**Statistical Data Mining – Homework 1 & 2**

**Submitted By: Sheetal Aggarwal**

From the “Catalog” data:

1. **Create a column for single mothers (gender = 0, married = 0, children = 1) and a second column for single fathers (gender = 0, married = 0, children >= 1). Perform a regression using these two columns, as well as Salary and Catalogs to predict AmountSpent. Are either of these new columns statistically significant in predicting AmountSpent? Are they positively or negatively related?**

**Solution:** columns single mother and single father are very less statistically significant when the children >=1 as the p value for both the columns are 0.0449 and 0.0307 respectively which is very near to 0.05 and both the columns have negative coefficients so they are negatively related.

**#Code**

Catalog = read.csv(file.choose())

head(Catalog)

tail(Catalog)

Catalog$SingleMothers = ifelse(Catalog$Gender==0 & Catalog$Married==0 & Catalog$Children>=1 , 1 , 0)

Catalog$SingleFathers = ifelse(Catalog$Gender==1 & Catalog$Married==0 & Catalog$Children>=1 , 1 , 0)

head(Catalog)

CatalogRegr = lm(AmountSpent ~ SingleMothers + SingleFathers + Salary + Catalogs , data = Catalog)

summary(CatalogRegr)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.748e+02 6.363e+01 -9.033 <2e-16 \*\*\*

**SingleMothers -1.114e+02 5.545e+01 -2.008 0.0449 \***

**SingleFathers -1.533e+02 7.085e+01 -2.164 0.0307 \***

Salary 1.916e-02 7.026e-04 27.268 <2e-16 \*\*\*

Catalogs 5.108e+01 2.918e+00 17.504 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 597.7 on 995 degrees of freedom

Multiple R-squared: 0.6148, Adjusted R-squared: 0.6133

F-statistic: 397.1 on 4 and 995 DF, p-value: < 2.2e-16

1. **Consider the “Catalog” data. Are the columns Salary and Catalogs correlated**?

**Solution:** The correlation value of salary and catalog column is 0.18355 which is near to 0. This implies that both the columns do not have significant correlation.

**#Code**

Catalog= read.csv(file.choose())

tail(Catalog)

cor(Catalog$Salary, Catalog$Catalogs)

**Output:**

cor(Catalog$Salary, Catalog$Catalogs)

[1] 0.1835509

1. **Find the three columns such that when you perform multiple regression you end up with the minimal model standard error. What three columns of data did you use and what was the model standard error?**

**Solution:** After performing the multiple regression, the minimal model standard error for three columns Close, Salary and Catalog is 556.1

**#Code: For multiple regression**

Catalog = read.csv(file.choose())

head(Catalog)

tail(Catalog)

CatalogRegr = lm(AmountSpent ~ Customer + Age + Gender + Married + Close + Children +PrevCust + PrevSpent + Salary + Catalogs + OwnHome + AmountSpent , data = Catalog)

summary(CatalogRegr)

**## for selected three columns**

CatalogRegr = lm(AmountSpent ~ Close + Salary + Catalogs, data = Catalog)

summary(CatalogRegr)

Call:

lm(formula = AmountSpent ~ Customer + Age + Gender + Married +

Close + Children + PrevCust + PrevSpent + Salary + Catalogs +

OwnHome + AmountSpent, data = Catalog)

Residuals:

Min 1Q Median 3Q Max

-1946.04 -313.62 -37.67 229.50 2848.22

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.851e+02 7.937e+01 2.332 0.0199 \*

Customer -1.173e-02 5.399e-02 -0.217 0.8281

Age 6.288e-01 1.111e+00 0.566 0.5713

Gender -4.001e+01 3.289e+01 -1.217 0.2240

Married -1.564e+01 4.248e+01 -0.368 0.7129

**Close** -**4.044e+02 3.843e+01 -10.523 < 2e-16 \*\*\***

Children -1.517e+02 1.781e+01 -8.517 < 2e-16 \*\*\*

PrevCust -5.413e+02 5.427e+01 -9.974 < 2e-16 \*\*\*

PrevSpent 2.903e-01 4.470e-02 6.496 1.31e-10 \*\*\*

**Salary 1.765e-02 9.924e-04 17.789 < 2e-16 \*\*\***

**Catalogs**  **4.236e+01 2.452e+00 17.280 < 2e-16 \*\*\***

OwnHome 3.382e+01 3.546e+01 0.954 0.3404

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 490.1 on 988 degrees of freedom

Multiple R-squared: 0.7428, Adjusted R-squared: 0.7399

F-statistic: 259.4 on 11 and 988 DF, p-value: < 2.2e-16

**Output: for selected 3 columns**

Call:

lm(formula = AmountSpent ~ Close + Salary + Catalogs, data = Catalog)

Residuals:

Min 1Q Median 3Q Max

-1668.99 -323.88 2.81 291.05 3052.60

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.616e+02 5.882e+01 -4.447 9.69e-06 \*\*\*

Close -4.978e+02 3.915e+01 -12.714 < 2e-16 \*\*\*

Salary 2.037e-02 5.857e-04 34.782 < 2e-16 \*\*\*

Catalogs 4.691e+01 2.728e+00 17.195 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Residual standard error: 556.1** on 996 degrees of freedom

Multiple R-squared: 0.6662, Adjusted R-squared: 0.6652

F-statistic: 662.7 on 3 and 996 DF, p-value: < 2.2e-16

1. **Create a new customer who is of average age, average Salary, and has received the average number of catalogs. What is the predicted AmountSpent?**

**Solution:** The predicted amount spent for new customer is 1216.77

**#Code**

Catalog = read.csv(file.choose())

head(Catalog)

NAge=mean(Catalog$Age)

NSalary = mean(Catalog$Salary)

NCatalog = mean(Catalog$Catalogs)

CatalogRegr=lm(AmountSpent ~ Age + Salary + Catalogs, data = Catalog)

New\_customer<- data.frame(Age=NAge,Salary=NSalary,Catalogs=NCatalog)

predict(CatalogRegr,New\_customer)

summary(CatalogRegr)

**Output:** predict(CatalogRegr,New\_customer)

1

1216.77

1. **Create a new customer who is female, married, has no children, has a salary of $75,000, and does not live close. What is the predicted AmountSpent?**

**Solution:** The predicted amount spent for new customer is 2297.903

NewCustomer<- data.frame(Gender = 0, Married= 1, Children= 0, Salary = 75000, Close= 0 )

CatalogRegr=lm(AmountSpent ~ Gender + Married + Children + Salary + Close, data = Catalog)

predict(CatalogRegr,NewCustomer)

summary(CatalogRegr)

**Output:** predict(CatalogRegr,NewCustomer)

1

2297.903

1. **Create a new customer who is a 50-year old male, not married, with a salary of $75,000 and does not live close. What is the predicted AmountSpent?**

**Solution:** The predicted amount spent for new customer is 2101.558

NewCustomer<- data.frame(Gender = 1, Age= 50, Married= 0, Salary = 75000, Close= 0 )

CatalogRegr=lm(AmountSpent ~ Gender + Age + Married + Salary + Close, data = Catalog)

predict(CatalogRegr,NewCustomer)

summary(CatalogRegr)

**Output:** predict(CatalogRegr,NewCustomer)

1

2101.558

From the “BankData” data:

1. Perform a regression using EducLev, JobGrade, Gender, and YrsPrior to predict Salary. Which of those columns is significant?

**Solution:** Significant columns are JobGrade3, JobGrade4, JobGrade5, JobGrade6 as they have very low p values.

**#Code**

BankData = read.csv(file.choose())

head(BankData)

tail(BankData)

BankData$EducLev= as.factor(BankData$EducLev)

BankData$JobGrade = as.factor(BankData$JobGrade)

Reg1 = lm(Salary ~ EducLev + JobGrade + Gender+ YrsPrior, data= BankData)

summary(Reg1)

**Outpu**t: Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 66106.4 2361.3 27.996 < 2e-16 \*\*\*

EducLev2 -4143.9 3024.9 -1.370 0.1723

EducLev3 -3006.2 2720.8 -1.105 0.2706

EducLev4 -5038.8 5147.4 -0.979 0.3288

EducLev5 -1866.1 3172.1 -0.588 0.5570

JobGrade2 4433.4 2595.6 1.708 0.0892 .

**JobGrade3 12959.4 2738.2 4.733 4.23e-06 \*\*\***

**JobGrade4 23321.7 3107.9 7.504 2.11e-12 \*\*\***

**JobGrade5 35355.5 3780.8 9.351 < 2e-16 \*\*\***

**JobGrade6 71178.9 4337.4 16.410 < 2e-16 \*\*\***

GenderMale 922.6 2106.3 0.438 0.6619

YrsPrior 318.0 293.3 1.084 0.2796

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12450 on 196 degrees of freedom

Multiple R-squared: 0.7104, Adjusted R-squared: 0.6941

F-statistic: 43.7 on 11 and 196 DF, p-value: < 2.2e-16

1. **Perform a regression using all the input (except Employee – column 1). How much less do females make in Salary than males?**

**Solution:** The coefficient of GenderMale is 5108.95 this implies that females make 5108.95 amount of less salary than males.

**#Code**

BankData = read.csv(file.choose())

head(BankData)

tail(BankData)

BankData$EducLev= as.factor(BankData$EducLev)

BankData$JobGrade = as.factor(BankData$JobGrade)

Reg2 = lm(Salary ~ . - Employee, data= BankData)

summary(Reg2)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 150528.82 14412.72 10.444 < 2e-16 \*\*\*

EducLev2 -971.10 2797.31 -0.347 0.7289

EducLev3 1055.83 2715.04 0.389 0.6978

EducLev4 570.35 4809.45 0.119 0.9057

EducLev5 5381.60 3241.78 1.660 0.0985 .

JobGrade2 3128.99 2371.54 1.319 0.1886

JobGrade3 10438.72 2524.79 4.134 5.30e-05 \*\*\*

JobGrade4 17189.67 2992.04 5.745 3.53e-08 \*\*\*

JobGrade5 27318.82 3748.54 7.288 7.86e-12 \*\*\*

JobGrade6 47664.78 5599.78 8.512 4.75e-15 \*\*\*

YrHired -1031.17 195.96 -5.262 3.77e-07 \*\*\*

YrBorn 17.92 115.40 0.155 0.8767

**GenderMale 5108.95 2023.95 2.524 0.0124 \***

YrsPrior 335.45 280.88 1.194 0.2338

PCJobYes 9845.69 2947.65 3.340 0.0010 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11300 on 193 degrees of freedom

Multiple R-squared: 0.7652, Adjusted R-squared: 0.7482

F-statistic: 44.94 on 14 and 193 DF, p-value: < 2.2e-16

1. **Create a new column called “Experience” which should be YrsPrior + (95-YrHired). Create a new column “Age”, which should be (95-YrBorn). Perform a regression using Experience, Age and Gender to predict Salary. How much less salary do females make according to this regression.**

**Solution:** The coefficient of GenderMale is 17050.68 Hence, females make 17050.68 less than males.

BankData$Experience = BankData$YrsPrior + (95-BankData$YrHired)

BankData$Age = 95-BankData$YrBorn

Reg3 = lm(Salary ~ Experience+Age+Gender , data = BankData)

summary(Reg3)

**Output:** Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 52571.77 5268.64 9.978 < 2e-16 \*\*\*

Experience 1720.66 229.11 7.510 1.81e-12 \*\*\*

Age 23.94 162.77 0.147 0.883

GenderMale 17050.68 2513.90 6.783 1.25e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16950 on 204 degrees of freedom

Multiple R-squared: 0.4414, Adjusted R-squared: 0.4332

F-statistic: 53.73 on 3 and 204 DF, p-value: < 2.2e-16

1. **Data point 208 is a statistical outlier by most definitions. Repeat questions 8 and 9 without this row in the data. Is Gender statistically significant without this data point?**

**Solution:**

BankData2 <- BankData[-208,]

tail(BankData2)

Reg4 = lm(Salary ~ . -Employee , data=BankData2)

summary(Reg4)

Reg5 = lm(Salary ~ Experience+Age+Gender , data = BankData2)

summary(Reg5)

**Output: for question 8:**

GenderMale 2854.62 1724.62 1.655 0.099515 .

Without data point 208, GenderMale is not significant as the p value increases to 0.099515 and females make 2854.62 amount less than males.

**Output: for question 9:**

GenderMale 16695.96 2463.71 6.777 1.31e-10 \*\*\*

Without data point 208, GenderMale is statistically significant as the p value is 1.31e-10 and females make 16695.96 amount less than males.

For problem 11, leave point 208 out of the analysis.

1. **As all the PC jobs are female, we want to isolate the impact of that. Create a new column “GenderTech” which has the factor values “Male”, “Female no PC”, and “Female PC”. Perform a regression using JobGrade, GenderTech, Experience, and Age. What is the “cost” in terms of lost salary for a “Female no PC” vs a “Male” from this new GenderTech column?**

**Solution:** The cost in terms of lost salary is 2703.44 as this is the coefficient of GenderTechMale.

**#Code**

BankData2$GenderTech = ifelse(BankData2$Gender %in% c("Male"),"Male",

ifelse(BankData2$PCJob %in% c("No"), "Female no PC","FemalePC"))

Reg6 = lm(Salary ~ JobGrade + GenderTech + Experience + Age , data = BankData2)

summary(Reg6)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 56960.89 3470.40 16.413 < 2e-16 \*\*\*

JobGrade2 3994.00 2028.75 1.969 0.050390 .

JobGrade3 11680.33 1993.29 5.860 1.92e-08 \*\*\*

JobGrade4 20810.29 2336.88 8.905 3.54e-16 \*\*\*

JobGrade5 32904.92 2659.78 12.371 < 2e-16 \*\*\*

JobGrade6 64870.96 3914.55 16.572 < 2e-16 \*\*\*

GenderTechFemalePC 8777.60 2488.09 3.528 0.000521 \*\*\*

**GenderTechMale 2703.44 1763.66 1.533 0.126916**

Experience 785.18 146.94 5.343 2.51e-07 \*\*\*

Age -38.70 96.54 -0.401 0.688959

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9934 on 197 degrees of freedom

Multiple R-squared: 0.814, Adjusted R-squared: 0.8055

F-statistic: 95.78 on 9 and 197 DF, p-value: < 2.2e-16

Consider the dataset Table 2.7 Softdrink.csv.

1. **Perform a regression using t (time period) only. What is the forecast for the next 4 quarters (57-60)?**

**Soultion:** prediction for next four quarters from 57 to 60 is 5339.384, 5406.213, 5473.041, 5539.869

**#Code**

Softdrink = read.csv(file.choose())

reg = lm(Sales ~ t , data = Softdrink)

new = data.frame(t = c(57,58,59,60))

predict(reg,new)

**Output:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | predict(reg,new)  1 2 3 4  5339.384 5406.213 5473.041 5539.869 | |  | | |  | | --- | |  | | |
| |  | | --- | |  | |

1. **Perform a regression using t and Q to predict Sales. What is the forecast for the next 4 quarters (57-60)?**

**Solution:** the prediction for next for quarters is 5282.841, 5384.058, 5485.275, 5586.492

**#Code**

Softdrink = read.csv(file.choose())

reg1 = lm(Sales ~ t+Q , data = Softdrink)

new1 = data.frame(t = c(57,58,59,60), Q = c(1,2,3,4))

predict(reg1,new1)

**Output:** predict(reg1,new1)

1 2 3 4

5282.841 5384.058 5485.275 5586.492

Consider the dataset mtcars, which is built into the R environment.

1. **Predict mpg using the other columns of input data. One by one, reduce the columns that have the lowest t-values until you have only three columns left in your model. What are those three columns?**

**Soultion:** The three column left are: wt, qsec, am

**#Code**

head(mtcars)

reg9 = lm(mpg ~ .-cyl-vs-carb-gear-drat-disp-hp, data = mtcars)

summary(reg9)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.6178 6.9596 1.382 0.177915

wt -3.9165 0.7112 -5.507 6.95e-06 \*\*\*

qsec 1.2259 0.2887 4.247 0.000216 \*\*\*

am 2.9358 1.4109 2.081 0.046716 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 28 degrees of freedom

Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336

F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11

1. **With those three columns, what is the mpg for the average vehicle (create a new vehicle with the averages for those three columns and predict the mpg)?**

**Solution:** the mpg for average vehicle is 20.09062

#**Code**

new= data.frame(wt = mean(mtcars$wt) , qsec = mean(mtcars$qsec) , am = mean(mtcars$am))

reg = lm(mpg ~ wt+qsec+am , data = mtcars)

predict(reg, new)

**Output:**

predict(reg, new)

1

20.09062

Consider the Advertising.csv dataset.

1. **Perform a regression to predict Sales based on TV, Radio, and Newspaper. What is the equation of the regression line?**

**Solution:** The equation of the regression line is: Sales=2.938889 + 0.045765\*TV + 0.188530\*Radio – 0.001037\*Newspaper

Advertising = read.csv(file.choose())

head(Advertising)

reg= lm(Sales ~ TV + Radio + Newspaper, data=Advertising)

summary(reg)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.938889 0.311908 9.422 <2e-16 \*\*\*

TV 0.045765 0.001395 32.809 <2e-16 \*\*\*

Radio 0.188530 0.008611 21.893 <2e-16 \*\*\*

Newspaper -0.001037 0.005871 -0.177 0.86

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom

Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956

F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

1. **Sometimes there is an interactive effect from advertising in more than one medium. For example, if there are both ads on TV and Radio, then it can be more effective. Create a new column for TVRadio which would be TV\*Radio. For example, the first row would have 230.1 \* 37.8 = 8697.78. Perform a regression on TV, Radio, and TVRadio to predict Sales. Is there any interactive effect between the two medium? Is it statistically significant?**

**Solution:** Yes, the interactive effect between two medium is statistically significant as the p value is very low (<2e-16) and the t value is very large.

**#Code**

Advertising$TVRadio = Advertising$TV \* Advertising$Radio

reg2 = lm(Sales ~ TV+Radio+TVRadio , data = Advertising)

summary(reg2)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.750e+00 2.479e-01 27.233 <2e-16 \*\*\*

TV 1.910e-02 1.504e-03 12.699 <2e-16 \*\*\*

Radio 2.886e-02 8.905e-03 3.241 0.0014 \*\*

**TVRadio 1.086e-03 5.242e-05 20.727 <2e-16 \*\*\***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom

Multiple R-squared: 0.9678, Adjusted R-squared: 0.9673

F-statistic: 1963 on 3 and 196 DF, p-value: < 2.2e-16

1. **Do the same for Radio and Newspaper. Answer the same questions.**

**Solution:** No, the interactive effect between two medium is statistically not significant as the p value is very high (<2e-16) and the t value is very low.

**#Code**

Advertising$RadioNewspaper = Advertising$Radio \* Advertising$Newspaper

reg2 = lm(Sales ~ Radio+Newspaper +RadioNewspaper , data = Advertising)

summary(reg2)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.7904734 1.0224848 8.597 2.58e-15 \*\*\*

Radio 0.2145684 0.0382985 5.603 7.08e-08 \*\*\*

Newspaper 0.0220611 0.0345866 0.638 0.524

RadioNewspaper -0.0005259 0.0010642 -0.494 0.622

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.292 on 196 degrees of freedom

Multiple R-squared: 0.3335, Adjusted R-squared: 0.3233

F-statistic: 32.7 on 3 and 196 DF, p-value: < 2.2e-16

1. **Do the same for TV and Newspaper. Answer the same questions.**

**Solution:** No, the interactive effect between two medium is statistically not significant as the p value is very high (0.221) and the t value is very low.

Advertising$TVNewspaper = Advertising$TV \* Advertising$Newspaper

reg2 = lm(Sales ~ TV+Newspaper +TVNewspaper , data = Advertising)

summary(reg2)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.4042175 0.7333818 8.732 1.1e-15 \*\*\*

TV 0.0426585 0.0043105 9.896 < 2e-16 \*\*\*

Newspaper 0.0241103 0.0192716 1.251 0.212

**TVNewspaper 0.0001324 0.0001079 1.228 0.221**

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.117 on 196 degrees of freedom

Multiple R-squared: 0.6485, Adjusted R-squared: 0.6432

F-statistic: 120.6 on 3 and 196 DF, p-value: < 2.2e-16

1. **Create a column for TVRadioNewspaper that represents the product of all three mediums. Perform a regression on all the columns (3 medium, 3 interactive, and one big interactive) to predict Sales. Which of the interactive effects are significant in this model? Is the three-way interactive effect statistically significant?**

**Solution:** No, the interactive effect between 3 way model is insignificant. Only TVRadio has statistical significance. Because of low pvalue (<2e-16) and high t value.

Advertising$TVRadioNewspaper = Advertising$TV \* Advertising$Radio \* Advertising$Newspaper

reg2 = lm(Sales ~ TV+Radio +Newspaper +TVRadio+TVNewspaper +RadioNewspaper +TVRadioNewspaper, data = Advertising)

summary(reg2)

**Output:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.556e+00 4.655e-01 14.083 < 2e-16 \*\*\*

TV 1.971e-02 2.719e-03 7.250 9.95e-12 \*\*\*

Radio 1.962e-02 1.639e-02 1.197 0.233

Newspaper 1.311e-02 1.721e-02 0.761 0.447

**TVRadio 1.162e-03 9.753e-05 11.909 < 2e-16 \*\*\***

TVNewspaper -5.545e-05 9.326e-05 -0.595 0.553

RadioNewspaper 9.063e-06 4.831e-04 0.019 0.985

TVRadioNewspaper -7.610e-07 2.700e-06 -0.282 0.778

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9406 on 192 degrees of freedom

Multiple R-squared: 0.9686, Adjusted R-squared: 0.9675

F-statistic: 847.3 on 7 and 192 DF, p-value: < 2.2e-16