National Technical
University of Athens

HOMEWORK 3

Kreativstorm Data Analysis Course

Full Name: Antonios Mitsis

Email: anmitsis@hotmail.com

Group: B

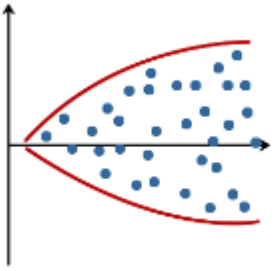
In this homework every question will be answered with the help of R and SPSS. Each line of code that will be used in R will be provided in blue color.

Q1. In your own words, describe what a residual is in linear regression.

A1. In linear regression the we have a set of observations that help us create a model that best predicts the value of the dependent variable based on the independent variables. For each observation we have a predicted value and the real or observed value, the difference of these two values is a residual. In other words, a residual is the error in each prediction of our model.

Q2. If you know that your residual data follow the below pattern, are your data better approximated with a linear model for the lower values of independent variable or higher values of independent variable and why?

A2. Our data is better approximated with a linear model for the lower values of the independent variable as the residuals on the left side are significantly smaller than the residuals on the right side. Smaller residuals means that the predictions of the model are closer to the real values.



Q3. What is the difference between R^2 and adjusted R^2?

A3. R^2 is a metric used to evaluate the goodness of fit of a model. It essentially is the proportion of the variance in the dependent variable that is predictable from the independent variables in the model. The adjusted R^2 has the same role but penalizes the inclusion of unnecessary variables that do not contribute much to the model. Another difference is that R^2 ranges from 0 to 1 whereas adjusted R^2 ranges from $-\infty$ to 1.

Q4. Is there independence of observations if you are trying to predict baby length with mother's height?

- **Yes**
- **No**

A4. Yes

Q5. Justify the above answer.

A5. We first fit the regression model with baby length as dependent variable and mother's height as independent by selecting Analyze -> Regression -> Linear and check the Durbin-Watson in the Statistics menu. The value of the Durbin-Watson statistic is 1.724 which means that there is independence of observations. The value should be between 1.5 and 2.5.

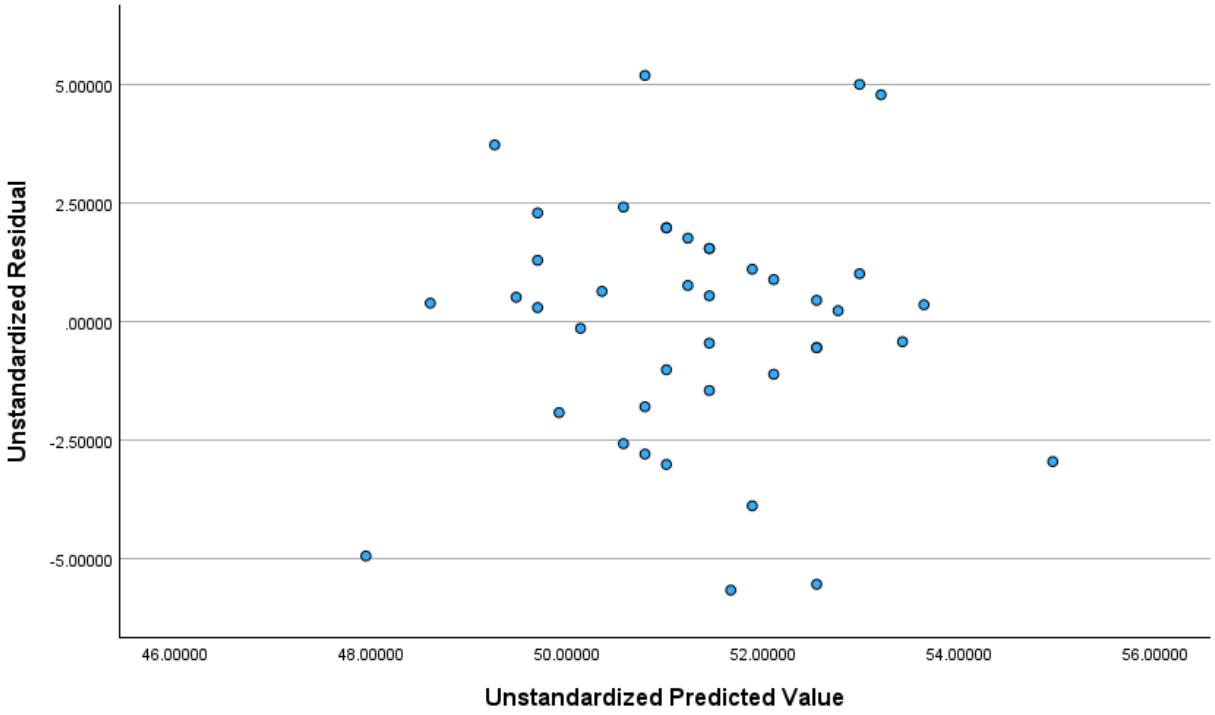
Q6. Do residual data show homoscedasticity?

- **Yes**
- **No**

A6. Yes

Q7. Justify the above answer.

A7. The residuals need to show homoscedasticity, in other words they should be randomly scattered in the plot and not follow any pattern. In the regression save menu we select to save the unstandardized predicted values and residuals and then we scatterplot with residuals on the y axis and predicted values on the x axis. The plot is presented below and we do not find any pattern in the points of the plot, so our residual data shows homoscedasticity.



Q8. What is the value of R^2 and what does this tell you?

A8. The value of R^2 is 0.235 which means that 23,5% of the variance of baby length can be explained by mother's height. This value is relatively small and we need more independent variables to be added to the model in order for it to have a better overall fitness.

Q9. Can you consider the relationship between mother's height and baby length a statistically significant linear relationship and why?

A9. Yes, the relationship of these two variables is statistically significant linear and we can check it in the ANOVA table where the significance value for the regression is 0.01, smaller than 0.05 so it is statistically significant.

Q10. Having the ANOVA table for the linear regression in mind, what is the null and alternative hypothesis in this case?

A10. The regression equation for one independent variable is $y = b_0 + b_1*x$ where y is dependent and x is independent variable, while a and b are coefficients. The null hypothesis is that the

coefficient of the independent variable is 0, or b=0 meaning that the independent variable cannot explain the dependent. The alternative hypothesis is that b≠0 and that the independent variable contributes to the model. Since the significance value is smaller than 0.05, we reject the null hypothesis and the linear regression is statistically significant.

Q11. In your own words, describe what the b_1 is.

A11. The value b_1 is the coefficient of the independent variable in the linear regression model. It explains the change in the dependent variable, for a one unit of measurement change in the independent variable. It is also the slope of the regression line, when talking about one independent variable regression.

Q12. What does the value of b_1 tell you in practical terms?

A12. The coefficient of mother's height, b_1, tells us that if the mother's height increases/decreases by 1cm then the baby's length will increase/decrease by b_1 = 0.219cm.

Q13. Could you claim the same for the mother's height in the range between 140cm and 145cm and why?

A13. No, we cannot claim the same for the mother's height in the range between 140cm and 145cm because it is out of our data's range.

Q14-Q15. According to this model, what is the prediction of baby length for mother's height of 170cm? Report on your findings for predicting baby length with mother's height.

A14-A15. The predicted value for the baby's length for mother's height of 170cm is 52.54772cm. To get this value first we save our variables as x and y because calling them with the $ operator creates some problems. Then we fit the linear model and use the predict function to get the predicted value for x=170. The arguments of the predict function is the fitted model and the new value inside a dataframe.

```
x<-df$mheight
y<-df$Length
l1<-lm(y~x)
predict(l1,newdata = data.frame(x=170))
```

Q16. Can you predict baby length with father's age? Why?

A16. No, we cannot predict baby length with father's age as when we fit the regression model, we see that the significance value of the regression is 0.386, therefore it is not statistically significant and the regression does not provide us with accurate information about our variables.

Q17. What does homogeneity of variance mean and why is it important assumption of an independent t-test?

A17. In the context of an independent t-test, the homogeneity or homoscedasticity of variance is an important assumption as we assume that the variances of the two groups are equal to perform the test. If the variances differ then our results are not reliable make conclusions.

Q18. Is there homogeneity of variance between head circumference for babies of smoking mothers and head circumference for babies of non-smoking mothers?

- **Yes**
- **No**

A18. Yes

Q19. Justify your choice.

A19. To check for homogeneity of variance we will perform Levene's test. In SPSS we select Analyze -> Compare means and proportions -> One-way ANOVA and we select the test for homogeneity of variance in the options menu. The Levene's statistic is 0.828 and the significance value is 0.368. This means that the assumption of homogeneity of variance is correct and we do not have enough evidence to say that variances differ significantly.

Q20. Do smokers have lighter babies? Justify your answer.

A20. To answer this question, we will perform an independent samples t-test. The assumptions for this test are normal distribution and homogeneity of variance, both of which are true. For normal distribution test the significance value is 0.968 and for homogeneity of variance it is 0.584. We are now able to perform the t-test. The birthweight for smoking mothers has shown to be lower (M= 3.51, SD=0.52) than for non-smoking mothers (M=3.13, SD=0.63) with a significance value of 0.43. Therefore, we have enough evidence to claim that smokers have lighter babies.

Q21. Do women over 35 have lighter babies? Justify your answer.

A21. No women over 35 do not have lighter babies. First, we need to compute a new variable that is 0 when mother's age is under 35 and 1 when it is over. To do this we select Transform -> Recode into different variable and select the values we want. After we have the new variable we perform the independent t-test and the significance value is 0.492. This means that the means of the two groups do not differ significantly and we cannot say that women over 35 have lighter babies.

Q22. Using the cholesterol dataset, was the diet effective in lowering cholesterol concentration after 8 weeks of use? Justify your answer.

A22. To answer this question, we will perform a dependent sample t-test. The assumptions on this test are that the difference of the two variables follows the normal distribution. To check this first we compute the difference of the two variables in the compute variable menu and then we select explore and check the normality testing. The significance value of the test is 0.987 so our data are definitely normally distributed. To perform the dependent t-test we select

Analyze -> Compare means and proportions -> Paired samples t-test and select the before and after 8 weeks variables. The cholesterol levels before are significantly higher before (M=6.41, SD=1.19) than after 8 weeks (M=5.78, SD=1.10). The significance value of the test is smaller than 0.01 so the means of the two groups are significantly different, therefore the diet was effective on lowering cholesterol.

Q23. For the above case, what is the null and alternative hypothesis?

A23. The null hypothesis on the above test is that the difference of the means of the two groups is equal to 0, while the alternative hypothesis is that the difference of the means of the two variables is not equal to 0.

Q24. Was the margarine diet more effective in the first 4 weeks of use or the last 4 weeks of use? Justify your answer.

A24. No, the diet was not more effective in the first 4 weeks of use. We create two new variables, the firstdiff and secdiff where the first one is the difference of 4weeks and before and the second one is the difference of 8weeks and 4 weeks. Essentially, we want to see if the mean of the first period is greater (in absolute value) that the mean of the second period. We will again perform a dependent samples t-test. Frist, we check for normality and both variables follow the normal distribution. Then in the same way as the last question we perform the paired samples t-test and the significance value of the test is 0.626. This means that the difference of the means of the two groups is 0 so there is no significance difference in the first 4 weeks or the last 4.

Q25. If you know that the average cholesterol concentration in healthy adults is 3 mmol/L, would you consider your sample (N=18) significantly better or worse than average adult population? Justify your answer.

A25. By performing a one-sample t-test both in the before and after 8 weeks variable we can see that the difference of the mean of the variable and 3 is significantly different. The mean difference of the before variable and 3 is 3.41 with a significance value of less than 0.01 and for the after 8weeks the mean difference is 2.78 with a significance value again smaller than 0.01. Therefore, we can say that our sample is significantly worse than the average adult population.