CS555B1 Data Analysis and Visualization

Lecture 7

Multiple Linear Regression

Kia Teymourian

Multiple Linear Regression

- MLR can be used to describe the relationships between a set of explanatory or **independent variables (x_1, x_2,…, x_k)** and a dependent variable (y)

- We are interested in the relationship between each independent variable and the dependent variable **after accounting for remaining independent variables**.

- MLR allows quantifying the relationship between our response variable and our explanatory variables as well as providing a tool for predicting the response of a new observation for a given set of values for x_1, x_2,… … , and x_k.

Multiple linear regression

The equation for the simple linear regression line is given by

$$y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

y is the response or dependent variable
x_1, x_2,…, x_k are the explanatory or independent variables
Beta$_0$ is the **intercept** (the value of y when x_1, x_2,…, x_k are set to 0)
Beta$_1$ is the **slope** (the expected change in y for each one-unit change in **x_1** after adjusting for **x_2,…, x_k**)
Beta$_2$ is the **slope** (the expected change in y for each one-unit change in **x_2** after adjusting for **x_1 , x_3,…, x_k**)
Beta$_k$ is the **slope** (the expected change in y for each one-unit change in **x_k** after adjusting for **x_1 , x_2,…, x_{k-1}**)

e is the **random error** which we assume is normally distributed with a mean of 0 and a variance of σ^2

Regression Equation

The equation for the multiple linear regression line is given by

$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \ldots + + \widehat{\beta_k}x_k$

$\hat{y}$ is the expected or predicted value of y for a given value of x_1, x_2,…, x_k
$\widehat{\beta_0}$ is the least-squares estimates of β_0 (the intercept)
$\widehat{\beta_1}$, $\widehat{\beta_2}$, … and $\widehat{\beta_k}$, is the least-squares estimates of β_1, β_2, …, β_k respectively

In the least-squares regression, the estimates are selected in such a way that the following quantity is minimized::

$$(y - \hat{y})^2 = (y - (\widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \ldots + + \widehat{\beta_k}x_k))^2$$

In multiple regression, the least squares estimates of β_1, β_2, …, β_k do not have simple closed form equations.

A MLR example - Book Blink

In the book Blink by Malcolm Gladwell, Gladwell states that a study of **CEOs of Fortune 500 companies** found that these individuals tend to be taller than the average US population. In order to study this phenomenon in more detail and to see if **height** is associated with **increased success** in business (as measured by salary), **100 men between the ages of 25 and 65** were polled for their **heights** (in inches) and **annual salaries**.

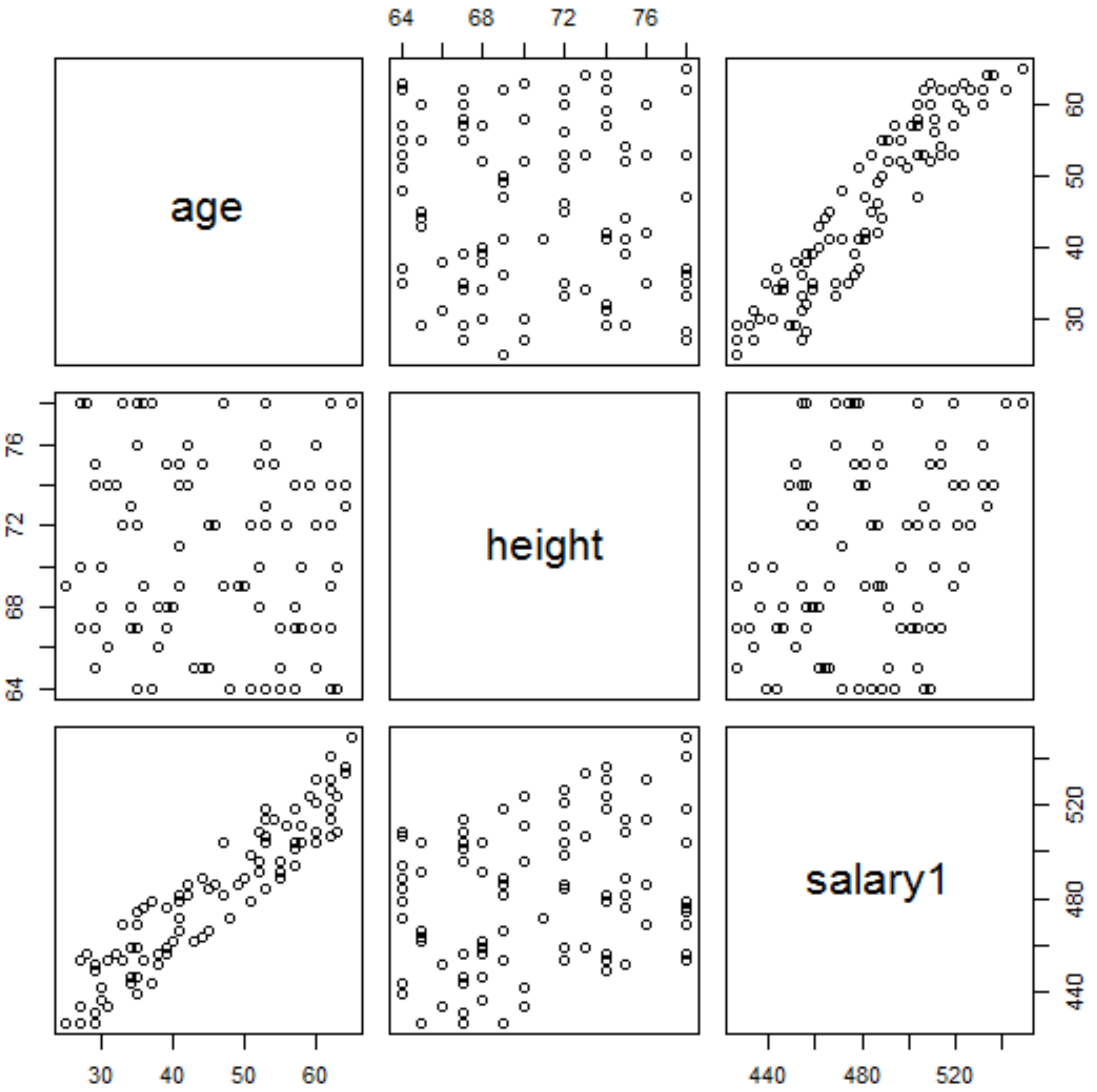
When fitting a multiple regression model, it is important to understand the associations between each of the **independent variables with the dependent variable** as well as the **associations between independent variables**

A MLR example – R commands

Scatterplot Matrix

```
> data <- read.csv("CEO_salary.csv")
> attach(data)
> salary1 <- salary/1000
> data1 <- data.frame(age, height, salary1)

> cor(data1)
> pairs(data1)
```



R command function

```
> m <- lm(salary1~age+height)
> m

Call:
lm(formula = salary1 ~ age + height)

Coefficients:
(Intercept)          age       height
    190.697        2.503        2.507

> summary(m)
```

Interpretation of a regression equation

The interpretation of a regression equation **focuses** on the **slope parameters for the independent variables**.

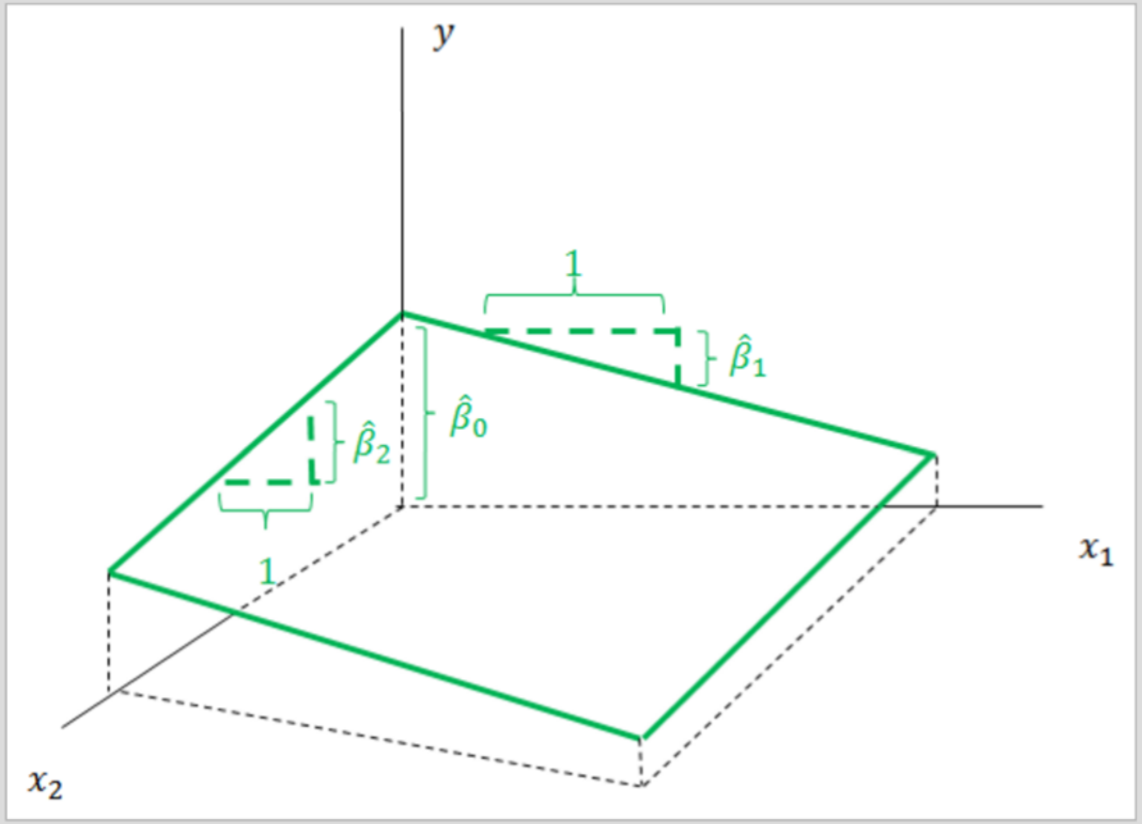
The estimate of the slope parameter **(^Beta$_i$)** in a MLR gives the expected or average change in the **response variable (y)** for a one unit increase in the independent variable **(x_i)** after **controlling for the other independent variables**.

The beta estimates, **(^Beta$_1$)** and **(^Beta$_2$)** from the regression of **x_1 and x_2 on y** **is not** the same as the beta estimates obtained from the **separate regression of $x_1$on y** and the **separate regression of x_2 on y**, respectively.

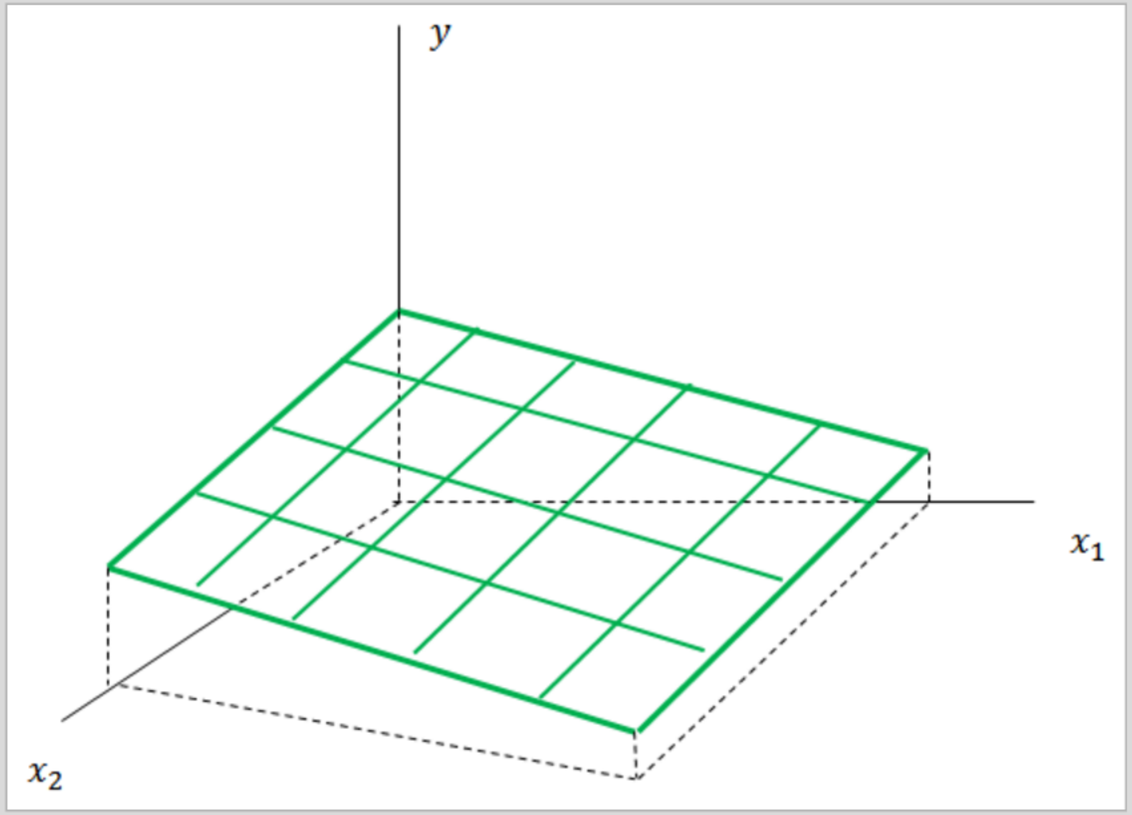
The equation of the least-squares regression line can be used **to predict the expected value of the response variable** for new values of the explanatory variables **x_1, x_2,…, x_k.**

Interpretation of the regression line

In MLR, the regression line corresponds to a **plane (k=2)** or **a hyperplane (k>2, a k-dimensional plane in a k+1-dimensional space)**.



The beta estimates define the surface of the plane. $\hat{\beta}_0$ is the expected value of y when x_1 and x_2 are 0. $\hat{\beta}_1$ is the slope of the surface projected onto the x_1, y plane. $\hat{\beta}_2$ is the slope of the surface projected onto the x_2, y plane.



The three dimensional plane is highlighted here. The perpendicular lines on the plane help reinforce the fact that for any given value of x_1 (for example), the regression asserts the same straight line relationship between x_2 and y. Similarly, for any given value of x_2 (for example), the regression asserts the same straight line relationship between x_1 and y.

Assessing the Fit of the Regression Line

The **coefficient of determination** represents the **proportion (percentage)** of the variation in the **response variable explained by the multiple regression** model.

The calculation of the coefficient of determination and the interpretation is the same in the **MLR setting as it was in the SLR setting** - the only difference now is that the predicted value **(^y) is based on the more complex regression equation** which involves more than one independent variable.

Given that there are more than one independent variable in this setting, the coefficient of determination is not simply the squared correlation coefficient.

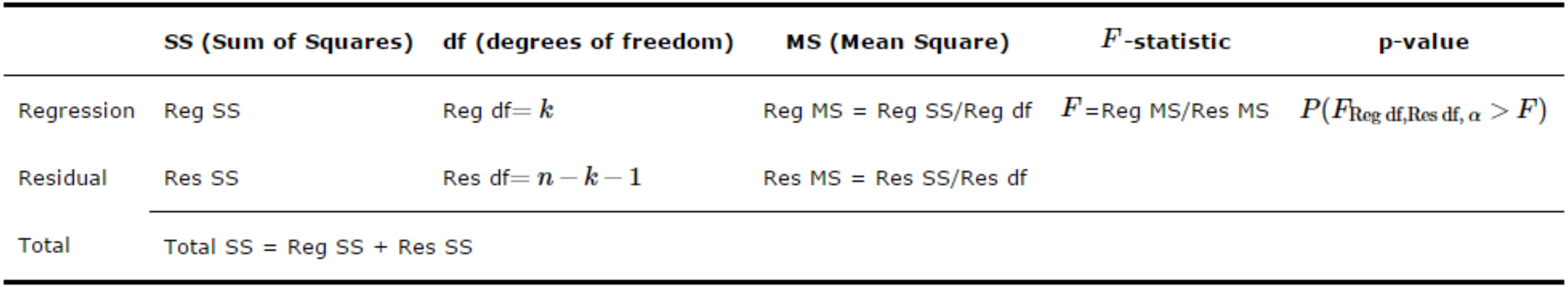
In the MLR setting, the **coefficient of determination** is often referred to as the **multiple R-squared.**

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} = \frac{\text{Reg SS}}{\text{Total SS}}$$

Inference – F test

- In MLR, the F-test for the model **sometimes is referred to as the global test**.
- We use the **ANOVA table** and are interested in testing the alternative hypothesis that **at least one of the slope parameters is different than 0**.
- If this test confirms that **there is at least one slope parameter that is different than 0,** then subsequent **F-tests or t-tests** can be used to assess whether each individual slope parameter is different than 0.

The general form of the ANOVA table: The **main difference from SLR is that, in this setting, k>1**. The exact value of **k depends on the number of variables in the model**.



	SS (Sum of Squares)	**df (degrees of freedom)**	**MS (Mean Square)**	F**-statistic**	**p-value**
Regression	Reg SS	Reg df$= k$	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	$P(F_{\text{Reg df},\text{Res df},\alpha} > F)$
Residual	Res SS	Res df$= n - k - 1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

Inference – F-test (continued)

$$Reg\,SS=\sum_{i=i}^{n}\left(\hat{y}_i-\bar{y}\right)^2 \qquad Res\,SS=\sum_{i=i}^{n}\left(y_i-\hat{y}_i\right)^2 \qquad Total\,SS=\sum_{i=i}^{n}\left(y_i-\bar{y}\right)^2$$

- **Reg df = k**, the degrees of freedom of Reg SS. It equals to the number of predictors in the model (that is, the number of parameters being estimated besides the intercept).

- Res df =n−k−1= the degrees of freedom of Res SS. It equals to the number of data points **minus the number of predictors in the model** (that is, the number of parameters being estimated besides the intercept) minus 1.

- **Reg MS = Reg SS/Reg df (the regression mean square)**

- **Res MS = Res SS/Res df (the residual mean square)**

- **F=Reg MS/Res MS** (the statistic which is the ratio of the regression mean square to the residual mean square)

- **p-value** = the probability that the observed value of test statistic or a more extreme value could have been observed by chance

An example: calculate R^2

- As displayed in the scatterplot matrix, CEO's salary is strongly associated with their age, and somewhat associated with their height.

- The least-squares regression equation for the data looking at the association between CEO's salary and these factors was calculated to be

  $\boldsymbol{y^{\wedge}}$ **=191+2.5*age+2.5*height.**

- Calculate the multiple R-squared and give its interpretation.

- **Reg df =2** and **Res df = n–k–1=n−3=100−3=97**.

An example: calculate R² (continued)

```
> totalss <- sum((salary1 - mean(salary1))^2)
> regss <- sum((fitted(m) - mean(salary1))^2)
> resiss <- sum((salary1-fitted(m))^2)
> fstatistic <- (regss/2)/(resiss/97)
> pvalue <- 1-pf(fstatistic, df1=2, df2=97)
> R2 <- regss/totalss
```

in R
fitted is a generic function which extracts fitted values from objects returned by modeling functions

- 99% of the variability in CEO’s salary can be explained by age and height. (Based on results of R^2)
- In **MLR**, the first formal tests of hypotheses is for the overall model. They test the **null hypothesis that all slope coefficients are equal to 0 (H0:$\beta_1=\beta_2=\cdots=\beta_k=0$)** versus the alternative that **at least one of the slope coefficients is different from zero (H1:$\beta_i\neq 0$ for at least one i)**.
  The null hypothesis is the same as asserting that there is no linear relationship between the response and explanatory variables.
- The general principle behind the test is that the **null hypothesis is rejected if there is at least one ^β_i** that is **sufficiently far from 0** or (equivalently) a large majority of the total sum of squares is explained by the regression.
- The global test for the regression is an **F-test using information from the ANOVA table**.

F-test for MLR

$$F = \frac{Reg\,MS}{Res\,MS}$$

F-distribution **with k and n−k-1 degrees of freedom** under H_0.

The decision rule for a level α test is:

- Reject H_0 : $H0{:}\beta_1=\beta_2=\cdots=\beta_k=0$ if $F \geq F_{k,n-k\text{-}1,\alpha}$
- Otherwise, do not reject H_0
  where $F_{k,n-k\text{-}1,\alpha}$ is the value from the F-distribution table with **k degree of freedom (numerator) and n−k-1 degrees of freedom (denominator)** and associated with a right-hand tail probability of α.

Quantities from the F-distribution

- **Calculating probability from F-statistics**

Use pf() function to calculate the area to the left of a given F-statistic

> **pf**([F statistic], df1=[degree of freedom of the numerator], df2=[degree of freedom of the denominator])

- **Calculating F-statistics from probability**

Use qf() function to calculate F-statistic with the specifies area to the left

> **qf**([probability], df1=[degree of freedom of the numerator], df2=[degree of freedom of the denominator])

A example: F-test for MLR

Is there a linear relationship between COE's salary and their age and height? Perform this test at the ɑ=0.01 level.

1. Set up the hypotheses and select the alpha level
H_0 :β_{age}=β_{height}=0 (age and height are not significant predictors of annual salary)
H_1 :β_{age}≠0 and/orβ_{height}≠0 (at least one of the slope coefficients is different than 0; age and/or height are significant predictors/is a significant predictor of annual salary))
ɑ=0.01

2. Select the appropriate test statistic
$F = \frac{Reg\ MS}{Res\ MS}$ with 2 and n−3=97 degrees of freedom

3. State the decision rule
F-distribution with 2, 97 degrees of freedom and associated with ɑ=0.01.
> qf(.99, df1=2, df2=97)
$F_{2,97,0.01}$=4.83
Decision Rule: Reject H_0 if F≥4.83
Otherwise, do not reject H_0

A example: F-test for SLR (continued)

4. Compute the test statistic

```
> fstatistic <- (regss/2)/(resiss/97)
3.56445e+10
```

Or

```
> summary(m)
```

Or

```
> pvalue <- 1-pf(fstatistic, df1=2, df2=97)
```

5. Conclusion

Reject H_0 since f statistic≥ 4.83. We have significant evidence at the $\alpha=0.01$ level that $\beta_{age} \neq 0$ and/or $\beta_{height} \neq 0$. That is, there is evidence of a linear association between annual salary and age and/or height (here, $p<0.001$ as calculated using software program).

Inference – t test

If the overall model is significant (that is, if the **F-test testing rejects the null hypothesis** in favor of the alternative), then **the significance could be attributed to any one of the independent variables**.

The **next step is to perform testing on each individual parameter to identify** the relative contribution of each independent variable.

In order to test each if **Beta$_i$=0** after controlling for the other independent variables in the model, we use a t statistic:

$$t = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

where SE$_{\wedge Bi}$ the standard error of the estimate of **(in the regression model with the other independent variables included)** which follows a **t-distribution with n−k−1** degrees of freedom under H$_0$.

Inference – t test (continued)

The decision rule for a two-sided level α test is:

Reject $H_0 : \beta_i = 0$ if $|t| \geq t_{n-k\text{-}1,\alpha/2}$

Otherwise, do not reject $H_0 : \beta_i = 0$

where $t_{n-k\text{-}1,\alpha/2}$ is the value from the t-distribution table with n−k−1 degrees of freedom and associated with a right hand tail probability of α/2.

We can also calculate the two-sided 100% × (1−α) confidence interval for β_i using the following formula:

$$\hat{\beta}_i \pm t_{n-k-1,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_i}$$

We can say with 100% × (1−α) confidence that the true value of β_i is between $\widehat{\beta_i} - t_{n-k\text{-}1,\alpha/2} * SE_{\widehat{\beta_i}}$ and $\widehat{\beta_i} + t_{n-k\text{-}1,\alpha/2} * SE_{\widehat{\beta_i}}$ after controlling for the other independent variables in the model.

A example: t-test for MLR

Is age a significant predictors of annual salary after controlling for height?

```
>summary(m)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.907e+02  1.884e-03  101223   <2e-16 ***
age         2.503e+00  9.862e-06  253808   <2e-16 ***
height      2.507e+00  2.541e-05   98679   <2e-16 ***
```

Perform a t-test at the ɑ=0.01 level and calculate the 99% confidence interval for β_{age}.

1. Set up the hypotheses and select the alpha level

H_0 : β_{age} =0 (after controlling for height)

H_1 : $\beta_{age} \neq 0$ (after controlling for height)

ɑ=0.01

2. Select the appropriate test statistic

$t = \frac{\hat{\beta}_{age}}{SE_{\hat{\beta}_{age}}}$ with df = n−3=100-3 = 97 degrees of freedom

A example: F-test for SLR (continued)

3. State the decision rule

- Determine the appropriate value from the t-distribution with 97 degrees of freedom and associated with a right hand tail probability of α/2=0.01/2=0.005
- $t_{n-k\text{-}1,\alpha/2}=t_{97,0.005}=2.63$

```
> qt(0.995, df=97)
```

- Decision Rule: Reject H0 if |t|≥2.63
- Otherwise, do not reject H0

4. Compute the test statistic

Using R function summary(m), we get the table:

$$t=\frac{\widehat{\beta}_{age}}{SE_{\widehat{\beta}_{age}}}=\frac{2.503}{9.862e\text{-}06}=253802.5 \text{ with df=97}$$

$$\widehat{\beta}_{age}\pm t_{97,\,0.005}\,SE_{\widehat{\beta}_{age}}=2.503 \pm 2.63*9.862e\text{-}06 = (2.5030,\ 2.5029)$$

Regression Diagnostics

- After fitting a regression model it is important to determine **whether all the necessary model assumptions are valid before performing inference**.

- If there are **any violations, subsequent inferential procedures** may be invalid resulting in faulty conclusions.

- **Regression Diagnostics** is an activity that we perform after fitting a regression model to check ***if the model fits the data well*** and ***if the assumptions underlying the theory were met*** in order to have confidence in the inferences we made from the regression model.

- These techniques often involve **visualizing the fit of the regression model via examination of the residuals** in order to identify data points that **don’t seem to fit the trend and issues with violations** of the assumptions of the regression model.

Residual Plots

- One of the most powerful tools in regression diagnostics are **residual plots** which help visualize how well a regression equation fits the sample data.

- **Residual plots are scatterplots** of the **regression residuals** [plotted on the y-axis] against the **explanatory variable** [plotted on the x-axis]).

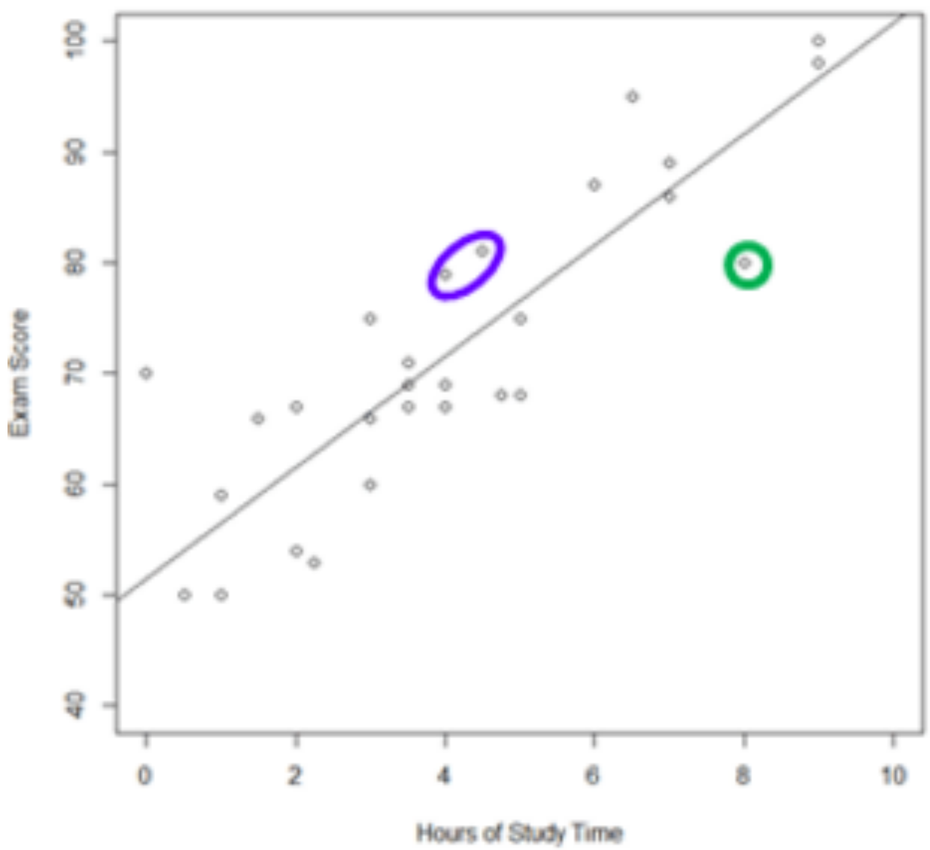
- The residual plot **turns the regression line on the horizontal** so it is easier to see patterns and **pick out unusual observations**.

- Residual plots can also be **generated using the predicted values** on the x-axis as opposed to the explanatory variable (which is especially **helpful in multiple regression analysis** when there are more than one explanatory variables).

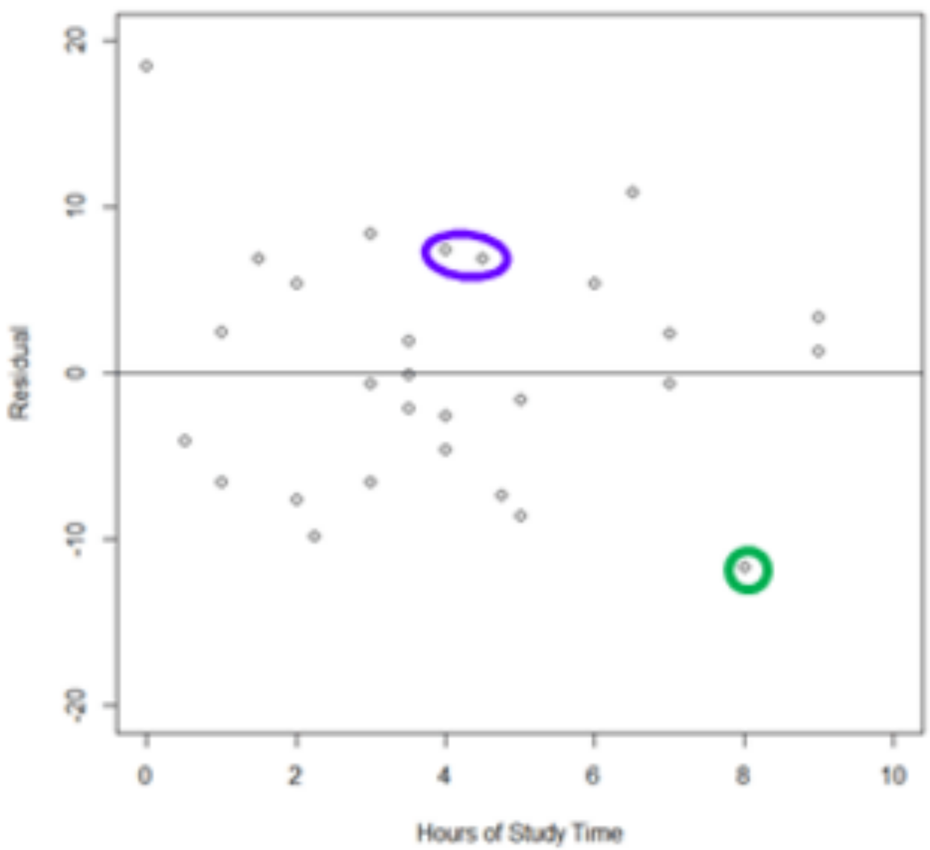
- Residual plots **can also be generated by plotting standardized or studentized residuals** (which involves **dividing** the residual by an **estimate of the variability of the residuals**).

Residual Plots

- The residual plot turns the regression line on the horizontal so it is easier to see patterns and pick out unusual observations.



Scatterplot of Exam Score versus Hours of Study Time



Residual Plot for the Regression of Exam Score on Hours of Study Tim

Residual Plots

```
> resid(m)
> par(mfrow=c(2,2))
> plot(fitted(m), resid(m), axes=TRUE, frame.plot=TRUE, xlab='fitted values',
ylab='residue')
> plot(age, resid(m), axes=TRUE, frame.plot=TRUE, xlab='age', ylab='residue')
> plot(height, resid(m), axes=TRUE, frame.plot=TRUE, xlab='height', ylab='residue')
> hist(resid(m))
```

in R

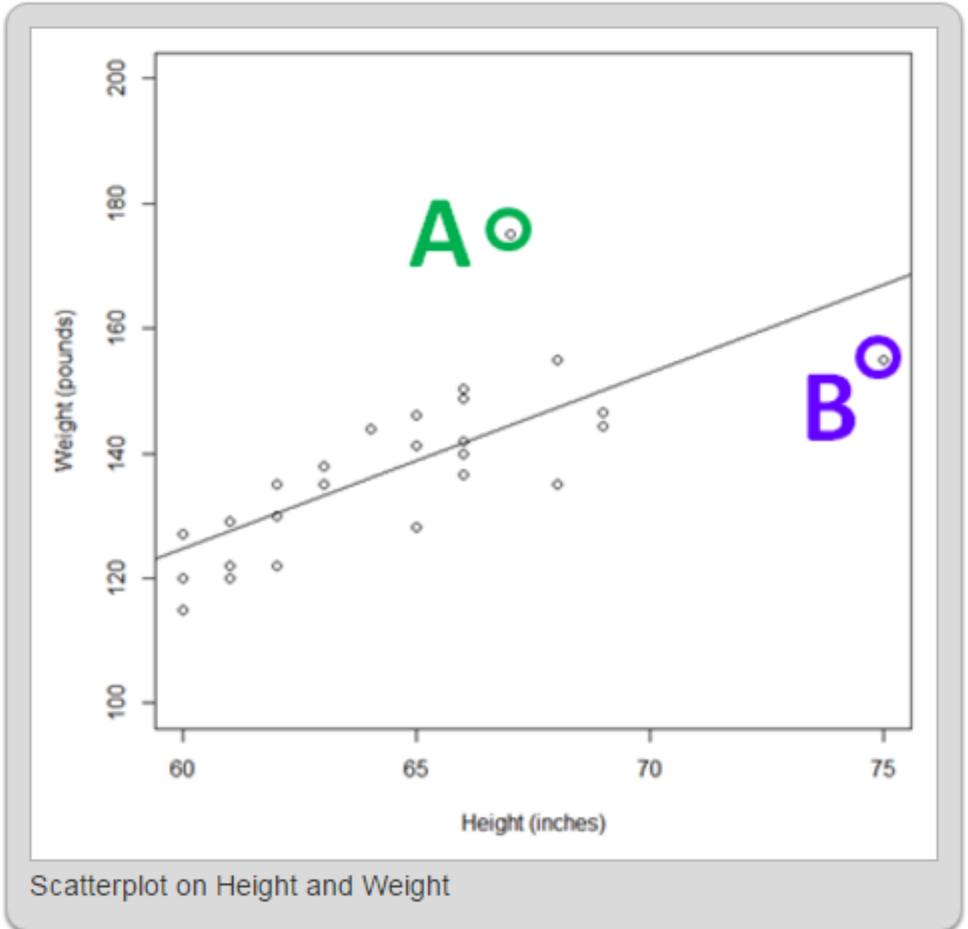
fitted is a generic function which extracts fitted values from objects returned by modeling functions

In R resid()

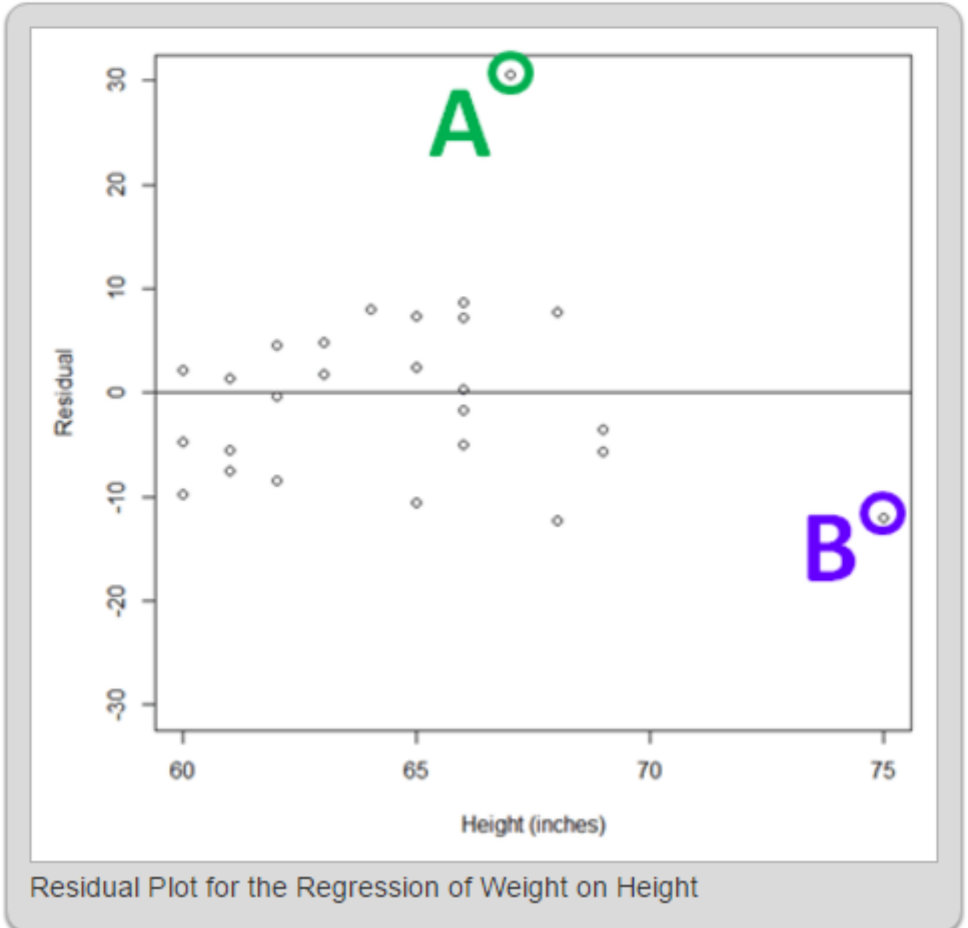
residuals is a generic function which extracts model residuals from objects returned by modeling functions.

Outliers and Influence Points

- Outliers in the y-direction tend to have large residuals.
- Outliers in the x-direction may or may not have large residuals but have the potential to be influential.
- An influence point is an observation that markedly changes the result of the regression if it were to be removed from the calculation.



Scatterplot on Height and Weight



Residual Plot for the Regression of Weight on Height

Outliers and Influence Points

- Outliers are observations that lie outside the overall pattern of the other observations.

- Outliers in the y-direction tend to have large residuals. Outliers in the **x-direction may** or may not have large residuals **but have the potential to be influential.**

- Outliers can be identified **via review of the scatterplot**.

- An **influence point is an observation that markedly changes** the result of the regression if it were to be removed from the calculation.

Outliers and Influence Points

- Given that least-squares regression is based on **minimizing the squared vertical distances between the observations and the regression line**, extreme points in the **x-direction tend to pull the regression line close** to itself.

- In these cases, the regression line equation may be quite different with or without the points and thus the points influences the regression equation.

- The influence of a particular point should be examined by removing it from the regression calculation and checking how the equations, inference, and conclusions change with its removal.

- When there appears to be observations out of range, one should always check to ensure that there was not an issue with data entry/recording.

- If an outlier in the x-direction, for example, is kept, it may be desirable to collect additional data within the same range to better characterize the relationship and so that the regression doesn’t depend so heavily on the data from a single observation.

Assumptions of the least-square regression

The following conditions must all be met before it is appropriate to make inference from a least-squares regression:

- The true ***relationship is linear***.
- The observations are **independent**.
- The variation of the **response variable around the regression line is constant**.
- The **residuals are normally distributed**.

Check the linear model you built

Is the model statistically significant?

- Check the F statistic (at the bottom of the summary)

Are the coefficients significant?

- Check the coefficient’s t statistics and p-values in the summary, or check their confidence intervals

Is the model useful?

- Check the R^2 near the bottom of the summary

Does the model fit the data well?

- Plot the residuals and check the regression diagnostics

Does the data satisfy the assumptions behind linear regression?

- Check if the diagnostics confirm that a linear model is reasonable for your data

Page 267-8, **R cookbook**