Statistical Modelling

Sinthu Marksamy

5/20/2022

Table of Contents

# Question 1.

df <- data.frame(x = c(3,2,6,5,3,8,4,2,6,3,2,7),  
 y = c(535, 425, 575, 639, 450, 630, 435, 498, 534, 530, 457, 559))  
  
mody <- lm(y~x, data = df)  
summary(mody)

Call:  
lm(formula = y ~ x, data = df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-80.969 -34.437 0.031 33.999 97.907   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 415.472 34.769 11.949 3.04e-07 \*\*\*  
x 25.124 7.399 3.396 0.00682 \*\*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 51.39 on 10 degrees of freedom  
Multiple R-squared: 0.5356, Adjusted R-squared: 0.4891   
F-statistic: 11.53 on 1 and 10 DF, p-value: 0.00682

anova(mody)

Analysis of Variance Table  
  
Response: y  
 Df Sum Sq Mean Sq F value Pr(>F)   
x 1 30457 30457.0 11.531 0.00682 \*\*  
Residuals 10 26413 2641.3   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Question 1a.

* by looking at the summary output, write down the fitted model.

The fitted model is given below:

## Question 1b.

* Write down the formula to compute the 95% confidence interval for .

The formula to compute the 95% confidence interval for is given below:

confint(object = mody, parm = 'x', level = 0.95)

## Question 1c.

* Compute the 95% confidence interval for .

confint(object = mody, parm = 'x', level = 0.95)

## 2.5 % 97.5 %  
## x 8.63876 41.60994

* The estimated 95% confidence interval is **(8.64, 41.6)**.

## Question 1d.

Fill in the blanks in the table.

* Residual (DF): = 10
* SSE = 15366
* MSE = = 1536.6
* F-Value = = = 10.4509306

## Question 1e

### Question 1e(i)

* Write down the null hypothesis, that there is no effect on mean sales from increasing the amount of shelf space, versus a suitable alternative hypothesis.

The test null hypothesis is stated below:

An appropriate alternative hypothesis is stated below:

### Question 1e(ii)

* Compare the above value of F with the table value of .

The tabulate value for is 10.044.

### Question 1e(iii)

* Comment on your findings.

The computed F-value exceeds the tabulated F-statistic. This implies that the null hypothesis cannot be accepted at 1% level of significance.

# Question 2.

## Question 2a.

* Looking at the value of above, is this linear model a reasonable fit?
* The computed is 0.9277. This implies that the linear model is a reasonable fit.

## Question 2b.

* Viewing the residual plot, is there a possible problem with the constancy of variance? Give reasons for you answer.
* There is a possible problem with the constancy of variance as the values of the independent variable changes, the residual changes. This is obvious for the higher values of the independent variable where the residual drifted higher as compared with other values.

## Question 2c.

* Using the Q-Q plot and the Shapiro-Wilk test, check if there is a possible problem with the assumption of normality.
* The null hypothesis of the Shapiro-Wilk test is that the distribution of the sample is normal. However, since the p-value of the test is lower than 0.05, it could be concluded that the sample i.e. stdres1 is not normal. Likewise, it’s evident from the Q-Q plot as many of the points drifted away from the straight line. Thus, the residual failed the normality test.

## Question 2d.

* Looking at the plots above, is there any other transformation that you would like to consider? Give reasons for your answer.
* I would like to consider the logarithm transformation. This is to standardize the variation in the value. Also, this could ensure constancy of the residual variance.

# Question 4

## Question 4a

* The forward method for selecting the optimal set of predictors can use the AIC ot BIC. Describe the model selection method, including the definition of the AIC and BIC.
* AIC and BIC are two criteria that could be used to estimate how well a model fit a dataset such that the larger difference in either AIC or BIC indicates stronger evidence for one model over the other (the lower the better)the lower the AIC or BIC the better the fit.
* The procedure for a forward stepwise model is such that starts from the null model and adds a variable that improves the model the most, one at a time, until the stopping criterion is met.

## Question 4b

* The model corresponds to the AIC criteria
* Whereas the model corresponds to the BIC criteria. This is because unlike the AIC, the BIC penalizes free parameters more strongly.

## Question 4c

* Write down the model that corresponds to the R command below.

fit.full <- lm(mpg ~ ., data = mtcars)  
forward\_aic <- step(lm(mpg~ 1, data = mtcars), direction = "forward", scope = formula(fit.full), k = 2, trace = 0)  
  
summary(forward\_aic)

The corresponding equation is:

## Question 1d

* In the command above, what does k mean?
* k is the multiple of the number of degrees of freedom used for the penalty. Only k = 2 gives the genuine AIC: k = log(n) is sometimes referred to as BIC or SBC.

## Question 4e.

* By looking at the output above, which procedure the statistician has employed. Describe the procedure. State the reason as to why the procedure stops at the point it does.
* The statistician employed a backward stepwise technique.
* The procedure goes thus:
  + Begins with a model that contains all variables under consideration (called the Full Model)
  + Then starts removing the least significant variables one after the other
  + Until a pre-specified stopping rule is reached or until no variable is left in the model.
* The procedure stops because the AIC can no longer be improved upon.

## Question 4f

* Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.
* Multicollinearity is a problem because it undermines the statistical significance of an independent variable. Other things being equal, the larger the standard error of a regression coefficient, the less likely it is that this coefficient will be statistically significant.

## Question 4g

* We have calculated vif (variance inflation factor). State what conclusion we can draw with respect to collinearity of predictors in the model.
* The vif estimates of the predictors are all less than 10. Thus, going by a rule-of-thumb that vif estimates less than 10 doesn’t imply the presence of collinearity. It can be concluded that there’s no presence of multicollinearity in the reduced model.

# Question 5

## Question 5a

* By looking at the R output, state whether one should include extra parameters in the model.
* The analysis of variance table suggested that the extra parameters in the model are necessary as the residual for the model 2 is much lower than that of model 1. Also, the analysis of variance table shows that there’s a significant difference between the two models.

## Question 5b

* Let us consider just Hip, Fore-arm, and Wrist as our predictors. How many possible linear models that predict BodyFat can one build?
* We can build at least 4 models with the predictors.

## Question 5c

* Consider the two models:

and

* 1. Interpret the coefficient for Male in each model.
* In the first model, the coefficient of Male i.e. -329.56 implies that the salary of a male individual would be 329.56 lower than that of a female individual.
* However, in the second model the coefficient of Male i.e. -0.321 implies that the salary of a male individual would be 0.321 lower than that of a female individual.
  1. Would it be correct to say that the second model is preferred over the first? Explain you reasoning.
* Yes, it would be correct to say that the second model is preferred over the first as the estimate improved and also the estimated variance i.e.  reduced significantly.
  1. Consider another model

Is model (3) better than model (2)? why?

* Yes, model 3 is better than model 2. This is because the estimate of model 3 improved significantly and also the estimated variance of model 3 i.e.  reduced.