Multiple Linear Regression - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Summer 2021

Multiple Linear Regression (MLR)

- MLR can be used to describe the relationships between a set of explanatory or independent variables $(x_1, x_2, \ldots, x_k)$ 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 $x_1, x_2, \ldots,$ and x_k.

Multiple Linear Regression (MLR)

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
- $x_1, x_2, \ldots, x_k$ are the explanatory or independent variables
- β_0 is the intercept (the value of y when the $x_1, x_2, \ldots, x_k$ are set to 0)
- β_1 is the slope (the expected change in y for each one-unit change in x_1 after adjusting for $x_2, \ldots, x_k$)
- β_k is the slope (the expected change in y for each one-unit change in x_k after adjusting for $x_1, x_2, \ldots, x_{k-1}$)
- e is the random error which we assume is normally distributed with **a mean of 0 and a variance of σ^2**

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_n x_n$$

- The ideal scenario, all the parameters are independent / uncorrelated
  - Each parameter's impact can be assessed separately
- Correlation of predictors cause problems
  - Interpretation of the impact become hazardous – be careful!

Example:

- Y: # of tackles of a football player
  - W: Weight
  - H: Height
- Y = b + 0.5 W - 0.1 H -> which means lower the weight higher the tackles!!

Regression Equation

The equation for the multiple linear regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

where

- $\hat{y}$ (read y hat) is the expected or predicted value of y for a given values of $x_1, x_2, \ldots, x_k$
- $\hat{\beta}_0$ is the least-squares estimates of (the intercept)
- $\hat{\beta}_1, \hat{\beta}_2, \ldots$ and $\hat{\beta}_k$ are the least-squares estimates of $\beta_1, \beta_2, \ldots$ and β_k respectively

Regression Equation

The equation for the multiple linear regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

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

$$(y - \hat{y})^2 = \left[y - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k\right)\right]^2$$

A MLR Example

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.

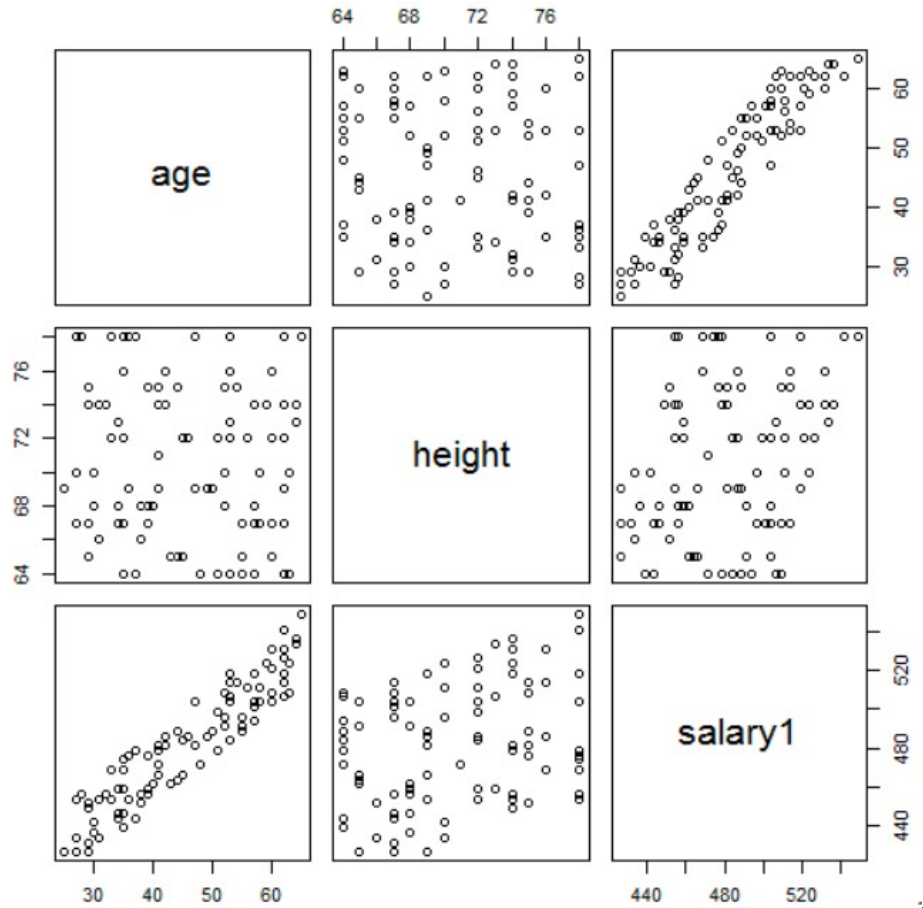
In order to study this phenomenon in more detail and to see if height is associated with increased success in business (as measured by salary), 100 men between the ages of 25 and 65 were polled for their heights (in inches) and annual salaries.

A MLR Example R Commands - Scatterplot Matrix

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

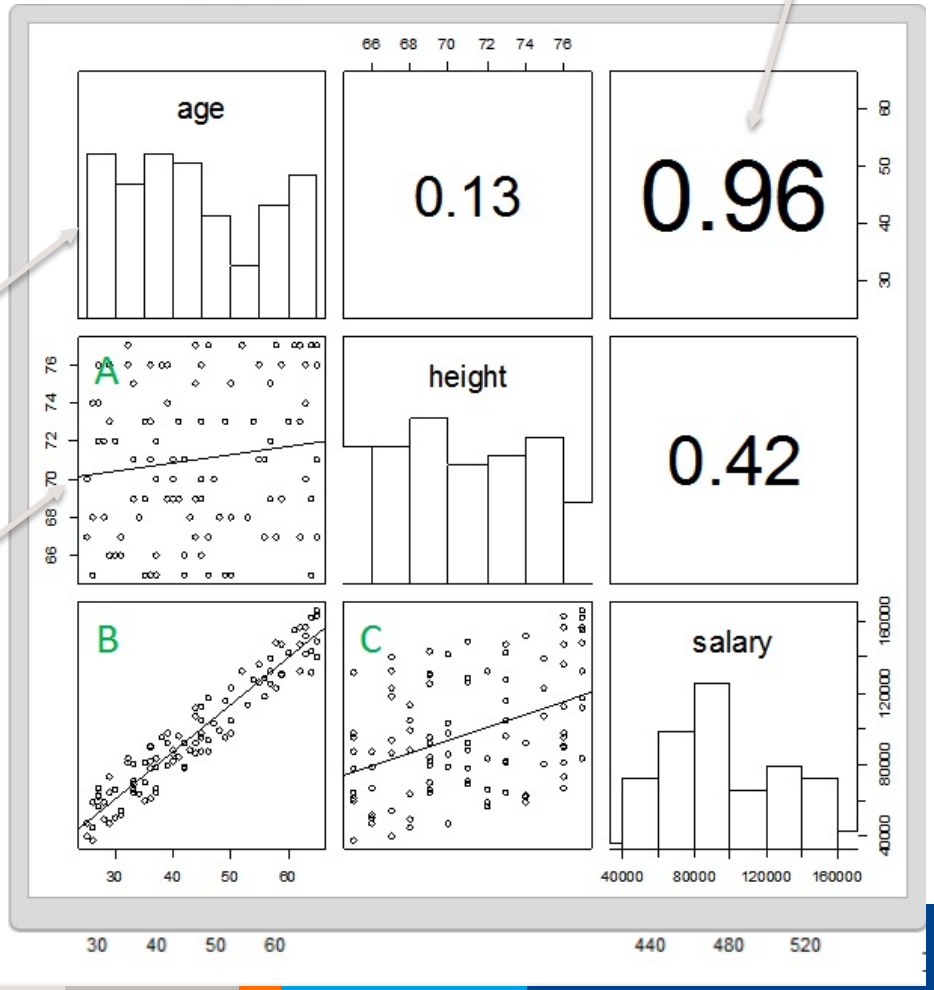
A MLR Example R Commands - Scatterplot Matrix

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



A MLR Example R Commands - Scatterplot Matrix

```
# 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

Finding the regression coefficients.

```
# Build a multiple linear regression model.

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

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

Coefficients:
(Intercept)          age       height
    190.697        2.503        2.507

# pass the model to summary function
> summary(m)
```

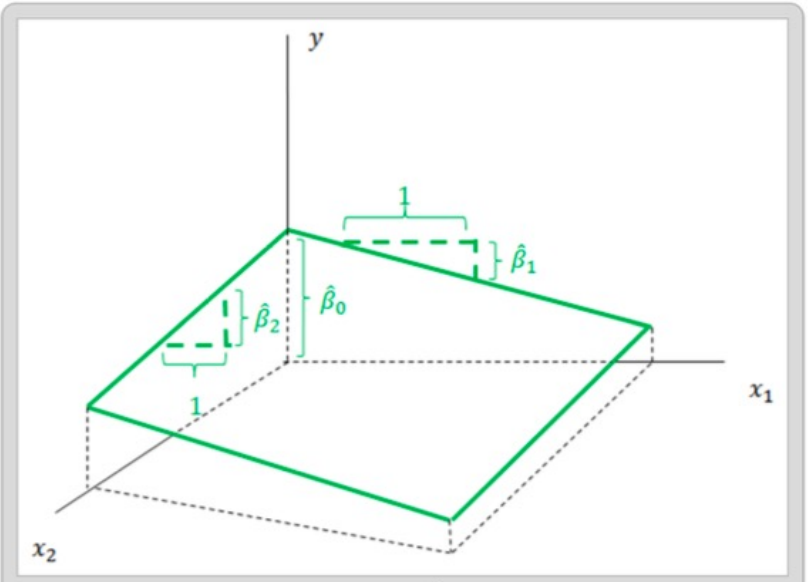
Interpretation of a regression equation

- **Focus is more on slope parameters**
- The slope parameter ($\hat{\beta}_i$) 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.
- **Beta estimates from MLR regression is not the same as beta for separate SLR**

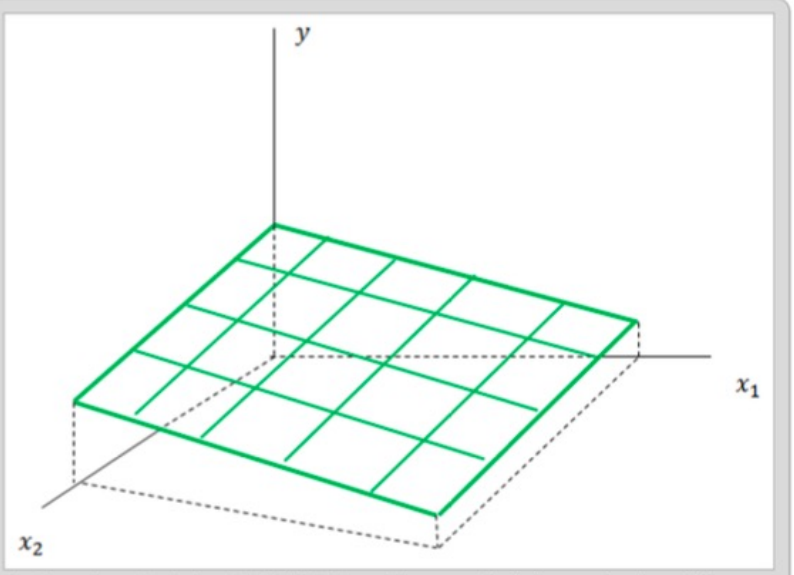
We use the equation of the least-squares regression line **to predict the expected value of the response variable for new values of the explanatory variables.**

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.

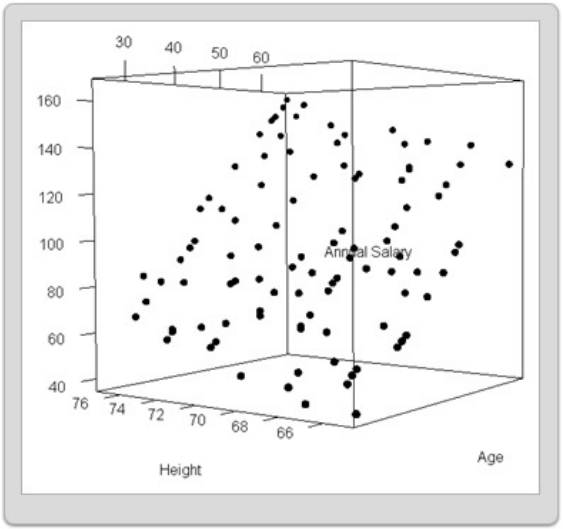
Example – CEO Salaries

- The least-squares regression line of annual salary (in dollars) and height (in inches) and age (in years) is as follows

$$\hat{y} = -190{,}700 + \$2507X_{height} + \$2503X_{age}$$

3D view of data

- Example: What is your brother's salary
  - Age:33
  - Height: 5'9"



Example – CEO Salaries

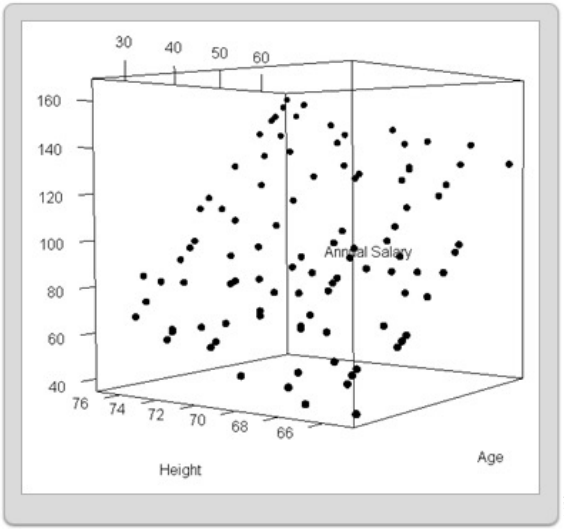
- The least-squares regression line of annual salary (in dollars) and height (in inches) and age (in years) is as follows

$$\hat{y} = -190{,}700 + \$2507X_{height} + \$2503X_{age}$$

3D view of data

- Example: What is your brother's salary
  - Age:33
  - Height: 5'9"
  - Answer:

$$\hat{y} = -\$190{,}700 + \$2507(69) + \$2503(33) \approx \$65{,}000$$



Assessing the Fit of the Regression Line - coefficient of determination

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

- Coefficient of determination (R^2) is the same as SLR - in MLR we have more variables.
- Given that there are more than one independent variable in this setting, the **coefficient of determination is not simply the squared correlation coefficient.**

In MLR, R^2 is referred to as the multiple R-squared.

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

Inference Global F 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.
- In MLR, the F-test for the model is referred to as the **global test.**

Inference Global F 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.
- In MLR, the F-test for the model is referred to as the **global test.**

Difference to SLR is that, here $k > 1$. 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 - Anova Table Components

	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	$P(F_{\text{Reg df,Res df},\alpha} > F)$
Residual	Res SS	Res df= $n-k-1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

- Reg SS=$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, Res SS=$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, Total SS=$\sum_{i=1}^{n}(y_i - \bar{y})^2$
- **Reg *df* = *k* equals to the number of predictors in the model.**
- **Res *df* = *n* – *k* – 1** equals to sample size minus the number of predictors in the model 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
- 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, CEOs 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 CEOs salary and these factors was calculated to be

$\hat{y} = 191 + 2.5 * age + 2.5 * height.$

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 R^2 (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
```

99% of the variability in CEOs salary can be explained by age and height.

Global F-Test

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
  $(H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0)$
  The null hypothesis is the same as asserting that there is **no linear relationship between the response and explanatory variables.**
- alternative that at least one of the slope coefficients is different from zero
  $(H_1 : \beta_i \neq 0$ **for at least one** $i)$.

The null hypothesis is rejected if there is at least one that is sufficiently far from 0 or (equivalently) a large majority of the total sum of squares is explained by the regression.

F-Test for MLR

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

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

The decision rule for a level α test is:

Reject $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ if $F \geq F_{k,n-k-1,\alpha}$

Otherwise, do not reject H_0

where **$F_{k,n-k-1,\alpha}$ is the value from the F-distribution** 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 α=0.01 level.

1. Set up the hypotheses and select the alpha level

$H_0 : \beta_{age} = \beta_{height} = 0$ (age and height are not significant predictors of annual salary)

$H_1 : \beta_{age} \neq 0$ and/or$\beta_{height} \neq 0$ (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))

$\alpha = 0.01$

2. Select the appropriate test statistic

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

with 2 and n-3=97 degrees of freedom

An Example: F-test for MLR

3. State the decision rule

F-distribution with 2, 97 degrees of freedom and associated with $\alpha = 0.01$.

```
> qf(.99, df1=2, df2=97)
```

$F_{2,97,0.01} = 4.83$

Decision Rule: Reject H_0 if $F \geq 4.83$

Otherwise, do not reject H_0

An Example: F-test for MLR

4. Compute the test statistic

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

# Or
# pass the model reference to summary function
> 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$\beta_{height} \neq 0$

There is evidence of a linear association between annual salary and age and/or height (here, $p < 0.001$ as calculated using software program).

MLR Inference t-test

If the overall model is significant, then the significance could be attributed to any one of the independent variables.

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

$SE_{\hat{\beta}_i}$ 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.

MLR Inference t-test

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

- **Reject $H_0 : \beta_i = 0$ if $|t| \geq t_{n-k-1,\alpha/2}$**
- Otherwise, do not reject $H_0 : \beta_i = 0$

where

$$t_{n-k-1,\alpha/2}$$

is the value from the t-distribution with $n - k - 1$ degrees of freedom and associated with a right hand tail probability of $\alpha/2$.

MLR Inference t-test Confidence Interval

We can calculate the two-sided $100\%(1-\alpha)$ confidence interval for using the following formula:

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

MLR Inference t-test Confidence Interval

We can calculate the two-sided $100\%(1-\alpha)$ confidence interval for using the following formula:

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

We can say with $100\% \times (1-\alpha)$ confidence that the true value of is between

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

after controlling for the other independent variables in the model.

An Example: t-test for MLR

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

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

An Example: t-test for MLR

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

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

1. Set up the hypotheses and select the alpha level

$H_0 : \beta_{age} = 0$ (after controlling for height)

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

$\alpha = 0.01$

2. Select the appropriate test statistic

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

with $df = n - 3 = 100 - 3 = 97$ degrees of freedom

An Example: t-test for MLR

```
>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 ***
```

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

$\alpha/2 = 0.01/2 = 0.005$

$t_{n-k-1,\alpha/2} = t_{97,0.005} = 2.63$

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

Decision Rule: Reject H_0 if $|t| \geq 2.63$

Otherwise, do not reject H_0

An Example: t-test for MLR

4. Compute the test statistic Using R function summary(m), we get the table:

$$t = \frac{\hat{\beta}_{\text{age}}}{SE_{\hat{\beta}_{\text{age}}}} = \frac{2.503}{9.862e{-}06} \approx 253808 \text{ with } df = 97$$

Reject H_0 since $253808 \geq 2.627$. We have significant evidence at the $\alpha = 0.01$ level that $\beta_{age} \neq 0$ after controlling for height. We can calculate the confidence interval.

$$\hat{\beta}_{\text{age}} \pm t_{n-k-1,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_{\text{age}}} = 2.503 \pm 2.627 \cdot 9.862e-06 = (2.5029702.503022)$$

Computation in R:

```
> m <- lm(salary1~age+height)
> confint(m, level =0.99)
                 0.5 %     99.5 %
(Intercept) 190.692341 190.702241
age           2.502970   2.503022
height        2.506966   2.507099
```

Multiple Linear Regression Interpretation

- Ideally independent parameters are uncorrelated
- Note that the independent variables might be correlated
  - As a result variance of all coefficients tends to increase. Example is having two highly correlated coefficient. In this case coefficient of the two parameters can get exchanged with no change in accuracy of results.
  - Interpretation is hard and hazardous
- Avoid concluding relationship between parameters and the output, because of impact of other parameters

Example

- Predicting number of tackles by a football player in a season:

$$y = b_0 + 0.5\, Weight - 0.1\, Height$$

- This means it is better to be short football player!

Qualitative Parameters

- Example of group one vs group 2, like male vs female.
- $x_i = \begin{cases} 1, if\ ith\ sample\ is\ female \\ 0, if\ ith\ sample\ is\ male \end{cases}$
- The model becomes
- $y = \beta_0 + \beta_1 x_i + e =$
  $\begin{cases} \beta_0 + \beta_1, & if\ ith\ sample\ is\ female \\ \beta_0, & if\ ith\ sample\ is\ male \end{cases}$

Qualitative Parameters

- Example of group one vs 2 cs 3, race
- $x_{1i} = \begin{cases} 1, if\ ith\ sample\ is\ in\ group\ 1 \\ 0, if\ ith\ sample\ is\ not\ in\ group 1 \end{cases}$
- $x_{2i} = \begin{cases} 1, if\ ith\ sample\ is\ in\ group\ 2 \\ 0, if\ ith\ sample\ isn't\ in\ group\ 2 \end{cases}$
- The model becomes
- $y = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} = \begin{cases} \beta_0 + \beta_1, & if\ ith\ sample\ is\ group\ 1 \\ \beta_0 + \beta_2, & if\ ith\ sample\ is\ group 2 \\ \beta_0, & if\ ith\ sample\ is\ group\ 3 \end{cases}$

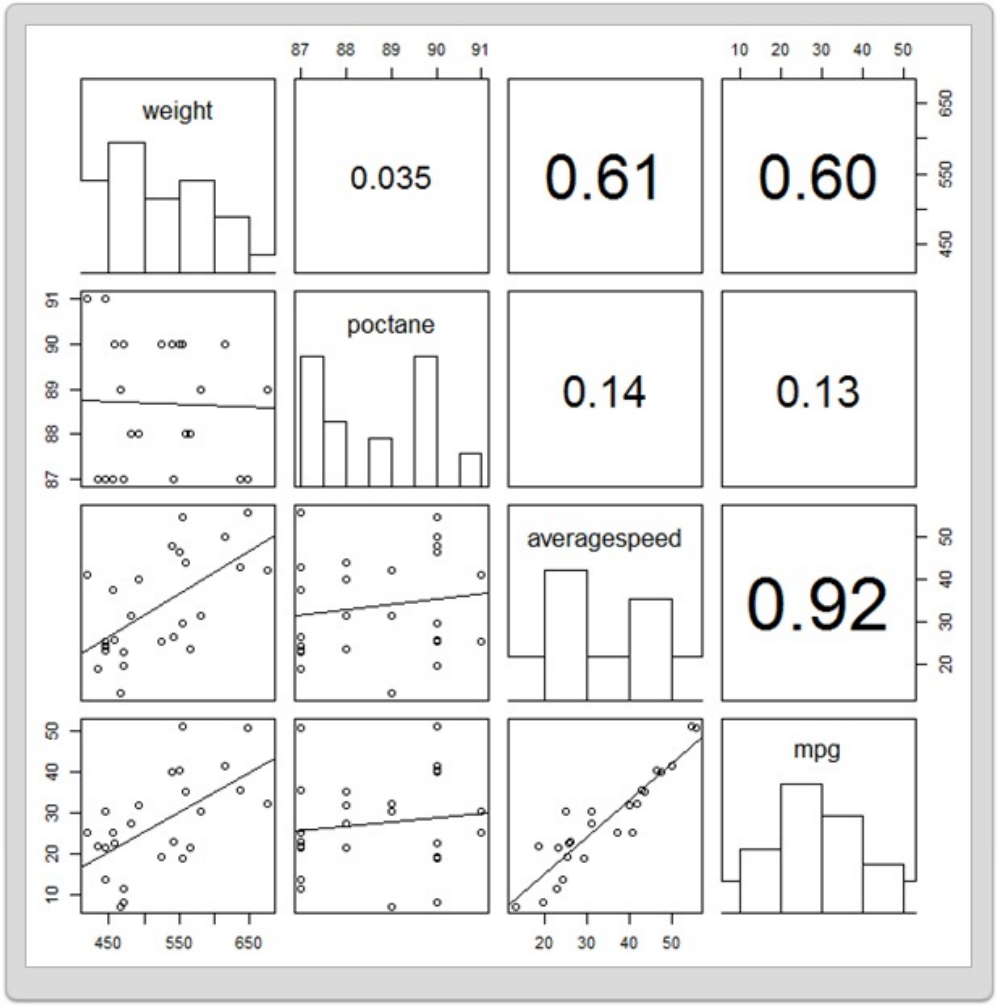
Example – Fuel Consumption

- In order to best understand what factors are related to fuel consumption, a truck driver collected data on
  - weight of carrying load,
  - percent octane of the gasoline used,
  - average speed

  resulting miles per gallon

  for 25 trips he took last month.

Fuel Example – Scatter plot



Example – Fuel Consumption

- Step 1: Finding the MLR model

$$\hat{y} = -9.52 + 0.01X_{weight} + 0.03X_{poctane} + 0.86X_{avespeed}$$

Example – Fuel Consumption

- Step 1: Finding the MLR model

$$\hat{y} = -9.52 + 0.01X_{weight} + 0.03X_{poctane} + 0.86X_{avespeed}$$

- Step 2: Calculate R^2

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic
Regression	Reg SS = 2810.06	Reg df = $k = 3$	Reg MS = 2810.06/3 = 936.7	F = Reg MS/Res MS = 936.7/24.9 = 37.6
Residual	Res SS = 522.20	Res df = $n - k - 1 = 25 - 3 - 1 = 21$	Res MS = 522.20/21 = 24.9	
Total	Total SS = Reg SS + Res SS = 3332.26			

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{2810.06}{3332.26} \approx 84.3\%$$

Example – Fuel Consumption

- Global F-test
  - As in simple linear regression the FF statistic is calculated as:

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

  - which follows an F-distribution with k and n−k−1 degrees of freedom under H0.
  - The decision rule for a two-sided level α test is:
    - Reject H0:$\beta 1=\beta 2=\cdots=\beta k=0$, if $F \geq F_{k,n-k-1,\alpha}$
    - Otherwise, do not reject H0 (there is at least one $\beta i \neq 0$)
  - where $F_{k,n-k-1,\alpha}$ is the value from the F-distribution table with k,n−k−1 degrees of freedom and associated with a right hand tail probability of α.

Example – Fuel Consumption

- Is weight of carrying load a significant predictor of miles per gallon after controlling for percent octane and average speed?
  - Perform a t-test at the α=0.05 level.

	Estimate	SE	t	$Pr(>\|t\|)$
Intercept	–9.53	65.5	–0.145	0.886
Weight	0.01087	0.01792	0.607	0.550
Percent Octane	0.02914	0.72565	0.040	0.968
Average Speed	0.85635	0.10882	7.869	< 0.0001

Five steps

- Set up the hypotheses and select the alpha level
- Select the appropriate test statistic
- State the decision rule
- Compute the test statistic
- Conclusion

Five steps

- Set up the hypotheses and select the alpha level
  - H0: $\beta_{weight}=0$ (after controlling for average speed and percent octane)
  - H1: $\beta_{weight}\neq 0$ (after controlling for average speed and percent octane)
  - $\alpha = 0.05$
- Select the appropriate test statistic
- State the decision rule
- Compute the test statistic
- Conclusion

Five steps

- Set up the hypotheses and select the alpha level
  - H0: $\beta_{weight}=0$ (after controlling for average speed and percent octane)
  - H1: $\beta_{weight}\neq 0$ (after controlling for average speed and percent octane)
  - $\alpha = 0.05$
- Select the appropriate test statistic

$$t = \frac{\hat{\beta}_{weight}}{SE_{\hat{\beta}_{weight}}} \quad df = n - k - 1$$

- State the decision rule
- Compute the test statistic
- Conclusion

Five steps

- Set up the hypotheses and select the alpha level
  - H0: $\beta_{weight}=0$ (after controlling for average speed and percent octane)
  - H1: $\beta_{weight}\neq 0$ (after controlling for average speed and percent octane)
  - $\alpha = 0.05$
- Select the appropriate test statistic

$$t = \frac{\hat{\beta}_{weight}}{SE_{\hat{\beta}_{weight}}} \quad df = n - k - 1$$

- State the decision rule
  - Determine the appropriate value from the t-distribution table with
  - $n-k-1=25-3-1=21$
  - A right hand tail probability of $\alpha/2=0.052=0.025$
  - $t_{n-k-1,\alpha/2}=t_{21,0.025}=2.08$
  - Decision Rule: Reject H0 if $t\geq 2.08$ or if $t\leq -2.08$
  - Otherwise, do not reject H0
- Compute the test statistic
- Conclusion

Five steps

- Set up the hypotheses and select the alpha level
  - H0: $\beta_{weight}=0$ (after controlling for average speed and percent octane)
  - H1: $\beta_{weight}\neq 0$ (after controlling for average speed and percent octane)
  - $\alpha = 0.05$
- Select the appropriate test statistic

$$t = \frac{\hat{\beta}_{\text{weight}}}{SE_{\hat{\beta}_{\text{weight}}}} \quad \text{df} = n - k - 1$$

- State the decision rule
  - Determine the appropriate value from the t-distribution table with
  - $n-k-1=25-3-1=21$
  - A right hand tail probability of $\alpha/2=0.052=0.025$
  - $t_{n-k-1,\alpha/2}=t_{21,0.025}=2.08$
  - Decision Rule: Reject H0 if $t\geq 2.08$ or if $t\leq -2.08$
  - Otherwise, do not reject H0
- Compute the test statistic

$$t = \frac{\beta_{\text{weight}}}{SE_{\hat{\beta}_{\text{weight}}}} = \frac{0.01087}{0.01792} \approx 0.607$$

- Conclusion

Five steps

- Set up the hypotheses and select the alpha level
  - H0: $\beta_{weight}=0$ (after controlling for average speed and percent octane)
  - H1: $\beta_{weight}\neq 0$ (after controlling for average speed and percent octane)
  - $\alpha = 0.05$
- Select the appropriate test statistic

$$t = \frac{\hat{\beta}_{\text{weight}}}{SE_{\hat{\beta}_{\text{weight}}}} \quad \text{df} = n - k - 1$$

- State the decision rule
  - Determine the appropriate value from the t-distribution table with
  - $n-k-1=25-3-1=21$
  - A right hand tail probability of $\alpha/2=0.052=0.025$
  - $t_{n-k-1,\alpha/2}=t_{21,0.025}=2.08$
  - Decision Rule: Reject H0 if $t\geq 2.08$ or if $t\leq -2.08$
  - Otherwise, do not reject H0
- Compute the test statistic

$$t = \frac{\beta_{\text{weight}}}{SE_{\hat{\beta}_{\text{weight}}}} = \frac{0.01087}{0.01792} \approx 0.607$$

- Conclusion
  - Do not reject H0 since $0.607 < 2.08$. Do not have significant evidence at the $\alpha=0.05$. (here, $p=0.550$ as calculated using software program).

Regression Notes

- Extrapolation vs interpolation
- Lurking variables
- Causation and association
- co-linearity

Multi Linear Regression –Numerical Solution

- Error Surface
- Gradient Descent
- Multi-Linear Regression in matrix form

$$y_i=\beta_0+\beta_1 x_i+\epsilon_i \qquad \text{for } i=1,\ldots,n$$

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

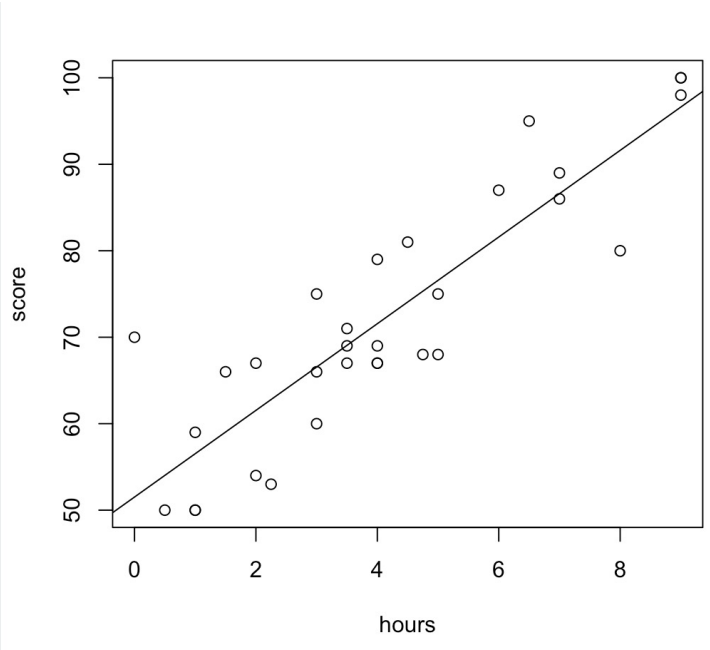
$$Y = X\beta + \varepsilon$$

Overfitting

Example of Overfitting

Data: number of hours studied and score of 31 students.

Linear regression estimate



	name	hours	score
1	Allen	8.0	80
2	Brown	2.0	67
3	Cole	9.0	98
4	Collins	9.0	100
5	Cooper	7.0	86
6	Cox	6.5	95

Learning From the Entire Data

- Perfect training with zero error

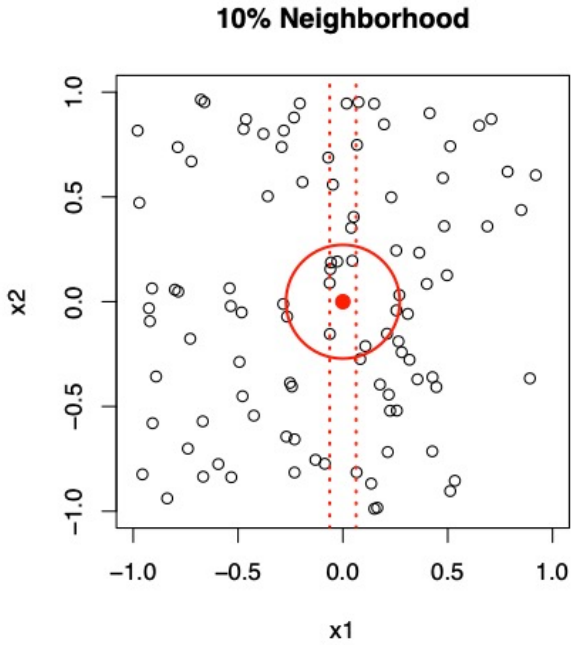
  $R^2 = 1!$ PERFECT!

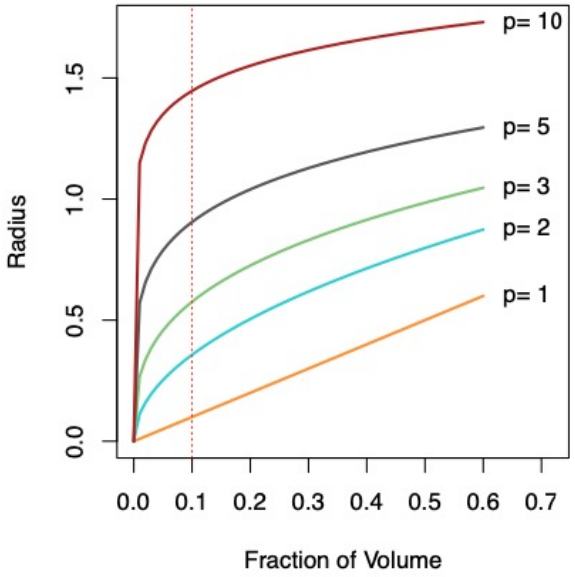
```
Residuals:
ALL 31 residuals are 0: no residual degrees of freedom!

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)        80         NA      NA       NA
nameBrown         -13         NA      NA       NA
nameCole           18         NA      NA       NA
nameCollins        20         NA      NA       NA
nameCooper          6         NA      NA       NA
nameCox            15         NA      NA       NA
nameHall          -13         NA      NA       NA
nameHans          -21         NA      NA       NA
nameHoward          9         NA      NA       NA
nameJeffers       -10         NA      NA       NA
nameJohnson       -30         NA      NA       NA
nameJones         -14         NA      NA       NA
nameKing            1         NA      NA       NA
nameKnight        -11         NA      NA       NA
nameLee            -9         NA      NA       NA
nameMartin        -14         NA      NA       NA
nameMiller         -5         NA      NA       NA
nameMoore         -20         NA      NA       NA
nameMorris        -13         NA      NA       NA
nameMurphy          7         NA      NA       NA
nameReed           -5         NA      NA       NA
nameSmith         -30         NA      NA       NA
nameStewart       -12         NA      NA       NA
nameTaylor        -27         NA      NA       NA
nameThomas        -26         NA      NA       NA
nameThompson      -11         NA      NA       NA
nameWalker        -13         NA      NA       NA
nameWard           20         NA      NA       NA
nameWilliams      -30         NA      NA       NA
nameWright        -12         NA      NA       NA
nameYoung          -1         NA      NA       NA
hours              NA         NA      NA       NA
id                 NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
```

The curse of dimensionality





Trade offs

- Prediction accuracy versus interpretability
- Parsimony/Occam's razon vs black box
- Simple model involving less variable is more preferred compare to a black-box model involving all the possible parameters
  - Linear regression models are easy to interpret but some other complex methods are not