158
F-statistic: 80.34 on 2 and 97 DF,  p-value: < 2.2e-16
```

Interpretation:

We can observe from the above table that coefficients of both the regressor variables x1 & x2 are significant along with the intercept having p value less than 0.05. The standard error of coefficients both for x1 & x2 is also pretty small indicating less variance which is good for model stability. The p value of the F-statistic is also less than 0.05 indicating that the fitted model is significant. Both R square & Adjusted R square value is around 0.62 indicating the amount of variability explained by the fitted regression model. Also, we see that the root mean square error is 2.223 which is not very high indicating that the regression model majorly explains variability of the response variable. Both RMSE and R square are inversely related and we can see the same relation here as well.

3.2 We will compute the residue and use the R function pairs() to draw a matrix scatterplot of the computed residue, x1 and x2.

The residual value is computed using R code. The matrix scatterplot is as below



From above figure, we see that the plot between residuals and x1 represents a slight curved band indicating a non-linear relationship between regressor variable and the response variable. As such adding polynomial variables or transforming the existing variables could be considered for better modelling. The plot between residuals and x2 represents a double bow pattern indicating inequality of residual variance i.e problem of heteroskedasticity. We cannot see any apparent pattern between x1 & x2 indicating that both these variables are apparently linearly independent.

3.3 We will build another model (model2) using second order polynomials of x1 and x2. From the summary, we will discuss about the significance or value of these three new variables and also how does the R2 for model2 compare to model1

Model Summary:

```
lm(formula = y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1 * x2))

Residuals:
    Min      1Q  Median      3Q     Max
-8.6741 -0.5745 -0.0108  0.4070  8.4338

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.71071    0.30004   9.035 2.08e-14 ***
x1           1.79299    0.25506   7.030 3.27e-10 ***
x2          -1.87215    0.34062  -5.496 3.31e-07 ***
I(x1^2)      0.92957    0.14652   6.344 7.78e-09 ***
I(x2^2)     -0.07608    0.15675  -0.485    0.629
I(x1 * x2)   0.25002    0.20665   1.210    0.229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.873 on 94 degrees of freedom
Multiple R-squared:  0.7411,    Adjusted R-squared:  0.7274
F-statistic: 53.83 on 5 and 94 DF,  p-value: < 2.2e-16
```

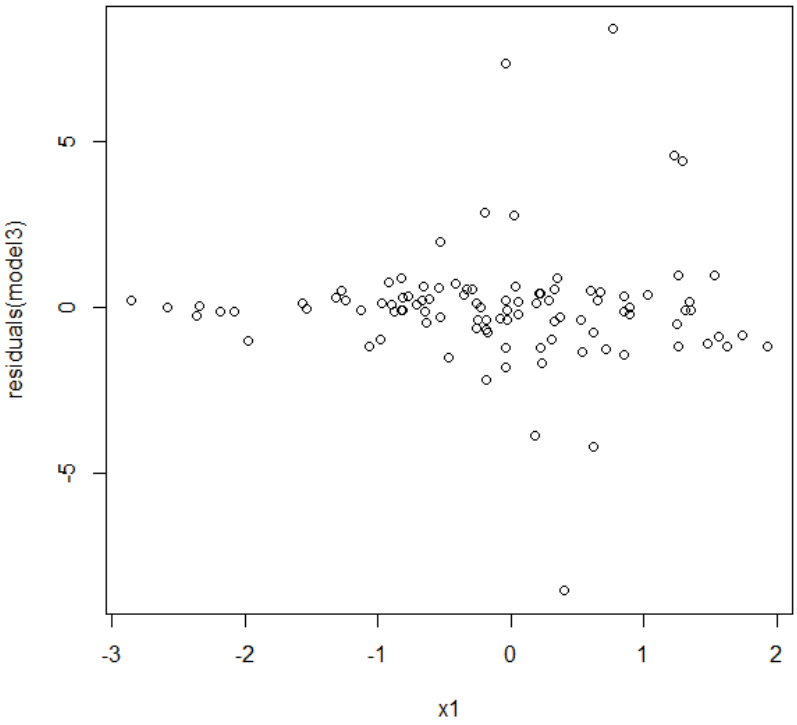
We can observe from the above table that coefficients of regressor variables x1,x2 & x1^2 are significant along with the intercept having p value less than 0.05. However, the regressor variables x2^2 and interaction term x1*x2 are not significant having p value greater than 0.05. We also observe that the R square value has significantly improved from 0.62 in model1 to 0.74 in this model. This indicates the fact that adding a polynomial term x1^2 to the data helps the model to fit better and explain more variability of the response variable. The added polynomial term also helps reduce the residual error in model2 again indicating better model fit.

3.4 We will build a new model3 for backwards feature selection by starting with model2 and using the function stepAIC().

```
Start:  AIC=131.31
y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1 * x2)

              Df Sum of Sq    RSS    AIC
- I(x2^2)      1     0.826 330.55 129.56
- I(x1 * x2)   1     5.134 334.86 130.85
<none>                     329.72 131.31
- x2           1   105.965 435.69 157.18
- I(x1^2)      1   141.187 470.91 164.95
- x1           1   173.335 503.06 171.55

Step:  AIC=129.56
y ~ x1 + x2 + I(x1^2) + I(x1 * x2)

              Df Sum of Sq    RSS    AIC
- I(x1 * x2)   1      4.95 335.50 129.05
<none>                     330.55 129.56
- I(x1^2)      1    142.25 472.80 163.35
- x1           1    176.54 507.09 170.35
- x2           1    401.71 732.26 207.10

Step:  AIC=129.05
y ~ x1 + x2 + I(x1^2)

           Df Sum of Sq    RSS    AIC
<none>                   335.50 129.05
- I(x1^2)   1    144.01 479.51 162.76
- x1        1    331.83 667.33 195.81
- x2        1    432.20 767.70 209.82
```

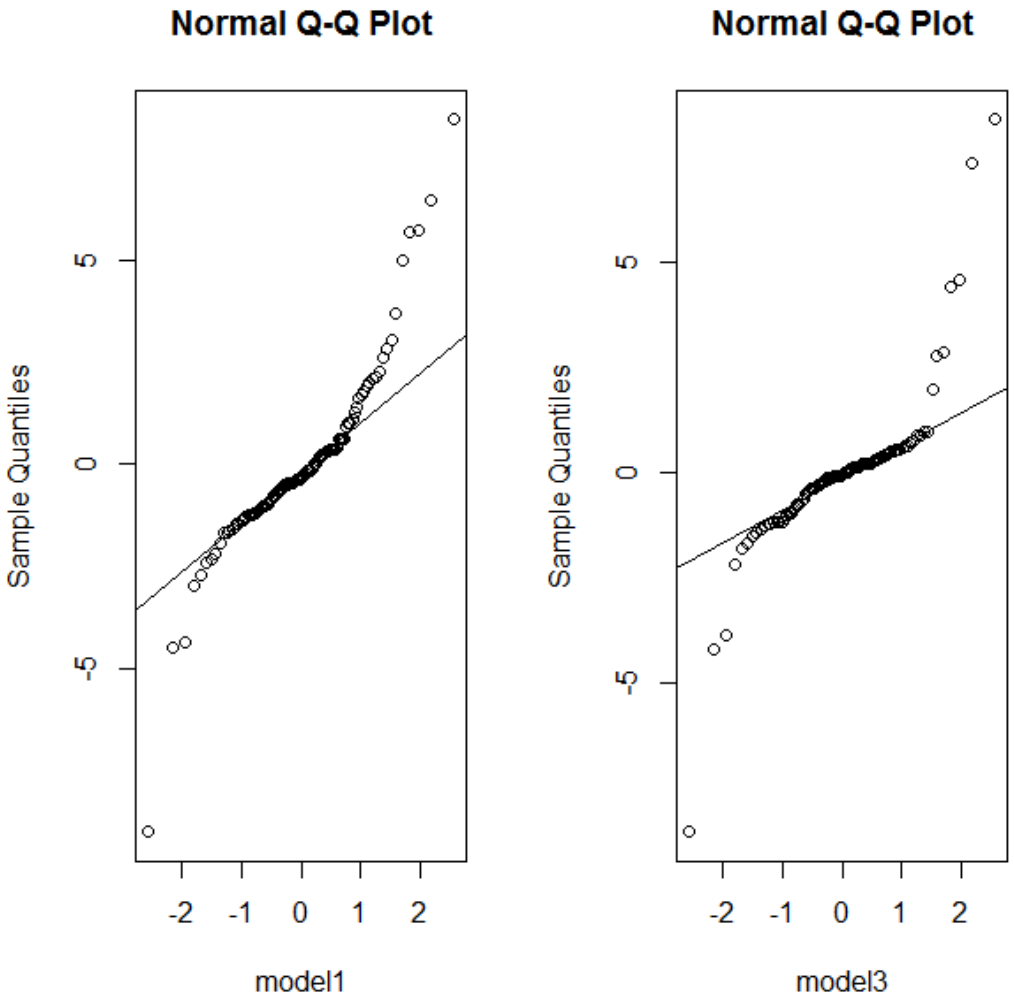
From the above summary table we can see that when we started with model2 the AIC was 131.31. We have implemented backward feature selection here as such at each step, AIC value is calculated and the variable which upon removal decreases the AIC value the maximum is deleted from the model in the subsequent step. We can see that removal of variable x2^2 in the first step reduces AIC value to 129.56 and subsequent removal of x1*x2 reduced the AIC to 129.05 upon which the iteration converges.

3.5 We will Plot the residue of model3 against x1 and interpret the results.



Interpretation: We can see from the above plot that there is no significant pattern between residuals and x1. We can also see some potential outliers having high absolute residual values. This plot indicates that there is no apparent inequality in residual variance and neither there is misspecification of the regressors thus no transformation or higher order terms are as such required.

3.6We will draw the QQ-Plots for both the residues of model1 and model3 and Interpret our results.



The residuals for model 1 shows deviation from normality. We can see that the tails at the extreme ends of the distribution are moderately sharp indicating that distribution is heavy tailed. We can also see some outliers in the plot at the far ends of the distribution.

The residuals for model 2 again shows deviation from normality. We can see that the tails at the extreme ends of the distribution are sharp indicating that distribution is heavy tailed. We can also see some outliers in the plot at the far ends of the distribution.

3.7 Now let's look at 100 different 80/20 cross-validation splits of the data. We will Compare the performance of model1, model2 and model3 across the splits. We will report the mean of the training and test error ± the standard deviation of the error. We will use the MSE metric. Interpret our results. We will also discuss which model has the best prediction accuracy and what do the differences in training and test performance imply

```
Model1    Training Error:  4.84965580326098 +/- 0.653305934585058"
          Test Error:  4.81780176322769 +/- 2.68162986097394"
Model2    Training Error:  3.37554720892232 +/- 0.631496666282127"
          Test Error:  3.23267672542338 +/- 2.44523043563081"
Model3    Training Error:  3.44170106757032 +/- 0.621253087463596"
          Test Error:  3.18884176184986 +/- 2.3972597763175"
```

We can see the performance of the three models as above.
Here, Training Error= Training MSE +/- Standard deviation
Test Error= Test MSE +/- Standard deviation

We can see that both for training and testing Model 1 has the highest error and also the highest standar d deviation among the three models. When we compare Model2 with Model 3 we observe that the train ing error for Model 3 is slightly more than that of Model2. However, the test error for Model 3 is lesser t han that of Model 2. Also, Model 3 has lower standard deviation both for training and testing. Conseque ntly, Model 3 has the best prediction accuracy which is apparent from the least test error and standard d eviation among all the three models. In case the training error is lesser than the test error the difference in the training and testing error implies that the model has less bias and comparatively more variance. H owever, in case the test error is lesser than the training error it indicates that model has more bias and c omparatively less variance. We see the latter in case of Model 3 as such it has lesser prediction error as i s evident from low test error.

PROBLEM 4: Training and Testing Models

Data Description: We have different values for town, district, street, family and gender. We have a number of categorical variables in this dataset, so we need to convert them. For example, we will create 4 different town categories, one each for A, B, C and D. These are Boolean variables. Similarly we will convert the districts into 2 variables, streets into 3, etc.

STEPS:

4.1 Read in the training data into a data frame.

The mentioned tasks have been done using R code.

4.2 Build a linear model with all the variables. For the sake of simplicity, let's assume there are no non-linear relationships in the data, so you don't need to plot out the residues against all the variables.

```
Call:
lm(formula = subject ~ ., data = train_data)

Residuals:
   Min     1Q Median     3Q    Max
-9.845 -3.455 -0.284  3.565 10.149

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.17874    1.17079   6.986 1.55e-11 ***
townA       -0.34086    0.78525  -0.434 0.664515
townB        1.66763    0.76594   2.177 0.030165 *
townC        3.39408    0.79987   4.243 2.86e-05 ***
townD        0.29547    0.82965   0.356 0.721963
d1          -0.77360    0.61334  -1.261 0.208086
d2           0.54589    0.61009   0.895 0.371552
fam1         0.26587    0.61156   0.435 0.664039
fam2         0.04285    0.63620   0.067 0.946342
st1         -1.69395    0.72113  -2.349 0.019406 *
st2          0.31508    0.70692   0.446 0.656100
st3         -0.47795    0.71174  -0.672 0.502355
GenM        -1.70765    0.50004  -3.415 0.000717 ***
replicate   -0.13310    0.50165  -0.265 0.790918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.628 on 333 degrees of freedom
Multiple R-squared:  0.1395,    Adjusted R-squared:  0.1059
F-statistic: 4.153 on 13 and 333 DF,  p-value: 2.134e-06
```

4.3 Use the anova function to explore the significance of all these variables via an analysis-ofvariance. Manually review the p-values and pick a set of variables and interaction variables which seem significant.

```
> summary(aov(model2))
                Df Sum Sq Mean Sq F value   Pr(>F)
replicate        1      0     0.1   0.006 0.936848
townA            1    147   146.9   8.624 0.003608 **
townB            1     11    10.6   0.623 0.430558
townC            1    466   466.4  27.377 3.40e-07 ***
townD            1      4     4.0   0.234 0.629038
d1               1     75    75.4   4.426 0.036333 *
d2               1     24    23.7   1.391 0.239361
st1              1    153   153.4   9.002 0.002953 **
st2              1     16    15.8   0.928 0.336193
st3              1      7     6.9   0.404 0.525396
fam1             1      3     3.2   0.186 0.667009
fam2             1      0     0.0   0.001 0.970997
GenM             1    250   249.8  14.660 0.000161 ***
replicate:townA  1     67    67.1   3.940 0.048185 *
```

replicate:townB	1	24	24.4	1.432	0.232534	
replicate:townC	1	27	27.0	1.584	0.209315	
replicate:townD	1	4	3.9	0.229	0.632464	
replicate:d1	1	17	16.7	0.980	0.323203	
replicate:d2	1	0	0.0	0.000	0.997016	
replicate:st1	1	2	1.5	0.089	0.766203	
replicate:st2	1	58	58.3	3.422	0.065453	.
replicate:st3	1	0	0.1	0.003	0.954446	
replicate:fam1	1	5	5.3	0.312	0.576788	
replicate:fam2	1	7	6.9	0.405	0.525097	
replicate:GenM	1	0	0.0	0.001	0.975315	
townA:d1	1	3	3.5	0.204	0.651905	
townD:d1	1	35	34.6	2.032	0.155236	
townA:d2	1	200	200.3	11.759	0.000702	***
townA:st1	1	137	136.5	8.014	0.004997	**
townA:st2	1	3	3.3	0.192	0.661263	
townA:st3	1	0	0.0	0.002	0.965436	
townA:fam1	1	1	0.9	0.050	0.822527	
townA:fam2	1	1	0.9	0.055	0.813971	
townA:GenM	1	8	8.1	0.475	0.491464	
townB:d1	1	1	1.5	0.087	0.768533	
townB:d2	1	7	7.3	0.428	0.513749	
townB:st1	1	95	95.0	5.576	0.018928	*
townB:st2	1	23	23.5	1.377	0.241713	
townB:st3	1	1	0.7	0.041	0.839702	
townB:fam1	1	177	176.8	10.378	0.001434	**
townB:fam2	1	2	1.6	0.092	0.762322	
townB:GenM	1	2	1.5	0.091	0.763398	
townC:d1	1	41	41.3	2.422	0.120852	
townC:d2	1	70	70.0	4.109	0.043643	*
townC:st1	1	38	38.1	2.239	0.135721	
townC:st2	1	86	85.6	5.024	0.025821	*
townC:st3	1	17	16.7	0.978	0.323707	
townC:fam1	1	3	2.6	0.153	0.695538	
townC:fam2	1	4	3.6	0.210	0.646993	
townC:GenM	1	286	286.3	16.805	5.51e-05	***
townD:d2	1	138	137.6	8.078	0.004827	**
townD:st1	1	63	63.4	3.719	0.054865	.
townD:st2	1	49	49.0	2.874	0.091197	.
townD:st3	1	0	0.2	0.011	0.916082	
townD:fam1	1	11	10.8	0.631	0.427529	
townD:fam2	1	150	150.4	8.827	0.003239	**
townD:GenM	1	24	23.7	1.391	0.239303	
d1:st1	1	19	18.6	1.091	0.297170	
d1:st2	1	35	34.9	2.051	0.153313	
d1:st3	1	1	1.4	0.083	0.774143	
d1:fam1	1	9	9.2	0.541	0.462717	
d1:fam2	1	5	4.6	0.270	0.603528	
d1:GenM	1	17	17.4	1.019	0.313579	
d2:st1	1	54	54.2	3.179	0.075711	.
d2:st2	1	43	43.3	2.543	0.111958	
d2:st3	1	278	277.9	16.314	7.02e-05	***
d2:fam1	1	8	7.7	0.452	0.502071	
d2:fam2	1	4	3.5	0.208	0.648613	
d2:GenM	1	0	0.0	0.000	0.996985	
st1:fam1	1	61	60.5	3.553	0.060514	.
st1:fam2	1	36	36.2	2.124	0.146227	
st1:GenM	1	0	0.0	0.000	0.999999	

```
st2:fam1          1     51    50.7   2.974 0.085787 .
st2:fam2          1      0     0.0   0.002 0.962202
st2:GenM          1      9     9.4   0.552 0.458308
st3:fam1          1     11    10.6   0.622 0.430964
st3:fam2          1      6     5.6   0.331 0.565617
st3:GenM          1      6     5.8   0.341 0.559655
fam1:GenM         1     46    46.1   2.706 0.101168
fam2:GenM         1     16    15.7   0.923 0.337534
Residuals       266   4532    17.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Looking at the p-values pertaining to the variables in the above table below are the variables which seem significant:

```
townA             1    147   146.9   8.624 0.003608 **
townC             1    466   466.4  27.377 3.40e-07 ***
d1                1     75    75.4   4.426 0.036333 *
st1               1    153   153.4   9.002 0.002953 **
GenM              1    250   249.8  14.660 0.000161 ***
replicate:townA   1     67    67.1   3.940 0.048185 *
townA:d2          1    200   200.3  11.759 0.000702 ***
townA:st1         1    137   136.5   8.014 0.004997 **
townB:st1         1     95    95.0   5.576 0.018928 *
townB:fam1        1    177   176.8  10.378 0.001434 **
townC:d2          1     70    70.0   4.109 0.043643 *
townC:st2         1     86    85.6   5.024 0.025821 *
townC:GenM        1    286   286.3  16.805 5.51e-05 ***
townD:d2          1    138   137.6   8.078 0.004827 **
townD:fam2        1    150   150.4   8.827 0.003239 **
d2:st3            1    278   277.9  16.314 7.02e-05 ***
```

4.4 Now build a linear model with the variables (independent and interaction) that you picked. Then d o forward and backward feature selection using stepAIC with direction=”both”. Give the summary of t his post-feature selection model.

Model Building on the selected variables in the previous question:

model3 = lm(subject~GenM+st1+d1+townC+townA+townD*fam2+ d2*st3+townD*d2+townC*GenM+to wnC*st2+townC*d2+townB*fam1+townB*st1+townA*st1+townA*d2+replicate*townA,train_data)
summary(model3)

Forward and backward feature selection using stepAIC with direction=”both”. Give the summary of this post-feature selection model.

Below is the summary pertaining to the final iteration of Step AIC:

```
Step:  AIC=1044.35
subject ~ GenM + st1 + d1 + townC + townA + townD + fam2 + d2 +
    st3 + st2 + townB + fam1 + replicate + townD:fam2 + d2:st3 +
```

```
    townD:d2 + GenM:townC + townC:st2 + townB:fam1 + st1:townB +
    st1:townA + townA:d2 + townA:replicate

                  Df Sum of Sq    RSS    AIC
<none>                         6128.1 1044.3
- townD:fam2       1    40.853 6169.0 1044.7
- townC:st2        1    47.638 6175.8 1045.0
- townD:d2         1    47.821 6176.0 1045.0
+ townC:d2         1    19.880 6108.3 1045.2
- d1               1    52.751 6180.9 1045.3
- d2:st3           1    83.890 6212.0 1047.1
- st1:townB        1    94.052 6222.2 1047.6
- townA:replicate  1    95.386 6223.5 1047.7
- townB:fam1       1   112.215 6240.4 1048.7
- st1:townA        1   151.448 6279.6 1050.8
- townA:d2         1   162.958 6291.1 1051.5
- GenM:townC       1   234.688 6362.8 1055.4
```

4.5 Next build a final linear model with the features selected. What is the MSE of this model? List all of your features.

Final linear model

```
lm(formula = subject ~ GenM + st1 + d1 + townC + townA + townD +
    fam2 + d2 + st3 + st2 + townB + fam1 + townD * fam2 + d2 *
    st3 + townD * d2 + GenM * townC + townC * st2 + townB * fam1 +
    st1 * townB + st1 * townA + townA * d2 + townA * replicate +
    replicate, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0443  -2.9645  -0.4741   3.2791   9.6771
```

MSE of this model

MSE = SSE/n-p

MSE=18.97259

List all of your features

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.10713    1.21880   7.472 7.45e-13 ***
GenM            -0.83410    0.53022  -1.573 0.116665
st1             -3.03569    0.83113  -3.652 0.000303 ***
d1              -0.96848    0.58081  -1.667 0.096393 .
townC            4.73936    0.99642   4.756 2.98e-06 ***
townA           -6.24810    1.98284  -3.151 0.001779 **
townD            0.29853    1.05747   0.282 0.777892
fam2             0.63510    0.65505   0.970 0.332996
d2              -1.32500    0.75203  -1.762 0.079034 .
st3             -1.30679    0.78904  -1.656 0.098657 .
st2             -0.06774    0.72220  -0.094 0.925323
townB            1.91303    0.88419   2.164 0.031229 *
```

```
fam1             1.01804    0.63639   1.600 0.110642
replicate       -0.61076    0.53549  -1.141 0.254896
townD:fam2      -1.90396    1.29750  -1.467 0.143238
d2:st3           2.36964    1.12691   2.103 0.036259 *
townD:d2         2.11554    1.33253   1.588 0.113353
GenM:townC      -4.22214    1.20047  -3.517 0.000499 ***
townC:st2        2.14221    1.35192   1.585 0.114042
townB:fam1      -2.93498    1.20682  -2.432 0.015559 *
st1:townB        3.17025    1.42388   2.226 0.026671 *
st1:townA        4.10456    1.45277   2.825 0.005017 **
townA:d2         3.78534    1.29161   2.931 0.003623 **
townA:replicate  2.64674    1.18041   2.242 0.025625 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.356 on 323 degrees of freedom
Multiple R-squared:  0.2605,       Adjusted R-squared:  0.2079
F-statistic: 4.948 on 23 and 323 DF,p-value: 1.126e-11
```

4.6 Load in the test data and apply your model to this data. What is the MSE in the test set?

MSE= 23.43232