STAT332

2023

Assignment 1

Ruben Traicevski 6790021

1. (Note, for the hand written answers, I submitted in a separate PDF file to keep the scanned image clear)

A simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i\,, \qquad i = 1,2,\ldots,n,$$

In matrix format will have the form of

$$Y_{n\,x\,1} = X_{n\,x\,2}\beta_{2\,x\,1} + e_{n\,x\,1}$$

,which is a special case of multiple linear regression when there is only 1 explanatory variable.

Thus, we will have the following below.

$$Design\ Matrix: X_{n\,x\,2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

$$Vector\ of\ Parameters: \beta_{2\,x\,1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$Error\ Term\ Vector: e_{n\,x\,1} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \sim^{iid} N(0, \sigma^2)$$

$$Response\ Vector: Y_{n\,x\,1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$n = \#\ of\ observations, \qquad and\ p = \#\ of\ explanatory\ variables\ (In\ our\ case = 1)$$

$$And\ assuming\ that\ E[e_{n\,x\,1}] = Zero\ Vector, \qquad then$$

$$Var(e_{n\,x\,1}) = E[e_{n\,x\,1}\, e_{n\,x\,1}{}^T] = \begin{bmatrix} \sigma^2{}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2{}_n \end{bmatrix} = \sigma^2 I_n$$

2.

For multiple linear regression in its matrix form we have,

$$Y = X\beta + e$$

The RSS is given by,

$$RSS(\beta) = \; e^T e = (Y - X\beta)^T(Y - X\beta)$$

$$= Y^T Y - X\beta Y^T - X^T\beta^T Y + X^T\beta^T X\beta, \qquad As\ (A+B)^T = A^T + B^T$$

Then using the identity, $(X^T X)\beta = X^T Y$,

$$RSS(\beta) = Y^T Y - 2X^T\beta^T Y + X^T\beta^T Y = Y^T Y - X^T\beta^T Y$$

Which in our case, changes to

$$RSS(\beta) = Y^T Y - \mathbf{1^T}{\beta_0}^T Y$$

As our design matrix degenerates to a column vector of 1s (n x 1)

Likewise, the Sum of Squares due to Regression changes from

$$\beta^T X^T Y - n\overline{y}^2$$

to,

$${\beta_0}^T \mathbf{1^T} Y - n\overline{y}^2$$

Thus, if we use the relation that,

$$SS_{reg} = SS_T - RSS$$

Then,

$$SS_{reg} + RSS = SS_T$$

$${\beta_0}^T \mathbf{1^T} Y - n\overline{Y}^2 + Y^T Y - \mathbf{1^T}{\beta_0}^T Y = Y^T Y - n\overline{y}^2$$

Note that also our matrix of estimators reduces to a single coefficient, hence ${\beta_0}^T = \beta_0$

3.

a)

β_0	β_1	β_2
8.0136	-1.3548	0.4626
Pr(>|t|) = 5.62e-05	Pr(>|t|) += 0.0271	Pr(>|t|) = 0.3039

F - Statistic	Multiple R - Squared	P value	Residual SE
3.668	0.3284	0.05049	1.999

The intercept, β_0, was found to be 8.0136 while β_1 was -1.3548 and β_2 was 0.4626.

Approximately 32.84% of the variation of the Profit (Dependent Variable) is being explained by the Contract Size and Supervisor Experience (Independent Variables).

On average the true values are approximately 1.999 points away from the predicted values by the model.

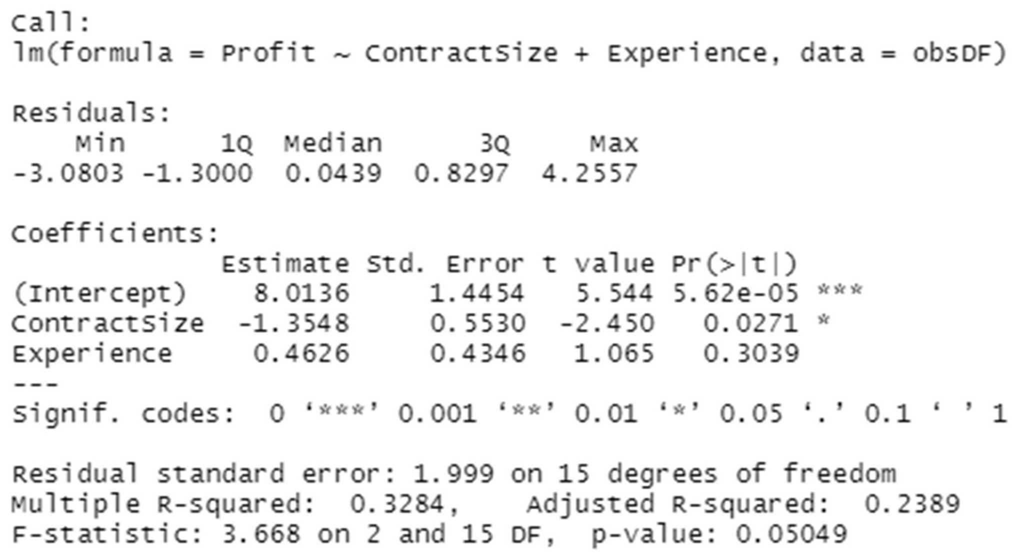
The F – Statistic is not equal to 1, but is close to. The P – value is almost exactly on the bound for a 95% significance level.

Also note that the Pr(>|t|) for the β_2 coefficient is 0.3039 which is significantly greater than 0.05, indicating that Experience does not have a statistically significant relationship with Profit. This may mean it is advisable to remove the coefficient from the model.

All together this is indicating that Profit is 'dependent' on contract size, not so much on Supervisor Experience, alongside that there is an inverse relationship, indicating a bigger contract size will reduce profits.

SOURCE CODE

```
Regression <- lm(Profit ~ ContractSize + Experience, data = obsDF)

summary(Regression)

confint(Regression, level =0.95)

DataPoint = data.frame(ContractSize = 3.0, Experience = 5 )

predict(Regression, DataPoint, interval="predict")

res = residuals(Regression)

shapiro.test(res)

autocorrelation <- acf(ContractSize, lag.max=20, plot=TRUE)

install.packages("lmtest")

library(lmtest)

dwtest(formula = Regression)
```



```
Call:
lm(formula = Profit ~ ContractSize + Experience, data = obsDF)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0803 -1.3000  0.0439  0.8297  4.2557 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    8.0136     1.4454   5.544 5.62e-05 ***
ContractSize  -1.3548     0.5530  -2.450   0.0271 *  
Experience     0.4626     0.4346   1.065   0.3039    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.999 on 15 degrees of freedom
Multiple R-squared:  0.3284,	Adjusted R-squared:  0.2389 
F-statistic: 3.668 on 2 and 15 DF,  p-value: 0.05049
```

b)

The coefficient β_1 which pertains to Contract Size fell within the 95% confidence interval of (-2.5335, -0.17603), according the output.

Computing this manually would require the following equation.

$$\widehat{\beta_1} \pm se(\widehat{\beta_1})t_{\frac{0.95}{2},(18-2-1)}$$

Where,

$$se(\widehat{\beta_1}) = \sqrt{\widehat{Var}(\widehat{\beta_1})}$$

c)

$\tilde{y}^*$	x_1^*	x_2^*
6.26	3.0	5.0

Lower	Upper
1.598	10.92509

This value can be either given from the following R code

DataPoint = data.frame(ContractSize = 3.0, Experience = 5)

predict(Regression, DataPoint, interval="predict")

or using the following equation

$$\tilde{y}^* \pm \widehat{sepred}(\tilde{y}^*|x^*)t_{\frac{0.95}{2},(18-2-1)}$$

Where,

$$\widehat{sepred}(\tilde{y}^*|x^*) = \hat{\sigma}\sqrt{(1 + x^{*T}(X^TX)^1x^*)}$$

Even though our response (Y) may be unobservable, we can make a prediction, $\tilde{y}^*$, given some input, $x_1^*, and\ x_2^*$, by using the linear model. However, this prediction comes along with it associated standard error. Our 95% prediction interval yielded, (1.598, 10.92509) which in our case has a large margin of 9.327 points, or 6.26 $\pm$ 4.662 points. Considering the context of the data, ideally, we would like the interval to be narrower, considering some actual data is $\sim |\pm 4.662|$.

d)

For the model to satisfy all the assumptions of using the OLS method, while regression is very robust, the data cannot violate the following

i) There is no collinearity, i.e the variance of the residuals is independent of one another, $Var(e) = \sigma^2 I_n$

ii) There is homogeneity of the variance.

iii) $E[e] = 0$

iv) $e \sim^{iid} N(0, \sigma^2)$

Using R, we can immediately perform tests such as the Shapiro Wilk Test for normality of residuals and the Durbin Watson test for one step auto correlation.

A Shapiro Wilk Test returned a P value $= 0.683 > 0.05$ significance level, meaning that the null hypothesis that the residuals are normally distributed is not rejected.

The Durbin Watson test returned a value of $2.0407 \sim 2$, and a p value of $0.6863 > 0.05$ significance level, which means the null hypothesis is not rejected, indicating that the residuals are uncorrelated (one step).

Bartlett's test returned a P – value $= 0.3392 > 0.05$ significance level, meaning we do not reject the null hypothesis that our two data groups (ContractSize , Experience) are homogenous in variance.

Taking the expected value of the residuals gave a value of -1.110223e-15 $\sim$ 0 which is extremely close to the OLS model assumption.

Therefore, combining these tests together points towards the conclusion that the model is valid regarding the assumptions for an OLS estimator. This means that the method of producing OLS estimator's is applicable to such data.