Multiple Regression Analysis

Dhafer Malouche

Introduction

- We extend the concept of simple linear regression as we investigate a response y which is affected by several independent variables: $x_1, x_2, \ldots, x_k$
- Our objective is to use the information provided by the x_i to predict the value of y.
- y is called response variable and $x_1, \ldots, x_k$ are called predictors or independent variables

Example

- Let y be a student's college achievement, measured by his/her GPA
- We want to predict y using knowledge using the following variables:
  - x_1 rank in high school class
  - x_2 high school's overall rating
  - x_3 high school GPA
  - x_4 SAT scores

Example

- Let y be the monthly sales revenue for a company.
- We want to predict y using knowledge
  - x_1 advertising expenditure
  - x_2 time of year
  - x_3 state of economy
  - x_4 size of inventory

Some questions

- How well does the model fit?
- how strong is the relationship between y and the predictor variables?
- have any assumptions been violated?
- how good are the estimates and predictions?

Data is collected using n n observations on the response y and the independent variables, $x_1, x_2, x_3, \ldots, x_k$:

For $i = 1, \ldots, n$, we have: $y_i, x_{1,i}, \ldots, x_{k,i}$

The General Multivariate Linear Model

For $i = 1, \ldots, n$,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{i,k} + \epsilon_i$$

Or in a generic way:

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_i}_{\text{Deterministic}} + \underbrace{\epsilon}_{\text{Random}}$$

And $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the unknown parameters to estimate and ϵ_i are the errors

Assumptions

- A Linear relationship between y and $x_1, \dots, x_k$
- The errors ϵ_i
  - are Independent
  - have a zero mean
  - have a common variance σ^2
  - A Normal distribution

The Ordinary Least Square (OLS) method

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are estimated using the Ordinary Least Square method: $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$

- The best-fitting prediction of y_i are computed as follows:

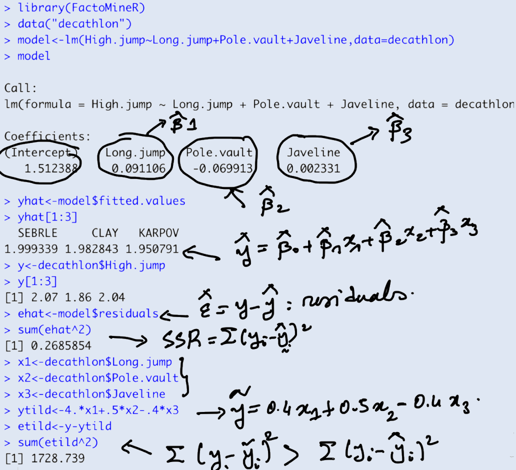
$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_k x_i$$

  where $e_i = y_i - \widehat{y}_i$ are the estiamtion of the errors ϵ_i, called the residuals.

- $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ are estimators that minimize the Sum of Squares of the Residuals (SSR):

$$\text{SSR} = \sum_i e_i^2 = \sum (y_i - \widehat{y}_i)^2$$

Example:



```
> library(FactoMineR)
> data("decathlon")
> model<-lm(High.jump~Long.jump+Pole.vault+Javeline,data=decathlon)
> model

Call:
lm(formula = High.jump ~ Long.jump + Pole.vault + Javeline, data = decathlon

Coefficients:
(Intercept)    Long.jump   Pole.vault     Javeline
   1.512388     0.091106    -0.069913     0.002331

> yhat<-model$fitted.values
> yhat[1:3]
  SEBRLE     CLAY   KARPOV
1.999339 1.982843 1.950791
> y<-decathlon$High.jump
> y[1:3]
[1] 2.07 1.86 2.04
> ehat<-model$residuals
> sum(ehat^2)
[1] 0.2685854
> x1<-decathlon$Long.jump
> x2<-decathlon$Pole.vault
> x3<-decathlon$Javeline
> ytild<-4.*x1+.5*x2-.4*x3
> etild<-y-ytild
> sum(etild^2)
[1] 1728.739
```

$\hat\beta_0$ $\hat\beta_1$ $\hat\beta_2$ $\hat\beta_3$

$$\hat y = \hat\beta_0 + \hat\beta_1 x_1 + \hat\beta_2 x_2 + \hat\beta_3 x_3$$

$\hat e = y - \hat y$: residuals.

$$SSR = \Sigma (y_i - \hat y_i)^2$$

$$\tilde y = 0.4 x_1 + 0.5 x_2 - 0.4 x_3 .$$

$$\Sigma (y_i - \tilde y_i)^2 > \underbrace{\Sigma (y_i - \hat y_i)^2}_{SSR}$$

The Analysis of Variance (ANOVA: Testing linearity)

- We perform two kinds of ANOVA
- ANOVA I: it can be used to test the overall linear relationship between y and the used variables $x_1, \ldots, x_k$

  H_0 There is no linear relationship between $x_1, \ldots, x_k$ and y
  H_1 There is a linear relationship between $x_1, \ldots, x_k$ and y

- ANOVA II: it can be used to detect the presence of each variable in the linear regression model. Testing for all $j = 1, \ldots, k$

  H_0^j x_j is not useful in the regression model
  H_1^j x_j is useful in the regression model

ANOVA I

- We tested in ANOVA

$$\begin{array}{ll} H_0 & \text{Null model is true} \iff x_1, \ldots, x_k \text{ aren't useful} \\ H_1 & \text{Full model is true} \end{array}$$

- Null model: Linear regression without independent variables:

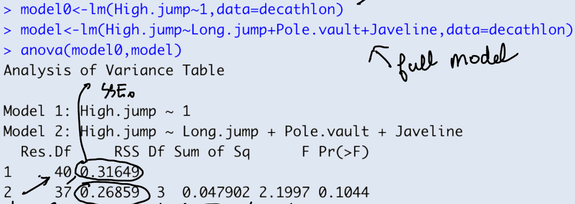
$$y_i = \beta_i + \epsilon_i$$

- Full model: Linear regression without independent variables:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{i,k} + \epsilon_i$$

- Since the $p - value > 0.1044$, H_0 is not reject and then we deduce that the variables: `Long.jump`, `Pole.vault`, `Javeline` aren't useful (together) to predict `High.jump`

ANOVA I



```
> model0<-lm(High.jump~1,data=decathlon)
> model<-lm(High.jump~Long.jump+Pole.vault+Javeline,data=decathlon)
> anova(model0,model)
Analysis of Variance Table

Model 1: High.jump ~ 1
Model 2: High.jump ~ Long.jump + Pole.vault + Javeline
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     40 0.31649
2     37 0.26859  3  0.047902 2.1997 0.1044
```

→ null model

↖ full model

SSE_0

$n-1$

$n-k-1$

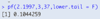
SSE_1

k

$\underbrace{SSE_0 - SSE_1}_{\Delta SE}$

F

$F(k,n-k-1)$



```
> pf(2.1997,3,37,lower.tail = F)
[1] 0.1044259
```

$$F = \frac{\frac{\Delta SE}{k}}{\frac{SSE_1}{n-k-1}} = \frac{n-k-1}{k} \times \frac{\Delta SE}{SSE_1}.$$

SSE_1 : Sum Sq Residual of full model

SSE_0 : ______ null model.

ANOVA II

- We will test now the presence of each variable in the model using an Analysis of the Variance (ANOVA):
- He will perform k ANOVA by testing the following Hypothesis: for a given $j = 1, \dots, k$

$$\begin{array}{ll} H_0 & \text{The true is the one without } x_j \\ H_0 & \text{Full model is true} \end{array}$$

- $F = \dfrac{SSE_j - SSE_{full}}{(n-k-1)SSE_{full}} \sim F(1, n-k-1)$ Under H_0^j
- ANOVA on Nested models

ANOVA II

H_0 : `Long.jump` is not in the model vs H_1 : Full model

```
> model1<-lm(High.jump~Pole.vault+Javeline,data=decathlon)
> anova(model1,model)
Analysis of Variance Table

Model 1: High.jump ~ Pole.vault + Javeline
Model 2: High.jump ~ Long.jump + Pole.vault + Javeline
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     38 0.29991
2     37 0.26859  1  0.031328 4.3156 0.04476 *
```

Conclusion: `Long.jump` should stay in the model

ANOVA II

H_0 : Javeline is not in the model vs H_1 : Full model

```
> model2<-lm(High.jump~Long.jump+Pole.vault,data=decathlon)
> anova(model2,model)
Analysis of Variance Table

Model 1: High.jump ~ Long.jump + Pole.vault
Model 2: High.jump ~ Long.jump + Pole.vault + Javeline
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     38 0.27356
2     37 0.26859  1 0.0049774 0.6857 0.4129
```

Conclusion: Javeline shouldn't be in the model

ANOVA II

H_0 : Pole.vault is not in the model vs H_1 : Full model

```
> model3<-lm(High.jump~Long.jump+Javeline,data=decathlon)
> anova(model3,model)
Analysis of Variance Table

Model 1: High.jump ~ Long.jump + Javeline
Model 2: High.jump ~ Long.jump + Pole.vault + Javeline
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     38 0.28302
2     37 0.26859  1  0.014436 1.9886 0.1668
```

Conclusion: Pole.vault shouldn't be in the model

The Coefficient of Determination R^2

Definition

- Coefficient of Determination:

$$R^2 = \frac{SSR}{TSS} = \frac{SSR}{SSE + SSR} \in [0, 1]$$

- R^2 increases when the number of variables increases
- Adjusted R^2:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-k-1} (1 - R^2) \right)$$

Testing the coefficients (one by one)

```
> summary(model)

Call:
lm(formula = High.jump ~ Long.jump + Pole.vault + Javeline, data = decathlon)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15775 -0.04843 -0.01003  0.07066  0.13777

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.512388   0.379072   3.990 0.000301 ***
Long.jump    0.091106   0.043856   2.077 0.044760 *
Pole.vault  -0.069913   0.049577  -1.410 0.166837
Javeline     0.002331   0.002816   0.828 0.412948
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0852 on 37 degrees of freedom
Multiple R-squared:  0.1514,Adjusted R-squared:  0.08255
F-statistic:   2.2 on 3 and 37 DF,  p-value: 0.1044
```

$(1-\alpha)\%$ Confidence intervals of the coefficients

```
> model$coefficients
 (Intercept)    Long.jump   Pole.vault     Javeline
 1.512388475  0.091106422 -0.069912639  0.002331461
> confint(model,level = .95)
                  2.5 %      97.5 %
(Intercept)  0.744315633 2.280461316
Long.jump    0.002246204 0.179966640
Pole.vault  -0.170365009 0.030539732
Javeline    -0.003373451 0.008036374
> qt(.025,37,lower.tail = F)
[1] 2.026192
> SEa=(2.280461316-0.744315633)/(2*2.026192)
> SEa
[1] 0.3790721
```

Nested Multiple linear regression models

Two Multiple linear regression models $\mathcal{M}_0$ and $\mathcal{M}_1$ are **nested** if by removing some variables from $\mathcal{M}_1$ we can retrieve the model $\mathcal{M}_0$.

We will denote $\mathcal{M}_0 \subseteq \mathcal{M}_1$

Nested Models used to test a group of coefficients

Nested Multiple linear regression models

- Assume we would like to perform a Multiple linear regression model from a data containing one response variable y and 10 independent variables $x_1, x_2, \ldots, x_{10}$
- Indicate which of the following pairs of Models are nested. Specify $\mathcal{M}_0$ and $\mathcal{M}_1$
  - $y \sim x_1 + x_2 + x_3$ and $y \sim x_1 + x_2 + x_4 + x_5$
  - $y \sim x_1 + x_3$ and $y \sim x_1 + x_2 + x_4 + x_5 + x_3$
  - $y \sim x_1 + x_4 + x_5$ and $y \sim x_1 + x_2 + x_4 + x_5 + x_3$

Testing nested models

- Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two nested models such that $\mathcal{M}_0 \subseteq \mathcal{M}_1$
- We always test
  - H_0 : $\mathcal{M}_0$ is true
  - H_1 : $\mathcal{M}_1$ is true
- To test H_0 vs H_1 we perform an ANOVA.
- We will use R to test the following Hypothesis

$$H_0 : \texttt{High.jum} \sim \texttt{Long.jump} + \texttt{Javeline}$$
$$H_1 : \texttt{High.jum} \sim \texttt{Long.jump} + \texttt{Javeline} + \texttt{Pole.vault} + \texttt{Discus}$$

Example with R

```
> library(FactoMineR)
> data(decathlon)
> colnames(decathlon)
 [1] "100m"          "Long.jump"    "Shot.put"     "High.jump"    "400m"
 [7] "Discus"        "Pole.vault"   "Javeline"     "1500m"        "Rank"
[13] "Competition"
> model0<-lm(High.jump~Long.jump+Javeline,data=decathlon)
> model1<-lm(High.jump~Long.jump+Javeline+Pole.vault+Discus,
+            data=decathlon)
```

Example with R

```
> anova(model0,model1)
Analysis of Variance Table

Model 1: High.jump ~ Long.jump + Javeline
Model 2: High.jump ~ Long.jump + Javeline + Pole.vault + Discus
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     38 0.28302
2     36 0.24622  2  0.036805 2.6907 0.08146 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example with R

```
> model_null<-lm(High.jump˜1,data=decathlon)
> anova(model_null,model0)
Analysis of Variance Table

Model 1: High.jump ˜ 1
Model 2: High.jump ˜ Long.jump + Javeline
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     40 0.31649
2     38 0.28302  2  0.033467 2.2467 0.1196
```

Example with R

```
> anova(model_null,model1)
Analysis of Variance Table

Model 1: High.jump ~ 1
Model 2: High.jump ~ Long.jump + Javeline + Pole.vault + Discus
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     40 0.31649
2     36 0.24622  4  0.070272 2.5687 0.05444 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A polynomial regression model

- A response y is related to a single independent variable x, but not in a linear manner. The polynomial model is:

$$y = \beta_0 + \beta_1 x + \ldots + \beta_k x^k + \epsilon$$

- When $k = 2$, the model is **quadratic**

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- When $k = 3$, the model is **cubic**

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

Case Study: Gas Consumption the US?

- We will consider the data in the file `US_gas.csv`
- It contains three variables: Price, Consumption, and Production
- We aim to build a Regression model that predicts the gas consumption using the gas prices
- Check the R code to follow the whole analysis procedure

Linear regression with Qualitative variable

- Assume we want to predict the fuel consumption (`mpg`) in terms of the following variables using a Multiple linear regression model:
  - wt Weight (1000 lbs) *quantitative variable*
  - hp Gross horsepower *quantitative variable*
  - vs Engine (0 = V-shaped, 1 = straight) *qualitative variable*
- How to proceed?