Chapter 3

Multiple Linear Regression

Chapter 3c

Decomposition of variance: $SS_T = SS_{Res} + SS_R$ (pages 3-4)

$$y'y = SS_{Res} + \hat{\beta}'X'y$$

Test of overall fit of model – analysis of variance (pages 5-8)

Example - Delivery time data (pages 9-12)

Example – Cholesterol age data (page 13)

R^2 and adjusted R^2 for assessing the overall adequacy of a model (page 14)

Example - Delivery time data (page 15)

Example – Cholesterol age data (page 16)

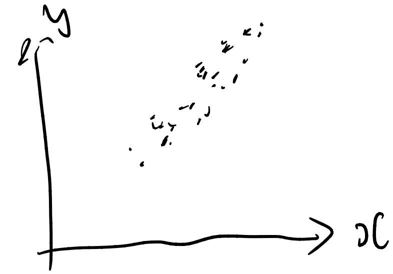
Tests and confidence intervals on individual regression coefficients (page 17)

Example - Delivery time data (pages 18-20)

Decomposition of variance: $SS_T = SS_{Res} + SS_R$

$$SS_{Res} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

$$\| \quad \sum_{i=1}^{n}(y_i - \bar{y})^2$$



$$\boldsymbol{y}'\boldsymbol{y} = SS_{Res} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

$$y'y = \sum_{i=1}^{n} y_i^2$$

$$\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2 = SS_{Res} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2$$

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = SS_{Res} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2$$

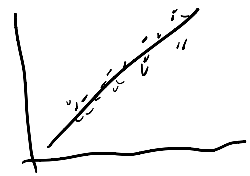
$$\underset{SS_T}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = SS_{Res} + \underbrace{\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2}_{SS_R}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = SS_{Res} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2$$

constant large small

$$SS_T = SS_{Res} + SS_R$$

small large



where

$$SS_T \equiv \sum_{i=1}^{n}(y_i-\bar{y})^2 = \text{total sum of squares or total variation in } y$$

$$SS_{Res} \equiv \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} = \text{residual sum of squares}$$

$$SS_R \equiv \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2 = \text{sum of squares due to the regression model}$$

$$F = \frac{SS_R/k}{SS_{Res}/(n-p)}$$

Reject $H_0: \beta_1 = \beta_2 = \dots \beta_k = 0$

if large SS_R,
y have to depend on at least 1

Test of overall fit of model - analysis of variance

Multiple linear regression model:

$\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$ where $\boldsymbol{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\boldsymbol{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$

Assume $\boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

Note that $\boldsymbol{\beta}' = (\beta_0, \beta_1, ..., \beta_k)$.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

H_1 : At least one β_j's is not equal to zero.

1. The total variation in y can be decomposed as $SS_T = SS_{Res} + SS_R$. For a given data set, SS_T is a constant. How large SS_R is with respect to SS_{Res} depends on how close the points are to the regression model.

2. If all the points fall exactly on the model $y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon$, $\overset{=0}{SS_{Res}}$ will be zero, and $SS_R = SS_T$. In other words, all the variation in y is contributed by the regression model.

   $SS_R = SS_T$ large, reject H_0
   ↳ depend on regressor

3. If all the points fall close to the regression model $y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon$, SS_{Res} will be small, and SS_R will be large, hence SS_R/SS_{Res} is large. In other words, most of the variation in y is contributed by the regression model.

4. If the points do not fall close to the model $y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon$, this means y does not depend on the regressor variables $x_1, x_2, \ldots, x_k$, consequently SS_{Res} will be large and SS_R will be small, hence SS_R/SS_{Res} is small.

   do not reject H_0

distribution

5. It makes sense to use the ratio SS_R/SS_{Res} to test the hypotheses

   $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

   H_1 : At least one β_j's is not zero.

6. If H_0 is true, the ratio will be small. On the other hand, if H_0 is not true, the ratio will be large, so we reject H_0 if the ratio is large.

7. Instead of SS_R/SS_{Res}, we will use $F = \dfrac{SS_R/k}{SS_{Res}/(n-p)} = \dfrac{MS_R}{MS_{Res}}$ because according to linear model theory (can be learned in a graduate course on multiple linear regression analysis), F follows the F distribution with degrees of freedom k and $n-p$ when H_0 is true. For a given level of significance level α, we reject H_0 if $F > F_{\alpha,k,n-k-1}$.

8. The decomposition of variance and test of $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ can be summarized in the following table. This table is commonly known as an analysis of variance (ANOVA) table.

Analysis of variance table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SS_R	k	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-(k+1)$	MS_{Res}	
Total	SS_T	$n-1$		

Example - Delivery time data

```r
# Test of overall fit of model
bxy <- t(betahat) %*% t(X) %*% y
bxy
SSR <- bxy - sum(y)^2/length(time)
SSR
SSRes <- t(y) %*% y - bxy
SSRes
F0 <- (SSR/2)/(SSRes/(length(time)-3))
F0
pvalue <- pf(F0, df1=2, df2=length(time)-3, lower.tail=FALSE)
pvalue
# ANOVA table
anova(lm(time ~ case + dist))
anova(lm(time~1),lm(time ~ case + dist))
```

$H_0: \beta_1 = \beta_2 = 0$ ✓

$H_0 = \beta_0 = \beta_1 = \beta_2 = 0$ ✗

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
> # Test of overall fit of model
> bxy <- t(betahat) %*% t(X) %*% y
> bxy
        [,1]
[1,] 18076.9
> SSR <- bxy - sum(y)^2/length(time)
> SSR
         [,1]
[1,] 5550.811
> SSRes <- t(y) %*% y - bxy
> SSRes
         [,1]
[1,] 233.7317
> F0 <- (SSR/2)/(SSRes/(length(time)-3))
> F0
         [,1]
[1,] 261.2351
> pvalue <- pf(F0, df1=2, df2=length(time)-3, lower.tail=FALSE)
> pvalue
             [,1]
[1,] 4.687422e-16
```

```
> # ANOVA table
> anova(lm(time ~ case + dist))
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value    Pr(>F)    
case       1 5382.4  5382.4 506.619 < 2.2e-16 ***
dist       1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(lm(time~1),lm(time ~ case + dist))
Analysis of Variance Table

Model 1: time ~ 1
Model 2: time ~ case + dist
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     24 5784.5                                  
2     22  233.7  2    5550.8 261.24 4.687e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSR

p-value is small, reject Ho

Example - Delivery time data

$$SS_{\mathrm{T}} = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 18,310.6290 - \frac{(559.60)^2}{25} = 5784.5426$$

$$SS_{\mathrm{R}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 18,076.9030 - \frac{(559.60)^2}{25} = 5550.8166$$

$$SS_{\mathrm{Res}} = SS_{\mathrm{T}} - SS_{\mathrm{R}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = 233.7260$$

$$H_0\colon \beta_1 = \beta_2 = 0 \qquad F_0 = \frac{MS_{\mathrm{R}}}{MS_{\mathrm{Res}}} = \frac{2775.4083}{10.6239} = 261.24$$

Analysis of variance

Source Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P Value
Regression	5550.8166	2	2775.4083	261.24	4.7×10^{-16}
Residual	233.7260	22	10.6239		
Total	5784.5426	24			

Example - Cholesterol age data

```
# test overall fit of straight line model
anova(lm(cholesterol~1), lm(cholesterol~age))

# test overall fit of quadratic model
anova(lm(cholesterol~1), lm(cholesterol~age + I(age^2)))

```

53:1 (Top Level)

Console | Background Jobs

R 3.4.1

```
> # test overall fit of straight line model
> anova(lm(cholesterol~1), lm(cholesterol~age))
Analysis of Variance Table

Model 1: cholesterol ~ 1
Model 2: cholesterol ~ age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     32 36.636
2     31  5.899  1    30.737 161.52 7.856e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # test overall fit of quadratic model
> anova(lm(cholesterol~1), lm(cholesterol~age + I(age^2)))
Analysis of Variance Table

Model 1: cholesterol ~ 1
Model 2: cholesterol ~ age + I(age^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     32 36.636
2     30  3.514  2    33.122 141.38 5.356e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$H_0: \beta_1 = 0$

p-value is small, reject Ho

p-value is small, reject Ho

R^2 and adjusted R^2 for assessing overall adequacy of a model

1. The R^2 and adjusted R^2 are two measures for assessing the overall adequacy of a model.

2. $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$

   Adjusted $R^2 = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} = 1 - \frac{MS_{Res}}{SS_T/(n-1)}$

3. In general, R^2 almost increases when a regressor variable is added to the model, regardless of the contribution of that variable. Therefore, it is difficult to judge whether an increase in R^2 is actually due to the contribution of the variable.

4. Note that $SS_T = SS_{Res} + SS_R$, and SS_T is a constant for a given data set no matter how many variables are there in the model. Adjusted R^2 will only increase if the addition of a regressor variable reduces MS_{Res}, therefore adjusted R^2 is more useful in assessing the contribution of an additional variable.

Example - Delivery time data

```r
#R-square and adjusted R-square
summary(lm(time ~ dist + case))

```

75:41 (Top Level)

Console Background Jobs

R 3.4.1

```
> #R-square and adjusted R-square
> summary(lm(time ~ dist + case))

Call:
lm(formula = time ~ dist + case)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7880 -0.6629  0.4364  1.1566  7.4197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.341231   1.096730   2.135 0.044170 *  
dist        0.014385   0.003613   3.981 0.000631 ***
case        1.615907   0.170735   9.464 3.25e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,	Adjusted R-squared:  0.9559 
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

Example - Cholesterol age data

```
# coefficient of multiple determination
summary(lm(cholesterol~age))

# test overall fit of quadratic model
summary(lm(cholesterol~age + I(age^2)))

```

74:1 (Top Level)

Console Background Jobs

R 3.4.1

```
> # coefficient of multiple determination
> summary(lm(cholesterol~age))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.994130   0.184877   5.377 7.28e-06 ***
age         0.056330   0.004432  12.709 7.86e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4362 on 31 degrees of freedom
Multiple R-squared:  0.839,     Adjusted R-squared:  0.8338
F-statistic: 161.5 on 1 and 31 DF,  p-value: 7.856e-14

> # test overall fit of quadratic model
> summary(lm(cholesterol~age + I(age^2)))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6053871  0.3830121  -1.581    0.124
age          0.1512522  0.0213221   7.094 6.89e-08 ***
I(age^2)    -0.0011555  0.0002561  -4.512 9.20e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3423 on 30 degrees of freedom
Multiple R-squared:  0.9041,    Adjusted R-squared:  0.8977
F-statistic: 141.4 on 2 and 30 DF,  p-value: 5.356e-16
```

Tests and confidence intervals on individual regression coefficients

1. Linear model theory:

   $$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$ follows the t distribution with $n-p$ degrees of freedom.

2. For testing $H_0 : \beta_j = c$ versus $H_1 : \beta_j \neq c$, we reject H_0 if

   $$\frac{\hat{\beta}_j - c}{\sqrt{\hat{\sigma}^2 C_{jj}}} < -t_{\alpha/2,n-p} \quad \text{or} \quad \frac{\hat{\beta}_j - c}{\sqrt{\hat{\sigma}^2 C_{jj}}} > t_{\alpha/2,n-p}$$

3. A $100(1-\alpha)\%$ confidence interval for β_j is given as

   $$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

4. For example, to perform tests and construct confidence intervals on individual regression coefficients, the following R codes can be used for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$,

   ```
   summary(lm(y~x1+x2+x3))
   ```

   ```
   confint(lm(y~x1+x2+x3), level=0.95)
   ```

Example - Delivery time data

```r
#Test individual coefficient X1=case X2=dist
XPX <- t(X) %*% X
XPXI <- solve(XPX)
XPXI
betahat <- solve(XPX, XPy)
betahat
SSRes <- t(y) %*% y - bxy
SSRes
sigma2hat <- SSRes/(length(time)-3)
sigma2hat
t2 <-betahat[2]/(sigma2hat*XPXI[2,2])^0.5
t2
t3 <-betahat[3]/(sigma2hat*XPXI[3,3])^0.5
t3
summary(lm(time ~ case + dist))
confint(lm(time ~ case + dist), level=0.95)
```

$X'X$

$(X'X)^{-1}$

$\hat{\beta} = (X'X)^{-1}X'y$

```
> #Test individual coefficient X1=case X2=dist
> XPX <- t(X) %*% X
> XPXI <- solve(XPX)
> XPXI
              [,1]          [,2]          [,3]
[1,]  1.132152e-01 -4.448593e-03 -8.367257e-05
[2,] -4.448593e-03  2.743783e-03 -4.785709e-05
[3,] -8.367257e-05 -4.785709e-05  1.228745e-06
> betahat <- solve(XPX, XPy)
> betahat
           [,1]
[1,] 2.34123115
[2,] 1.61590721
[3,] 0.01438483
> SSRes <- t(y) %*% y - bxy
> SSRes
         [,1]
[1,] 233.7317
> sigma2hat <- SSRes/(length(time)-3)
> sigma2hat
         [,1]
[1,] 10.62417
> t2 <-betahat[2]/(sigma2hat*XPXI[2,2])^0.5
> t2
         [,1]
[1,] 9.464421
> t3 <-betahat[3]/(sigma2hat*XPXI[3,3])^0.5
> t3
         [,1]
[1,] 3.981313
```

```
> summary(lm(time ~ case + dist))

Call:
lm(formula = time ~ case + dist)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7880 -0.6629  0.4364  1.1566  7.4197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.341231   1.096730   2.135 0.044170 *  
case        1.615907   0.170735   9.464 3.25e-09 ***
dist        0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,	Adjusted R-squared:  0.9559 
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16

> confint(lm(time ~ case + dist), level=0.95)
                  2.5 %     97.5 %
(Intercept) 0.066751987 4.61571030
case        1.261824662 1.96998976
dist        0.006891745 0.02187791
```

The End