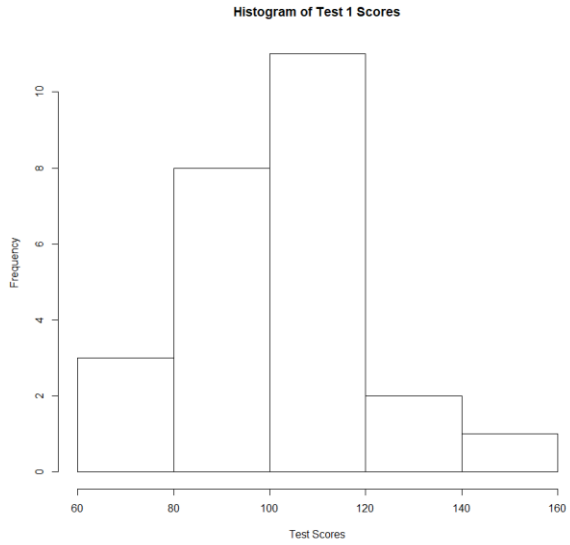
Time Series

STA 137

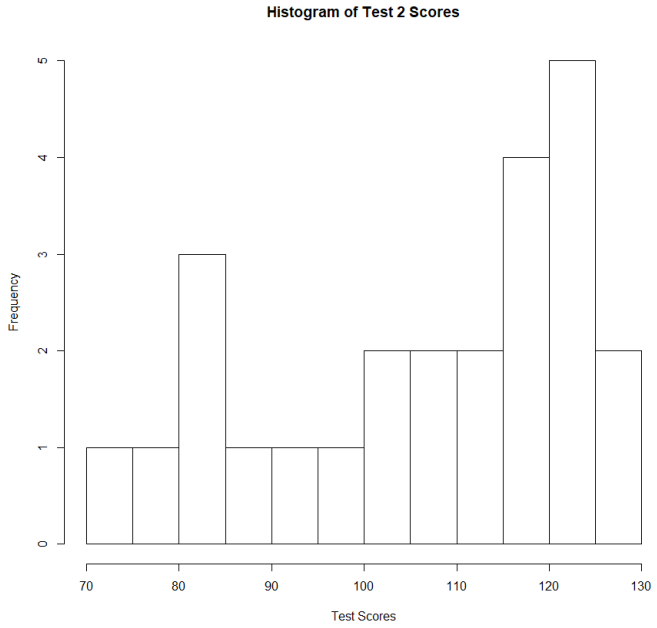
Homework 1

Jared Yu, Danli Zheng
1-16-2019

1. A) The first histogram of scores for Test 1 show that the distribution is quite unimodal. The scores concentrate more towards the lower end than the higher end. It shows that not many test takers performed well, and most of them were either average or below average.

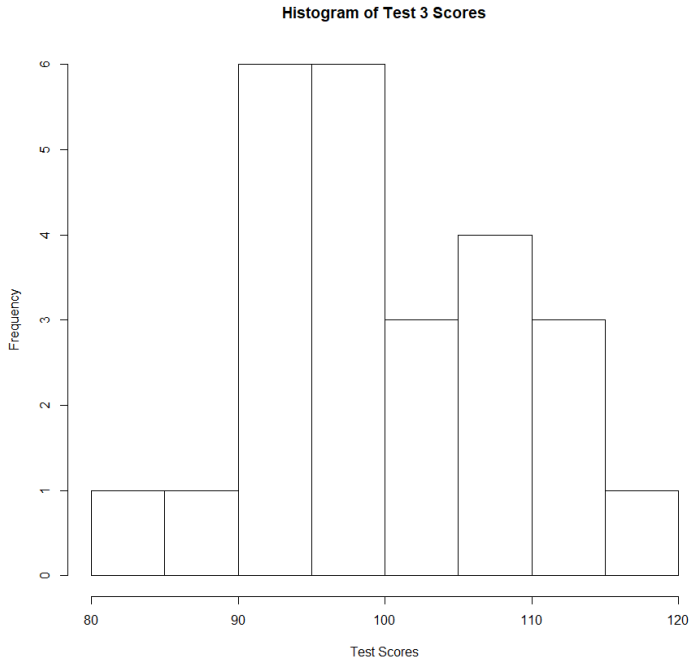


The second histogram of scores for Test 2 show that there is possibly a bimodal distribution. There is a high concentration towards the high end, where many people did well (assuming that 130 is a good score). The distribution appears to be bimodal at the default number of breaks, by increasing it to 10 there seems to be a spike in the lower range. At this level of breaks, there seems to be the concentration around 120, and a small spike around the low 80's.

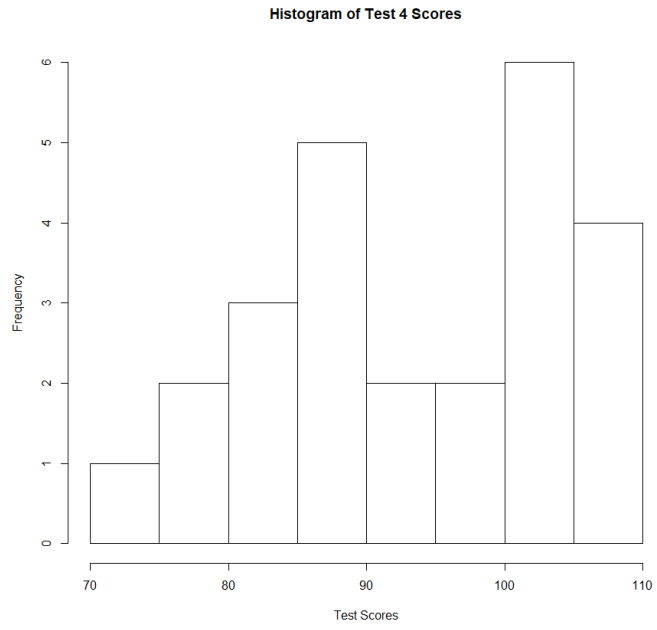


The third histogram of scores for Test 3 show that there is the highest concentration of scores between 90-100. The distribution is also heavier towards the right, implying that people tended

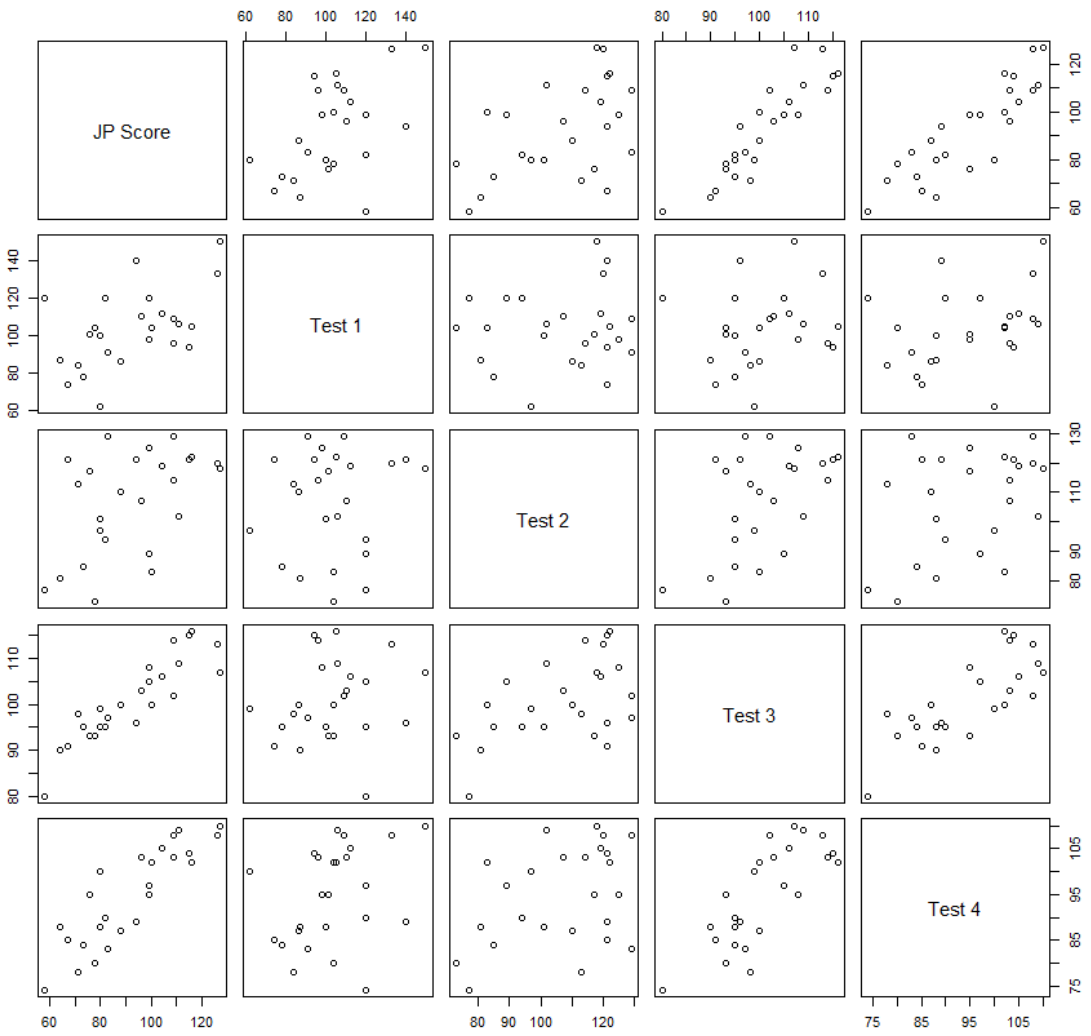
to do better, with only a small portion that did particularly worse than others. It is entirely clear, but the distribution appears more unimodal than bimodal.



The fourth histogram of scores for Test 4 show that there is a bimodal distribution. There is a concentration around the 100-110 range, and a concentration leading up to the high 80's. It shows that although many people scored around the high end for the test, many people did not do as well.



1. B) The following is a pairs matrix for the data:



The pairs matrix seems to show that the strongest linear relationship for Job Proficiency scores and test scores is between Test 3 and Job Proficiency score and possibly Test 4 and Job Proficiency score. It is also interesting that Test 3 and Test 4 also share a strong linear relationship. Although Test 1 and Test 2 seems to show some hints of linearity between each of them and Job Proficiency (more so in Test 1), there is quite a large spread that makes it difficult to say they are useful in prediction. At the same time, Test 1 and Test 2 don't seem to have a significant linear relationship based on their plot. The most notable other relationship between variables is between Test 2 and Test 3. There seems to be a linear relationship, but it does have problems as there is an uneven spread.

Below is a correlation matrix for the data.

	Job Proficiency	Test 1 Score	Test 2 Score	Test 3 Score	Test 4 Score
Job Proficiency	1	0.5144	0.4970	0.8971	0.8694
Test 1 Score	0.5144	1	0.1023	0.1808	0.3267
Test 2 Score	0.4970	0.1023	1	0.5190	0.3967
Test 3 Score	0.8971	0.1808	0.5190	1	0.7820
Test 4 Score	0.8694	0.3267	0.3967	0.7820	1

The correlation matrix also confirms what was at first noticed in the pairs matrix. There is a strong relationship between Job Proficiency with Tests 3 and 4 (0.8971 and 0.8694). The two tests 3 and 4 also share a noticeable correlation (0.7820). The correlation between other variables such as between Job Proficiency and tests 1 and 2 are noticeably weaker (0.5144 and 0.4970). Tests 2 and 3 also share a correlation similar to the level that tests 1 and 2 share with Job Proficiency (0.5190).

In this case it is evident in both the pairs matrix and correlation matrix that multicollinearity is an issue for tests 3 and 4. Although tests 3 and 4 share a strong relationship with Job Proficiency, the two tests also share a significant relationship with each other. The relationship between the two variables could make it problematic to utilize both to predict Job Proficiency due to the lack of independence. When the model tries to understand how important the two variables are in predicting Job Proficiency, a problem will arise due to their relative importance being tied to the connection that the two variables already have.

1. C) Below is a table of parameter estimates and standard errors after fitting the multiple linear regression model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\beta}_4 X_{i4}$, for $i = 1, \cdots, n = 25$.

$Coefficients$	$Estimate$	$Standard\ Error$
β_0	-124.38182	9.94106
β_1	0.29573	0.04397
β_2	0.04829	0.05662
β_3	1.30601	0.16409
β_4	0.51982	0.13194

The corresponding R^2: 0.9629 and the R^2_{adj}: 0.9555

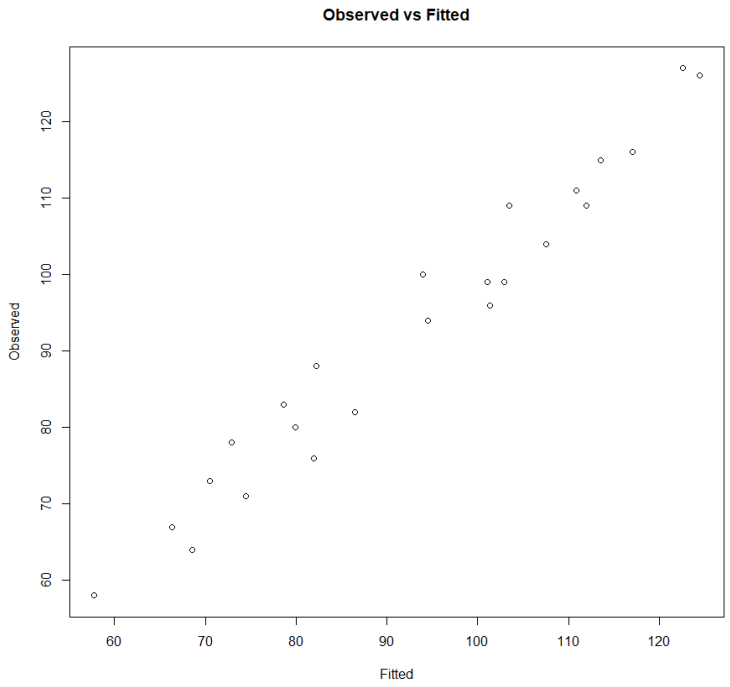
Below is the Analysis of Variance Table

Analysis of Variance Table

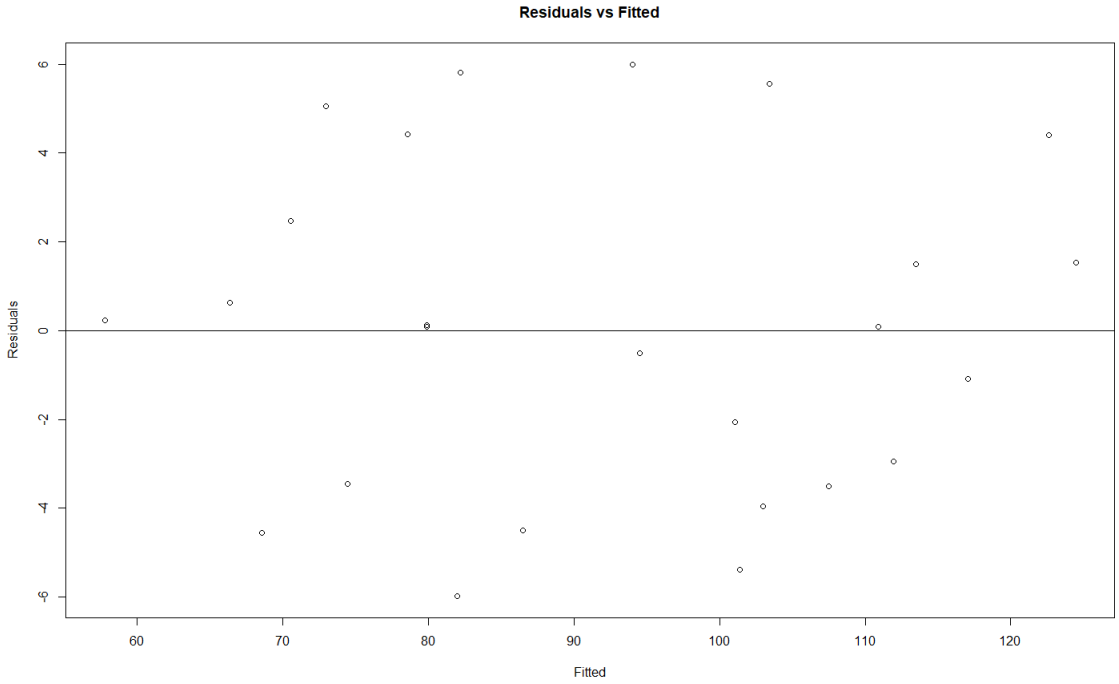
$Source$	df	SS	MS	F	P
$Regression$	$p-1=4$	8,718	$\frac{SSR}{df(SSR)}=2179.5$	$\frac{MSR}{MSE}=129.7321$	$F^{*}=5.262457\times 10^{-14}$
$Error$	$n_T-p=20$	336	$\frac{SSE}{df(SSE)}=16.8$		
$Total$	$n_T-1=24$	9,054			

1. D) The ANOVA model in general seems to reject the hypothesis that the test scores are not at all effective in predicting the job proficiency scores. This can be seen by the F statistic from summary() which has a p-value of less than 0.000. It seems from the summary output that the variable for Test 2 can be dropped from the model. The p-value for the Test 2 score is the largest, at 0.40383. The next nearest p-value is Test score 4, which is still smaller than 0.000.

2. A) Below is a plot of observed against fitted values.

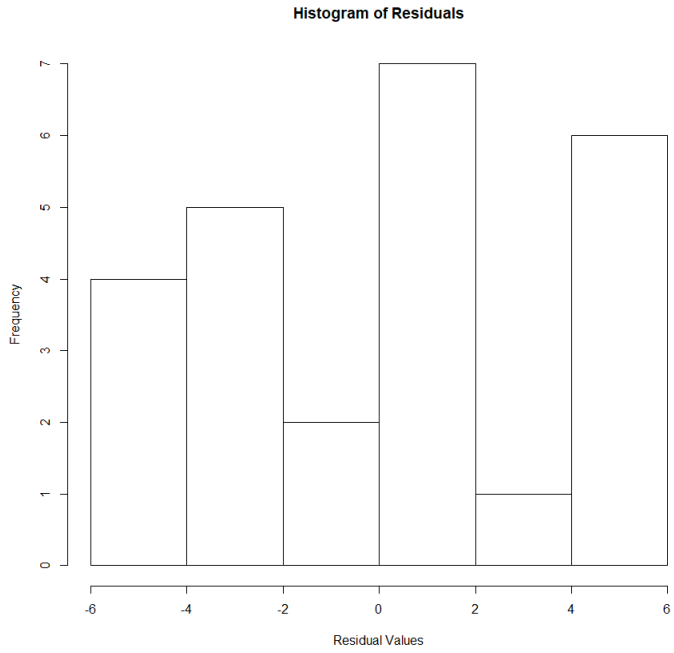


It seems from the graph that the fitted values are quite close to the actual observed values. This helps with the idea that the fitted model is appropriate for the data. If there were an alternate appearance in the plot, such as some sort of curvilinear relationship, it would suggest that the fitted values are quite distant from the observed values. This would mean that the fitted values didn't quite reasonably match the appearance of the data.



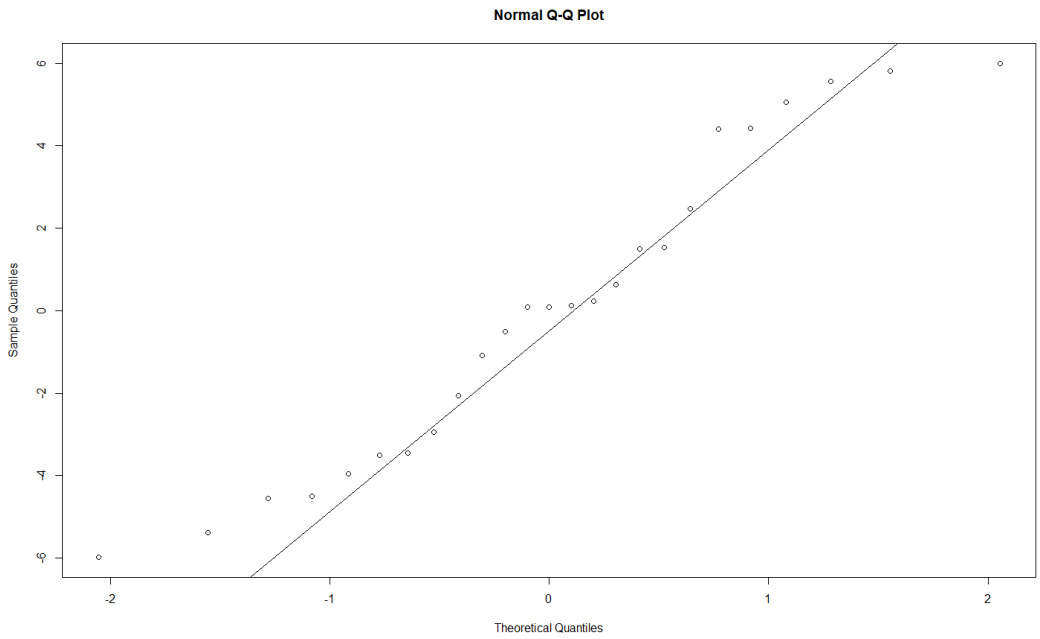
The residuals vs fitted plot overall looks good, the residuals seem to be quite random. There is not much of a noticeable pattern in the plot, with the residuals being spread across 0. This suggests that the assumption for a linear relationship is reasonable given the data. Also, there is a horizontal band throughout the residuals as they go about the 0 line. This suggests that the variance of the error terms is roughly equal throughout. There is also not a great deal of data in the dataset, so there may be issues with an adequate interpretation.

2. B) Below is a histogram of the residuals:



The epsilon error terms are estimated using residuals. Here they are plotted in a histogram, and it lacks the appearance of a bell-shaped pattern. There seems to be a spike in the center, but there are also high frequencies at the edges. This implies that there is an issue with the assumption of normality of errors. If the errors were normal, then the histogram would appear bell-shaped, like the normal distribution. However, it is important to note that the sample size is quite small, which does make it difficult to gauge whether there is a violation of the assumption from the histogram alone.

Below is a normal probability plot:



The normal probability plot shows that there is some spread around the QQ line, with a gap included close to the Theoretical Quantile 1. The tails are also light tailed. This indicates that there is some degree of non-linearity in the normal probability plot. The correlation between the residuals and normal scores is quite high at 0.9772966. It is possible that the tails are balancing out the correlation so that it is so high. Overall the normal probability plot looks safe when trying to decide whether the error terms are normal.

3. A) Below are the 5 models that are chosen by eliminating one parameter at a time.

$Model$	AIC	BIC
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$	147.9011	155.2144
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$	146.7942	152.8886
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i$	158.6741	163.5496
$Y_i = \beta_0 + \beta_3 X_{i3} + \varepsilon_i$	183.4155	187.0721
$Y_i = \beta_0 + \varepsilon_i$	222.2491	224.6868

Both AIC and BIC have chosen the model with parameters X_1, X_3, X_4. The corresponding AIC and BIC values are 146.7942 and 152.8886.

$Coefficients$	$Estimate$	$Standard\ Error$
β_0	-124.20002	9.87406
β_1	0.29633	0.04368
β_3	1.35697	0.15183
β_4	0.51742	0.13105

The corresponding R^2: 0.9615 and the R^2_{adj}: 0.956.

3. B) Below are all 16 possible models along with the corresponding AIC and BIC

$Model$	AIC	BIC
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$	147.9011	155.2144
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$	160.2613	166.3556

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \varepsilon_i$	181.583	187.6774
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$	146.7942	152.8886
$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$	175.4562	181.5506
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$	210.6495	215.525
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i$	158.6741	163.5496
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_4 X_{i4} + \varepsilon_i$	184.0282	188.9037
$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$	185.2422	190.1177
$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_4 X_{i4} + \varepsilon_i$	188.0189	192.8944
$Y_i = \beta_0 + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$	173.8075	178.683
$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$	216.5649	220.2216
$Y_i = \beta_0 + \beta_2 X_{i2} + \varepsilon_i$	217.1563	220.813
$Y_i = \beta_0 + \beta_3 X_{i3} + \varepsilon_i$	183.4155	187.0721
$Y_i = \beta_0 + \beta_4 X_{i4} + \varepsilon_i$	189.0015	192.6581
$Y_i = \beta_0 + \varepsilon_i$	222.2491	224.6868

The model chosen by AIC and BIC are the same, it is the mode with X_1, X_3, X_4, same as before in part A. The AIC values are also the same at 146.7942 for AIC and 152.8886 for BIC. Therefore, the parameter estimates, standard errors, R^2 and R^2_{adj} are the same as before:

$Coefficients$	$Estimate$	$Standard\ Error$
β_0	-124.20002	9.87406
β_1	0.29633	0.04368
β_3	1.35697	0.15183
β_4	0.51742	0.13105

The corresponding R^2: 0.9615 and the R^2_{adj}: 0.956.