

# APPLIED STATISTICS

## Inferential Tools for Multiple Linear Regression

Dr Tao Zou

Research School of Finance, Actuarial Studies & Statistics  
The Australian National University

Last Updated: Mon Aug 21 14:40:09 2017

# Overview

- Sampling Distribution of Estimation
- Standard Error of Estimation
- Hypothesis Testing
  1.  $t$ -Test
  2.  $F$ -Test
- Confidence Intervals and Prediction Intervals

# References

1. **F.L. Ramsey and D.W. Schafer** (2012)  
Chapter 10 of *The Statistical Sleuth*
2. The slides are made by **R Markdown**.  
<http://rmarkdown.rstudio.com>

## Review: Estimation of MLR Parameters

For MLR

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \text{ where } X = (X_1, \cdots, X_k),$$

the LS estimates of  $\beta_0, \cdots, \beta_k$  are

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y},$$

where the  $n \times (k+1)$  matrix

$$\mathbb{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{k,1} \\ 1 & X_{1,2} & \cdots & X_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & \cdots & X_{k,n} \end{pmatrix} \text{ is called design matrix, and } \mathbb{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

# Fitted Values and Residuals

The estimated mean function is given by:

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k \text{ (plug-in idea).}$$

- The estimated mean is called the fitted or predicted value:

$$\text{fit}_i = \hat{Y}_i = \hat{\mu}\{Y_i|X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_k X_{k,i}.$$

- Residual:  $\text{res}_i = \hat{\mathcal{E}}_i = Y_i - \hat{Y}_i.$

## Sampling Distributions of $\hat{\beta}_j$ , $j = 0, \dots, k$

MLR model assumptions 1 & 2 & 3 can be described by

$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \mathcal{E}$ , where  $\mathcal{E} \sim N(0, \sigma^2)$ . It follows  
 $Y \sim N(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k, \sigma^2)$ .

Similarly to SLR, the sampling distribution for  $\hat{\beta}_j$  can be described by

$$\frac{\hat{\beta}_j - \beta_j}{\text{SD}(\hat{\beta}_j)} \sim N(0, 1), \text{ where}$$

$$\text{SD}(\hat{\beta}_j) = \sigma \sqrt{\mathbf{e}_{j+1}^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{e}_{j+1}}, \text{ and}$$

$$\mathbf{e}_{j+1} = \text{Row } j+1 \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \text{ is a } (k+1) \times 1 \text{ vector.}$$

## Standard Errors of $\hat{\beta}_j$ , $j = 0, \dots, k$

However, for a real dataset,  $\sigma$  is unknown. But we can estimate it by

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \text{res}_i^2}{n - k - 1}},$$

where  $n - k - 1$  is called the number of degrees of freedom. Here we can understand it as a number such that  $E(\hat{\sigma}^2) = \sigma^2$ .

- $k + 1$  is the number of regression coefficients in the model.
- The number of degrees of freedom arises from the fact we are now estimating  $k + 1$  regression coefficients instead of 2 in SLR.

**Standard deviation of  $\hat{\beta}_j$ :**

$$\text{SD}(\hat{\beta}_j) = \sigma \sqrt{\mathbf{e}_{j+1}^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{e}_{j+1}}.$$

**Standard error of  $\hat{\beta}_j$ :**

$$\text{SE}(\hat{\beta}_j) = \hat{\sigma} \sqrt{\mathbf{e}_{j+1}^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{e}_{j+1}}.$$

Plug-in idea.

## Practical Sampling Distributions of $\hat{\beta}_j$ , $j = 0, \dots, k$

Under the condition of the normal MLR model, it can be shown mathematically that

$$\frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \sim t_{n-k-1}, \text{ for } j = 0, \dots, k.$$

Here,  $\text{SE}(\hat{\beta}_j)$  is known, where  $n - k - 1$  here is from the number of degrees of freedom in the estimate  $\hat{\sigma}$ .

This theory leads directly to tests and confidence intervals for individual regression coefficients, in the familiar way.



## Hypothesis Testing for $\beta_j$ : $t$ -Test

$$H_0 : \beta_j = 0 \leftrightarrow H_a : \beta_j \neq 0.$$

$$\text{Test Statistic} = TS = \frac{\hat{\beta}_j - 0}{\text{SE}(\hat{\beta}_j)}$$

which should be compared to the  $t_{n-k-1}$  distribution.

The  $p$ -value is

$$p\text{-value} = 2 \times P(T > |TS|), \text{ where } T \sim t_{n-k-1}.$$

If  $p\text{-value} < \alpha \Rightarrow \text{reject } H_0$ ;  $p\text{-value} \geq \alpha \Rightarrow \text{not reject } H_0$ .

## Significance Depends on Other Explanatory Variables in the Model

**Example:** Suppose we are interested in predicting ANU students' 2nd year GPA ( $Y$ ) given their 1st year GPA ( $X_1$ ) and UAC score ( $X_2$ ). The following regression line is fit:

$$\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \quad (1)$$

Based on the data, the  $p$ -values for the  $t$ -tests of whether  $\beta_j = 0$  versus  $\beta_j \neq 0$  for  $j = 1, 2$  are 0.15 and 0.20, respectively. (This example is very similar to “Size of Wing” in Lecture Notes 5.)

Does this mean that both  $X_1$  and  $X_2$  are not needed in the model? NO!

The test for  $\beta_2$  tells us whether  $X_2$  is needed in the model that already contains  $X_1$ , i.e., does  $X_2$  offer any information about mean GPA over and above that of  $X_1$ ?

The meaning of the coefficient of an explanatory variable depends on what other explanatory variables have been included in the regression.

## Significance Depends on Other Explanatory Variables in the Model (Con'd)

If we fit the following two models:

$$\mu\{Y|X_1\} = \alpha_0 + \alpha_1 X_1 \text{ and } \mu\{Y|X_2\} = \gamma_0 + \gamma_2 X_2.$$

For both models, the  $p$ -values for the  $t$ -tests of  $\alpha_1 = 0$  versus  $\alpha_1 \neq 0$  and  $\gamma_2 = 0$  versus  $\gamma_2 \neq 0$  can be computed. Based on the data, the results of the  $p$ -values are 0.01 and 0.02, respectively.

Hence at least one of  $X_1$  and  $X_2$  is needed in the model.

In this example  $X_1$  and  $X_2$  are probably highly correlated so we might expect this to be the case. The following  $F$ -test of model (1) avoids this problem.

$H_0$  : none of  $X_1$  and  $X_2$  is needed in the model  $\leftrightarrow$

$H_a$  : at least one of  $X_1$  and  $X_2$  is needed in the model.

$\leftrightarrow$

$H_0 : \beta_1 = \beta_2 = 0 \leftrightarrow H_a : \text{at least one of } \beta_1 \text{ and } \beta_2 \text{ is not } 0.$

## *F*-Test

The *F*-test is used to test whether or not a **subgroup** of  $\beta_j$ ,  $j = 1, \dots, k$  (**several** regression coefficients) in MLR are all zeros, e.g., for the model

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

we consider the test

$$H_0 : \beta_1 = \beta_3 = 0 \leftrightarrow H_a : \text{at least one of } \beta_1 \text{ and } \beta_3 \text{ is not } 0.$$

**Remark:** compare to the *t*-test  $H_0 : \beta_j = 0$  (whether or not a **single**  $\beta_j$  in MLR is zero).

*t*-tests cannot be used to test a hypothesis involving more than one parameter.

In order to propose the test statistic for the *F*-test, we introduce the following terminology.

# Sum of Squared Errors (SSE)

The sum of squared errors (SSE) for a MLR

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \text{ where } X = (X_1, \cdots, X_k),$$

is defined by

$$\text{SSE} = \sum_{i=1}^n \text{res}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Sometimes, we call

$$\text{deviance} = \text{SSE}$$

in multiple linear regression model.

SSE (deviance) measures the goodness of fit for MLR.

Based on the definition of SSE (deviance), the smaller the SSE (deviance) is, the better fitting of a model.

## Example: Size of Wing (Con'd)

```
rm(list=ls())
setwd('~\\Desktop\\Research\\AppliedStat2017\\L6')
#install.packages('Sleuth3')
library(Sleuth3)
wing=ex0918
wingsize=wing$Females      #response variable
con=wing$Continent         #explan variable (categorical)
lat=wing$Latitude          #explan variable quantitative
#creating the indicator for North America
indNA=ifelse(con=="NA",1,0)
#fitting the MLR allowing for an interaction
wingint.reg=lm(wingsize~lat+indNA+indNA*lat)
sum((wingint.reg$residuals)^2)
```

```
## [1] 2107.605
```

```
deviance(wingint.reg)
```

```
## [1] 2107.605
```

## Full Model and Reduced Model

For MLR

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \text{ where } X = (X_1, X_2, X_3),$$

the following  $F$ -test is considered

$$H_0 : \beta_1 = \beta_3 = 0 \leftrightarrow H_a : \text{at least one of } \beta_1 \text{ and } \beta_3 \text{ is not } 0;$$

$$\Leftrightarrow$$

$$H_0 : \text{the **reduced model** } \mu\{Y|X_2\} = \beta_0 + \beta_2 X_2 \text{ is appropriate} \Leftrightarrow$$

$$H_a : \text{the **full model** } \mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \text{ is appropriate.}$$

We use  $d$  to denote the number of  $\beta$ s being tested. Here  $d = 2$ . Note that

$d$  = number of regression coefficients in the full model

– number of regression coefficients in the reduced model.

We further use  $k + 1$  to denote the number of regression coefficients in the full model. Here  $k + 1 = 4$ .

## Extra Sum of Squares and Drop in Deviance

We can compute  $SSE_{\text{full}} = \text{deviance}_{\text{full}}$  for the full model and  $SSE_{\text{reduced}} = \text{deviance}_{\text{reduced}}$  for the reduced model. We also compute  $\hat{\sigma}_{\text{full}}$  for the full model.

One can verify that

$$SSE_{\text{full}} \leq SSE_{\text{reduced}} \text{ and } \text{deviance}_{\text{full}} \leq \text{deviance}_{\text{reduced}}.$$

That means the full model has a better fitting compared to the reduced model.

Extra sum of squares is defined by

$$SSE_{\text{reduced}} - SSE_{\text{full}}.$$

Drop in deviance is defined by

$$\text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}}.$$



## Extra-Sums-of-Squares / Drop-in-Deviance $F$ -Test

The  $F$ -test statistic is

$$TS = \frac{(\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}})/d}{\hat{\sigma}_{\text{full}}^2} = \frac{(\text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}})/d}{\hat{\sigma}_{\text{full}}^2}.$$

This test statistic should be compared to  $F_{d,n-k-1}$  distribution, where  $n - k - 1$  is from the number of degrees of freedom in the estimate  $\hat{\sigma}_{\text{full}}$  for the full model.

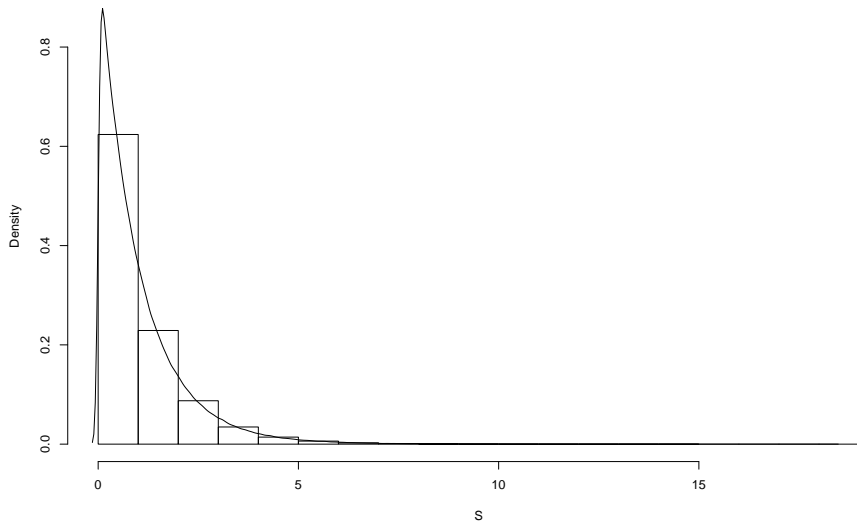
The  $p$ -value is

$$p\text{-value} = P(F > TS), \text{ where } F \sim F_{d,n-k-1}.$$

If  $p\text{-value} < \alpha \Rightarrow \text{reject } H_0$ ;  $p\text{-value} \geq \alpha \Rightarrow \text{not reject } H_0$ .

# F Distribution

F Distribution



# Special Cases of $F$ -Tests

Consider MLR

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \text{ where } X = (X_1, X_2, X_3).$$

**Testing the “overall significance”:**

The following  $F$ -test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \leftrightarrow H_a : \text{otherwise}$$

involves testing whether all the coefficients except  $\beta_0$ , equal to 0 or not.

This hypothesis proposes that none of the considered explanatory variables is useful in explaining the mean of response.

The test statistic and  $p$ -value for this test are given in the R summary(lm()) output.

# Special Cases of $F$ -Tests (Con'd)

## Testing a single coefficient:

$F$ -test can also be used to test

$$H_0 : \beta_j = 0 \leftrightarrow H_a : \beta_j \neq 0.$$

$$\text{Test Statistic} = TS = \frac{(\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}})/d}{\hat{\sigma}_{\text{full}}^2}$$

which should be compared to the  $F_{d,n-k-1}$  distribution, where  $d = 1$ .

For testing that a single coefficient is zero, the  $F$ -test and  $t$ -test are the same, i.e., they return the same  $p$ -value.

The test statistic in this case will be used for sequential variable selection.

## Confidence Intervals (CI) for $\beta_j$

Recall the practical sampling distributions of  $\hat{\beta}_j$ :

$$\frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \sim t_{n-k-1}, \text{ for } j = 0, \dots, k.$$

Using this information, a  $(1 - \alpha)$  CI for  $\beta_j$  is

$$\hat{\beta}_j \mp t_{n-k-1, \alpha/2} \times \text{SE}(\hat{\beta}_j)$$

where  $t_{n-k-1, \alpha/2}$  is the  $1 - \alpha/2$  quantile of  $t_{n-k-1}$ , namely

$$P(T \leq t_{n-k-1, \alpha/2}) = 1 - \alpha/2 \text{ or } P(T > t_{n-k-1, \alpha/2}) = \alpha/2$$

for  $T \sim t_{n-k-1}$ .

# Confidence Interval (CI) and Prediction Interval (PI)

## Confidence Interval

**Target:**  $\beta_0 + \beta_1 x_{1,0} + \cdots + \beta_k x_{k,0}$ ; **single estimation:**

$$\hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \cdots + \hat{\beta}_k x_{k,0}.$$

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \cdots + \hat{\beta}_k x_{k,0} - (\beta_0 + \beta_1 x_{1,0} + \cdots + \beta_k x_{k,0})}{\text{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \cdots + \hat{\beta}_k x_{k,0})} \sim t_{n-k-1}.$$

## Prediction Interval

**Target:**  $Y_{\text{new}}$ ; **single prediction:**  $\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_k x_{k,\text{new}}$ .

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_k x_{k,\text{new}}) - Y_{\text{new}}}{\text{SE}\{(\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_k x_{k,\text{new}}) - Y_{\text{new}}\}} \sim t_{n-k-1}.$$

The expressions of SEs here for MLR can be given in complicated matrix forms, and hence we do not provide.

## Confidence Interval (CI) and Prediction Interval (PI) (Con'd)

A  $(1 - \alpha)$  CI for

$\mu\{Y|X_1 = x_{1,0}, \dots, X_k = x_{k,0}\} = \beta_0 + \beta_1 x_{1,0} + \dots + \beta_k x_{k,0}$  is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \dots + \hat{\beta}_k x_{k,0}) \mp t_{n-k-1, \alpha/2} \times \text{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \dots + \hat{\beta}_k x_{k,0}).$$

A  $(1 - \alpha)$  PI for  $Y_{\text{new}}$  at  $(X_{1,\text{new}}, \dots, X_{k,\text{new}})$  is

$$\begin{aligned} &(\hat{\beta}_0 + \hat{\beta}_1 X_{1,\text{new}} + \dots + \hat{\beta}_k X_{k,\text{new}}) \mp t_{n-k-1, \alpha/2} \\ &\times \text{SE}\{(\hat{\beta}_0 + \hat{\beta}_1 X_{1,\text{new}} + \dots + \hat{\beta}_k X_{k,\text{new}}) - Y_{\text{new}}\}. \end{aligned}$$

Here,  $t_{n-k-1, \alpha/2}$  is the  $1 - \alpha/2$  quantile of  $t_{n-k-1}$ , namely

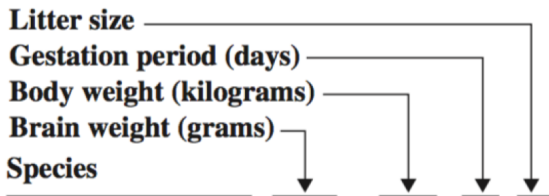
$$P(T \leq t_{n-k-1, \alpha/2}) = 1 - \alpha/2 \text{ or } P(T > t_{n-k-1, \alpha/2}) = \alpha/2$$

for  $T \sim t_{n-k-1}$ .

## Example: Brain Weight

(example from "The Statistical Sleuth")

The data are the average values of brain weight, body weight, gestation lengths (length of pregnancy), and litter size for 96 species of mammals. Since brain size is obviously related to body size, the question of interest is this: Which, if any, variables are associated with brain size, after accounting for body size?



<b>Litter size</b>				
<b>Gestation period (days)</b>				
<b>Body weight (kilograms)</b>				
<b>Brain weight (grams)</b>				
<b>Species</b>				
Quokka	17.5	3.5	26	1.0
Hedgehog	3.50	0.93	34	4.6
Tree shrew	3.15	0.15	46	3.0
Elephant shrew I	1.14	0.049	51	1.5
Elephant shrew II	1.37	0.064	46	1.5
Lemur	22	2.1	135	1.0

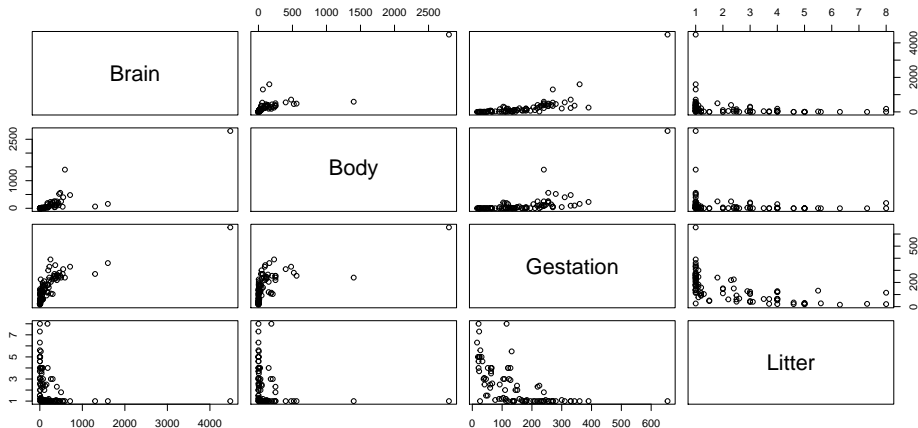


# R Code

```
brain<-case0902  
names(brain)
```

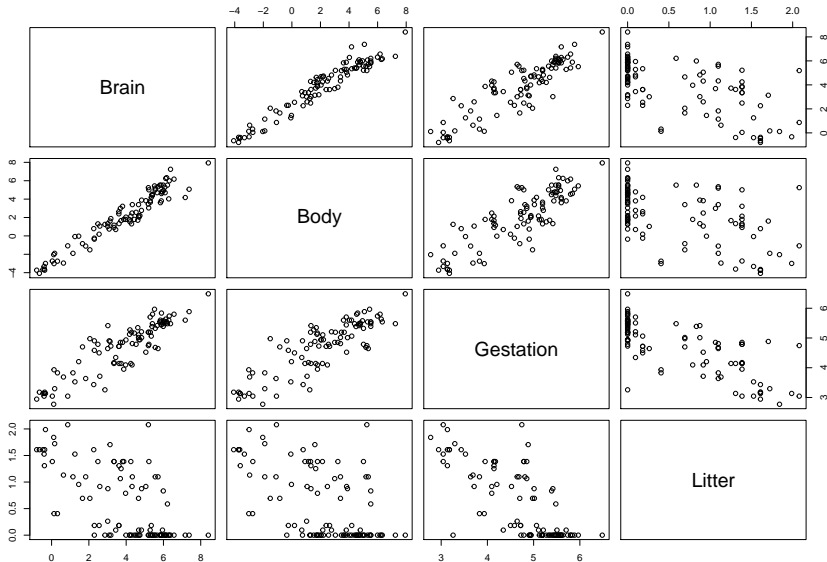
```
## [1] "Species" "Brain" "Body" "Gestation" "Litter"
```

```
pairs(brain[,-1]) #pairwise scatterplots
```



# R Code (Con'd)

```
pairs(log(brain[, -1])) #using the log transformation
```



# R Code (Con'd)

```
Y=log(brain$Brain)
X1=log(brain$Gestation)
X2=log(brain$Body)
X3=log(brain$Litter)
brain.reg = lm(Y ~ X1 + X2 + X3) #full model
summary(brain.reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95415 -0.29639 -0.03105  0.28111  1.57491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.85482     0.66167   1.292  0.19962
## X1           0.41794     0.14078   2.969  0.00381 **
## X2           0.57507     0.03259  17.647 < 2e-16 ***
## X3          -0.31007     0.11593  -2.675  0.00885 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4748 on 92 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9522
## F-statistic: 631.6 on 3 and 92 DF,  p-value: < 2.2e-16
```

## R Code (Con'd)

$$H_0 : \beta_1 = \beta_3 = 0 \leftrightarrow H_a : \text{otherwise.}$$

```
#extra-sums-of-squares test  
brain.regr=lm(Y~X2) #reduced model  
anova(brain.regr,brain.reg,test='F')
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: Y ~ X2
```

```
## Model 2: Y ~ X1 + X2 + X3
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)  
## 1      94 31.411  
## 2      92 20.736   2    10.675 23.681 5.053e-09 ***  
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## R Code (Con'd)

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \leftrightarrow H_a : \text{otherwise.}$$

```
#extra-sums-of-squares test  
brain.regr=lm(Y~1) #reduced model  
anova(brain.regr,brain.reg,test='F')
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: Y ~ 1
```

```
## Model 2: Y ~ X1 + X2 + X3
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1      95 447.81
```

```
## 2      92  20.74   3    427.08 631.6 < 2.2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## R Code (Con'd)

$$H_0 : \beta_3 = 0 \leftrightarrow H_a : \text{otherwise.}$$

```
#extra-sums-of-squares test  
brain.regr=lm(Y~X1+X2) #reduced model  
anova(brain.regr,brain.reg,test='F')
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: Y ~ X1 + X2
```

```
## Model 2: Y ~ X1 + X2 + X3
```

```
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1      93 22.349
```

```
## 2      92 20.736  1    1.6125 7.1541 0.008852 **
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## R Code (Con'd)

```
#CI  
x0=data.frame(X1=5,X2=2,X3=0.6)  
predict(brain.reg,x0,interval='confidence',level=0.95)
```

```
##           fit      lwr      upr  
## 1 3.908632 3.778657 4.038608
```

```
#PI  
Xnew=data.frame(X1=5,X2=2,X3=0.6)  
predict(brain.reg,Xnew,interval='prediction',level=0.95)
```

```
##           fit      lwr      upr  
## 1 3.908632 2.956812 4.860453
```