MULTIPLE LINEAR REFRESSION MODEL

Mohammed Swaned M

05-12-2022

Contents

Contents

Name : Mohammed Swaned M

Enrollment_no: GN4994

Faculty_no : 22DSMSA104

1 Multiple Linear Regression Model

1.1 Introduction

The Multiple linear regression model is defined as $y = \beta_0 + \beta_1 \times X_1 + \cdots + \beta_p X_p + \epsilon$

Now we consider the wage12 data and the multiple linear regression model for that data is given as

$wage = \beta_0 + \beta_1 \times education + \beta_2 \times tenure + \beta_3 \times nonwhite + \beta_4 \times female + \beta_5 \times married + \epsilon$

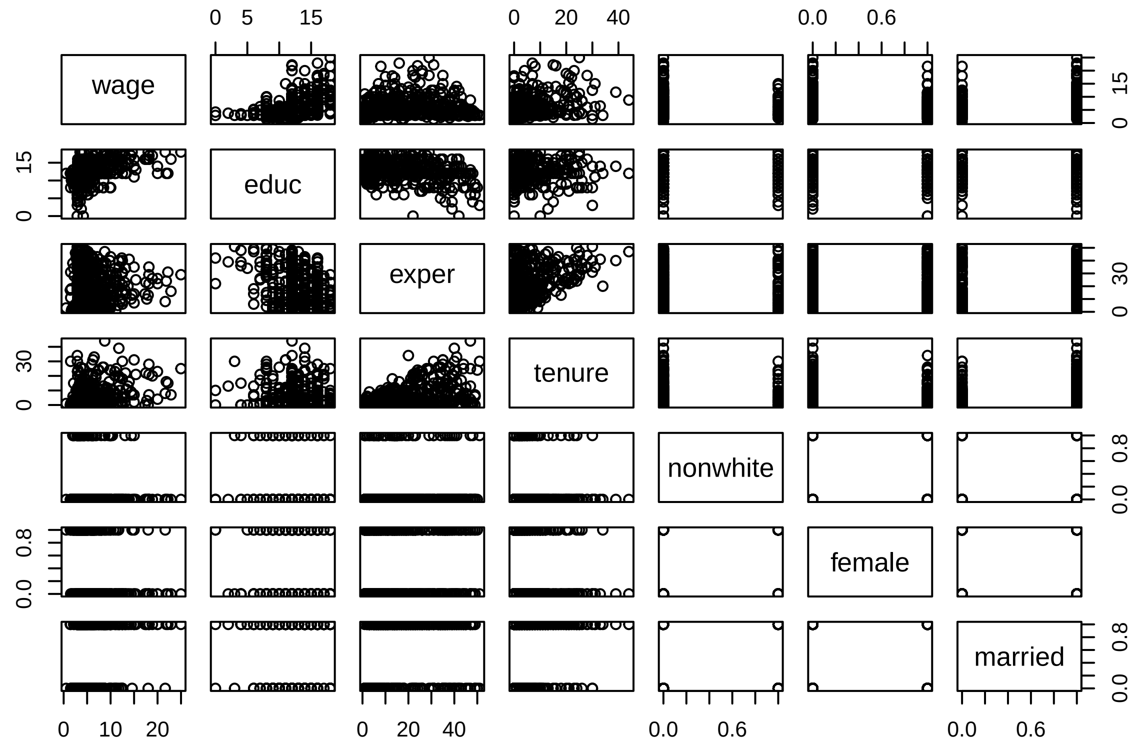
Now read data

```
wage12<-wooldridge::wage1[c(1,2,3,4,5,6,7)]
head(wage12)
```

```
##   wage educ exper tenure nonwhite female married
## 1 3.10   11     2      0        0      1       0
## 2 3.24   12    22      2        0      1       1
## 3 3.00   11     2      0        0      0       0
## 4 6.00    8    44     28        0      0       1
## 5 5.30   12     7      2        0      0       1
## 6 8.75   16     9      8        0      0       1
```

1.2 Make the scatter plot

```
pairs(wage12)
```



From the above plot, we may see that there is a linear relationship between TV and sales, radio“andsales’ and newspaper and sales. To figure out more we obtain the correlation coefficient among the variables.

1.3 Calculate the correlation coefficient

```
cor(wage12)
```

```
##                 wage        educ       exper      tenure    nonwhite      female     married
## wage      1.00000000  0.40590333  0.11290344  0.34688957 -0.03851959 -0.34009786  0.22881718
## educ      0.40590333  1.00000000 -0.29954184 -0.05617257 -0.08465433 -0.08502941  0.06888104
## exper     0.11290344 -0.29954184  1.00000000  0.49929145  0.01435563 -0.04162597  0.31698428
## tenure    0.34688957 -0.05617257  0.49929145  1.00000000  0.01158880 -0.19791027  0.23988874
## nonwhite -0.03851959 -0.08465433  0.01435563  0.01158880  1.00000000 -0.01091747 -0.06225929
## female   -0.34009786 -0.08502941 -0.04162597 -0.19791027 -0.01091747  1.00000000 -0.16612843
## married   0.22881718  0.06888104  0.31698428  0.23988874 -0.06225929 -0.16612843  1.00000000
```

1.4 Fitting of multiple Linear Regression Model

To estimate the coefficients of the variables educ,exper,tenure,nonwhite,female and married, we fit the following model

```
M1=lm(wage~educ+exper+tenure+nonwhite+female+married, data = wage12)
summary(M1)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure + nonwhite + female +
##     married, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6716 -1.8239 -0.4967  1.0403 13.9209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.60221    0.73107  -2.192   0.0289 *
## educ         0.55510    0.05006  11.090  < 2e-16 ***
## exper        0.01875    0.01204   1.557   0.1201
## tenure       0.13883    0.02116   6.562 1.29e-10 ***
## nonwhite    -0.06581    0.42657  -0.154   0.8775
## female      -1.74241    0.26682  -6.530 1.57e-10 ***
## married      0.55657    0.28674   1.941   0.0528 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.952 on 519 degrees of freedom
## Multiple R-squared:  0.3682, Adjusted R-squared:  0.3609
## F-statistic: 50.41 on 6 and 519 DF,  p-value: < 2.2e-16
```

1.5 Checking Significance of Model

To check the significance of the model, we check F statistic and for that we set the hypotheses as follows: Null Hypothesis: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Alternative Hypothesis: $H_1 : Atleast\ one\ \beta_i \neq 0, i = 1, 2, \cdots, k$. F-statistic: 50.41 on 6 and 519 DF, p-value: $< 2.2e-16$ Since the F statistic is 50.41 on 6 and 519 df with p-value < 2.2e-16 i.e. almost zero. Hence we *reject the null hypothesis* that means there is at least oneβ_i that is not equal to zero. Therefore, our model is significant.

1.6 Checking Significance of Variables

To check the significance of variable(s), we check the $t-ratios$ and its corresponding $p-values$. We set the hypotheses as follows: $H_0 : \beta_i = 0\ \ VsH_1 : \beta_i \neq 0\ where\ i = 0, 1, 2, 3$ Now we check the t-statistics and p value of the corresponding variables one by one and decide that which one is significant. SO the p-value for *intercept* term is .0289 that is almost zero, hence we reject the null hypothesis and accept that $\beta_0 \neq 0$. The p-value of *Education* is 2×10^{-16}that is almost zero, hence we reject the null hypothesis and accept that $\beta_1 \neq 0$ The p-value of *Tenure* is 1.29×10^{-10} that is almost zero, hence we reject the null hypothesis and accept that $\beta_3 \neq 0$. The p-value of *Female* is 1.57×10^{-10} that is almost zero, hence we reject the null hypothesis and accept that $\beta_5 \neq 0$. The p-value of *Experience* is 0.1201 that is greater than 0.05, hence we fail to reject the null hypothesis and accept that $\beta_3 = 0$. The p-value of *Nonwhite* is 0.8775 that is greater than 0.05, hence we fail to reject the null hypothesis and accept that $\beta_2 = 0$. The p-value of *married* is 0.0525 that is greater than 0.05, hence we fail to reject the null hypothesis and accept that $\beta_6 = 0$. This means that the variables *Experience,Nonwhite and Married*is not significant in this model. Since these variables is not significant, so we remove these variable from the model. Hence our final model is: $wage = \beta_0 + \beta_1 \times education + \beta_3 \times tenure + \beta_5 \times female + \epsilon$

```r
M2=lm(wage~educ+tenure+female,data=wage12)
summary(M2)
```

```
##
## Call:
## lm(formula = wage ~ educ + tenure + female, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5184 -1.8074 -0.4477  1.0270 14.1229
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84503    0.64774  -1.305    0.193
## educ         0.53799    0.04709  11.425  < 2e-16 ***
## tenure       0.16441    0.01835   8.962  < 2e-16 ***
## female      -1.78839    0.26559  -6.734 4.38e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.968 on 522 degrees of freedom
## Multiple R-squared:  0.3577, Adjusted R-squared:  0.354
## F-statistic: 96.88 on 3 and 522 DF,  p-value: < 2.2e-16
```

Here,The p-value of *intercept* is 0.193 that is greater than 0.05, hence we fail to reject the null h*ypothesis and accept that* $\beta_0 = 0$. *since* intercept* is not significant,so we remove the variable from the model Hence our final model is: $wage = \beta_1 \times education + \beta_3 \times tenure + \beta_5 \times female + \epsilon$

```r
M3=lm(wage~educ+tenure+female-1,data=wage12)
summary(M3)
```

```
##
## Call:
## lm(formula = wage ~ educ + tenure + female - 1, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4129 -1.8273 -0.6037  0.9576 14.0708
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## educ    0.48004    0.01563  30.706  < 2e-16 ***
## tenure  0.15836    0.01776   8.916  < 2e-16 ***
## female -1.89775    0.25219  -7.525 2.32e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.97 on 523 degrees of freedom
## Multiple R-squared:  0.8187, Adjusted R-squared:  0.8176
## F-statistic:   787 on 3 and 523 DF,  p-value: < 2.2e-16
```

1.7 Adequacy of Model

1.7.1 Residual Standerd Error

$RSE = \sqrt(\frac{RSS}{n-p-1})$

where $RSS = \sqrt(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and is called residual sum of squares(RSS)

```
wage=wage12$wage
educ=wage12$educ
tenure=wage12$tenure
female=wage12$female
wagehat=M3$coefficients[1]*educ+M3$coefficients[2]*tenure+M3$coefficients[3]*female
#alternaively
wagehat2=M3$fitted.values
residManual=wage-wagehat
rssMan=sum(residManual^2)
rssMan
```

```
## [1] 4614.452
```

```
n=length(educ)
p=3
rseman=sqrt(rssMan/n-p-1)
rseman
```

```
## [1] 2.184656
```

```
resid=M3$residuals
rss=sum(resid^2)
rss
```

```
## [1] 4614.452
```

```
n=length(wage12$educ)
p=3
rse=sqrt(rss/n-p-1)
rse
```

```
## [1] 2.184656
```

Now Calculate RSE from the rectified model

```
resid2=M3$residuals
rss2=sum(resid2^2)
rss2
```

```
## [1] 4614.452
```

```
p=3
rse2=sqrt(rss/n-p-1)
rse2
```

```
## [1] 2.184656
```

1.7.2 R Squared

Multiple R-squared: 0.3682, Adjusted R-squared: 0.3609. The reported R squared is 0.3682 that is approximately 0.37. So we see that 37% variability of wage is explained by education, experience, tenure,nonwhite,female and married and the adjusted R- squared is 0.8956 when insignificant variables (experience,nonwhite and married) is attached. After omitting these insignificant variables from the model and we examine the R-squared and adjusted R squared. They are as follows: Multiple R-squared: 0.3577, Adjusted R-squared: 0.3544.Here the intercept is not a significant variable .After ommitting the insignificant variables from the variables and we examine the R-squared and adjusted R-squared.They are as follows: Multiple R-squared: 0.8187, Adjusted R-squared: 0.8176.We find that there is no difference in R squared but adjusted R-squared has increased a little bit. That is theevidence that if we remove any insignificant variable from the model, them adjusted R-squared increased.

1.8 Dummy Variable

1.8.1 Female as dummy variable

Here we model the data as follows: $Wage = \beta_0 + \beta_1 \times female$

```
M4=lm(wage~female,data = wage12)
summary(M4)
```

```
##
## Call:
## lm(formula = wage ~ female, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5995 -1.8495 -0.9877  1.4260 17.8805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
## female       -2.5118     0.3034  -8.279 1.04e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 524 degrees of freedom
## Multiple R-squared:  0.1157, Adjusted R-squared:  0.114
## F-statistic: 68.54 on 1 and 524 DF,  p-value: 1.042e-15
```

Table displays the coefficient estimates and other information associated with the model. So the Model of the data as follows: $Wage = 7.0995 - 2.5118 \times female$ However, we notice that the p-value for the dummy variable that is 1.042×10^{-15}. This means that the variable *female* is significant.

1.8.2 Nonwhite as dummy variable

Here we model the data as follows: $Wage = \beta_0 + \beta_1 \times nonwhite$

```
M5=lm(wage~nonwhite,data = wage12)
summary(M5)
```

```
##
## Call:
## lm(formula = wage ~ nonwhite, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.414 -2.526 -1.259  1.026 19.036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9442     0.1700  34.961   <2e-16 ***
## nonwhite     -0.4682     0.5306  -0.882    0.378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.694 on 524 degrees of freedom
## Multiple R-squared:  0.001484,   Adjusted R-squared:  -0.0004218
## F-statistic: 0.7786 on 1 and 524 DF,  p-value: 0.378
```

Table displays the coefficient estimates and other information associated with the model. So the Model of the data as follows: $Wage = 5.9442 - 0.4682 \times nonwhite$ However, we notice that the p-value for the dummy variable is high that is 0.378 This means that the variable *nonwhite* is insignificant. ### Married as dummy variable Here we model the data as follows: $Wage = \beta_0 + \beta_1 \times married$

```
M6=lm(wage~married,data = wage12)
summary(M6)
```

```
##
## Call:
## lm(formula = wage ~ married, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.144 -2.181 -1.094  1.406 18.407
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8439     0.2507  19.320  < 2e-16 ***
## married       1.7296     0.3214   5.381 1.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.599 on 524 degrees of freedom
## Multiple R-squared:  0.05236,    Adjusted R-squared:  0.05055
## F-statistic: 28.95 on 1 and 524 DF,  p-value: 1.121e-07
```

Table displays the coefficient estimates and other information associated with the model. So the Model of the data as follows: $Wage = 4.8439 + 1.7296 \times married$ However, we notice that the p-value for the dummy variable is 1.12×10^{-7} This means that the variable *married* is significant.