Regression and Anova

Problem Set 3

Mohamed Salem

Abstract

This paper takes a look at the assumptions underlying the simple linear regression model. Assumptions are examined sequentially and are often followed by statistical tests to verify their non-violatio in the presented case study. We find that, for the given case, the assumptions of our linear model hold, and we can confidently fit a linear model to represent the relationship between our independent variable and our response variable.

Introduction:

Throughout this work, we will be examining the properties and workings of the simple linear regression model. More specifically, we will construct multiple small datasets, and fit our simple linear model to those datasets. Along the way, we will examine the assumptions underlying the simple linear model, and test for whether or not these assumptions hold. We will present a number of tests, describing how they work, how to apply them and their associated hypotheses. Should any of our model assumptions be violated, we will also show how to remedy such violations. This work is presented in the form of four problems. The problems have some overlapping elements such as the model used. Notation will be consistent throughout this work.

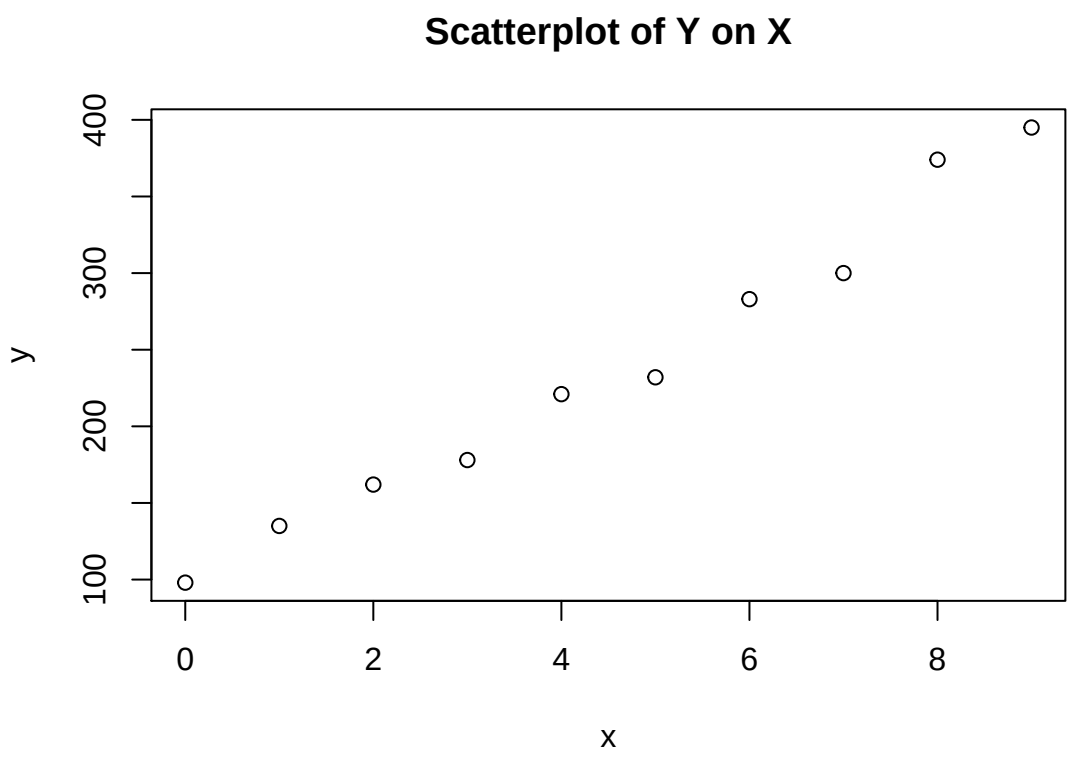
Model and Methodology:

Problem 1:

We begin by constructing a dataset representing annual sales of a product over 10 years, notice that there is an element of time in this problem. The data is as follows:

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
x	0	1	2	3	4	5	6	7	8	9
y	98	135	162	178	221	232	283	300	374	395

We first proceed by constructing a scatterplot of our data, this allows us to visually examine the relationship between our independent and response variables, and therefore we might be able to discern whether a linear model would be a good fit for the problem at hand.



From the scatterplot, a linear assumption seems to be reasonable. However, this is only a visual test. We can be more rigorous in testing our linearity assumption using some statistical tests such as the ANOVA linear lack of fit test. We could also take a look at the residual plot, but to construct a residual plot, we first need to fit a linear model to our data. We will use the following simple linear regression model:

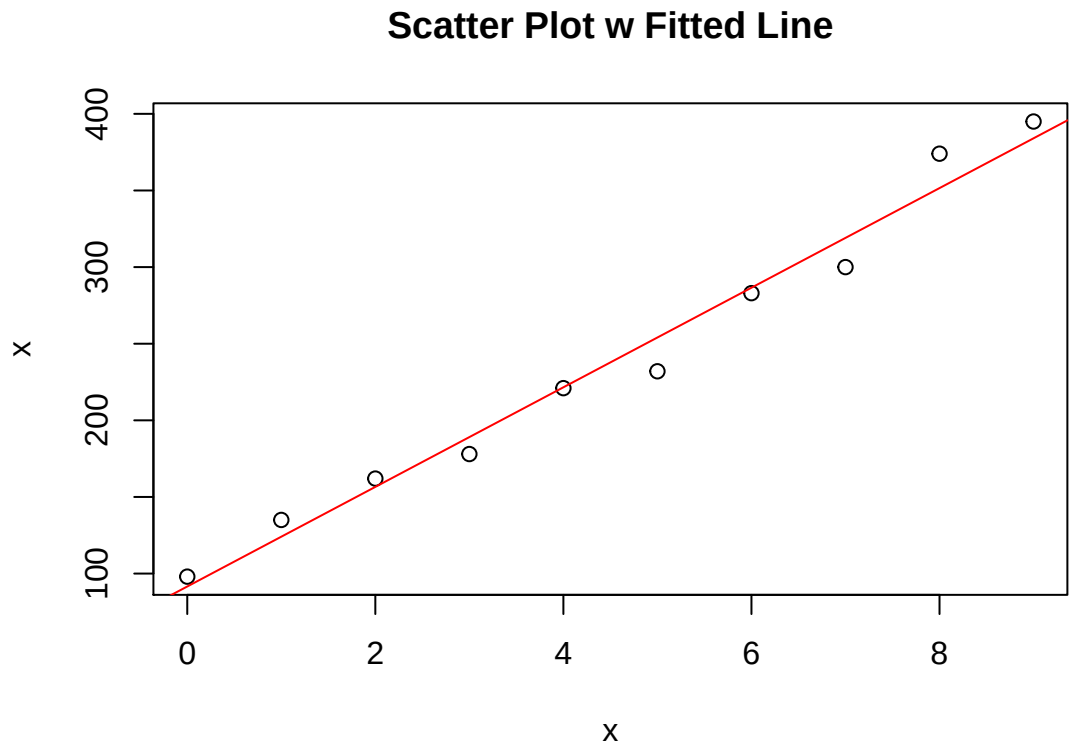
$$Y_i = \beta_0 + \beta_1.x_i + \epsilon_i$$

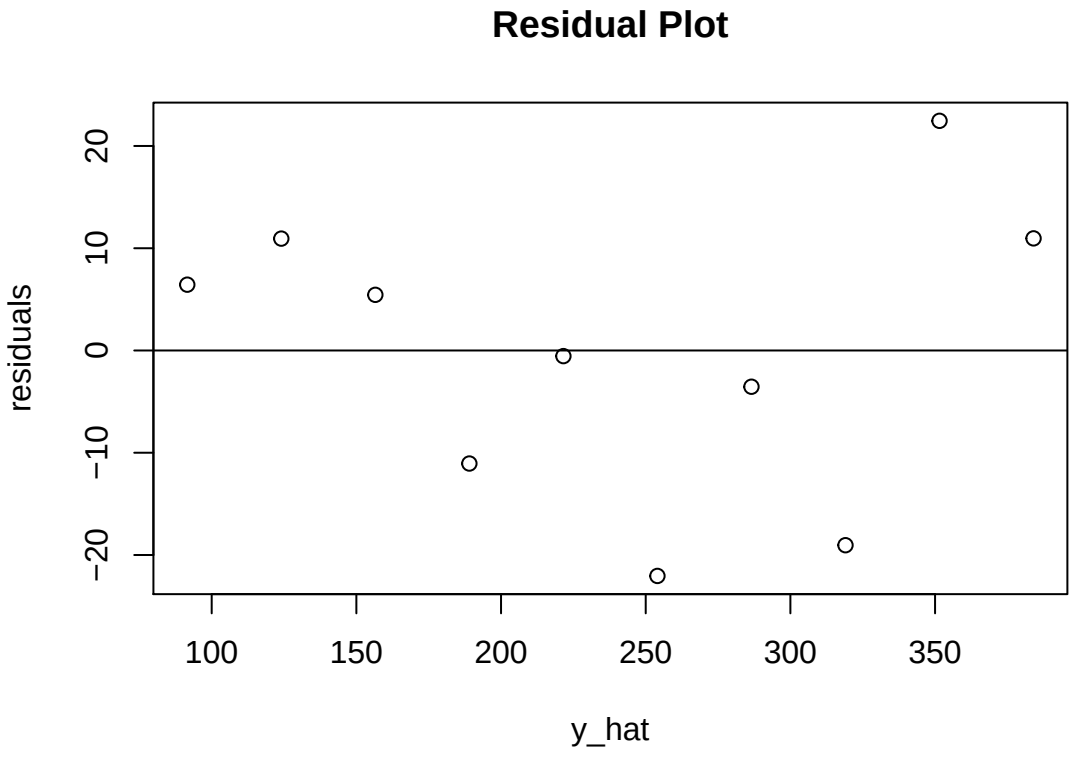
and we will estimate our β's by the following formulas:

$$\hat{\beta}_{[2x1]} = (X'X)^{-1}X'Y$$

Applying the above, we estimate $\hat{\beta}_0 = 91.5636364$ and $\hat{\beta}_1 = 32.4969697$. Using these estimates we can construct our residuals e_i:

$$e_i = Y - \hat{Y} \quad where \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1.x_i$$


Scatter Plot w Fitted Line
x
100
200
300
400
0
2
4
6
8
x


Residual Plot
residuals
−20
−10
0
10
20
100
150
200
250
300
350
y_hat

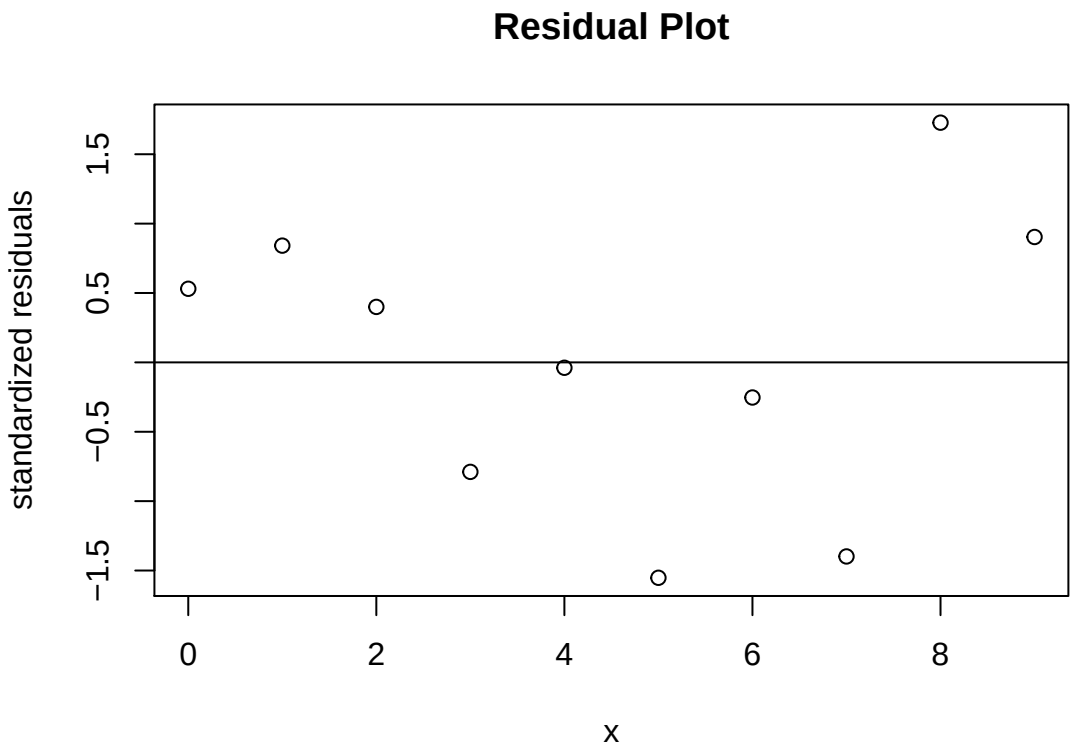
Now we have our scatterplot again, with our fitted linear model line, as well as our residual plot. Visually examining both plots indicate that a linear relationship may exist between X & Y, and since we have no duplicates in our data, we are unable to perform an ANOVA test for linear lack of fit (which requires duplication in the independent variable). Therefore we do not have compelling evidence against an assumption of linearity. For further measure, we will also look at the standardized and studentized residuals. Standardized residuals can be defined as:

$$std.res = \frac{e_i}{\hat{\sigma}.\sqrt{1-h_{ii}}}$$

While studentized residuals are defined as:

$$stu.res = std.res.\sqrt{\frac{n-p-1}{n-p-(std.res)^2}}$$

Estimating and plotting our standardized residual yields the below residual plot:



The residual plot also shows no strong evidence for violation of the linearity assumption, we also do not have any outliers (no residuals are more than 2 standard deviations away from zero). Next we'll sequentially examine each of the remaining assumptions associated with fitting a linear model. We continue using the linear model results to test the assumptions. We notice from the residual plot that there seems to be some increase in the variability of the residuals in the value of the predictor variable x. This may indicate a violation of the linear model's constant variance assumption. To verify that, we'll apply both the Breusch-Pagan, and Brown-Forsythe tests for constant variance.

The Breusch-Pagan test was performed using two test statistics the BP_F statistic follows an $F_{p,n-p-1}$ distribution where $p = 1$, $n = 10$, in our example. While the BP_{LM} statistic follows a χ^2_p distribution. The

tests yielded p-values of 0.1509018 and 0.1215546, respectively. Both p-values indicate that we should fail to reject the null hypothesis, which in this test, represents the hypothesis of constant variance. We present the test statistics in detail below:

$$BP_F = \frac{SSReg/p}{SSRes/n-p-1} = \frac{\sum_i (\hat{e}_i^2 - \bar{e}^2)^2/p}{\sum_i (e_i^2 - \hat{e}_i^2)^2/n-p-1}$$

Where $\hat{e}_i^2$ is obtained from the below linear model:

$$e_i^2 = \delta_0 + \delta_1.x_i + \xi_i$$

While for the BP_{LM} statistic we have:

$$BP_{LM} = n.\frac{SSReg}{SST} = n.R^2$$

Where the hypotheses are:

$$H_0 : constant\ variance$$

$$H_a : non - constant\ variance$$

We will carry out another test, the Brown-Forsythe test, for further verification, where the BF test statistic is:

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where we have separated our residuals into two groups, each group containing the residuals that are closest to each other and farthest from the residuals in the other group. $\bar{d}_1$ and $\bar{d}_2$ represent the mean distances from the median for the values of these two groups. Next, we computed the median of each group and found the distances from the respective median for each observation, we also use the mean distance to find the sum of squared differences between distance from the median and average distance from the median, for each group.We used those sum of squared distances to compute a pooled variance estimate as follows:

$$s^2 = \frac{\sum_i (d_{i1} - \bar{d}_1)^2 + \sum_i (d_{i2} - \bar{d}_2)^2}{n-p-1}$$

The Brown-Forsythe test also provides evidence against non-constant variance, where we observe a p-value of 0.1148953 which supports the null hypothesis of constant variance; therefore we will go ahead and assume that the constant variance assumption holds. Next we will test the assumption that the error terms are independent and randomly distributed. For this assumption, since our data is collected over ten years, it may have time dependence as we noted earlier. This allows us to use the Durbin-Watson test to check for autocorrelation in the error terms. Our Durbin-Watson test statistic is constructed as follows:

$$DW = \frac{\sum_t (e_t - e_{t-1})^2}{\sum_t e_t^2}$$

Comparing our Durbin-Watson test statistic value of 1.8520634 to the Durbin-Watson critical values table, we conclude that our data offers evidence against autocorrelation of the error terms, where our hypothesis are as follows:

$$H_0 : No\ autocorrelation$$

$$H_a : Autocorrelation$$

We can also use a version of the Runs test, to test that our errors are randomly distributed. To do that we will start by sorting our residuals along with their associated fitted vlue. The Runs test is a non-parametric test that observes only binary outcomes. For that reason, we will check whether we randomly observe different signs for our residuals. Our test statistics are:

r: the number of runs
r_+: the number of positive values
r_-: the number of positive values

Since we have small sample size of 10, we will compare our Runs test statistic r = 3, with the appropriate critical values in the Runs test table. We note that if we had a large enough sample size (greater than 20), we would have constructed a standardized test statistic using r, which approximately follows a standard normal distribution. From the table, our lower critical value r_L is 2; our upper critical value r_U is 10. Since our test statistic r is between these two values we fail to reject the null hypothesis, where our hypotheses are as follows:

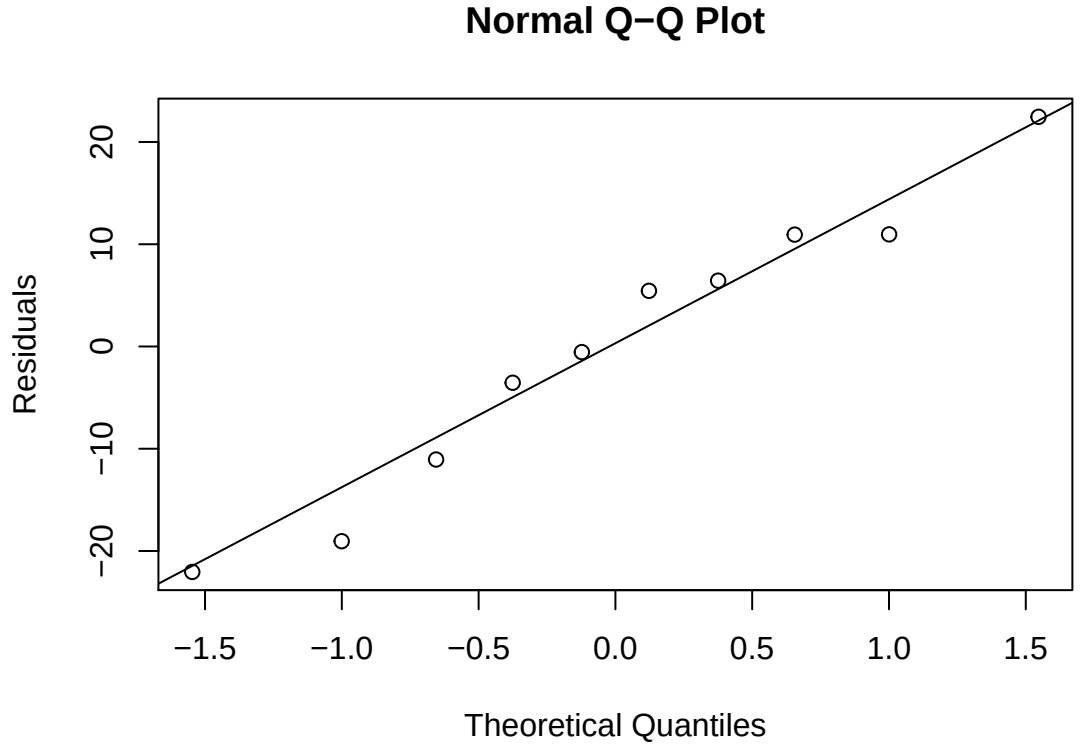
$$H_0 : \textit{Residual signs are randomly distributed}$$

$$H_a : \textit{Residual signs are not randomly distributed}$$

Finally, we take a look at the Normality of the Residuals assumption, we do this in two ways, first, by using a Normal QQ plot; and second, by doing a Shapiro-Wilk test. Our Shapiro-Wilk test statistic is:

$$SW = \frac{\sum_i (a_i.e_i)^2}{\sum_i e_i^2}$$

where the a_i represent prespecified Shapiro-Wilk coefficients.



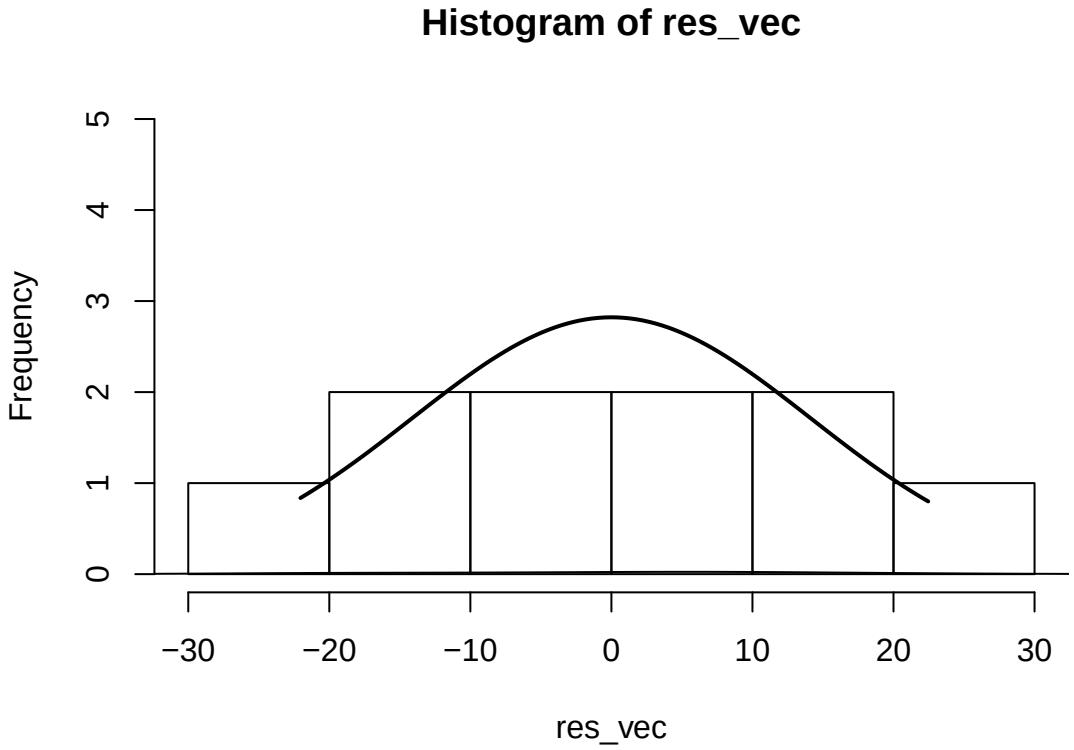
Our Normal QQ plot shows that the data may have a thinner left tail than we would observe in a standard normal distribution. It's not easy to make a judgment about normality based on visuals alone. We carry out the Shapiro-Wilk test to see the result.

Our Shapiro-Wilk test statistic is very close to 1 for both the residuals (0.9613) and the y values (0.9585), and the associated p-values are 0.80 and 0.77 for the residuals and the y values, respectively. To understand what this means, we look at the Shapiro-Wilk test statistic. With some manipulation, we can conclude that the Shapiro-Wilk test, is in essence, a squared correlation coefficient between residuals (or y values) and a theoretical standard normal distribution, therefore, the closer the observed statistic is to one, the stronger the evidence for normality. Therefore, we fail to reject the null hypothesis of normality of our residuals. Our Shapiro-Wilk hypotheses are:

$$H_0 : Residuals\ are\ from\ a\ normally\ distributed\ population$$

$$H_a : Residuals\ are\ not\ from\ a\ normally\ distributed\ population$$

For further measure, we can also construct a histogram of our residuals to check for normality:



Conclusion

After performing a number of tests, we failed to reject the hypotheses that an assumption is not violated for any of our assumptions. We conclude that given the data, the evidence points towards our linear model assumptions being satisfied, and therefore we can confidently say that a linear model seems to be a suitable method for capturing variation in the data we are given.

Problem 2:

(a) We are asked to fit a simple linear regression to the data and estimate the associated parameters. We use the estimators provided by the least squares method and we obtain the following results:

$$\hat{\beta}_0 = 10.2$$
$$\hat{\beta}_1 = 4$$

where our estimator is:

$$\hat{\beta}_{[2x1]} = (X'X)^{-1}X'Y$$

Using the above estimates, our confidence intervals for β are:

$$\beta_0 : [\ 8.67\ ,\ 11.73\ ]$$
$$\beta_1 : [\ 2.92\ ,\ 5.08\ ]$$

which were obtained using the formulas:

$$\hat{\beta}_0 \pm t_{n-p-1,\ \alpha/2}\sqrt{\hat{\sigma}^2.(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}})}$$
$$\hat{\beta}_1 \pm t_{n-p-1,\ \alpha/2}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}$$

(c) Assuming that our data satisfies the assumptions of independence of error terms, normal distribution of error terms, and constant variance; and since our data has multiple duplicates, we can use the ANOVA linear lack of fit test to test for the suitability of a linear model. We also report the F-statistic and its associated p-value for the proposed linear model and we carry out a t-test for β_1.

For our proposed ANOVA linear lack of fit test, we start by separating our observations into groups indexed by i, such that each individual j in the ith group has the same covariate value X_i. Next we construct our test statistics as follows:

$$SSR_{Lack\ of\ Fit} = \sum_{i=1}^{n\ of\ groups} \sum_{j=1}^{n\ ind.\ in\ i} (\bar{y}_i - \hat{y}_{ij})^2$$

$$SSPE = \sum_{i=1}^{n\ of\ groups} \sum_{j=1}^{n\ ind.\ in\ i} (y_{ij} - \bar{y}_i)^2$$

$$F_{LOF} = \frac{SSR_{LOF}/(c-p-1)}{SSPE/(n-c)} \sim F_{c-p-1,\ n-c}$$

Where c represents the number of groups. Our F-statistic here is 0.168, and the associated p-value is 0.8492. Based on our ANOVA linear lack of fit test, we fail to reject the null hypothesis of a linear fit where our hypotheses are:

$$H_0 : No\ lack\ of\ fit$$
$$H_a : Lack\ of\ fit$$

We also construct our model's regular ANOVA table:

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	160.0	160.0	72.727	0
Residuals	8	17.6	2.2		

The F-statistic presented in this table represents a comparison between the full and reduced models. Here the F-statistic is associated with a p-value very close to zero, which supports the existence of a linear relationship between x and y based on our hypotheses below:

$$H_0 : c\hat{\beta} = 0$$

$$H_a : c\hat{\beta} \neq 0$$

$$reject\ H_0\ if\ \ F^* > F_{1,\ n-p-1}$$

Finally, we perform a t-test on β_1 with the following hypothesis:

$$H_0 : \hat{\beta}_1 = 0$$

$$H_a : \hat{\beta}_1 \neq 0$$

$$reject\ H_0\ if\quad |t^*|\ >\ t_{n-2,\ \alpha/2}$$

Our t-test produces a t-statistic equal to 8.5280287, which is greater than the critical t-value at $\alpha = 0.05$ which is equal to 2.3060041, therefore we reject our null hypothesis of no existence of a linear relationship between x and y.

Problem 3: Bootstrapping Confidence Intervals

We find that our bootstrap confidence interval [3.176 , 5] is close, although not exactly similar to to the ones obtained under the normality assumption. This is due to the law of large numbers which implies that the empirical distribution obtained from the data will converge to the true distribution given a large enough sample size. This in turn implies that the distribution of parameters obtained by the bootstrap will be a good approximation for the sampling distribution of the parameter, and since the bootstrapped distribution converges in large sample size to the true distribution, the bootstrap confidence intervals will also converge to the true values. They will never be exactly the same because the empirical distribution is not exactly the same as the true distribution, but the difference will decrease as the sample size and the number of draws increases.

Problem 4:

We have previously checked all the linear model assumptions in **Problem 1**. We had also previously both constructed and carried out the Breusch-Pagan and the Brown-Forsyth tests. We will now apply those same tests, this time, however, we will construct multiple different groups for the Brown-Forsyth test. Since the Breusch-Pagan test is independent of any grouping, we will retain the value of the BP statistic from earlier.

	BP	BF1	BF2	BF3	BF4	BF5	BF6
statistic	2.3971753	-1.7688046	-0.7221835	-0.5807704	-1.7688046	-0.9369497	-0.5808102
p-value	0.1215546	0.1148953	0.4907499	0.5773851	0.1148953	0.3762034	0.5773596

We observe that the Brown-Forsyth test yields very variable results based on the choice of group splitting rule. Our rule of thumb choice of having the groups split by value of the independent variable where low values and high values of the independent variable are grouped together, yields the closest result to the Breusch-Pagan test. For all six runs of the Brown-Forsyth tests, we obtain evidence for constant variance, which is expected since the group splitting by extreme values of the dependent variable is naturally expected to provide the greatest change in variance if variance is linearly dependent on X.