Week3 Lecture - Multiple Linear Regression

Unit Coordinator - Dr Liwan Liyanage

School of Computing, Engineering and Mathematics

WESTERN SYDNEY UNIVERSITY

Multiple Linear Regression

Here the estimated model is:

The expected value of Y given X is

$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$

Y has a normal distribution with standard deviation σ. It is the random component of the model, which has a normal distribution.

We interpret β_j as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed. In the advertising example, the estimated model becomes

$E(Sales) = \alpha + \beta_1 TV + \beta_2 Radio + ... + \beta_n Newspaper$

WESTERN SYDNEY UNIVERSITY

This lecture introduces basic concepts and presents examples of various regression techniques.

- Multiple linear regression
- Non-linear regression
  - Interaction Terms of X Variables
  - Polynomial Regrssions
  - Transformations of the response and explanatory variables
- A collection of helpful R functions for regression analysis

WESTERN SYDNEY UNIVERSITY

Import the Data Set “Advertising”

```r
Advertising <- read.csv("Advertising.csv")
attach(Advertising)
names(Advertising)
```

```
## [1] "TV"        "Radio"     "Newspaper" "Sales"
```

```r
head(Advertising)
```

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

WESTERN SYDNEY UNIVERSITY

Multiple Linear Regression

R code

```r
model2=lm(Sales~TV+Radio+Newspaper)
summary(model2)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper)
##
## 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 '
##
## Residual standard error: 1.686 on 196 degrees of freedom
```

Degree of scatter

```
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
```

- Shows that linear relationship between Sales and TV, and Radio are significant while Sales and Newspaper are not linealy related.

ANOVA Table and critical value of F:

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: Sales
##            Df Sum Sq Mean Sq   F value Pr(>F)
## TV          1 3314.6  3314.6 1166.7308 <2e-16 ***
## Radio       1 1545.6  1545.6  544.0501 <2e-16 ***
## Newspaper   1    0.1     0.1    0.0312 0.8599
## Residuals 196  556.8     2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

```r
qf(0.95,1,198)
```

```
## [1] 3.888853
```

Multiple Regression Model and Summary

Model2

$E(Sales) =$
$2.938889 + 0.045765TV + 0.188530Radio - 0.001037Newspaper$

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

Significant linear relationship

95% confidence intervals for the estimated parameters

```r
confint(model2)
```

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

WESTERN SYDNEY UNIVERSITY

Accepted Model

Model3

```
model3=lm(Sales~TV+Radio)
summary(model3)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## TV           0.04575    0.00139  32.909   <2e-16 ***
## Radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
```

ANOVA Table and critical value of F:

```
anova(model3)
```

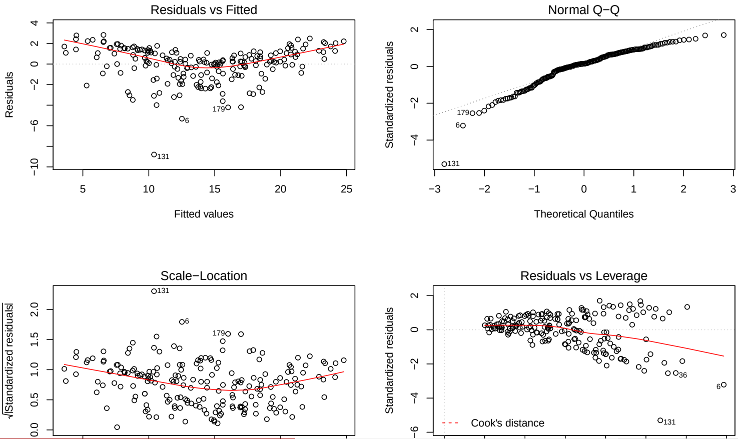
```
## Analysis of Variance Table
##
## Response: Sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## TV          1 3314.6  3314.6 1172.50 < 2.2e-16 ***
## Radio       1 1545.6  1545.6  546.74 < 2.2e-16 ***
## Residuals 197  556.9     2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

```
qf(0.95,1,198)
```

```
## [1] 3.888853
```

Model checking

```r
par(mfrow=c(2,2))
plot(model3)
```



WESTERN SYDNEY UNIVERSITY

Extensions of the Linear Model

Removing the additive assumption: interactions and nonlinearity

Interactions:

- In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

$$E(Sales) = \beta_0 + \beta_1\text{TV} + \beta_2\text{Radio} + \beta_3\text{Newspaper}$$

states that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

Interactions - continued

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- In this situation, given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
- In marketing, this is known as a synergy effect, and in statistics it is referred to as an interaction effect.

WESTERN SYDNEY UNIVERSITY

Modelling interactions - Advertising data

Model with Interactions takes the form

$E(Sales) = \alpha + \beta_1 TV + \beta_2 Radio + \beta_3 TV.Radio$

and the estimated line takes the form
$\hat{Sales} = \hat{\alpha} + \hat{\beta}_1 TV + \hat{\beta}_2 Radio + \hat{\beta}_3 TV.Radio$

```
model4=lm(Sales~TV+Radio+TV*Radio)
summary(model4)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + TV * Radio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## 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 **
## TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 0.9435 on 196 degrees of freedom
```

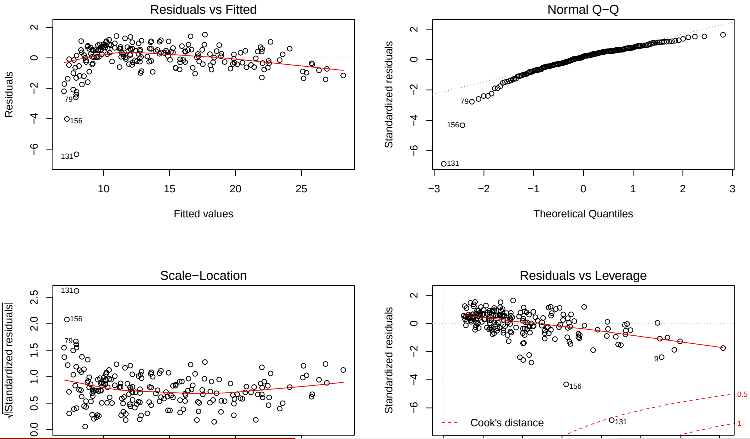
ANOVA

```r
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: Sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## TV          1 3314.6  3314.6 3723.36 < 2.2e-16 ***
## Radio       1 1545.6  1545.6 1736.22 < 2.2e-16 ***
## TV:Radio    1  382.4   382.4  429.59 < 2.2e-16 ***
## Residuals 196  174.5     0.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

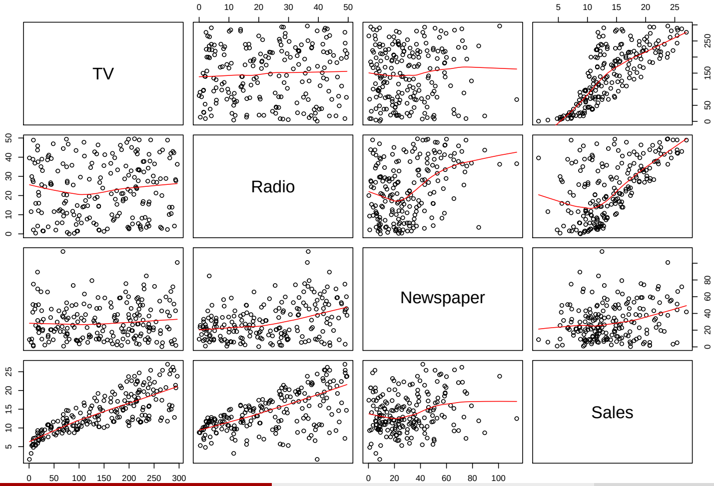
Model Checking

```
par(mfrow=c(2,2))
plot(model4)
```



Covariance and Correlations

```
pairs(Advertising,panel=panel.smooth)
```



Covariance and Correlations

```r
cov(Advertising,method="pearson")
```

```
##                   TV     Radio Newspaper     Sales
## TV        7370.94989  69.86249 105.91945 350.39019
## Radio       69.86249 220.42774 114.49698  44.63569
## Newspaper  105.91945 114.49698 474.30833  25.94139
## Sales      350.39019  44.63569  25.94139  27.22185
```

Covariance and Correlations

```r
cor(Advertising,method="pearson")
```

```
##                   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
cor(TV,Sales)
```

```
## [1] 0.7822244
```

Useful Codes

```
model3$residuals

plot(predict(model2),model2$residuals)

hist(model3$residuals)

predict(model3)

predict(model3,interval='confidence')

predict(model3,as.data.frame(
    cbind(TV=50,Radio=50,Newspaper=50)))
```

Non-linear effects of predictors - Polynomial Regression

The relationship between Y and X often turns out not to be a straight line.

How do we assess the significance of departures from linearity?

One of the simplest ways is to use polynomial regression.

As before, we have just one continuous explanatory variable, X, but we can fit higher powers of X, such as X^2 and X^3, to the model in addition to X to explain curvature in the relationship between Y and X.

Non-linear effects of predictors - Polynomial Regression

Consider the model $E(Sales) = \beta_0 + \beta_1 TV + \beta_2 TV^2$

Will this model provide a better fit?

```
model5=lm(Sales~TV+I(TV*TV))
summary(model5)
```

```
##
## Call:
## lm(formula = Sales ~ TV + I(TV * TV))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6844 -1.7843 -0.1562  2.0088  7.5097
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.114e+00  6.592e-01   9.275  < 2e-16 ***
## TV           6.727e-02  1.059e-02   6.349 1.46e-09 ***
## I(TV * TV)  -6.847e-05  3.558e-05  -1.924   0.0557 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 3.237 on 197 degrees of freedom
## Multiple R-squared:  0.619,  Adjusted R-squared:  0.6152
```

Anova

```r
anova(model5)
```

```
## Analysis of Variance Table
##
## Response: Sales
##              Df Sum Sq Mean Sq  F value  Pr(>F)
## TV            1 3314.6  3314.6 316.4072 < 2e-16 ***
## I(TV * TV)    1   38.8    38.8   3.7036 0.05574 .
## Residuals   197 2063.7    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Transformations of the response and explanatory variables

The use of transformation to linearize the relationship between the response and the explanatory variables:

- log y against x for exponential relationships
- log y against log x for power functions
- exp y against x for logarithmic relationships
- 1/y against 1/x for asymptotic relationships
- log p/1-p against x for proportion data

WESTERN SYDNEY UNIVERSITY

Other transformations are useful for variance stabilization:

- sqrt(y) to stabilize the variance for count data
- arcsin(y) to stabilize the variance of percentage data

TEXT BOOK

Lecture notes are based on the textbook.

For further reference refer;

Prescribed Textbook - Chapter 3

– James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R Springer.