STAT1201 - Summer Semester 2022

Lecture 8 -Statistical Models

Dr. Wasanthi Thenuwara

Learning Objectives

- Correlation between two variables
- Simple Linear Regression
- Multiple Linear Regression

Correlation between two variables

Correlation coefficient (r) measures the relative strength and direction of the linear relationship between two numerical variables.

Correlation does not imply a causal effect of the two varaibles. We will use simple linear regression to identify the causal effect.

- r is between -1 and $+1$ ($-1 \leq r \leq +1$)
- r close to -1 implies a strong negative linear relationship.
- r close to $+1$ implies a strong positive linear relationship.
- r close to 0 implies a weak linear relationship.
- r $=$ 0 implies no linear relationship

Inferences for correlation

Definitions

r - sample correlation coefficient

ρ - population correlation coefficient

We can use correlation coefficient to determine whether there is a statistically significant linear relationship between two variables (X and Y). The null and alternative hypothses to test whether there is a significant linear relationship between X and Y are follows.

$H_0 : \rho = 0$ Vs $H_1 : \rho \neq 0$

We first calculate the sample correlation coefficient using data and get an idea about the strength of the linear relationship between two varaibles. Then, we can use the t - test to test whether the linear relationship observed in the sample is statistically significant.

$t_{stat} = \frac{r-\rho}{se(r)}$ where se(r) $= \sqrt{\frac{1-r^2}{n-2}}$

Inferences for correlation - Example

A researcher is interested to see whether there is a linear association (correlation) between the breath holding time and height of adults aged between 18 and 20 years. A sample of size 20 (10 males and 10 females) used for the study and performed a breath holding test. The data was recorded in “M8Breath.csv”file for the following variables.

Sex - Male/Female

Height - Height of the subjects in centimeters

BreathHeld - Breath holding time in seconds

We can first calculate the sample correlation coefficient

Inferences for correlation - Example cont. . .

```r
breath = read.csv("M8Breath.csv")
cor(breath$Height, breath$BreathHeld)
```

The sample correlation coefficient (r) is 0.6642512 implies a fairly strong linear relationship between height and the breath holding time.

Inferences for correlation - Example cont. . .

To test whether this correlation between height and the breath holding time is statistically significant (i.e. r is significantly different from 0), we can perform a hypothesis test.

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$t_{stat} = \frac{r-\rho}{se(r)}$

se(r) $= \sqrt{\frac{1-r^2}{n-2}}$

se(r) $= \sqrt{\frac{1-0.6642512^2}{(20-2)}}$ —> se(r) $= 0.1761897$

$t_{stat} = \frac{0.6642512-0}{0.1761897}$

$t_{stat} = 3.770091$

t-stat has a *t*-distribution with degrees of freedom equals (*n*-2)

Inferences for correlation - Example cont...

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$t_{stat} = 3.770091$

What is the p-value? Using R:

```r
2*(1-pt(3.770091, df=18))
```

p-value $= 0.0014018$

p-value < 0.01. That is, we have strong evidence to conclude that there is a significant linear relationship between height and the breath holding time.

Inferences for correlation - Example cont. . .

We can do this correlation test directly in R.

```
cor.test(breath$Height, breath$BreathHeld)
```

```

    Pearson's product-moment correlation

data:  breath$Height and breath$BreathHeld
t = 3.7701, df = 18, p-value = 0.001402
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3140413 0.8553471
sample estimates:
      cor
0.6642512
```

Inferences for correlation - Example cont. . .

Poll Question 1

Suppose that you need to test whether there is a positive linear association between breath holding time and height. Assuming ρ is the population correlation coefficient, the appropriate null and alternative hypotheses are:

a) H_0 : r $=0$ Vs H_1 : r >0

b) H_0 : $\rho=0$ Vs H_1 : $\rho>0$

c) H_0 : $\rho=0$ Vs H_1 : $\rho<0$

d) H_0 : $\rho>0$ Vs H_1 : $\rho=0$

Simple Linear Regression Models (SLR)

SLR involves

- One dependent (or response variable): Y
- One independent (or explanatory variable): X

A single independent variable (X) is used to predict the numerical dependent variable (Y).

Examples:

Relationship between basal plasma oxytocin level and age (Y - Basal plasma Oxytocin level, X - Age)

Relationship between breath holding time and height (Y - Breath holding time, X - Height)

Relationship between body mass and height (Y - Body mass, X - Height)

Simple Linear Regression Models (SLR)

Population SL regression equation:

$Y_i = \beta_0 + \beta_1 X_i + U_i$ for i=1,2,. . . ,n

β_0 and β_1 are population parameters to be estimated using sample data. We will use least square estimation method.

$\beta_0 =$ population Y intercept

$\beta_1 =$ population slope (expected change in Y per unit change in X i.e. the mean amount that Y changes for a one unit change in X)

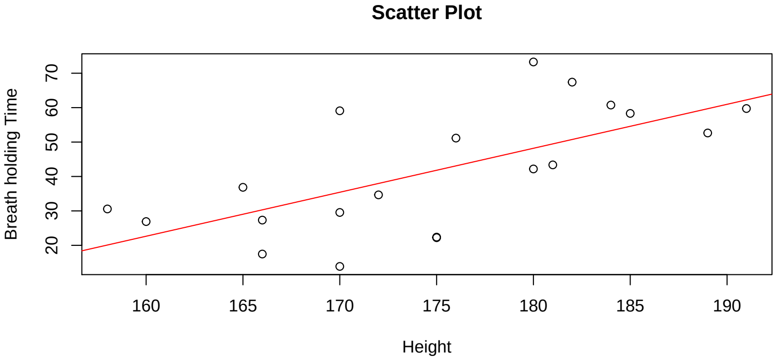
$U_i =$ random error in Y for observation i

$U_i \sim \text{N}(0, \sigma)$ and $\text{E}(U_i) = 0$, $\text{Var}(U_i) = \sigma^2$

Simple Linear Regression Models (SLR) - An example

Consider the BreathHeld - Height example again and we now estimate the relationship between breath holding time and height. More specifically, we can estimate the impact of height on breath holding time.

Simple Linear Regression Models (SLR) - An example



There can be many different Y values for each X. That is, there will be a distribution of Y for each value of X and we assume that this distribution is normal.

Simple Linear Regression Models (SLR) - An example

The setimated model is written as:

$\hat{Y} = b_0 + b_1 X$

b_0 and b_1 can be found using sample data and using lm() in RStudio.

```
lm(BreathHeld ~ Height, data=breath)
```

```

Call:
lm(formula = BreathHeld ~ Height, data = breath)

Coefficients:
(Intercept)       Height
   -181.740        1.277
```

Simple Linear Regression Models (SLR) - An example

Thus the estimated SLR equation is: $\hat{Y}$ = -181.7405 + 1.277X

Interpret coefficient of estimates

b_0 = -181.7405 indicates the mean breath holding time when height is zero is -181.7405second. This does not make sense.

b_1 = 1.277 - For each one centimeter increase in height, the mean breath holding time is estimated to increase by 1.277seconds.

Prediction about Y for a given value of X using the estimated equation

What is the predicted value for breath holding time when height equals to 160cm?

Substitute X = 160 in the estimated equation

$\hat{Y}$ = -181.7405 + 1.277X

$\hat{Y}$ = -181.7405 + 1.277*160 —> $\hat{Y}$ = 22.64519seconds

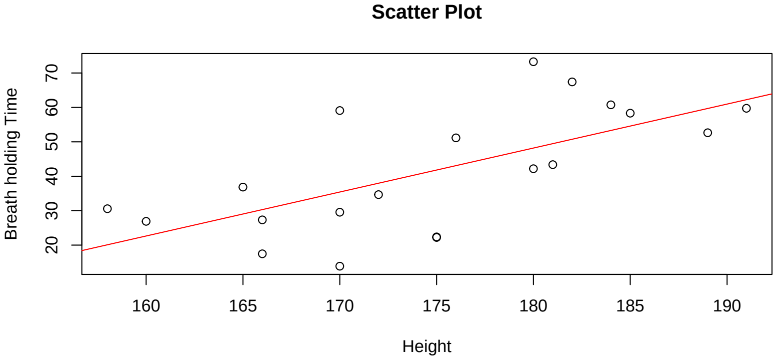
The following function can be used in R to find the predicted value of Y for a given value of X.

```r
predict(lm(BreathHeld ~ Height, data=breath),
          newdata=data.frame(Height=160))
```

Simple Linear Regression Models (SLR) - Introduce residuals

Residual is also called the estimated error (e_i)

$e_i = Y_i$-$\hat{Y}_i$



SLR - Measures of variation

- How much of the variation in the dependent variable, Y is explained by variation in the independent variable, X.
- To quantify this, we calculate the coefficient of determination (R^2)

$R^2 = \frac{SSL}{SST}$

SST = Total Sum of Squares —> SST $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

SSL = Line Sum of Squares (or explained variation). Also, call Regression sum of squares.

SSL $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

SSR = Residual Sum of Squares (or unexplained variation). Also call error sum of squares.

SSR $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

SST = SSL + SSR

SLR - Measures of variation

Using R, we can obtain SSL and SSR

```
summary(aov(BreathHeld ~ Height, data=breath))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Height       1   2656  2656.1   14.21 0.0014 **
Residuals   18   3364   186.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

SST $= 2656{+}3364 = 6020$

$R^2 = \frac{2656}{6020} = 0.441196 = 44.12\%$

Only 44.12% of variation in breath holding time is explained by variation in height.

SLR - Inferences for the slope coefficient

$Y = \beta_0 + \beta_1 X + U$

If $\beta_1 = 0$, there is no association between Y and X.

That is, we need to test whether β_1 is significantly different from zero. We can use *t*-test.

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

This is a two tail test.

$t_{stat} = \frac{b_1 - \beta_1}{se(b_1)}$

The t_{stat} has a t - distribution with df = (n-2)

SLR - Inferences for the slope coefficient

For our example of Breathheld and Height we can test whether there is a significant linear relationship between breath holding time and height.

$H_0 : \beta_1 = 0$ Vs $H_1 : \beta_1 \neq 0$

We need standard error of b_1 to calculate the t statistic. Using R:

```
summary(lm(BreathHeld ~ Height, data=breath))$coef
```

```
               Estimate Std. Error   t value    Pr(>|t|)
(Intercept) -181.740491 59.2889370 -3.065336 0.006665604
Height         1.277411  0.3388274  3.770092 0.001401765
```

$t_{stat} = \frac{1.2774 - 0}{0.3388}$

$t_{stat} = 3.770366$

SLR - Inferences for the slope coefficient

Find p-value using R function pt()

pt(3.770366, df=18) gives P(t<=3.770366)

$\frac{p-value}{2}$ = 1-pt(3.770366, df=18)

p-value = 2*(1-pt(3.770366, df=18))

p-value = 0.0014009.

That is p-value $<$ 0.01. We have strong evidence to suggest that there is a significant linear relationship between breath holding time and height.

SLR - Inferences for the slope coefficient

Poll Question 2

The p-value to test whether there is a positive linear relationship between breath holding time and height is:

a) 0.0007

b) 0.0014

c) 0.0004

d) 0.0028

SLR - Confidence Interval for the slope coefficient

Every confidence interval has the following form.

estimate ± margin of error

where; margin of error = critical value * se(estimate)

In our example, 95% CI for β_1 is

$b_1 \pm t^*.\text{se}(b_1)$

Use qt() in R to find t^*

```r
qt(0.975, 18)
```

$t^* = 2.100922$

95% CI for β_1 is

1.2774 ± 2.100922*0.3388 —> 1.2774 ± 0.7117924

(0.5656076, 1.989192)

SLR - Confidence and prediction Intervals for Y

The predicted value of Y for a given value of X is called the point estimate of the population average value.

Now we can compute the confidence interval for the mean response of Y for a given value of X.

95% CI for mean breath holding time for the population when the height equals 160cm can be found using R:

```r
predict(lm(BreathHeld~Height, data=breath),
newdata=data.frame(Height = 160), interval ="confidence")
```

(10.33719, 34.95320)

We are 95% confident that the mean breath holding time for the entire population of young adults aged between 18 and 20 years when the height is 160cm is between 10.34seconds and 34.95seconds.

SLR - Confidence and prediction Intervals for Y

Similarly, we can calculate the prediction interval of a specific value of Y for a given value of X =160cm. This is denoted by $\hat{Y}_{X=160}$.

Using R:

```r
predict(lm(BreathHeld~Height, data=breath),
newdata=data.frame(Height = 160), interval ="prediction")
```

(-8.60088, 53.89127)

However, the lower limit = -8.60088 does not make sense. Therefore, ignore the lower limit in this particular example.

We are 95% confident that breath holding time of a young adult aged between 18 and 20 years and who is 160cm tall is less than 53.89 seconds (ignoring the lower limit).

Notes: Prediction intervals are wider than confidence intervals as there is much more variability in predicting an individual value than in estimating a mean value.

Simple Linear Regression - Assumptions

1) Linearity - Relationship between X and Y is linear
2) Errors are independent
3) Normality of errors - Errors are normally distributed at each value of X
4) Equal variance of the errors (also called homoscedasticity) - Variance of the errors is constant for all values of X

- The fitted linear regression model is appropriate only if the assumptions are satisfied.
- Use residual analysis to check the assumptions

Simple Linear Regression - Residual Analysis to Check Assumptions

1) Assumption 1 - Linearity - Plot Residuals Vs X

If the linearity assmption is not violated, the residuals should evenly spread above and below zero.

.

.

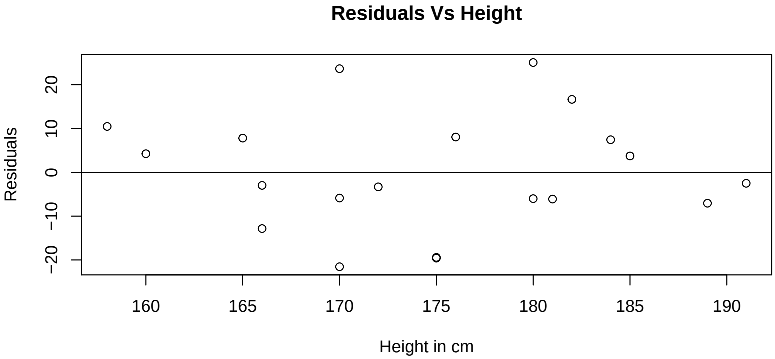
.

.

.

Assumption 1 - Linearity cont. . .

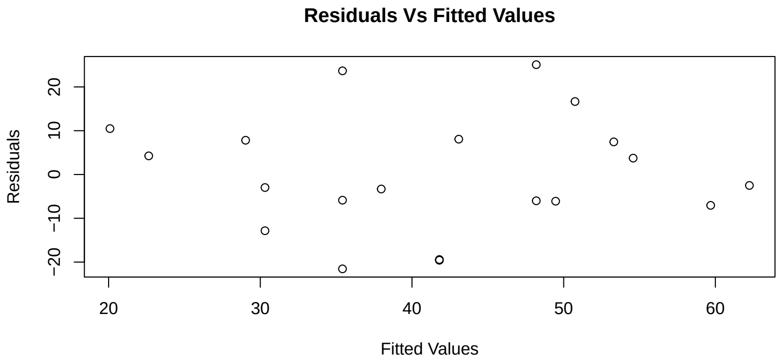
```
breath$Resid=resid (lm(BreathHeld ~ Height, data=breath))
plot(Resid ~ Height, data=breath,
     main = "Residuals Vs Height",
       xlab="Height in cm", ylab="Residuals")
abline(h=0)
```



Assumption 1 - Linarity cont…

We can also use Residuals Vs Fitted values to test for linearity.

```
breath$Fitted=predict(lm(BreathHeld ~ Height, data=breath))
plot(Resid ~ Fitted, data=breath, xlab="Fitted Values",
main = "Residuals Vs Fitted Values",  ylab="Residuals")
```



There is no apparent pattern or relationship between residuals and fitted values

Simple Linear Regression - Assumption 2

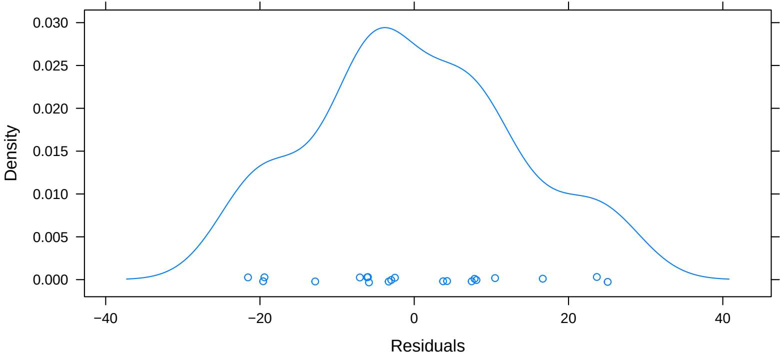
Errors are independent - Use Residuals Vs independent variable (Height)

This assumption is not very important in this example as data were collected at the same time period. That is, data is cross sectional. If we use time series data in the regression model (collected over time - yearly, quarterly, monthly, weekly etc.), then this assumption should be checked.

Simple Linear Regression - Assumption 3

Normality of errors- We can use frequency distribution of the residuals or a normal probability plot.

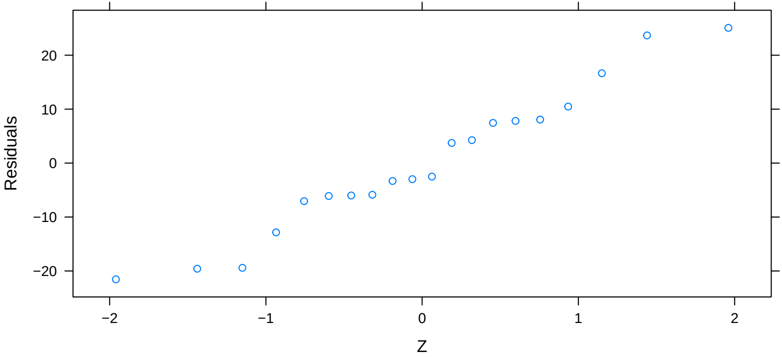
```r
library(lattice)
densityplot(breath$Resid, xlab="Residuals")
```



Check assumption 3 using the normal probability plot

If errors are normally distributed, the normal probabilty plot should be a straight line.

```
qqmath(breath$Resid, xlab="Z", ylab="Residuals")
```



Both density plot and normal probability polt suggest that errors are normally distributed.

Simple Linear Regression - Assumption 4

Equal variance of the errors - Use a plot of Residuals Vs independent variable, X (Height)

If the assumption is violated, variability of errors should increase or decrease when X is increasing.

.

.

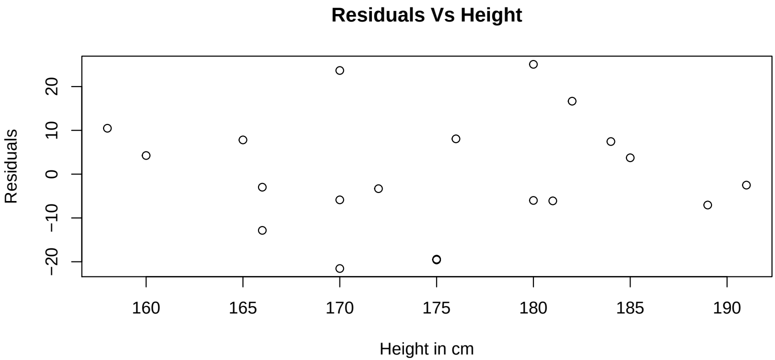
.

.

.

.

Simple Linear Regression - Assumption 4



There is no apparent increasing or decreasing pattern. Thus, the equal variance of the errors assumption is not violated.

Multiple Linear Regression Models

- One dependent (or response) variable
- More than one independent (or explanatory) variables

Examples

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$

Y - Breath Holding Time; X_1 - Height; X_2 - Weight

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$

Y - Breath Holding Time; X_1 - Height; X_2 - Weight

X_3 is a dummay variable. $X_3 = 1$ for males, 0 for females

Multiple Linear Regression Models - Example

Consider the breah holding time example that we considered in simple linear regression. Now we will use a multiple linear regression model to see the effect of height and sex on breath holding time.

The population regression equation:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$

Y = Breath Holding Time (in seconds)

X_1 = Height (in centimeters)

X_2 = 1 for Males and 0 for Females

For Males: $Y = (\beta_0 + \beta_2) + \beta_1 X_1 + U$

For Females: $Y = \beta_0 + \beta_1 X_1 + U$

We can use the least square estimation method to estimate the population parameters of β_0, β_1 and β_2.

Multiple Linear Regression Models - Example cont ...

```r
summary(lm(BreathHeld ~ Height + Sex, data=breath))
```

```
Call:
lm(formula = BreathHeld ~ Height + Sex, data = breath)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8269  -4.0341   0.8832   4.1264  16.2531 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  47.0452    58.8693   0.799    0.435    
Height       -0.1244     0.3506  -0.355    0.727    
SexMale      32.3662     6.3267   5.116 8.61e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 8.827 on 17 degrees of freedom
```

Multiple Linear Regression Models - Example cont ...

```
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 47.0451876 58.8692840  0.7991466 4.352294e-01
Height      -0.1244138  0.3506444 -0.3548146 7.270920e-01
SexMale     32.3662340  6.3266789  5.1158332 8.605604e-05
```

The estimated eqution is:

$Breat\hat{h}Held =$ 47.045 - 0.124Height + 32.366Sex

Note that β_1 is not statistically significant (i.e. p-value = 0.727). That is, there is no evidence to conlude that height has a significant impact on breath holding time, after taking into account the effect of sex.

Multiple Linear Regression Models - Example cont . . .

Interpret coefficient estimates

$b_1 =$ -0.124 —> After takig into account the effect of sex, for each 1cm increase in height, the mean breath holding time is estimated to be 0.124seconds lower.

$b_2 =$ 32.366 –> For a given height, the mean breath holding time for males is 32.366seconds higher than that of for females.

Multiple Linear Regression Models - Example cont ...

What is the estimated breath holding time for a 170cm height male?

The estimated eqution is:

$\hat{BreathHeld}$ = 47.045 - 0.124Height + 32.366Sex

Substitute Height = 170 and Sex=1

$\hat{BreathHeld}$ = 58.26

```r
predict(lm(BreathHeld ~ Height+ Sex, data=breath),
        newdata=data.frame(Height=170, Sex="Male"))
```

```
       1
58.26108
```

Multiple Linear Regression Models - Example cont . . .

Poll Question 3

What is the estimated breath holding time for a 160cm height female?

a) 27.14 seconds

b) 59.51 seconds

c) 47.05 seconds

d) 58.26 seconds

Multiple Linear Regression Models - Example cont . . .

Testing the significance of the individual coefficient estimates.

Consider the following multiple linear regression model we estimated before.

$BreathHeld = \beta_0 + \beta_1 Height + \beta_2 Sex + U$

```
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 47.0451876 58.8692840  0.7991466 4.352294e-01
Height      -0.1244138  0.3506444 -0.3548146 7.270920e-01
SexMale     32.3662340  6.3266789  5.1158332 8.605604e-05
```

We can use the *t*-test as we did in the simple linear regression to test the statistical significance of each variable individualy in the model.

Multiple Linear Regression Models - Example cont ...

Testing the significance of the individual coefficient estimates.

$BreathHeld = \beta_0 + \beta_1 Height + \beta_2 Sex + U$

Suppose we test the statistcal significance of variable Height.

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$t_{stat} = \frac{b_1 - \beta_1}{se(b_1)}$

$t_{stat} = \frac{-0.1244 - 0}{0.3506}$. Thus, $t_{stat} =$ -0.355

The t_{stat} has a t-distribution with df = (n- k-1) where k is the number of independent variables in the model. In this example, what is k? What is the df for corresponding to t_{stat}?

Multiple Linear Regression Models - Example cont . . .

Poll Question 4

Suppose that you extend the multiple linear regression model we discussed and add Weight to the model. What is the degrees of freedom to be used in the *t* test whether Weight is related to breath holding time in the presence of Height and Sex?

a) 20

b) 16

c) 17

d) 18

Multiple Linear Regression - R^2 and Adjusted R^2

Consider the following multiple linear regression model.

$$BreathHeld = \beta_0 + \beta_1 Height + \beta_2 Sex + U$$

```
summary(aov(BreathHeld ~ Height + Sex, data=breath))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
Height       1   2656  2656.1   34.09 1.97e-05 ***
Sex          1   2039  2039.2   26.17 8.61e-05 ***
Residuals   17   1324    77.9                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

$R^2 = \frac{SSL}{SST}$

$R^2 = \frac{2656+2039}{2656+2039+1324}$ —> $R^2 = \frac{4695}{6019}$

$R^2 = 0.78$.

That is, 78% of variation in breath holding time is explained by the variation in height and sex.

Multiple Linear Regression - R^2 and Adjusted R^2

- In multiple linear regression analysis, adjusted R^2 is more important.
- Adjusted R^2 reflects the effect of the number of varaibles and the sample size.

Adjusted $R^2 = 1$ - $[(1\text{-}R^2)(\frac{n-1}{n-k-1})]$

Where, k is the number of independent varaibles in the model.

Adjusted $R^2 = 0.751 = 75.41\%$

Multiple Linear Regression Models - Example cont . . .

Interaction effect

Include the interaction effect to the multiple linear regression model we considered.

$$BreathHeld = \beta_0 + \beta_1 Height + \beta_2 Sex + \beta_3 Height * Sex + U$$

```r
summary(lm(BreathHeld ~ Height*Sex, data=breath))
```

Multiple Linear Regression Models - Example cont ...

Interaction effect

```
                  Estimate  Std. Error    t value  Pr(>|t|)
(Intercept)     84.3330983  86.8827751  0.9706538 0.3461666
Height          -0.3467627   0.5178065 -0.6696762 0.5126135
SexMale        -41.7485688 125.0206482 -0.3339334 0.7427665
Height:SexMale   0.4249171   0.7158171  0.5936112 0.5610712
```

The estimated model is:

$\hat{BreathHeld} = 84.333\text{-}0.347 Height\text{-}41.749 Sex + 0.425 Height*Sex$

Multiple Linear Regression - Assumptions

- The fitted linear regression model is appropriate only if the assumptions are satisfied.
- Use residual analysis to check the assumptions as in SLR.

1) Linearity - Relationship between each X and Y is linear (Use plot of Residuals Vs Fitted values)
2) Errors are independent
3) Normality of errors - Errors are normally distributed (Use normal probability plot of residuals)
4) Equal variance of the errors (also called homoscedasticity) - (Use plot of Residuals Vs Fitted values)

Next

Reminders

Lecture 9 - Analysis of Variance

Tuesday, 03 January 2023 at 12:00 via Zoom (818 1453 7986)

Paper Review is due on 11 jan 2023 at 3:00 pm