R Lab - MLR (with adv data)

Tanvir

2023-12-01

This is a the running example with the advertisement data that we have in our slides. Now we will do multiple linear regression with this data. Recall the dependent variable is sales and the independent variables are TV, Radio and Newspaper.

Load the data

The first task is same, load the data.

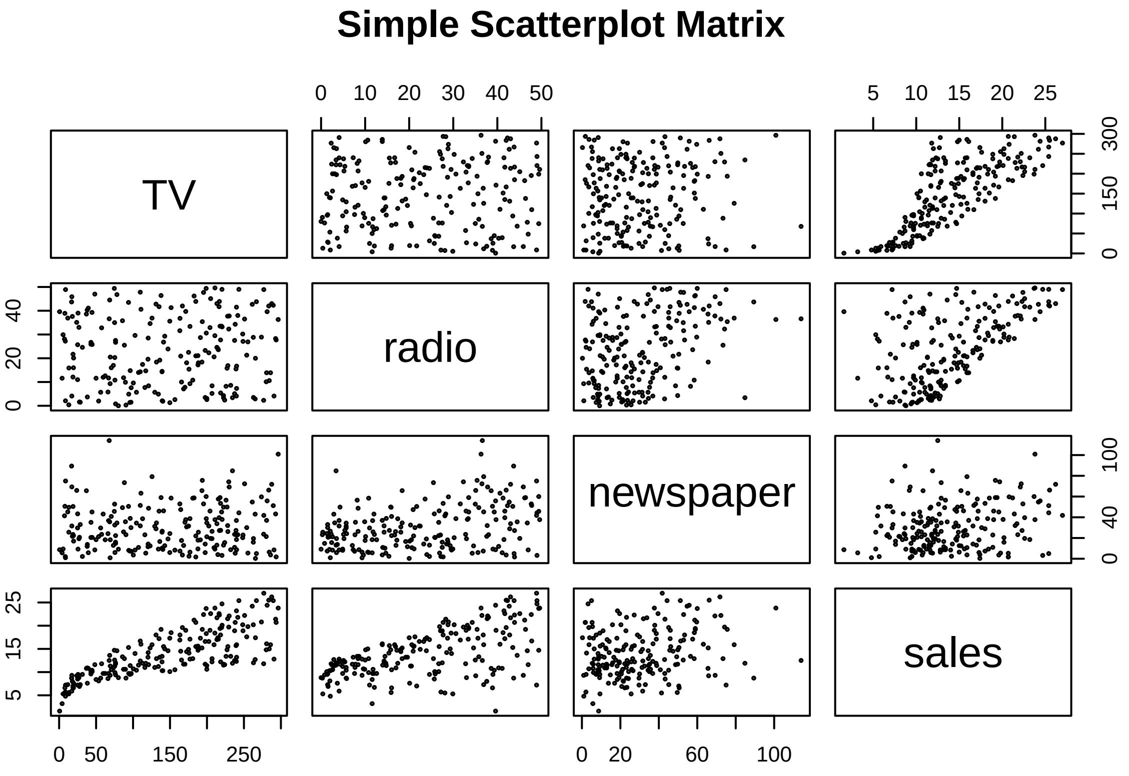
```r
## clear the env
rm(list = ls())


## set the directory
setwd("/home/tanvir/Documents/ownCloud/Git_Repos/EWU_repos/3_Fall_2023/eco_204/ewu-eco204.github.io/pdf


# load the library, load the data
library(readxl)
Advertising <- read_excel("Advertising.xlsx")
```

Here we are not going to do the summary statistics, we have already done in the SLR chapter, but you can do it if you want. But let's do the correlation matrix and correlation plot

```r
## correlation matrix
cor(Advertising)
```

```
##                   TV      radio  newspaper     sales
## TV        1.00000000 0.05480866 0.05664787 0.7822244
## radio     0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales     0.78222442 0.57622257 0.22829903 1.0000000
```

```r
# correlation plot
pairs(~TV+radio+newspaper+sales, data=Advertising,
   main="Simple Scatterplot Matrix", cex = .3)
```



Running Multiple Linear Regression

Now let's perform Multiple Linear Regression

```
# options(scipen = 999) # stop scientific printing
## multiple linear regression
mlr_fit <- lm(sales ~ TV + radio + newspaper, data = Advertising)
summary(mlr_fit)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## 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
```

A little bit of organized output

```r
#install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
stargazer(mlr_fit, type = "text", ci = TRUE)
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                sales
## -----------------------------------------------
## TV                           0.046***
##                           (0.043, 0.048)
##
## radio                        0.189***
##                           (0.172, 0.205)
##
## newspaper                     -0.001
##                           (-0.013, 0.010)
##
## Constant                     2.939***
##                           (2.328, 3.550)
##
## -----------------------------------------------
## Observations                    200
## R2                             0.897
## Adjusted R2                    0.896
## Residual Std. Error      1.686 (df = 196)
## F Statistic          570.271*** (df = 3; 196)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

With the results now we can write the equation of the sample regression function

$$\hat{y}_i = 2.9389 + 0.0458 \times TV + 0.1885 \times radio - 0.0010 \times newspaper$$

I am skipping the interpretation of the coefficients, but you can find the slides.

Confidence Ineteval for the model coefficients

Recall here the model is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where Y is sales and X_1 is TV, X_2 is Radio and X_3 is Newspaper. So the model coefficients are, $\beta_0, \beta_1, \beta_2, \beta_3$. Now we will find the confidence interval for these coefficients.

```r
## confidence interval for the coefficients
confint(mlr_fit)
```

```
##                   2.5 %     97.5 %
## (Intercept)  2.32376228 3.55401646
## TV           0.04301371 0.04851558
## radio        0.17154745 0.20551259
## newspaper   -0.01261595 0.01054097
```

Note that these confidence intervals used the standard errors that are calculated assuming homoskedasticity and also assuming that the CEF errors are normally distributed. You can also find for other significnace levels by changing the `level` argument.

```r
## confidence interval for the coefficients
confint(mlr_fit, level = 0.90)
```

```
##                     5 %        95 %
## (Intercept)  2.42340953 3.454369213
## TV           0.04345935 0.048069943
## radio        0.17429853 0.202761502
## newspaper   -0.01074031 0.008665319
```

Hypothesis Testing

There are different kinds of hypothesis testing that we can do with the MLR models. We learned 3 types,

1. Testing the individual significance, this means testing the significance of each of the coefficients.
2. Testing the overall significance, this means testing the significance of the model as a whole (this is a special case of 3).
3. Testing for certain restrictions, this means testing the significance of a group of coefficients.

Testing the individual significance

Here we will test the significance of each of the coefficients. For example for the TV we can do the test

$$H_0 : \beta_1 = 0 \text{ Vs. } H_a : \beta_1 \neq 0$$

Clearly from the p value we can reject this Null hypothesis. We can also do the same for Radio, but note for the newspaper we cannot since the p value is 0.86.

Testing the overall significance

Let's do the overall significance test. Here we will test the significance of the model as a whole. The test is,

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ Vs. } H_a : \text{At least one of the coefficients is not zero}$$

let's see the anova table also. An important point is in this case we need to fit a restricted model and an unrestricted model. The restricted model is the model where all the coefficients are zero. The unrestricted model is the full model, So let's run the regression again

```r
nullmodel <- lm(sales ~ 1, data = Advertising)
anova(nullmodel, mlr_fit)
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ 1
## Model 2: sales ~ TV + radio + newspaper
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    199 5417.1
## 2    196  556.8  3    4860.3 570.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can do this test using the F-statistic. The value of the F statistic is 570.3 and the p value is is also very small so we can reject the null hypothesis, this means that at least one of the coefficients is not zero.

It's important to understand that the output of the ANOVA table, here

- $\text{SSE}_R = \text{ SST} = 5417, \text{ Df} = 199$
- $\text{SSE} = 556.8, \text{ Df} = 196$
- $\text{SSE}_R - \text{SSE} = \text{SSR} = 4860 \text{ and } \text{ Df} = 3$

Testing on certain restrictions

Now let's do the joint significance test. Here we will test the significance of a group of coefficients. For example we can test the significance of the coefficients of TV and Radio. The test is,

$$H_0 : \beta_1 = \beta_2 = 0 \text{ Vs. } H_a : \text{At least one of the coefficients above is not zero}$$

We can do this test also using the F-statistic. Again here we have fit a restricted model and an unrestricted model. The restricted model is the model where the coefficients of TV and Radio are zero. The unrestricted model is the full model, So let's run the regression again

```r
## multiple linear regression
restrictedmodel <- lm(sales ~ newspaper, data = Advertising)
anova(restrictedmodel, mlr_fit)
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ newspaper
## Model 2: sales ~ TV + radio + newspaper
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    198 5134.8
## 2    196  556.8  2      4578 805.71 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we can see the the value of the F statistic is 805.71 and the p value is also very small so we fail to accept the Null that both the coefficients are zero. This means that at least one of the coefficients β_1, β_2 and β_3 is not zero.

Also note, here

- $\text{SSE}_R = 5134.8, \text{ Df} = 198$
- $\text{SSE} = 556.8, \text{ Df} = 196$
- $\text{SSE}_R - \text{SSE} = 4578 \text{ and } \text{ Df} = 2$

Finally we check what happens if we impose only one restriction, the answer is this will be similar to t-test or individual significance test. In fact in this case we will have $t^2 = F$ and the p value will be same. Let's see this.

```r
## multiple linear regression
restrictedTV <- lm(sales ~ newspaper + radio, data = Advertising)
anova(restrictedTV, mlr_fit)
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ newspaper + radio
## Model 2: sales ~ TV + radio + newspaper
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    197 3614.8
## 2    196  556.8  1      3058 1076.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction

We can do the prediction using the `predict` function. Let's predict the sales for the following values of TV, Radio and Newspaper. This is called point predicition, since we are predicting for a single point.

```r
## predict the sales for the following values of TV, Radio and Newspaper
newdata <- data.frame(TV = c(100), radio = c(20), newspaper = c(20))

predict(mlr_fit, newdata = newdata)
```

```
##       1
## 11.2652
```

For multiple points

```r
## predict the sales for the following values of TV, Radio and Newspaper
newdata <- data.frame(TV = c(100, 200), radio = c(20, 30), newspaper = c(20, 30))

predict(mlr_fit, newdata = newdata)
```

```
##        1        2
## 11.26520 17.71659
```

For interval predicition we can do two kinds of prediciton,

1. Confidence interval for the mean response

```r
## predict the sales for the following values of TV, Radio and Newspaper
newdata <- data.frame(TV = c(100, 200), radio = c(20, 30), newspaper = c(20, 30))
predict(mlr_fit, newdata = newdata, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 11.26520 10.97635 11.55405
## 2 17.71659 17.41835 18.01484
```

2. Prediction interval for a new observation

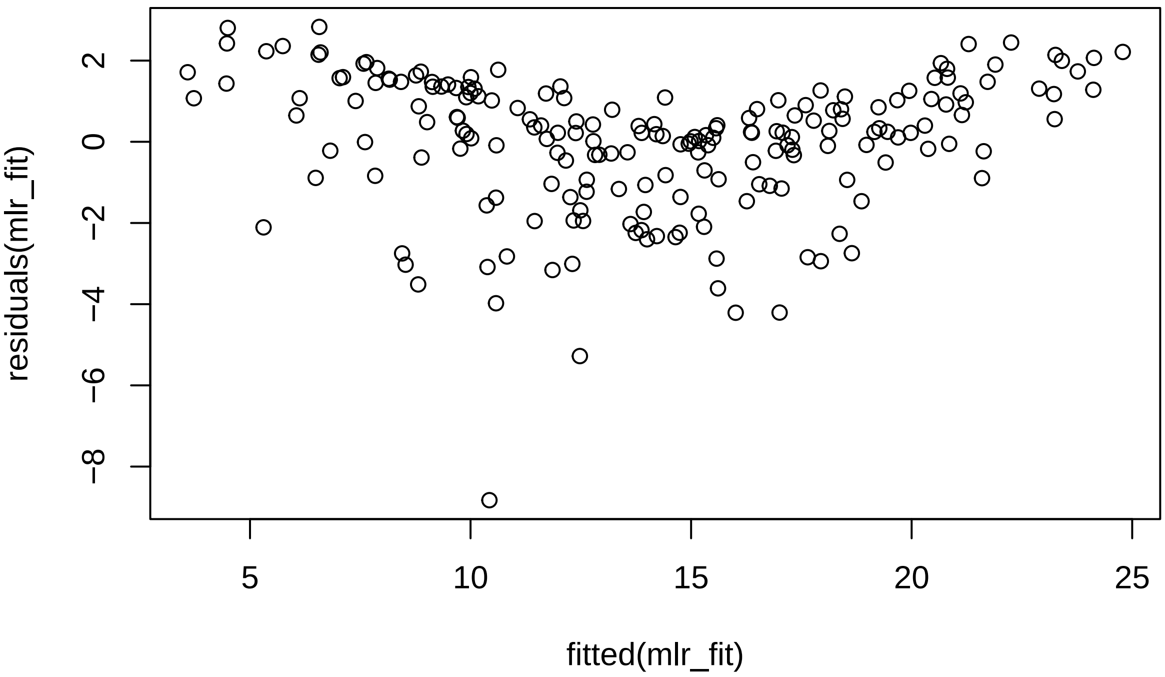
```r
## predict the sales for the following values of TV, Radio and Newspaper
newdata <- data.frame(TV = c(100, 200), radio = c(20, 30), newspaper = c(20, 30))
predict(mlr_fit, newdata = newdata, interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 11.26520  7.928613 14.60180
## 2 17.71659 14.379177 21.05401
```

Residual Analysis

We can do the residual analysis using the `plot` function. Let's plot the residuals vs fitted values.

```
## residual analysis
plot(fitted(mlr_fit), residuals(mlr_fit))
```



In this case we will plot the residuals vs the fitted values. We can see that there is a pattern in the residuals, so we can say that the assumption of linearity is probably not satisfied.

Interaction Term

Incorpating the interaction term is very easy. We just need to add the interaction term in the formula. For example let's add the interaction term between TV and Radio.

```
## multiple linear regression
mlr_fit_int <- lm(sales ~  TV*radio, data = Advertising)

stargazer(mlr_fit_int, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                sales
## -----------------------------------------------
## TV                            0.019***
##                               (0.002)
##
## radio                         0.029***
##                               (0.009)
##
## TV:radio                      0.001***
##                              (0.0001)
##
## Constant                      6.750***
##                               (0.248)
```

```
##
## -----------------------------------------------
## Observations                    200
## R2                             0.968
## Adjusted R2                    0.967
## Residual Std. Error      0.944 (df = 196)
## F Statistic         1,963.057*** (df = 3; 196)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Notice this `TV*radio` is equivalent to `TV + radio + TV:radio`. Here `TV:radio` is the interaction term. We can see that the interaction term is significant. So what does the interaction term show?. It shows that the effect of TV on sales depends on the value of radio. Note that we can write,