Regression analysis II

1. Multiple linear regression
2. Confounding, interactions

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
24.04.2024

Outline

Aalen chapter 11.4-11.6, Kirkwood and Sterne chapters 11 and 12

1. Morning: Regression II
   - Introduction to **Multiple linear regression** (briefly: multiple regression)
   - More details on linear regression models: **confounding**, **interactions**
2. Afternoon: Regression III
   - **categorical covariates** with more than 2 levels
   - Multiple regression **assumptions**, **leverage** effect
   - To explain, to predict or to describe? How the purpose of the analysis decides what is important

Schedule for today

08.30-10.15: Regression analysis II: multiple regression, confounding, interaction effects

10.15-11.15: R exercise for regression II

11.15-11.45: Discussion of the R exercise for regression II in class

- LUNCH

12.45-14.00: Regression analysis III: Multiple regression (continued), categorical variables, assumptions, leverage effect.
To explain, to predict or to describe?

14.00-15.00: R exercise for regression III

15.00-15.30: Discussion of the R exercises for regression III in class

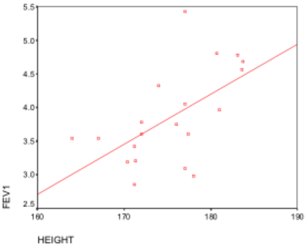
15.30-16.00: Course Summary

Yesterday: Simple linear regression

A **simple linear regression** describes the relationship between 1 independent variable (covariate, or predictor) and the dependent variable (response variable, or outcome) via a line.

Toy example: association between FEV1 and height.
Estimated regression line:

$$\text{FEV1} \approx -9.19 + 0.07 \cdot \text{height} \tag{1}$$



Relationship between simple linear regression and t-test

- There is a connection between the two approaches:
- Student's t-test (with equal variances) for the difference in the population mean between two independent groups is **equivalent** to a simple linear regression with the grouping as predictor variable.

Let us see this in a toy example:

Table 9.4 24 hour total energy expenditure (MJ/day) in groups of lean and obese women (Prentice *et al.*, 1986)

	Lean (n = 13)	Obese (n = 9)
	6.13	8.79
	7.05	9.19
	7.48	9.21
	7.48	9.68
	7.53	9.69
	7.58	9.97
	7.90	11.51
	8.08	11.85
	8.09	12.79
	8.11	
	8.40	
	10.15	
	10.88	
Mean	8.066	10.298
SD	1.238	1.398

R output for the t-test

R output for the Student's t-test (with equal variances) for the difference in `energy` between the lean and obese:

```
> t.test(energy ~ group, data=energy, var.equal=TRUE)

	Two Sample t-test

data:  energy by group
t = -3.9456, df = 20, p-value = 0.000799
alternative hypothesis: true difference in means between group Lean and group Obese is not equal to 0
95 percent confidence interval:
 -3.411451 -1.051796
sample estimates:
 mean in group Lean mean in group Obese
           8.066154           10.297778
```

R output for the simple linear regression

```
> fit <- lm(energy ~ group, data=energy)
> summary(fit)

Call:
lm(formula = energy ~ group, data = energy)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9362 -0.6153 -0.4070  0.2614  2.8138

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.0662     0.3618  22.297 1.34e-15 ***
groupObese    2.2316     0.5656   3.946 0.000799 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.304 on 20 degrees of freedom
Multiple R-squared:  0.4377,	Adjusted R-squared:  0.4096
F-statistic: 15.57 on 1 and 20 DF,  p-value: 0.000799
```

Multiple regression

- Is an extension of the simple linear regression with one independent variable (predictor / covariate)
- Still a continuous response (dependent) variable, but several explanatory (independent) variables (multiple predictors / covariates)
- The independent variables can be continuous, dichotomous or have more than two categories
- The **multiple linear regression model** is defined as

$$Y = b_0 + b_1 x_1 + \cdots + b_p x_p$$

Regression coefficients

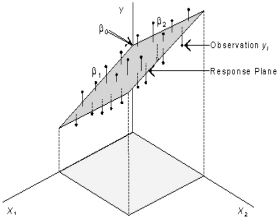
$$Y = b_0 + b_1 x_1 + \cdots + b_n x_n$$

- $b_1, \ldots, b_n$ are called regression coefficients
- b_i can be interpreted as the effect of one unit increase of the variable x_i when the other variables remain unchanged
- also called **adjusted effect**
- **Not necessarily a causal effect**

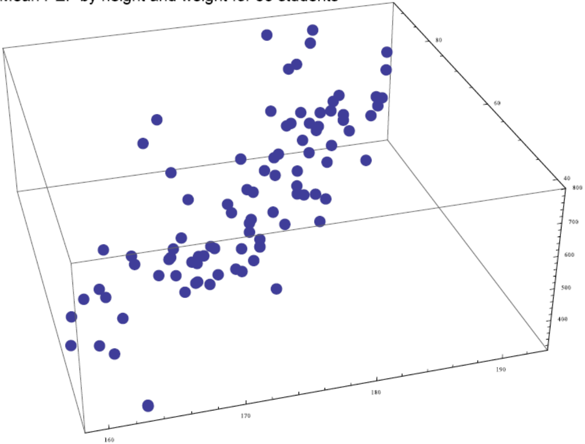
Interpretation

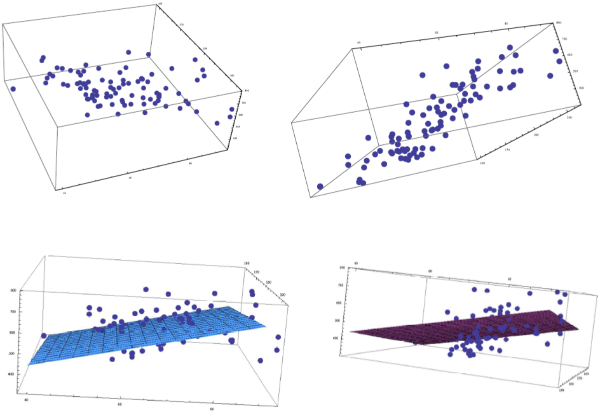
- Geometrically this corresponds to viewing data as points in a high-dimensional space.
- Beyond three dimensions we cannot picture such a space, but mathematically there is no difficulty with high-dimensional spaces.

Regression with two independent variables:



Mean PEF by height and weight for 95 students





Multiple regression via a toy example

Example: data on **systolic blood pressure**

Description	Name
Id	Id
Systolic blood pressure	SBP
Quetelet index (BMI)	QUET
Age	AGE
Smoking status	SMK

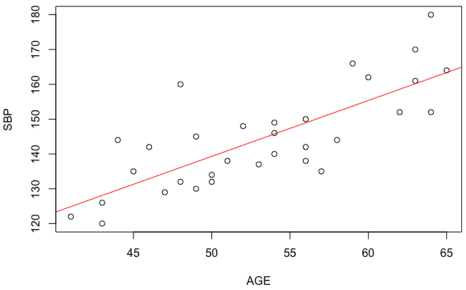
Simple linear regression: SBP vs AGE

```
> fit <- lm(SBP ~ AGE, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ AGE, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.548  -6.990  -2.481   5.765  23.892 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  59.0916    12.8163   4.611 6.98e-05 ***
AGE           1.6045     0.2387   6.721 1.89e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.245 on 30 degrees of freedom
Multiple R-squared:  0.6009,	Adjusted R-squared:  0.5876 
F-statistic: 45.18 on 1 and 30 DF,  p-value: 1.894e-07
```

- Note that $\hat{b}_0 = 59.09$ and $\hat{b}_1 = 1.61$,
- Confidence interval for b_1 $(1.12, 2.09)$ (calculate in R with `confint()`)
- $H_0 : b_1 = 0$ is rejected, as $p < 0.001$.
- SBP increases 1.6 units for **each year**.

Simple linear regression: SBP vs Age

```
> plot(SBP ~ AGE, data=bloodpressure)
> abline(reg=fit, col="red")
```



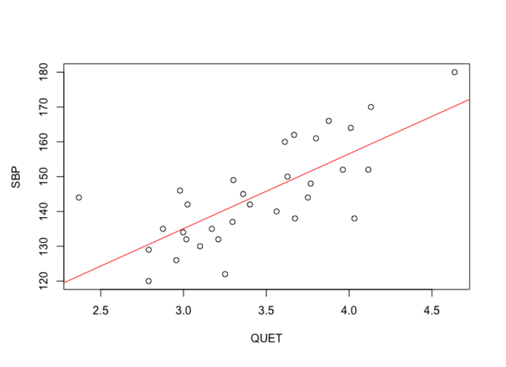
Simple linear regression: SBP vs QUET

```
> fit <- lm(SBP ~ QUET, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max
-19.231  -7.145  -1.604   7.798  22.531

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   70.576     12.322   5.728 2.99e-06 ***
QUET          21.492      3.545   6.062 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.812 on 30 degrees of freedom
Multiple R-squared:  0.5506,    Adjusted R-squared:  0.5356
F-statistic: 36.75 on 1 and 30 DF,  p-value: 1.172e-06
```

- Note that $\hat{b}_0 = 70.58$ and $\hat{b}_1 = 21.49$,
- Confidence interval for b_1 $(14.25, 28.73)$ (calculate in R with `confint()`)
- $H_0 : b_1 = 0$ is rejected, as $p < 0.001$.
- SBP increases 21.49 units for **each unit of QUET**.

Simple linear regression: SBP vs QUET

```
> plot(SBP ~ QUET, data=bloodpressure)
> abline(reg=fit, col="red")
```



Multiple regression: Combining AGE and QUET

```
> fit <- lm(SBP ~ QUET + AGE, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET + AGE, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.667  -6.793  -2.732   5.318  19.600 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  55.3234    12.5347   4.414 0.000129 ***
QUET          9.7507     5.4025   1.805 0.081489 .  
AGE           1.0452     0.3861   2.707 0.011253 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.916 on 29 degrees of freedom
Multiple R-squared:  0.6412,    Adjusted R-squared:  0.6165 
F-statistic: 25.92 on 2 and 29 DF,  p-value: 3.505e-07
```

- QUET does not have a significant effect on SBP, when adjusting for AGE,
- When AGE increases, then SBP will increase with 1.045 units,
- This is a significant increase $(p = 0.01)$, confidence interval $(0.26, 1.84)$ (calculate in R with `confint()`).

Confounding

What did we learn from the two previous models?

- Adjustment for AGE leads to a weaker relationship between SBP and QUET.
- AGE is associated with both SBP and QUET, and affects the association between them.

This implies that AGE is a **confounding variable.**

Confounders (more on this topic tomorrow)

Definition

A **confounder** is a variable that is a **common cause** of the exposure and the response (disease), and **NOT an effect** of the exposure or the disease.

- Confounding variables are important when we want to estimate (causal) effects from various exposures.
- As they cause both the exposure and the response, they are likely to cause biases.
- They can be dealt with by **adjusting in a multiple regression model**: always adjust for potential confounders by including them in the regression model!
- Multivariate regression models are thus important to include potential relevant variables.
- Be careful not to include common effects (also called *colliders*).

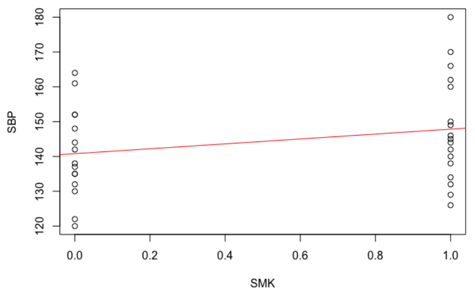
Simple linear regression: SBP vs SMK

```
> fit <- lm(SBP ~ SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ SMK, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.824  -9.056  -2.812  11.200  32.176 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  140.800      3.661  38.454   <2e-16 ***
SMK            7.024      5.023   1.398    0.172    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.18 on 30 degrees of freedom
Multiple R-squared:  0.06117,	Adjusted R-squared:  0.02988 
F-statistic: 1.955 on 1 and 30 DF,  p-value: 0.1723
```

- Note that $\hat{b}_0 = 140.80$ and $\hat{b}_1 = 7.02$,
- Confidence interval for b_1 $(-3.24, 17.28)$ (calculate in R with `confint()`)
- $H_0 : b_1 = 0$ is not rejected, as $p = 0.17$,
- Average difference between the two groups is 7.02.

Simple linear regression: SBP vs SMK

```
> plot(SBP ~ SMK, data=bloodpressure)
> abline(reg=fit, col="red")
```



Multiple regression: Combining AGE, QUET and SMK

```
> fit <- lm(SBP ~ QUET + AGE + SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET + AGE + SMK, data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5420  -6.1812  -0.7282   5.2908  15.7050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  45.1032    10.7649   4.190 0.000252 ***
QUET          8.5924     4.4987   1.910 0.066427 .  
AGE           1.2127     0.3238   3.745 0.000829 ***
SMK           9.9456     2.6561   3.744 0.000830 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.407 on 28 degrees of freedom
Multiple R-squared:  0.7609,	Adjusted R-squared:  0.7353 
F-statistic: 29.71 on 3 and 28 DF,  p-value: 7.602e-09
```

- Both AGE and SMK have significant effects,
- When AGE increases 1 unit, SBP increases with 1.2 units,
- Confidence interval: $(0.55, 1.88)$, $p = 0.001$,
- Smokers have 10 units higher SBP than non-smokers, confidence interval $(4.5, 15.4)$, $p = 0.001$.

Removing QUET from the model

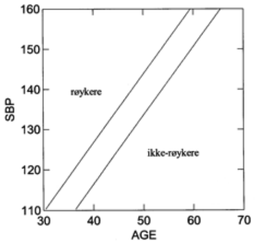
```
> fit <- lm(SBP ~ AGE + SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ AGE + SMK, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.639  -5.518  -1.637   4.900  19.616 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  48.0496    11.1296   4.317 0.000168 ***
AGE           1.7092     0.2018   8.471 2.47e-09 ***
SMK          10.2944     2.7681   3.719 0.000853 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.738 on 29 degrees of freedom
Multiple R-squared:  0.7298,	Adjusted R-squared:  0.7112 
F-statistic: 39.16 on 2 and 29 DF,  p-value: 5.746e-09
```

- Both AGE and SMK still have significant effects.
- Removing QUET lead to a slight decrease in the R^2: we might consider keeping it.

Closer look at the effect of AGE and SMK

$$\text{SBP} = 48.05 + 1.71 \cdot \text{AGE} + 10.29 \cdot \text{SMK}$$

- One year increase in age yields an increase of SBP 1.71 units,
- Non-smokers model: $\text{SBP} = 48.05 + 1.71 \cdot \text{AGE}$
- Smokers model: $\text{SBP} = 58.34 + 1.71 \cdot \text{AGE}$

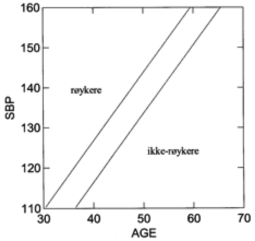


- The effect on SBP of the increase in AGE is the same regardless if one is a smoker or not. Is this realistic?
- NO $\rightarrow$ In reality, the effect of age could be larger for smokers.

Closer look at the effect of AGE and SMK

$$\text{SBP} = 48.05 + 1.71 \cdot \text{AGE} + 10.29 \cdot \text{SMK}$$

- One year increase in age yields an increase of SBP 1.71 units,
- Non-smokers model: $\text{SBP} = 48.05 + 1.71 \cdot \text{AGE}$
- Smokers model: $\text{SBP} = 58.34 + 1.71 \cdot \text{AGE}$



- The effect on SBP of the increase in AGE is the same regardless if one is a smoker or not. Is this realistic?
- NO $\rightarrow$ In reality, the effect of age could be larger for smokers.

Interaction between two explanatory variables

- If the effect of one variable might depend on another variable,
- we have to build a common model for main effects as well as interactions:

$$\mathsf{SBP} = b_0 + b_1 \cdot \mathsf{AGE} + b_2 \cdot \mathsf{SMK} + b_3 \cdot \mathsf{AGE} \cdot \mathsf{SMK}$$

- This is easily done in R with either the "*" or ":" operators:

```
lm(SBP ~ AGE*SMK, data=bloodpressure)
```

or

```
lm(SBP ~ AGE + SMK + AGE:SMK, data=bloodpressure)
```

Interaction between two explanatory variables

```
> fit <- lm(SBP ~ AGE*SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ AGE * SMK, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.036  -4.961  -1.958   5.552  20.665 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  58.5743    14.8048   3.956 0.000472 ***
AGE           1.5152     0.2703   5.605 5.32e-06 ***
SMK         -12.8460    21.7153  -0.592 0.558888    
AGE:SMK       0.4349     0.4048   1.074 0.291840    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.717 on 28 degrees of freedom
Multiple R-squared:  0.7405,    Adjusted R-squared:  0.7127 
F-statistic: 26.63 on 3 and 28 DF,  p-value: 2.369e-08
```

- Note that the interaction term is not significant, so we may drop this from the model if there are no particular biological/clinical reasons for keeping it,

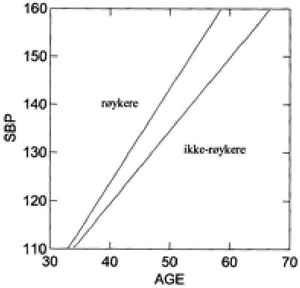
Interpretation

For the non-smokers (SMK $=0$):

$$\begin{aligned}\text{SBP} =&\hat{b}_0 + \hat{b}_1 \cdot \text{AGE} + \hat{b}_2 \cdot 0 + \hat{b}_3 \cdot \text{AGE} \cdot 0 \\ =&58.57 + 1.52 \cdot \text{AGE}\end{aligned}$$

For the smokers (SMK $= 1$):

$$\begin{aligned}\text{SBP} =&\hat{b}_0 + \hat{b}_1 \cdot \text{AGE} + \hat{b}_2 \cdot 1 + \hat{b}_3 \cdot \text{AGE} \cdot 1 \\ =&45.72 + 1.96 \cdot \text{AGE}\end{aligned}$$



Other possible interactions

```
> fit <- lm(SBP ~ QUET*SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET * SMK, data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max
-22.3713  -5.5705  -0.6357   7.4972  17.1051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.312     19.972   2.469   0.0199 *
QUET          26.303      5.703   4.612 8.01e-05 ***
SMK           29.944     24.164   1.239   0.2256
QUET:SMK      -6.185      6.932  -0.892   0.3799
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.948 on 28 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6137
F-statistic: 17.42 on 3 and 28 DF,  p-value: 1.408e-06
```

Other possible interactions

```
> fit <- lm(SBP ~ QUET*AGE, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET * AGE, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max
-13.385  -6.208  -2.284   6.243  21.926

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 207.3696    86.3654   2.401   0.0232 *
QUET        -34.1170    25.2168  -1.353   0.1869
AGE          -1.8468     1.6686  -1.107   0.2778
QUET:AGE      0.8224     0.4625   1.778   0.0863 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.601 on 28 degrees of freedom
Multiple R-squared:  0.6776,    Adjusted R-squared:  0.6431
F-statistic: 19.62 on 3 and 28 DF,  p-value: 4.742e-07
```

Model selection

- None of these interactions had significant effects, so in the light of a parsimony criterion (so to save degrees of freedom) we will skip the interactions in the final model.
- Automatic model selection is possible, but hard to use in practice.
- Models motivated by causal interpretations should be based on subject matter knowledge, not just an algorithm.

Final multiple regression model

No significant interactions, so we end up with the following model:

$$\mathsf{SBP} = b_0 + b_1 \cdot \mathsf{AGE} + b_2 \cdot \mathsf{QUET} + b_3 \cdot \mathsf{SMK}$$

```
> fit <- lm(SBP ~ QUET + AGE + SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET + AGE + SMK, data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5420  -6.1812  -0.7282   5.2908  15.7050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  45.1032    10.7649   4.190 0.000252 ***
QUET          8.5924     4.4987   1.910 0.066427 .  
AGE           1.2127     0.3238   3.745 0.000829 ***
SMK           9.9456     2.6561   3.744 0.000830 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.407 on 28 degrees of freedom
Multiple R-squared:  0.7609,    Adjusted R-squared:  0.7353 
F-statistic: 29.71 on 3 and 28 DF,  p-value: 7.602e-09
```

Interaction Effects

- Interaction means that the effect of a variable depends on a second variable,
- Not the same a confounding variable,
- Multivariate regression enables us to analyze interaction effects,
- We often need large data sets to get significant interaction effects.

- A variable Z that has an interaction effect on variable X is sometimes called an effect modifier of X.

Summary

Key words

- Multiple linear regression
- Confounder / collider (more tomorrow)
- Interaction effects