

# ASTA

*The ASTA team*

## Contents

<b>1</b>	<b>The regression problem</b>	<b>3</b>
1.1	We want to predict . . . . .	3
1.2	Initial graphics . . . . .	3
1.3	Simple linear regression . . . . .	4
1.4	Model for linear regression . . . . .	5
1.5	Least squares . . . . .	5
1.6	The prediction equation and residuals . . . . .	6
1.7	Estimation of conditional standard deviation . . . . .	6
1.8	Example in R . . . . .	7
1.9	Test for independence . . . . .	7
1.10	Example . . . . .	8
1.11	Confidence interval for slope . . . . .	8
1.12	Correlation . . . . .	9
<b>2</b>	<b>R-squared: Reduction in prediction error</b>	<b>10</b>
2.1	R-squared: Reduction in prediction error . . . . .	10
2.2	Graphical illustration of sums of squares . . . . .	10
2.3	$r^2$ : Reduction in prediction error . . . . .	11
<b>3</b>	<b>Multiple regression model</b>	<b>11</b>
3.1	Multiple regression model . . . . .	11
3.2	Example . . . . .	11
3.3	Correlations . . . . .	12
3.4	Several predictors . . . . .	13
3.5	Example . . . . .	13
3.6	Simpsons paradox . . . . .	14
<b>4</b>	<b>The general model</b>	<b>14</b>
4.1	Regression model . . . . .	14
4.2	Interpretation of parameters . . . . .	15
<b>5</b>	<b>Estimation</b>	<b>15</b>
5.1	Estimation of model . . . . .	15

<b>6</b>	<b>Multiple R-squared</b>	<b>15</b>
6.1	Multiple $R^2$ . . . . .	15
6.2	Example . . . . .	16
6.3	Example . . . . .	17
<b>7</b>	<b>F-test for effect of predictors</b>	<b>18</b>
7.1	F-test . . . . .	18
7.2	Example . . . . .	18
<b>8</b>	<b>Test for interaction</b>	<b>20</b>
8.1	Interaction between effects of predictors . . . . .	20
<b>9</b>	<b>The regression problem</b>	<b>20</b>
9.1	Example . . . . .	20
<b>10</b>	<b>Dummy coding</b>	<b>22</b>
10.1	Dummy coding . . . . .	22
10.2	Example . . . . .	22
10.3	Example: Prediction equations . . . . .	23
10.4	Agresti – summary . . . . .	25
<b>11</b>	<b>Model with interaction</b>	<b>25</b>
11.1	Interaction . . . . .	25
11.2	Interaction . . . . .	25
11.3	Example: Prediction equations . . . . .	26
11.4	Example: Individual tests . . . . .	26
<b>12</b>	<b>Test for no interaction</b>	<b>28</b>
12.1	Test for no interaction . . . . .	28
12.2	Hypothesis and test statistic . . . . .	30
12.3	Test for no interaction in R . . . . .	31
<b>13</b>	<b>Hierarchy of models</b>	<b>31</b>
13.1	Hierarchy of models . . . . .	31
13.2	Example . . . . .	32
13.3	Model with only a single categorical predictor . . . . .	33
13.4	Example with more predictors . . . . .	34

# 1 The regression problem

## 1.1 We want to predict

- We will study the dataset `trees`, which is on the course website (and actually also already available in R).

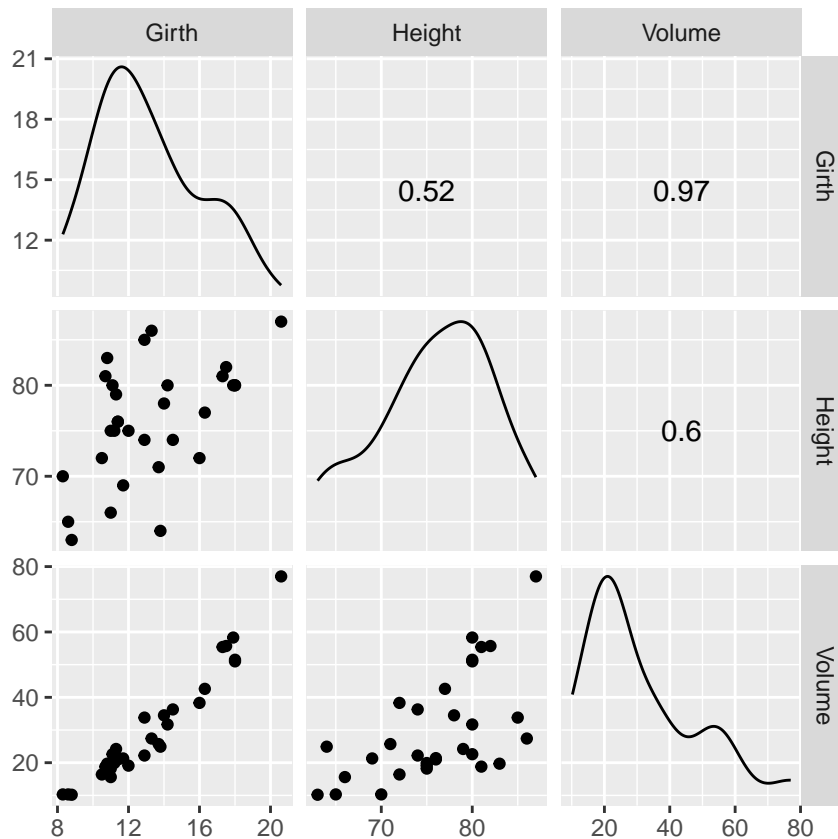
```
trees <- read.delim("https://asta.math.aau.dk/datasets?file=trees.txt")
```

- In this experiment we have measurements of 3 variables for 31 randomly chosen trees:
- [1] `Girth` numeric. Tree diameter in inches.
- [2] `Height` numeric. Height in ft.
- [3] `Volume` numeric. Volume of timber in cubic ft.
- We want to predict the tree volume, if we measure the tree height and/or the tree girth (diameter).
- This type of problem is called **regression**.
- Relevant terminology:
  - We measure a quantitative **response**  $y$ , e.g. `Volume`.
  - In connection with the response value  $y$  we also measure one (later we will consider several) potential **explanatory** variable  $x$ . Another name for the explanatory variable is **predictor**.

## 1.2 Initial graphics

- Any analysis starts with relevant graphics.

```
library(mosaic)
library(GGally)
ggscatmat(trees) # Scatter plot matrix from GGally package
```



- For each combination of the variables we plot the  $(x, y)$  values.
- It looks like **Girth** is a good predictor for **Volume**.
- If we only are interested in the association between two (and not three or more) variables we use the usual `gf_point` function.

### 1.3 Simple linear regression

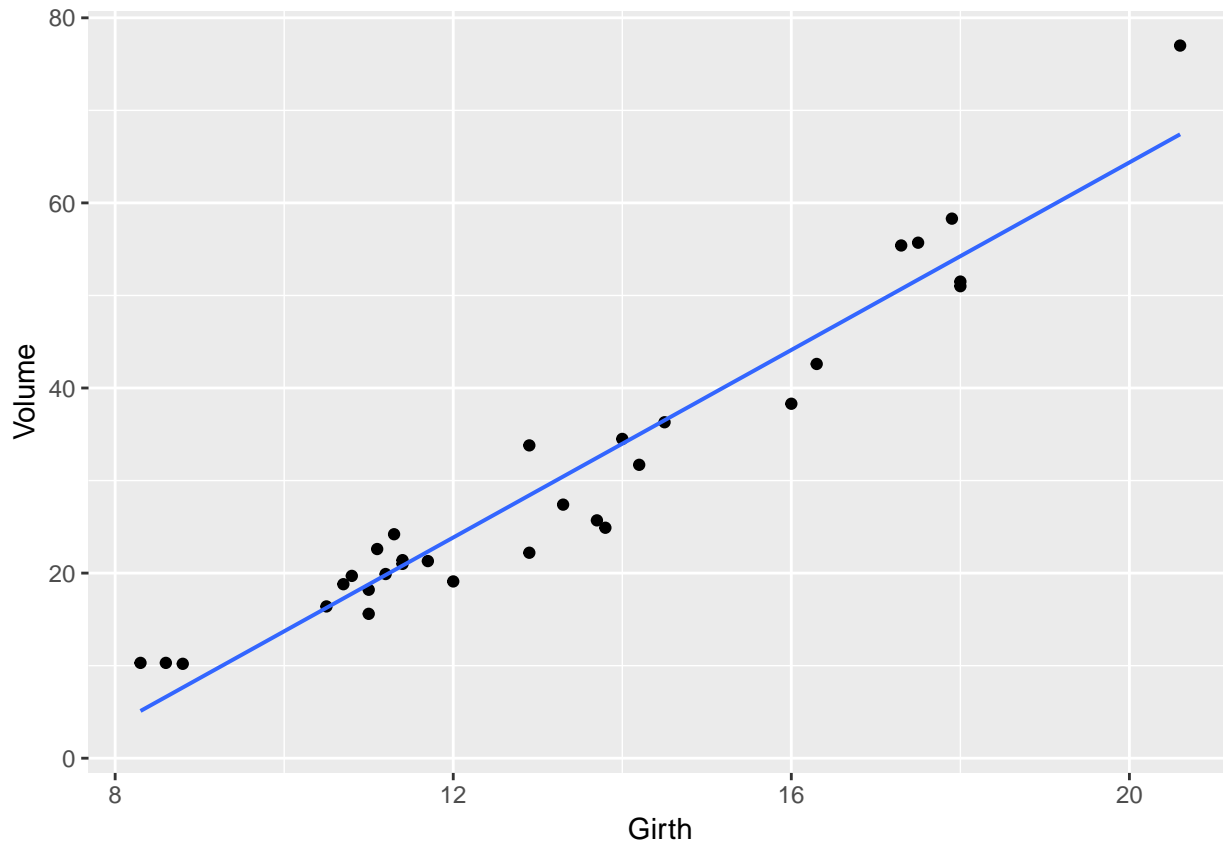
- We choose to use  $x=\text{Girth}$  as predictor for  $y=\text{Volume}$ . When we only use one predictor we are doing **simple regression**.
- The simplest **model** to describe an association between **response**  $y$  and a **predictor**  $x$  is **simple linear regression**.
- I.e. ideally we see the picture

$$y(x) = \alpha + \beta x$$

where

- $\alpha$  is called the **Intercept** - the line's intercept with the  $y$ -axis, corresponding to the response for  $x = 0$ .
- $\beta$  is called **Slope** - the line's slope, corresponding to the change in response, when we increase the predictor by one unit.

```
gf_point(Volume ~ Girth, data = trees) %>% gf_lm()
```



## 1.4 Model for linear regression

- Assume we have a sample with joint measurements  $(x, y)$  of predictor and response.
- Ideally the model states that

$$y(x) = \alpha + \beta x,$$

but due to random variation there are deviations from the line.

- What we observe can then be described by

$$y = \alpha + \beta x + \varepsilon,$$

where  $\varepsilon$  is a **random error**, which causes deviations from the line.

- We will continue under the following **fundamental assumption**:
  - The errors  $\varepsilon$  are normally distributed with mean zero and standard deviation  $\sigma_{y|x}$ .
- We call  $\sigma_{y|x}$  the **conditional standard deviation** given  $x$ , since it describes the variation in  $y$  around the regression line, when we know  $x$ .

## 1.5 Least squares

- In summary, we have a model with 3 parameters:
  - $(\alpha, \beta)$  which determine the line
  - $\sigma_{y|x}$  which is the standard deviation of the deviations from the line.
- How are these estimated, when we have a sample  $(x_1, y_1) \dots (x_n, y_n)$  of  $(x, y)$  values??

- To do this we focus on the errors

$$\varepsilon_i = y_i - \alpha - \beta x_i$$

which should be as close to 0 as possible in order to fit the data best possible.

- We will choose the line, which minimizes the sum of squares of the errors:

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

- If we set the partial derivatives to zero we obtain two linear equations for the unknowns  $(\alpha, \beta)$ , where the solution  $(a, b)$  is given by:

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

## 1.6 The prediction equation and residuals

- The equation for the estimates  $(\hat{\alpha}, \hat{\beta}) = (a, b)$ ,

$$\hat{y} = a + bx$$

is called **the prediction equation**, since it can be used to predict  $y$  for any value of  $x$ .

- Note: The prediction equation is determined by the current sample. I.e. there is an uncertainty attached to it. A new sample would without any doubt give a different prediction equation.
- Our best estimate of the errors is

$$e_i = y_i - \hat{y} = y_i - a - bx_i,$$

i.e. the vertical deviations from the prediction line.

- These quantities are called **residuals**.
- We have that
  - The prediction line passes through the point  $(\bar{x}, \bar{y})$ .
  - The sum of the residuals is zero.

## 1.7 Estimation of conditional standard deviation

- To estimate  $\sigma_{y|x}$  we need **Sum of Squared Errors**

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

which is the squared distance between the model and data.

- We then estimate  $\sigma_{y|x}$  by the quantity

$$s_{y|x} = \sqrt{\frac{SSE}{n-2}}$$

- Instead of  $n$  we divide  $SSE$  with **the degrees of freedom**  $df = n - 2$ . Theory shows, that this is reasonable.
- The degrees of freedom  $df$  are determined as the sample size minus the number of parameters in the regression equation.
- In the current setup we have 2 parameters:  $(\alpha, \beta)$ .

## 1.8 Example in R

```
model <- lm(Volume ~ Girth, data = trees)
summary(model)

##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

- The estimated residuals vary from -8.065 to 9.578 with median 0.152.
- The estimate of `Intercept` is  $a = -36.9435$
- The estimate of slope of `Girth` is  $b = 5.0659$
- The estimate of the conditional standard deviation (called residual standard error in **R**) is  $s_{y|x} = 4.252$  with  $31 - 2 = 29$  degrees of freedom.

## 1.9 Test for independence

- We consider the regression model

$$y = \alpha + \beta x + \varepsilon$$

where we use a sample to obtain estimates  $(a, b)$  of  $(\alpha, \beta)$ , an estimate  $s_{y|x}$  of  $\sigma_{y|x}$  and the degrees of freedom  $df = n - 2$ .

- We are going to test

$$H_0 : \beta = 0 \quad \text{against} \quad H_a : \beta \neq 0$$

- The null hypothesis specifies, that  $y$  **doesn't** depend linearly on  $x$ .
- In other words the question is: Is the value of  $b$  far away from zero?
- It can be shown that  $b$  has standard error

$$se_b = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

with  $df$  degrees of freedom.

- So, we want to use the test statistic

$$t_{\text{obs}} = \frac{b}{se_b}$$

which has to be evaluated in a t-distribution with  $df$  degrees of freedom.

## 1.10 Example

- Recall the summary of our example:

```
summary(model)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

- As we noted previously  $b = 5.0659$  and  $s_{y|x} = 4.252$  with  $df = 29$  degrees of freedom.
- In the second column(Std. Error) of the Coefficients table we find  $se_b = 0.2474$ .
- The observed t-score (test statistic) is then

$$t_{\text{obs}} = \frac{b}{se_b} = \frac{5.0659}{0.2474} = 20.48$$

which also can be found in the third column(t value).

- The corresponding p-value is found in the usual way by using the t-distribution with 29 degrees of freedom.
- In the fourth column( $\text{Pr}(>|t|)$ ) we see that the p-value is less than  $2 \times 10^{-16}$ . This is no surprise since the t-score was way above 3.

## 1.11 Confidence interval for slope

- When we have both the standard error and the reference distribution, we can construct a confidence interval in the usual way:

$$b \pm t_{\text{crit}} se_b,$$

where the t-score is determined by the confidence level and we find this value using `qdist` in **R**.

- In our example we have 29 degrees of freedom and with a confidence level of 95% we get  $t_{\text{crit}} = \text{qdist}("t", 0.975, df = 29) = 2.045$ .
- If you are lazy (like most statisticians are):

```
confint(model)
```



```
##           2.5 %    97.5 %
## (Intercept) -43.825953 -30.060965
## Girth       4.559914   5.571799
```

- i.e. (4.56, 5.57) is a 95% confidence interval for the slope of **Girth**.

## 1.12 Correlation

- The estimated slope  $b$  in a linear regression doesn't say anything about the strength of association between  $y$  and  $x$ .
- **Girth** was measured in inches, but if we rather measured it in kilometers the slope is much larger: An increase of 1km in **Girth** yield an enormous increase in **Volume**.
- Let  $s_y$  and  $s_x$  denote the sample standard deviation of  $y$  and  $x$ , respectively.
- The corresponding t-scores

$$y_t = \frac{y}{s_y} \quad \text{and} \quad x_t = \frac{x}{s_x}$$

are independent of the chosen measurement scale.

- The corresponding prediction equation is then

$$\hat{y}_t = \frac{a}{s_y} + \frac{s_x}{s_y} b x_t$$

- i.e. **the standardized regression coefficient** (slope) is

$$r = \frac{s_x}{s_y} b$$

which also is called **the correlation** between  $y$  and  $x$ .

- It can be shown that:
  - $-1 \leq r \leq 1$
  - The absolute value of  $r$  measures the (linear) strength of dependence between  $y$  and  $x$ .
  - When  $r = 1$  all the points are on the prediction line, which has positive slope.
  - When  $r = -1$  all the points are on the prediction line, which has negative slope.
- To calculate the correlation in **R**:

```
cor(trees)
```

```
##           Girth   Height   Volume
## Girth  1.0000000 0.5192801 0.9671194
## Height 0.5192801 1.0000000 0.5982497
## Volume 0.9671194 0.5982497 1.0000000
```

- There is a strong positive correlation between **Volume** and **Girth** ( $r=0.967$ ).
- Note, calling `cor` on a `data.frame` (like `trees`) only works when all columns are numeric. Otherwise the relevant numeric columns should be extracted like this:

```
cor(trees[,c("Height", "Girth", "Volume")])
```

which produces the same output as above.

- Alternatively, one can calculate the correlation between two variables of interest like:

```
cor(trees$Height, trees$Volume)
```

```
## [1] 0.5982497
```

## 2 R-squared: Reduction in prediction error

### 2.1 R-squared: Reduction in prediction error

- We want to compare two different models used to predict the response  $y$ .
- Model 1: We do not use the knowledge of  $x$ , and use  $\bar{y}$  to predict any  $y$ -measurement. The corresponding prediction error is defined as

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

and is called the **Total Sum of Squares**.

- Model 2: We use the prediction equation  $\hat{y} = a + bx$  to predict  $y_i$ . The corresponding prediction error is then the Sum of Squared Errors

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

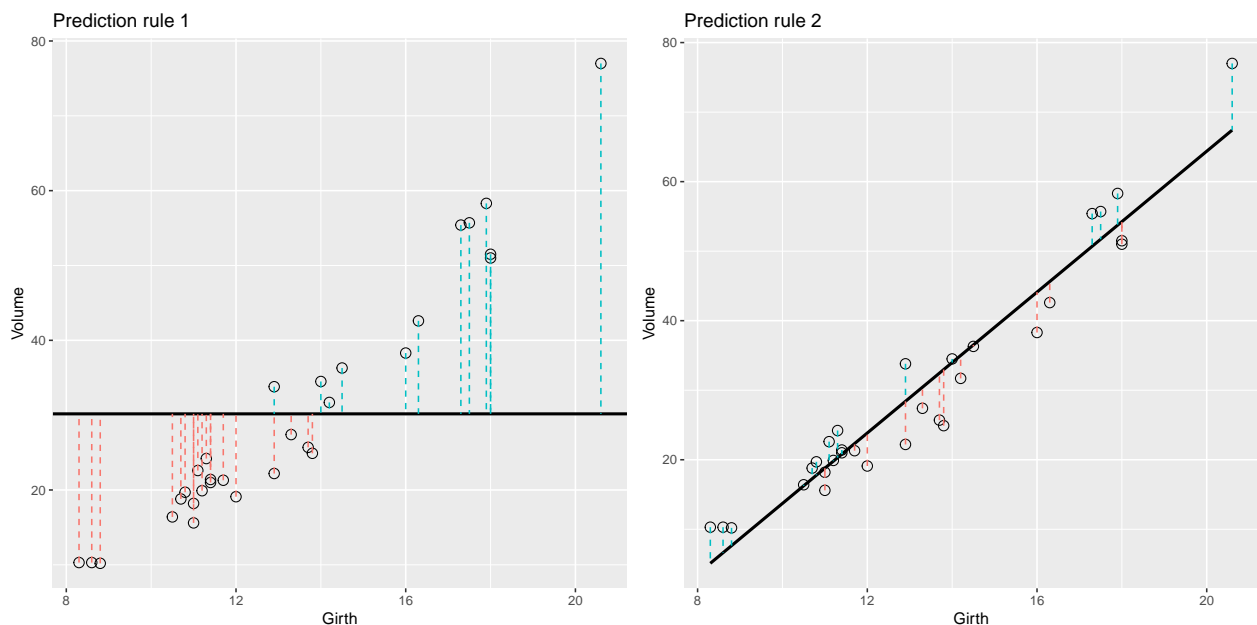
- We then define

$$r^2 = \frac{TSS - SSE}{TSS}$$

which can be interpreted as the relative reduction in the prediction error, when we include  $x$  as explanatory variable.

- This is also called the **fraction of explained variation, coefficient of determination** or simply **r-squared**.
- For example if  $r^2 = 0.65$ , the interpretation is that  $x$  explains about 65% of the variation in  $y$ , whereas the rest is due to other sources of random variation.

### 2.2 Graphical illustration of sums of squares



- Note the data points are the same in both plots. Only the prediction rule changes.
- The error of using Rule 1 is the total sum of squares  $E_1 = TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ .
- The error of using Rule 2 is the residual sum of squares (sum of squared errors)  $E_2 = SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ .

## 2.3 $r^2$ : Reduction in prediction error

- For the simple linear regression we have that

$$r^2 = \frac{TSS - SSE}{TSS}$$

is equal to the square of the correlation between  $y$  and  $x$ , so it makes sense to denote it  $r^2$ .

- Towards the bottom of the output below we can read off the value  $r^2 = 0.9353 = 93.53\%$ , which is a large fraction of explained variation.

```
summary(model)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

## 3 Multiple regression model

### 3.1 Multiple regression model

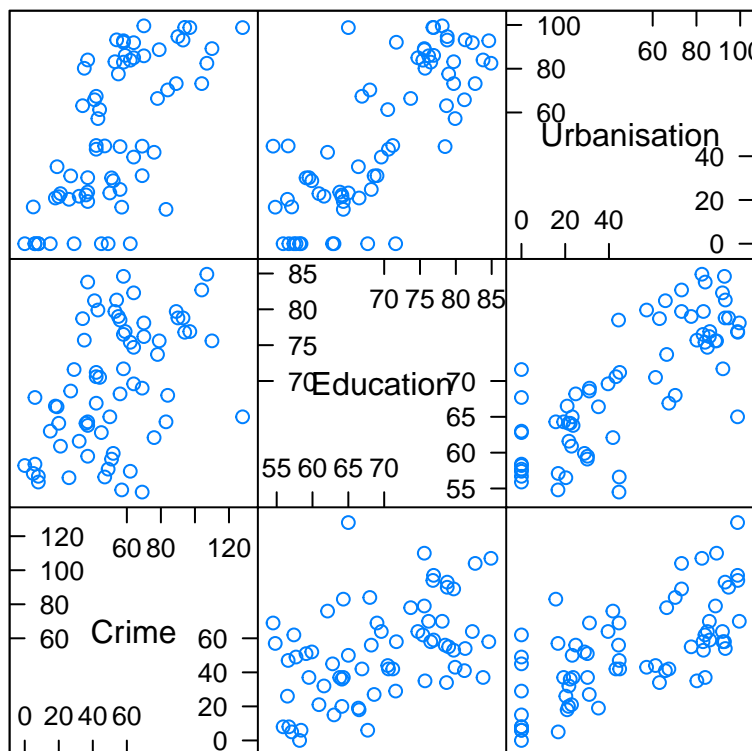
- We look at data from Table 9.15 in Agresti. The data are measurements in the 67 counties of Florida.
- Our focus is on
  - The response  $y$ : **Crime** which is the crime rate
  - The predictor  $x_1$ : **Education** which is proportion of the population with high school exam
  - The predictor  $x_2$ : **Urbanisation** which is proportion of the population living in urban areas

### 3.2 Example

```
FL <- read.delim("https://asta.math.aau.dk/datasets?file=fl-crime.txt")
head(FL, n = 3)
```

```
##   Crime Education Urbanisation
## 1   104      82.7      73.2
## 2    20      64.1      21.5
## 3    64      74.7      85.0
```

```
library(mosaic)
splom(FL) # Scatter PLOt Matrix
```



Scatter Plot Matrix

### 3.3 Correlations

- There is significant ( $p \approx 7 \times 10^{-5}$ ) positive correlation ( $r=0.47$ ) between **Crime** and **Education**
- Then there is also significant positive correlation ( $r=0.68$ ) between **Crime** and **Urbanisation**

```
cor(FL)
```

```
##           Crime Education Urbanisation
## Crime      1.0000000 0.4669119  0.6773678
## Education  0.4669119 1.0000000  0.7907190
## Urbanisation 0.6773678 0.7907190  1.0000000
```

```
cor.test(~ Crime + Education, data = FL)
```

```
##  
## Pearson's product-moment correlation  
##  
## data: Crime and Education  
## t = 4.2569, df = 65, p-value = 6.806e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2553414 0.6358104  
## sample estimates:  
## cor  
## 0.4669119
```

### 3.4 Several predictors

- Both Education ( $x_1$ ) and Urbanisation ( $x_2$ ) are pretty good predictors for Crime ( $y$ ).
- We therefore want to consider the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- The errors  $\epsilon$  are random noise with mean zero and standard deviation  $\sigma_{y|x}$ .
- The graph for the mean response is in other words a 2-dimensional plane in a 3-dimensional space.
- We determine estimates  $(a, b_1, b_2)$  for  $(\alpha, \beta_1, \beta_2)$  via the least squares method, i.e deviations from the plane.

### 3.5 Example

```
model <- lm(Crime ~ Education + Urbanisation, data = FL)  
summary(model)
```

```
##  
## Call:  
## lm(formula = Crime ~ Education + Urbanisation, data = FL)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -34.693 -15.742  -6.226  15.812  50.678   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   59.1181    28.3653   2.084   0.0411 *      
## Education     -0.5834     0.4725  -1.235   0.2214       
## Urbanisation  0.6825     0.1232   5.539 6.11e-07 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 20.82 on 64 degrees of freedom  
## Multiple R-squared:  0.4714, Adjusted R-squared:  0.4549   
## F-statistic: 28.54 on 2 and 64 DF,  p-value: 1.379e-09
```

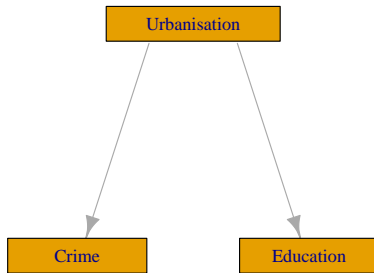
- From the output we find the prediction equation

$$\hat{y} = 59 - 0.58x_1 + 0.68x_2$$

- Not exactly what we expected based on the correlation.
- Now there appears to be a negative association between  $y$  and  $x_1$  (Simpsons Paradox)!
- We can also find the standard error (0.4725) and the corresponding t-score (-1.235) for the the slope of **Education**
- This yields a p-value of 22%, i.e. the slope is not significantly different from zero.

### 3.6 Simpsons paradox

- The example illustrates **Simpson's paradox**.
- When considered alone **Education** is a good predictor for **Crime** (with positive correlation).
- When we add **Urbanisation**, then **Education** has a negative effect on **Crime** (but not significant).



- A possible explanation is illustrated by the graph above.
  - **Urbanisation** has positive effect on both **Education** and **Crime**.
  - For a given level of **urbanisation** there is a (non-significant) negative association between **Education** and **Crime**.
  - Viewed alone **Education** is a good predictor for **Crime**. If **Education** has a large value, then this indicates a large value of **Urbanisation** and thereby a large value of **Crime**.

## 4 The general model

### 4.1 Regression model

- We have a sample of size  $n$ , where we have measured
  - the response  $y$ .
  - $k$  potential predictors  $x_1, x_2, \dots, x_k$ .

- Multiple regression model:

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \epsilon.$$

- The errors  $\epsilon$  are a sample from a population with mean zero and standard deviation  $\sigma_{y|x}$ .
- The **systematic** part of the model, i.e. when all errors are zero, says that **the mean response** is a linear function of the predictors:

$$E(y|x_1, x_2, \dots, x_k) = \alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k$$

- The symbol  $E$  is used here to denote expectation, i.e., mean value.

## 4.2 Interpretation of parameters

- In the multiple linear regression model

$$E(y|x_1, x_2, \dots, x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- The parameter  $\alpha$  is the **Intercept**, corresponding to the mean response, when all predictors are equal to zero.
- The parameters  $(\beta_1, \beta_2, \dots, \beta_k)$  are called **partial regression coefficients**.
- Imagine that all predictors but  $x_1$  are held fixed. Then  $y = \tilde{\alpha} + \beta_1 x_1$  is a line with slope  $\beta_1$ , which describes the rate of change in the mean response, when  $x_1$  is changed one unit. Here

$$\tilde{\alpha} = \alpha + \beta_2 x_2 + \dots + \beta_k x_k$$

is a constant number since we assumed all predictors but  $x_1$  was held fixed.

- The rate of change  $\beta_1$  does not depend on the value of the remaining predictors. In this case we say that there is **no interaction** between the effects of the predictors on the response.
- The above holds similarly for the other partial regression coefficients.
- An example of a model with interaction is

$$E(y|x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \alpha + \beta_2 x_2 + (\beta_1 + \beta_3 x_2) x_1$$

- When we fix  $x_2$  the line has slope  $\beta_1 + \beta_3 x_2$ , which depends on the chosen value of  $x_2$ .

## 5 Estimation

### 5.1 Estimation of model

- The estimate  $(a, b_1, b_2, \dots, b_k)$  for  $(\alpha, \beta_1, \beta_2, \dots, \beta_k)$  is determined by minimizing the sum of squared errors.
- Based on this estimate we write the prediction equation as

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

- The distance between model and data is measured by the sum of squared errors

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- We estimate  $\sigma_{y|x}$  by the quantity

$$s_{y|x} = \sqrt{\frac{SSE}{n - k - 1}}.$$

- Rather than  $n$  we divide  $SSE$  by **the degrees of freedom**  $df = n - k - 1$ . Theory shows, that this is reasonable.
- The degrees of freedom  $df$  are determined by the sample size minus the number of parameters in the regression equation.
- Currently we have  $k + 1$  parameters: 1 intercept and  $k$  slopes.

## 6 Multiple R-squared

### 6.1 Multiple $R^2$

- We want to compare two models to predict the response  $y$ . Analogous to simple linear regression we have the following setup:

- Model 1: We do not use the predictors, and use  $\bar{y}$  to predict any  $y$ -measurement. The corresponding prediction error is

- $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$  and is called the **Total Sum of Squares**.

- Model 2: We use the multiple prediction equation to predict  $y$ , i.e. the prediction error is

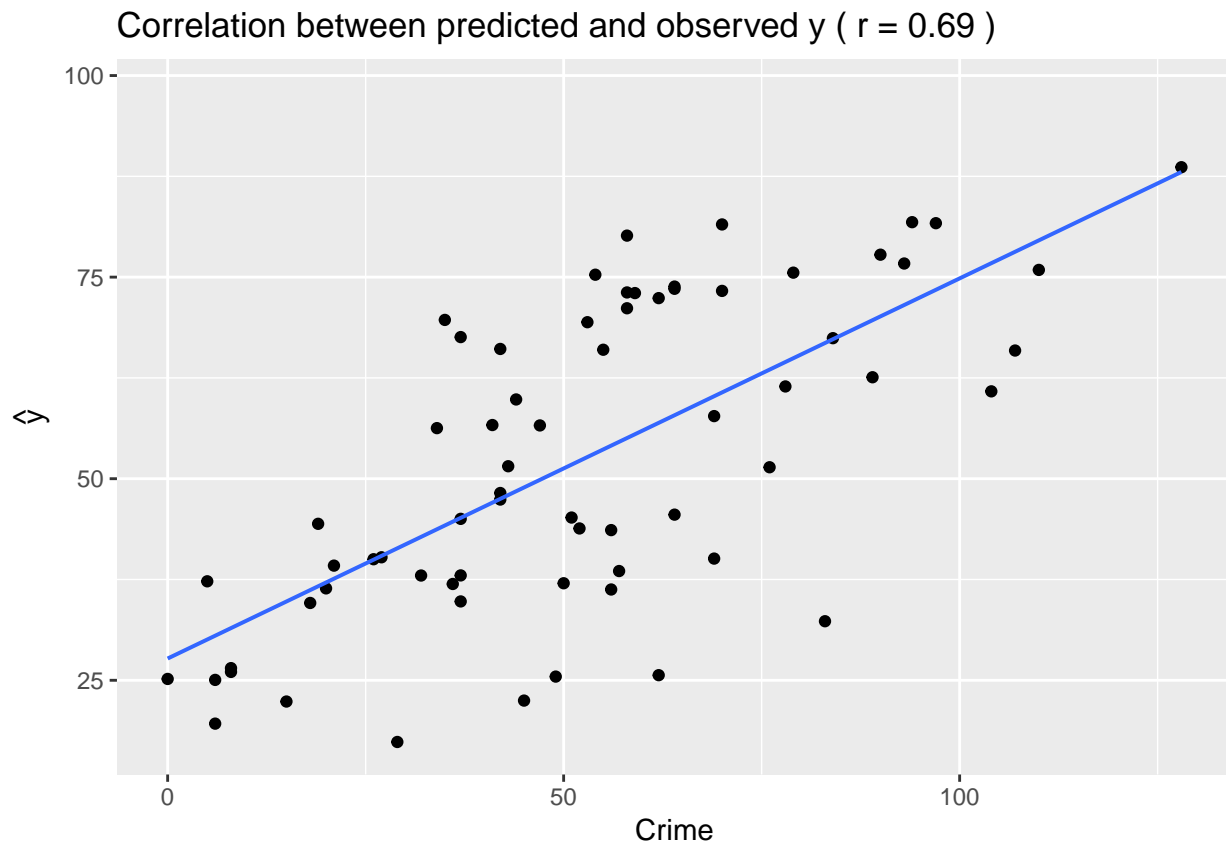
- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  and is called **Sum of Squared Errors**.

- We then define **the multiple coefficient of determination**

$$R^2 = \frac{TSS - SSE}{TSS}.$$

- Thus,  $R^2$  is the relative reduction in prediction error, when we use  $x_1, x_2, \dots, x_k$  as explanatory variables.
- It can be shown that the **multiple correlation**  $R = \sqrt{R^2}$  is the correlation between  $y$  and  $\hat{y}$ .

```
gf_point(predict(model) ~ FL$Crime) %>%
  gf_lm() %>%
  gf_labs(title = paste("Correlation between predicted and observed y ( r =", round(sqrt(summary(model)$r^2), 2), ")",
    x = "Crime",
    y = expression(hat(y)))
```



## 6.2 Example



```
summary(model)
```

```
##
## Call:
## lm(formula = Crime ~ Education + Urbanisation, data = FL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34.693 -15.742  -6.226  15.812  50.678
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.1181    28.3653   2.084  0.0411 *
## Education     -0.5834     0.4725  -1.235  0.2214
## Urbanisation   0.6825     0.1232   5.539 6.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.82 on 64 degrees of freedom
## Multiple R-squared:  0.4714, Adjusted R-squared:  0.4549
## F-statistic: 28.54 on 2 and 64 DF,  p-value: 1.379e-09
```

- The prediction equation is  $\hat{y} = 59 - 0.58x_1 + 0.68x_2$
- The estimate for  $\sigma_{y|x}$  is  $s_{y|x} = 20.82$  (Residual standard error in **R**) with  $df = 67 - 3 = 64$  degrees of freedom.
- Multiple  $R^2 = 0.4714$ , i.e. 47% of the variation in the response is explained by including the predictors in the model.
- The estimate  $b_1 = -0.5834$  has standard error (Std. Error)  $se = 0.4725$  with corresponding  $t$ -score ( $t$  value)  $t_{\text{obs}} = \frac{-0.5834}{0.4725} = -1.235$ .
- The hypothesis  $H_0 : \beta_1 = 0$  has the  $t$ -score  $t_{\text{obs}} = -1.235$ , which means that  $b_1$  isn't significantly different from zero, since the  $p$ -value ( $\text{Pr}(>|t|)$ ) is 22%. That means that we should exclude **Education** as a predictor.

### 6.3 Example

- Our final model is then a simple linear regression:

```
model2 <- lm(Crime ~ Urbanisation, data = FL)
summary(model2)
```

```
##
## Call:
## lm(formula = Crime ~ Urbanisation, data = FL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34.766 -16.541  -4.741  16.521  49.632
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.54125    4.53930   5.406 9.85e-07 ***
```

```
## Urbanisation 0.56220 0.07573 7.424 3.08e-10 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 65 degrees of freedom
## Multiple R-squared: 0.4588, Adjusted R-squared: 0.4505
## F-statistic: 55.11 on 1 and 65 DF, p-value: 3.084e-10
```

- The coefficient of determination always decreases, when the model is simpler. Now we have  $R^2 = 46\%$ , where before we had 47%. But the decrease is not significant.

## 7 F-test for effect of predictors

### 7.1 F-test

- We consider the hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

against the alternative, that at least one of these are non-zero.

- As test statistic we use

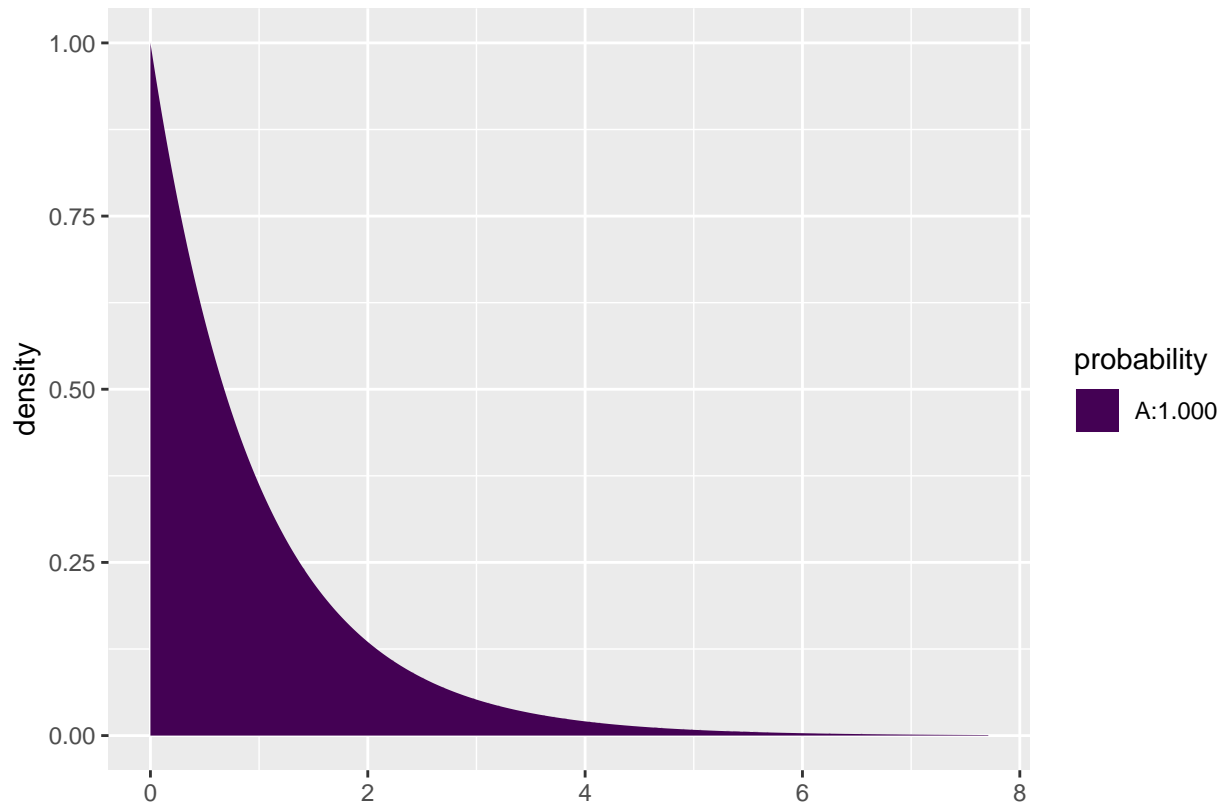
$$F_{obs} = \frac{(n - k - 1)R^2}{k(1 - R^2)}$$

- Large values of  $R^2$  implies large values of  $F$ , which points to the alternative hypothesis.
- I.e. when we have calculated the observed value  $F_{obs}$ , then we have to find the probability that a new experiment would result in a larger value.
- It can be shown that the reference distribution is (can be approximated by) a so-called **F-distribution** with **degrees of freedom**  $df_1 = k$  and  $df_2 = n - k - 1$ .

### 7.2 Example

- We return to **Crime** and the prediction equation  $\hat{y} = 59 - 0.58x_1 + 0.68x_2$ , where  $n = 67$  and  $R^2 = 0.4714$ . We have
  - $df_1 = k = 2$  since we have 2 predictors.
  - $df_2 = n - k - 1 = 67 - 2 - 1 = 64$ .
  - Then we can calculate  $F_{obs} = \frac{(n-k-1)R^2}{k(1-R^2)} = 28.54$
- To evaluate the value 28.54 in the relevant F-distribution:

```
1 - pdist("f", 28.54, df1=2, df2=64)
```



```
## [1] 1.378612e-09
```

- So  $p\text{-value} = 1.38 \times 10^{-9}$  (notice we don't multiply by 2 since this is a one-sided test; only large values point more towards the alternative than the null hypothesis).
- All this can be found in the summary output we already have:

```
summary(model)
```

```
##
## Call:
## lm(formula = Crime ~ Education + Urbanisation, data = FL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34.693 -15.742  -6.226  15.812  50.678
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.1181    28.3653   2.084  0.0411 *
## Education     -0.5834     0.4725  -1.235  0.2214
## Urbanisation   0.6825     0.1232   5.539 6.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.82 on 64 degrees of freedom
## Multiple R-squared:  0.4714, Adjusted R-squared:  0.4549
## F-statistic: 28.54 on 2 and 64 DF, p-value: 1.379e-09
```

## 8 Test for interaction

### 8.1 Interaction between effects of predictors

- Could it be possible that a combination of Education and Urbanisation is good for prediction? We want to investigate this using the model

$$E(y|x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

where we have extended with a possible effect of the product  $x_1 x_2$ :

```
model3 <- lm(Crime ~ Education * Urbanisation, data = FL)
summary(model3)
```

```
##
## Call:
## lm(formula = Crime ~ Education * Urbanisation, data = FL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -35.181 -15.207  -6.457  14.559  49.889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    19.31754   49.95871   0.387   0.700
## Education         0.03396    0.79381   0.043   0.966
## Urbanisation     1.51431    0.86809   1.744   0.086 .
## Education:Urbanisation -0.01205    0.01245  -0.968   0.337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.83 on 63 degrees of freedom
## Multiple R-squared:  0.4792, Adjusted R-squared:  0.4544
## F-statistic: 19.32 on 3 and 63 DF,  p-value: 5.371e-09
```

- When we look at the  $p$ -values in the table nothing is significant at the 5% level!
- But the F-statistic tells us that the predictors collectively have a significant prediction ability.
- Why has the highly significant effect of  $x_2$  disappeared? Because the predictors  $x_1$  and  $x_1 x_2$  are able to explain the same as  $x_2$ .
- Previously we only had  $x_1$  as alternative explanation to  $x_2$  - and that wasn't enough.
- The phenomenon is called **multicollinearity** and illustrates that we can have different models with equally good predictive properties.
- In this case we will choose the model with  $x_2$  since it is simpler.
- However, in general it can be difficult to choose between models. For example, if both height and weight are good predictors of some response, but one of them can be left out, which one do we choose?

## 9 The regression problem

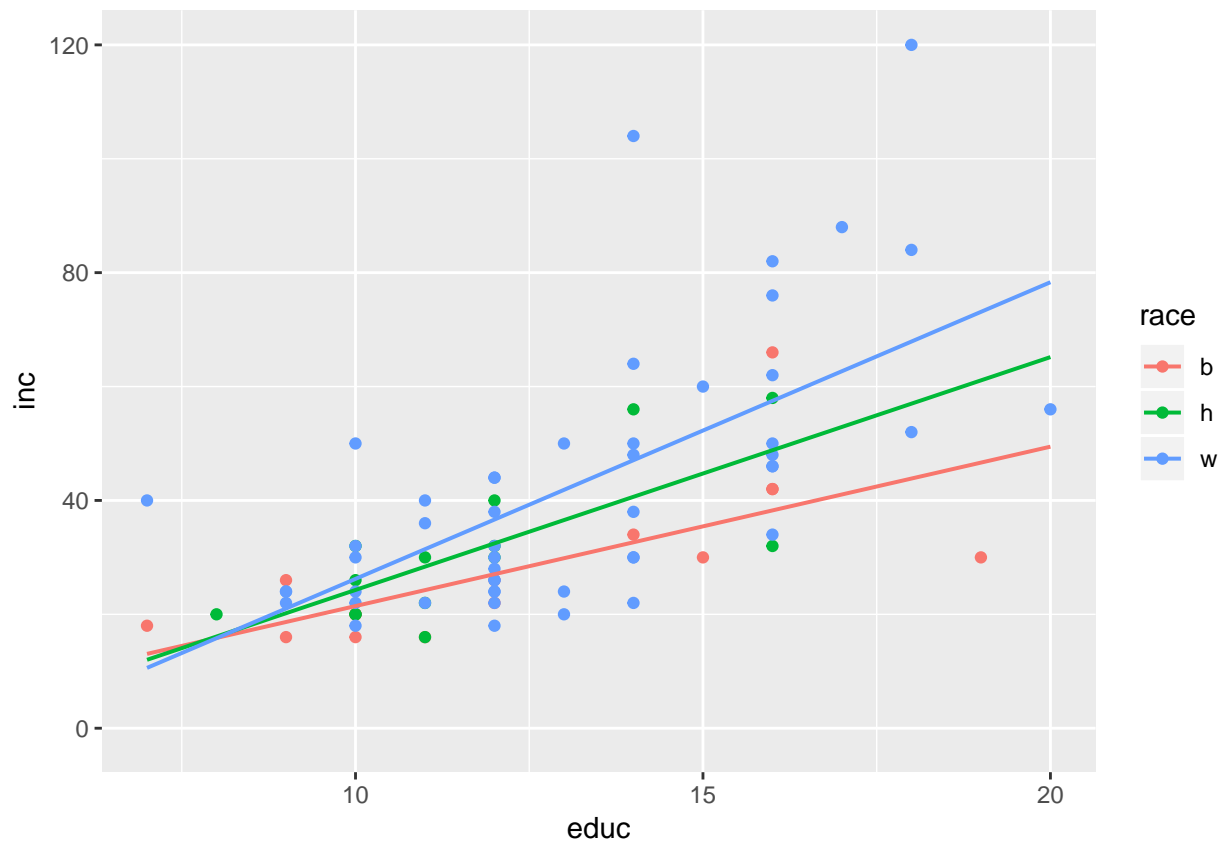
### 9.1 Example

- We will study the dataset in Agresti Table 13.1 available as `Income.txt` on the course website. We read in data in RStudio

```
Income <- read.delim("https://asta.math.aau.dk/datasets?file=Income.txt")
```

- We have a sample with measurements of 3 variables:
  - **y=income**: Quantitative variable, which is yearly income. This will be our response.
  - **x=education**: Quantitative predictor, which is the number of years of education.
  - **z=race**: Explanatory factor with levels **b**(black), **h**(hispanic) and **w**(white).
- We always start with some graphics (remember the function `gf_point` for plotting points and `gf_lm` for adding a regression line).

```
library(mosaic)  
gf_point(inc ~ educ, color = ~race, data = Income) %>% gf_lm()
```



- An unclear picture, but a tendency to increasing income with increasing education.
- The trend lines for the three races are different. But is the difference significant? Or can the difference be explained by sampling variation?
- Such a regression with both qualitative and quantitative predictors is called an analysis of covariance (ANCOVA). When the model only contains qualitative predictors, the problem is known as analysis of variance (ANOVA) which we briefly mention at the end.

## 10 Dummy coding

### 10.1 Dummy coding

- First, we will look at the model **without interaction**, i.e. the effect of `education` is the same for all races, which corresponds to parallel lines.
- We also have to introduce dummy coding of the factor `z`:
  - $z_1 = 1$  if `race=b` and zero otherwise
  - $z_2 = 1$  if `race=h` and zero otherwise
- This determines the regression model:

$$E(y|x, z) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2$$

which corresponds to **parallel** regressions lines for each race.

- **w**: ( $z_1 = 0, z_2 = 0$ )  $E(y|x) = \alpha + \beta x$
- **b**: ( $z_1 = 1, z_2 = 0$ )  $E(y|x) = \alpha + \beta_1 + \beta x$ .
- **h**: ( $z_1 = 0, z_2 = 1$ )  $E(y|x) = \alpha + \beta_2 + \beta x$ .
- $\beta_1$  is the difference in **Intercept** between black and white.
- $\beta_2$  is the difference in **Intercept** between Hispanic and white.

### 10.2 Example

- We want to tell R that we want `race` to be a factor (grouping variable) and we want `w` as reference level for race (default is lexicographical ordering, i.e. (`b`, `h`, `w`) and `b` would then be the reference):

```
Income$race <- factor(Income$race)
Income$race <- relevel(Income$race, "w")
```

- Then we use `+` in the model formula to only have additive effects of `educ` and `race`, i.e. a model without interaction:

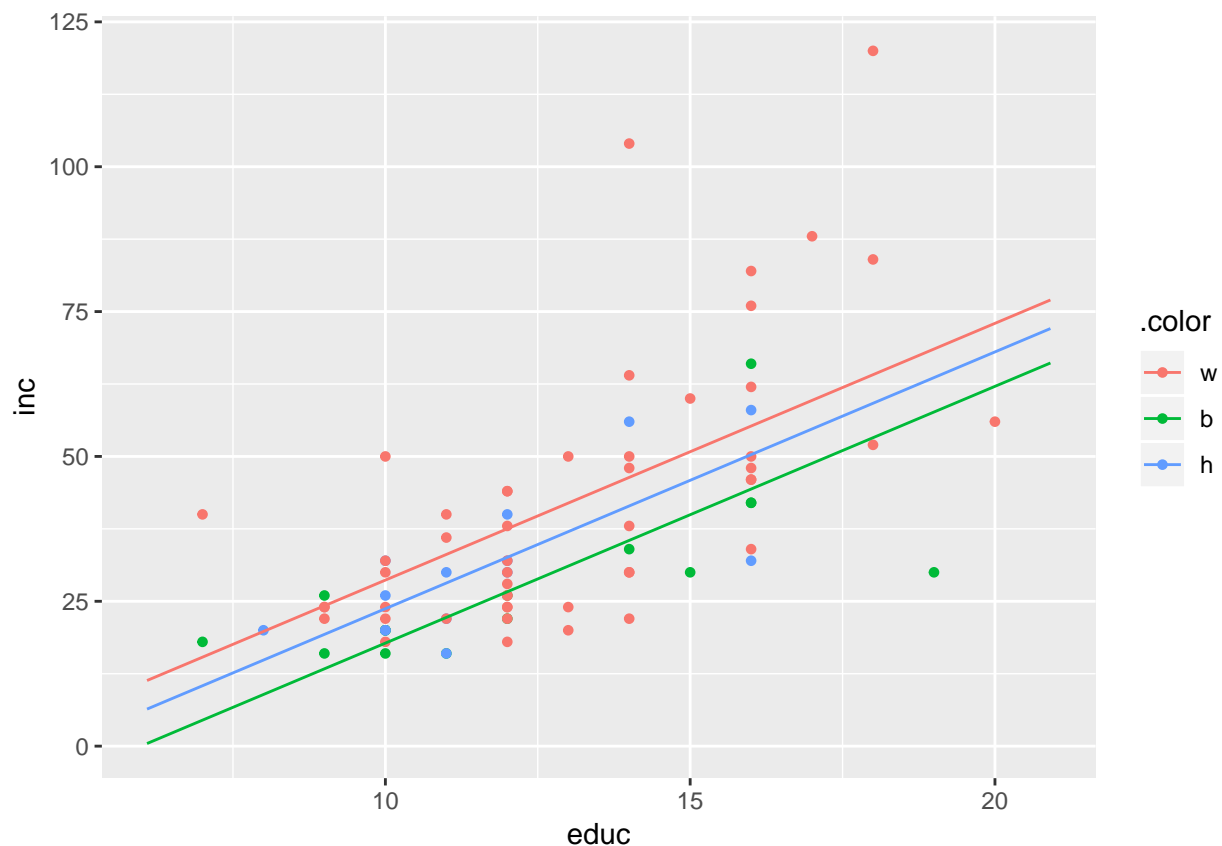
```
modell1 <- lm(inc ~ educ + race, data = Income)
summary(modell1)

##
## Call:
## lm(formula = inc ~ educ + race, data = Income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.664  -9.622  -1.642   6.552  57.620
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.6635     8.4121  -1.862  0.0665 .
## educ         4.4317     0.6191   7.158 4.42e-10 ***
```

```
## raceb      -10.8744    4.4730  -2.431   0.0174 *
## raceh      -4.9338    4.7632  -1.036   0.3036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.46 on 76 degrees of freedom
## Multiple R-squared:  0.462, Adjusted R-squared:  0.4408
## F-statistic: 21.75 on 3 and 76 DF,  p-value: 2.853e-10
```

- The common slope to `educ` is estimated to be  $\hat{\beta} = 4.4316685$ , with corresponding p-value= $4.42 \times 10^{-10}$  which is significantly different from zero.
- There is a clear positive effect of `educ` on `income`.
- The estimate for w-intercept is  $\hat{\alpha} = -15.6635$ , which isn't significantly different from zero if we test at level 5% (this test is not really of interest).
- **The difference** between b- and w-intercept (`raceb`) is  $\hat{\beta}_1 = -10.8744$ , which is significant with p-value=1.74%.
- There is no significant difference between h- and w-intercept.

```
plotModel(model1)
```



### 10.3 Example: Prediction equations

```
summary(model1)
```

```

##
## Call:
## lm(formula = inc ~ educ + race, data = Income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.664  -9.622  -1.642   6.552  57.620
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.6635     8.4121  -1.862  0.0665 .
## educ         4.4317     0.6191   7.158 4.42e-10 ***
## raceb       -10.8744     4.4730  -2.431  0.0174 *
## raceh        -4.9338     4.7632  -1.036  0.3036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.46 on 76 degrees of freedom
## Multiple R-squared:  0.462, Adjusted R-squared:  0.4408
## F-statistic: 21.75 on 3 and 76 DF, p-value: 2.853e-10

```

- Reference/baseline group (white):

$$\hat{y} = -15.66 + 4.43x$$

- Black:

$$\hat{y} = -15.66 - 10.87 + 4.43x = -26.54 + 4.43x$$

- Hispanic:

$$\hat{y} = -15.66 - 4.93 + 4.43x = -20.60 + 4.43x$$



## 10.4 Agresti – summary

TABLE 13.4: Summary of Regression Equations and Parameters for Model with No Interaction, when Categorical Predictor Has Three Categories

Category	y-Intercept	Slope	Mean $E(y)$ at Fixed $x$	Difference From Mean of Category 3, Controlling for $x$
1	$\alpha + \beta_1$	$\beta$	$(\alpha + \beta_1) + \beta x$	$\beta_1$
2	$\alpha + \beta_2$	$\beta$	$(\alpha + \beta_2) + \beta x$	$\beta_2$
3	$\alpha$	$\beta$	$\alpha + \beta x$	0

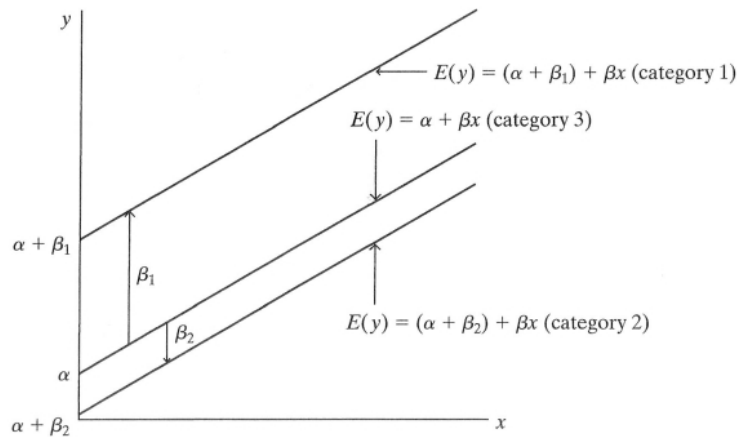


FIGURE 13.5: Graphic Portrayal of a Model with No Interaction, when the Categorical Predictor Has Three Categories

## 11 Model with interaction

### 11.1 Interaction

- In the following we will expand the model to include interaction between the effects of race and education on income. Before proceeding, let us recall what interaction means (and doesn't mean) in this context:
- Interaction between the effects of race and education on income does **not** mean that the values of education and race themselves are related or affect each other.
- Interaction between the effects of race and education on income means that the relationship between education and income depends on the value of race. I.e. for each fixed value of race the slope of the line relating education and income may have a different value.
- Often we just refer to this as “interaction between education and race” when it really should read “interaction between the effects of race and education on income”.

### 11.2 Interaction

- We will expand the regression model, so we include interaction between  $x$  and  $z_1$  respectively  $z_2$ :

$$E(y|x, z) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 x + \beta_4 z_2 x.$$

- This yields a regression line for each race:
- w ( $z_1 = 0, z_2 = 0$ ):  $E(y|x) = \alpha + \beta x$

- **b** ( $z_1 = 1, z_2 = 0$ ):  $E(y|x) = \alpha + \beta_1 + (\beta + \beta_3)x$ .
- **h** ( $z_1 = 0, z_2 = 1$ ):  $E(y|x) = \alpha + \beta_2 + (\beta + \beta_4)x$ .
- $\beta_1$  is **the difference** in Intercept between black and white, while  $\beta_3$  is **the difference** in slope between black and white.
- $\beta_2$  is **the difference** in Intercept between Hispanic and white, while  $\beta_4$  is the difference in slope between Hispanic and white.

### 11.3 Example: Prediction equations

- When we use `*` in the model formula we include interaction between `educ` and `race`:

```
model2 <- lm(inc ~ educ * race, data = Income)
summary(model2)

##
## Call:
## lm(formula = inc ~ educ * race, data = Income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.064  -9.448  -1.453   6.167  56.936
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.8688    10.4982  -2.464  0.0161 *
## educ         5.2095     0.7828   6.655 4.3e-09 ***
## raceb       19.3333    18.2928   1.057  0.2940
## raceh        9.2640    24.2797   0.382  0.7039
## educ:raceb  -2.4107     1.4177  -1.700  0.0933 .
## educ:raceh  -1.1208     2.0060  -0.559  0.5781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.37 on 74 degrees of freedom
## Multiple R-squared:  0.4825, Adjusted R-squared:  0.4475
## F-statistic: 13.8 on 5 and 74 DF,  p-value: 1.618e-09
```

- Reference/baseline group (white):

$$\hat{y} = -25.87 + 5.21x$$

- Black:

$$\hat{y} = -25.87 + 19.33 + (5.21 - 2.41)x = -6.54 + 2.80x$$

- Hispanic:

$$\hat{y} = -25.87 + 9.26 + (5.21 - 1.12)x = -16.60 + 4.09x$$

### 11.4 Example: Individual tests

```
summary(model2)
```

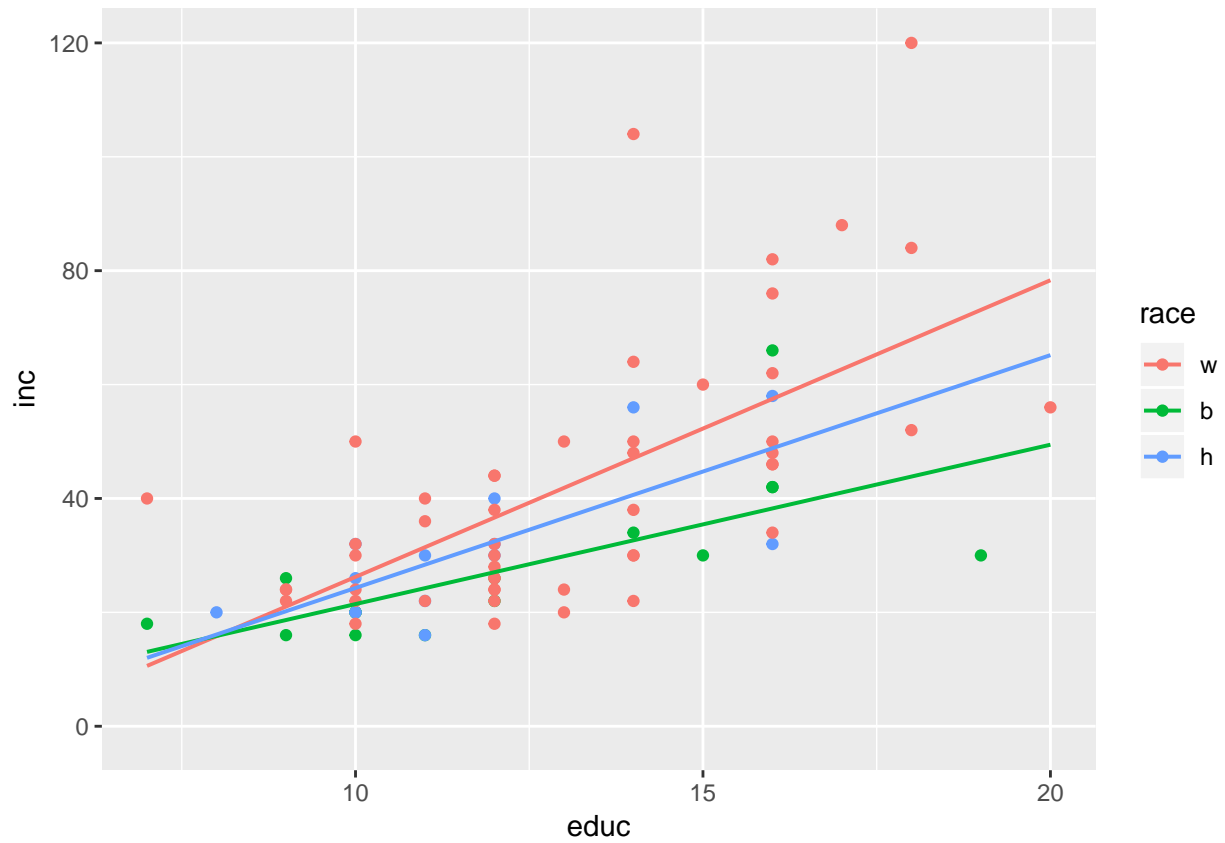
```

##
## Call:
## lm(formula = inc ~ educ * race, data = Income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.064  -9.448  -1.453   6.167  56.936
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.8688    10.4982  -2.464  0.0161 *
## educ         5.2095     0.7828   6.655 4.3e-09 ***
## raceb       19.3333    18.2928   1.057  0.2940
## raceh        9.2640    24.2797   0.382  0.7039
## educ:raceb  -2.4107     1.4177  -1.700  0.0933 .
## educ:raceh  -1.1208     2.0060  -0.559  0.5781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.37 on 74 degrees of freedom
## Multiple R-squared:  0.4825, Adjusted R-squared:  0.4475
## F-statistic: 13.8 on 5 and 74 DF, p-value: 1.618e-09

```

- **The difference** in slope between b and w (educ:raceb) is estimated to  $\hat{\beta}_3 = -2.4107$ . With p-value=9.33% there is no significant difference.
- Furthermore, there isn't any significant difference of slope between h and w. In other words there is probably not interaction between educ and race.

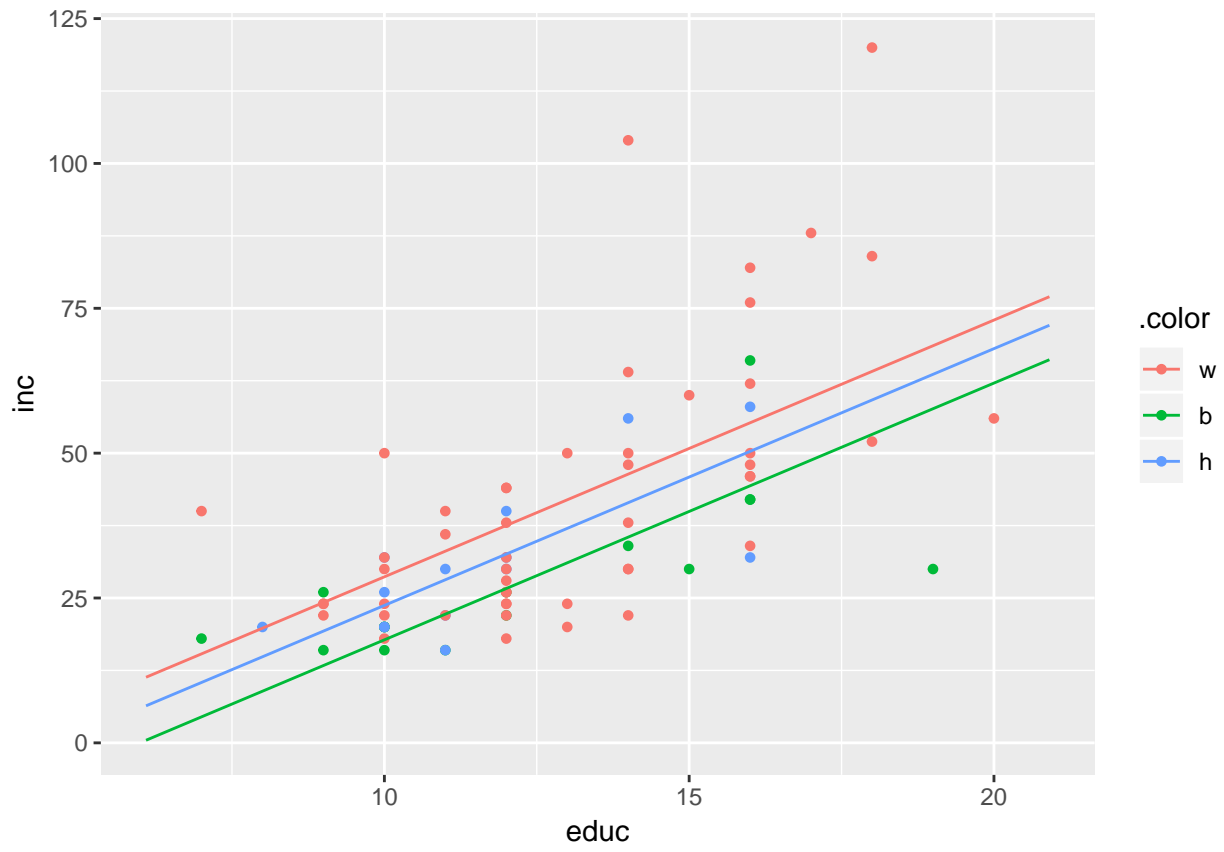
```
gf_point(inc ~ educ, color = ~ race, data = Income) %>% gf_lm()
```



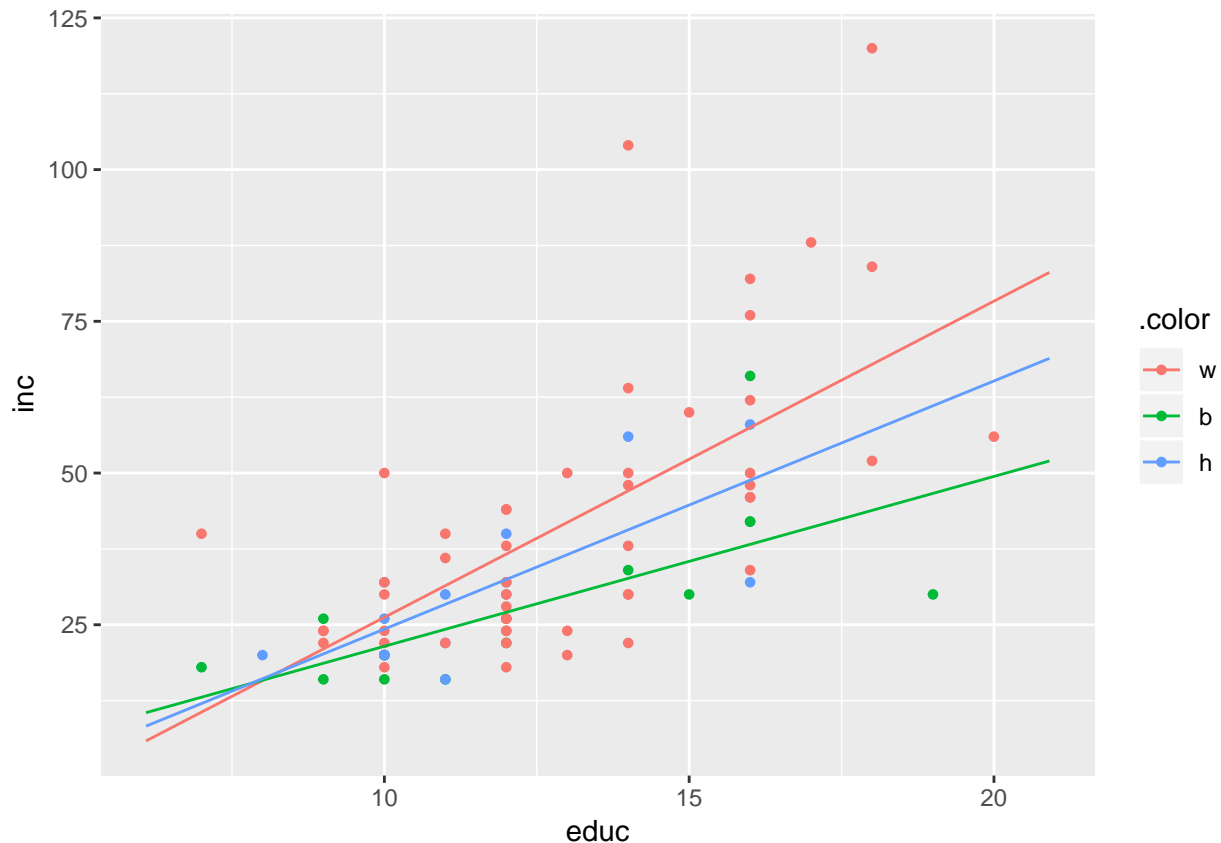
## 12 Test for no interaction

### 12.1 Test for no interaction

```
plotModel(model1)
```



```
plotModel(model2)
```



```
summary(model1)$r.squared
```

```
## [1] 0.4619906
```

```
summary(model2)$r.squared
```

```
## [1] 0.4824822
```

- Is `model2` significantly better than `model1`? I.e. is  $R^2$  significantly higher for `model2`?

## 12.2 Hypothesis and test statistic

- The simpler `model1` is obtained from the more complicated `model2` by setting  $\beta_3 = 0$  and  $\beta_4 = 0$ , so the null hypothesis “the simpler additive model describes data sufficiently well compared to the complicated interaction model” is really the simple mathematical hypothesis:

$$H_0 : \beta_3 = 0, \beta_4 = 0.$$

- We will look at the difference between  $R^2$  for the two models, but as before (for multiple linear regression) we have to convert this to an  $F$  statistic which we can then calculate a  $p$ -value for.
- Formula for  $F_{obs}$  (no need to learn this by heart):

$$F_{obs} = \frac{(R_2^2 - R_1^2)/(df_1 - df_2)}{(1 - R_2^2)/df_2}$$

where  $df_1$  and  $df_2$  are  $n$  minus the number of model parameters for the two models (i.e.  $80-4=76$  and  $80-6=74$  in our case).

- The formula for  $F_{obs}$  can be rewritten in terms of sums of squared errors (SSE) for each model (no need to memorize it):

$$F_{obs} = \frac{(SSE_1 - SSE_2)/(df_1 - df_2)}{(SSE_2)/df_2}.$$

- In the literature SSE is sometimes denoted by RSS for **Residual Sums of Squares**; i.e SSE = RSS.

## 12.3 Test for no interaction in R

- In R the calculations are done using `anova`:

```
anova(model1, model2)

## Analysis of Variance Table
##
## Model 1: inc ~ educ + race
## Model 2: inc ~ educ * race
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      76 18164
## 2      74 17472  2    691.84 1.465 0.2377
```

- The F-test for dropping the interaction `educ:race` has F-value=1.465, which in no way is significant with p-value=23.77%.

## 13 Hierarchy of models

### 13.1 Hierarchy of models

- **Interaction:** The most general model with main effects `educ` and `race` and interaction `educ:race`:

```
Interaction <- lm(inc ~ educ * race, data = Income)
```

- **MainEffects:** The model where there are additive effects of `educ` and `race`.

```
MainEffects <- lm(inc ~ educ + race, data = Income)
```

- **educEff:** Model where there only is an effect of `educ` (simple lin. reg.).

```
educEff <- lm(inc ~ educ, data = Income)
```

- **raceEff:** Model where there only is an effect of `race` (a different mean for each group – more on this later).

```
raceEff <- lm(inc ~ race, data = Income)
```

- We can, corresponding to Agresti Table 13.10, make F-tests for 3 pairwise comparisons of models.

## 13.2 Example

- Comparing `MainEffects` and `Interaction` is what we have already done.

```
anova(MainEffects, Interaction)
```

```
## Analysis of Variance Table
##
## Model 1: inc ~ educ + race
## Model 2: inc ~ educ * race
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      76 18164
## 2      74 17472  2    691.84 1.465 0.2377
```

- We recognize  $F = 1.465$  with  $p\text{-value} = 23.77\%$ , i.e. `model2` isn't significantly better than `model1`. So no `educ:race` interaction.
- In the same manner we can compare `educEff` and `MainEffects`. I.e. we investigate whether the effect of `race` can be left out.

```
anova(educEff, MainEffects)
```

```
## Analysis of Variance Table
##
## Model 1: inc ~ educ
## Model 2: inc ~ educ + race
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      78 19625
## 2      76 18164  2    1460.6 3.0556 0.05292 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- If any, the effect of `race` is weak with  $p\text{-value} = 5.292\%$ .
- Finally, we compare `raceEff` and `MainEffects`. Clearly `educ` cannot be left out ( $P\text{-value} = 4.422 \times 10^{-10}$ ).

```
anova(raceEff, MainEffects)
```

```
## Analysis of Variance Table
##
## Model 1: inc ~ race
## Model 2: inc ~ educ + race
##   Res.Df  RSS Df Sum of Sq    F    Pr(>F)
## 1      77 30410
## 2      76 18164  1    12245 51.235 4.422e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



### 13.3 Model with only a single categorical predictor

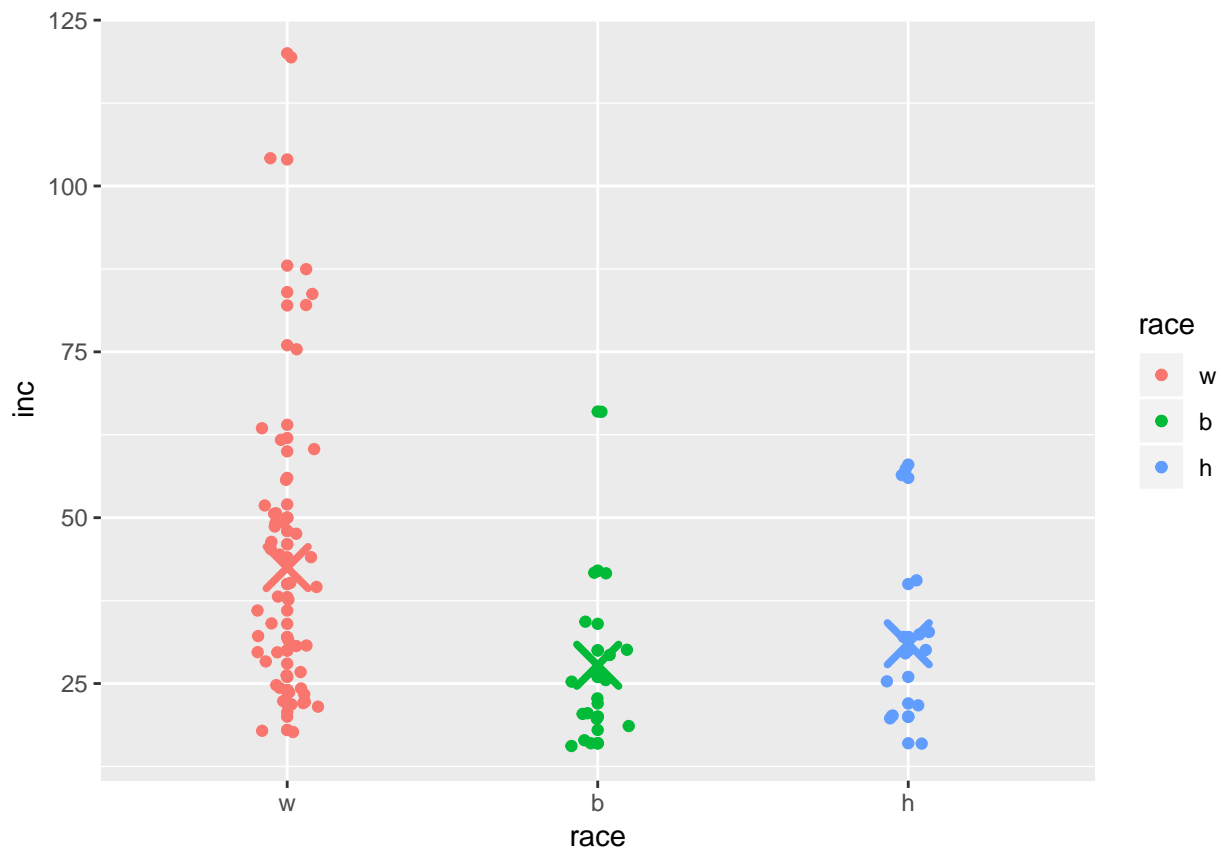
- The model called `raceEff` above has a different mean income for each race.
  - $\alpha$  is the mean income for the reference group (`w`)
  - $\beta_1$  is the difference between the `b` and `w` in mean income
  - $\beta_2$  is the difference between the `h` and `w` in mean income

- Mathematically

$$E(y|z) = \alpha + \beta_1 z_1 + \beta_2 z_2$$

- `w`: ( $z_1 = 0, z_2 = 0$ )  $E(y|x) = \alpha$
  - `b`: ( $z_1 = 1, z_2 = 0$ )  $E(y|x) = \alpha + \beta_1$ .
  - `h`: ( $z_1 = 0, z_2 = 1$ )  $E(y|x) = \alpha + \beta_2$ .
- Graphically (notice we don't have lines anymore since there is no quantitative predictor – predicted means are indicated with crosses on the figure):

```
gf_point(inc ~ race, data = Income, color = ~race) %>%  
gf_jitter(width = 0.1) %>%  
gf_point(mean ~ race, data = favstats(inc ~ race, data = Income),  
         size = 6, shape = 4, stroke = 2, show.legend = FALSE)
```



- Summary of the fitted model:

```
summary(raceEff)
```

```
##
## Call:
## lm(formula = inc ~ race, data = Income)
##
## Residuals:
##   Min     1Q   Median     3Q      Max
## -24.48 -12.48  -4.48   7.52  77.52
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.480      2.810  15.115 <2e-16 ***
## raceb         -14.730      5.708  -2.581  0.0118 *
## raceh         -11.480      6.009  -1.910  0.0598 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.87 on 77 degrees of freedom
## Multiple R-squared:  0.0993, Adjusted R-squared:  0.0759
## F-statistic: 4.244 on 2 and 77 DF,  p-value: 0.01784
```

- Here we see:
  - Significant overall effect of race since  $H_0 : \beta_1 = \beta_2 = 0$  is rejected based on p-value of 1.78%
  - Difference between w and b is significant (p-value is 1.18%)
  - Difference between w and h is nearly significant (p-value is 5.98%)

### 13.4 Example with more predictors

- The methods generalize to models with more than 2 predictors.
- We return to the dataset `Ericksen`, where we study the response `crime`:

```
Ericksen <- read.delim("https://asta.math.aau.dk/datasets?file=Ericksen.txt")
model <- lm(crime ~ city * highschool + city * poverty, data = Ericksen)
```

- The variables are:
  - `crime`: Quantitative variable
  - `city`: city or state
  - `highschool`: Quantitative variable
  - `poverty`: Quantitative variable
- The model has 3 predictors with main effects and includes
  - interaction between `city` and `highschool`
  - interaction between `city` and `poverty`.

```
summary(model)
```

```
##
## Call:
## lm(formula = crime ~ city * highschool + city * poverty, data = Ericksen)
##
## Residuals:
```

```

##      Min      1Q  Median      3Q      Max
## -38.741  -8.745  -1.557   7.820  47.470
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.1456    18.1254   3.373 0.001305 **
## citystate        18.1526    20.4131   0.889 0.377413
## highschool      -1.5711     0.6062  -2.592 0.011979 *
## poverty          5.3105     1.4333   3.705 0.000463 ***
## citystate:highschool  0.7025     0.7327   0.959 0.341523
## citystate:poverty  -5.1862     1.6619  -3.121 0.002773 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.16 on 60 degrees of freedom
## Multiple R-squared:  0.6577, Adjusted R-squared:  0.6292
## F-statistic: 23.06 on 5 and 60 DF,  p-value: 7.748e-13

```

- There isn't significant (p-value=34.1523%) interaction between `city` and `highschool`.
- I.e. the effect of `highschool` on crime is the same in metropolitan areas (`city=city`) and the non-metropolitan areas (`city=state`).
- There is clearly (p-value=0.2773%) interaction between `city` and `poverty`.
- I.e. the effect of `poverty` on crime is different in metropolitan and non-metropolitan areas.
- For `city=state`, the effect of `poverty` (on crime) is smaller than in the major cities.
- Hence, poverty has larger effect on crime in the major cities than in the states outside the major cities.