ASTA

The ASTA team

Contents

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).

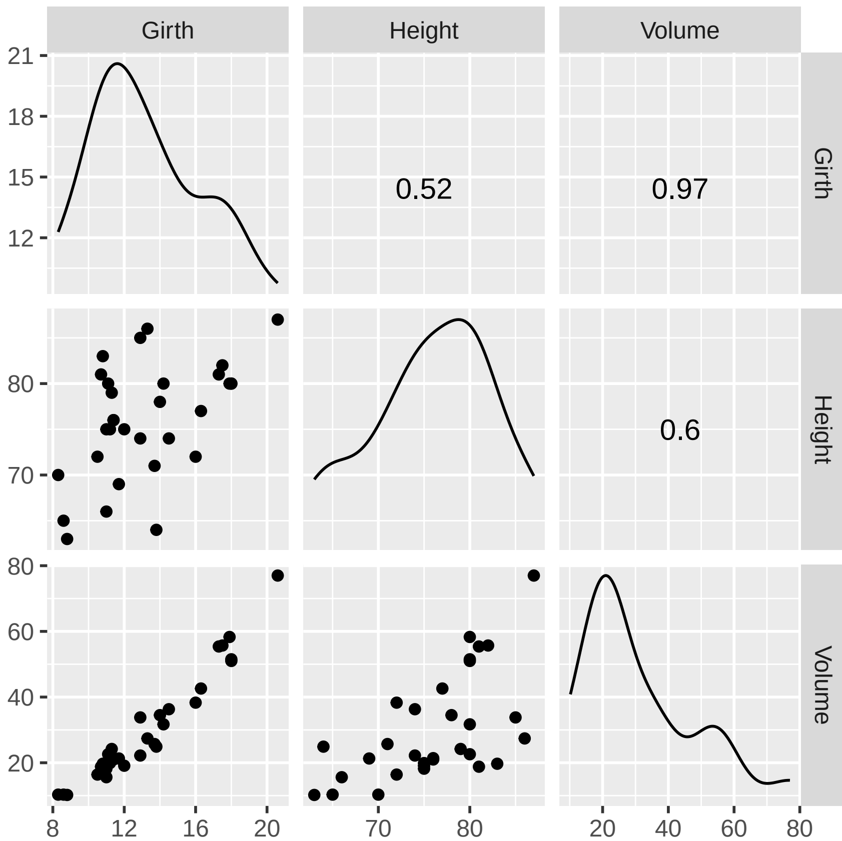
```
trees <- read.delim("https://asta.math.aau.dk/datasets?file=trees.txt")
```

- In this experiment we have measurements of 3 variables for 31 randomly chosen trees:
- `Girth` numeric. Tree diameter in inches.
- `Height` numeric. Height in ft.
- `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.

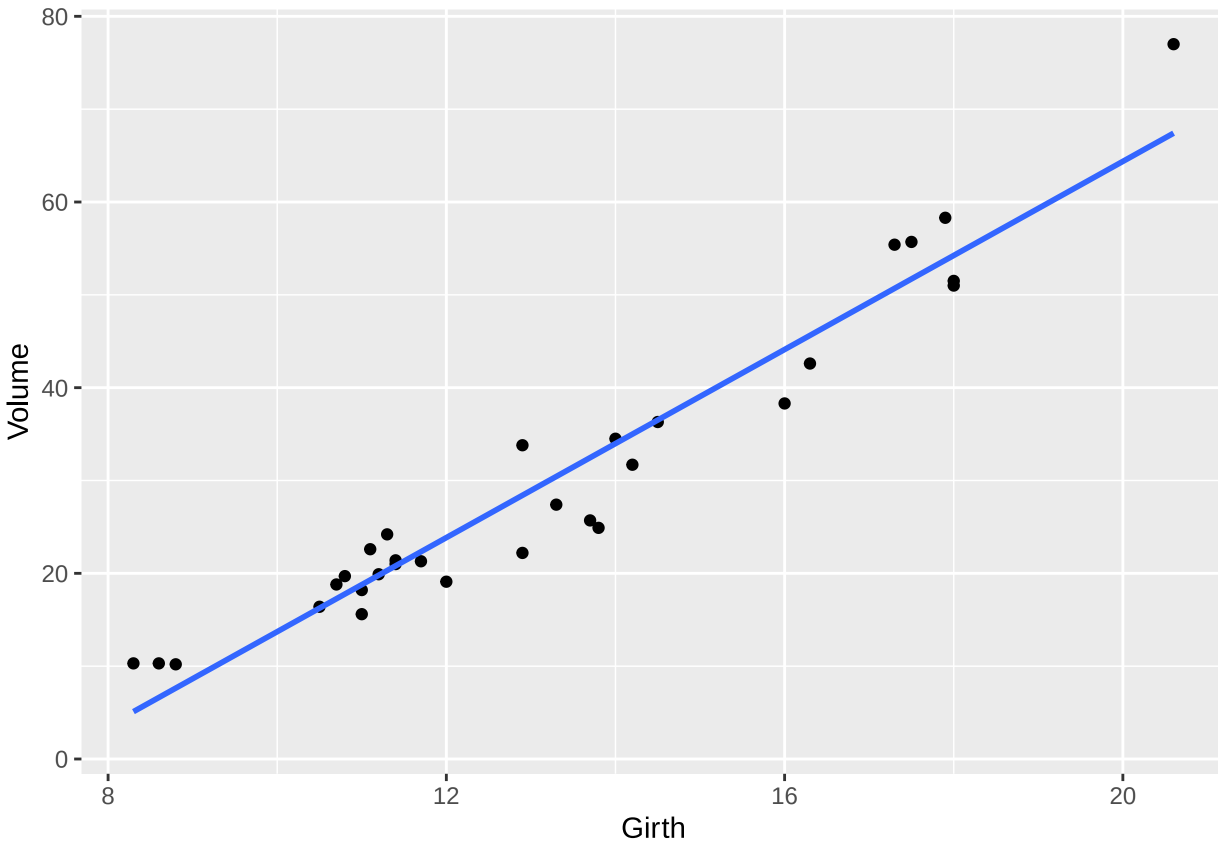
1.3 Simple linear regression

- We choose to use `x=Girth` as predictor for `y=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
  - α is called the `Intercept` - the line's intercept with the y-axis, corresponding to the response for $x = 0$.
  - β 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 ε is a **random error**, which causes deviations from the line.
- We will continue under the following **fundamental assumption**:
  - The errors ε are normally distributed with mean zero and standard deviation σ.
- We call σ 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:
  - (α, β) which determine the line
  - σ which is the standard deviation of the deviations from the line.
- How are these estimated, when we have a sample $(x_1, y_1), \ldots, (x_n, y_n)$ of pairs 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 (α, β), 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 given by the estimates $(\hat{\alpha}, \hat{\beta}) = (a, b)$,

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

  is called the **regression equation** or **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 σ we need the **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 σ by the quantity

$$s = \sqrt{\frac{SSE}{n-2}}$$

- 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: (α, β).

1.8 Example in R

```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 $s = 4.252$ 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 (α, β), the estimate s of σ 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.
- Observed values of b far away from zero are critical for the null-hypothesis?
- It can be shown that b has standard error
$$se_b = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$
with $df = n - 2$ 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:

```r
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 $s = 4.252$ 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(`Pr(>|t|)`) we see that the p-value is less than 2×10^{-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:

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

  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 $t_{crit} =$ `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 (sample) 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 sample 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

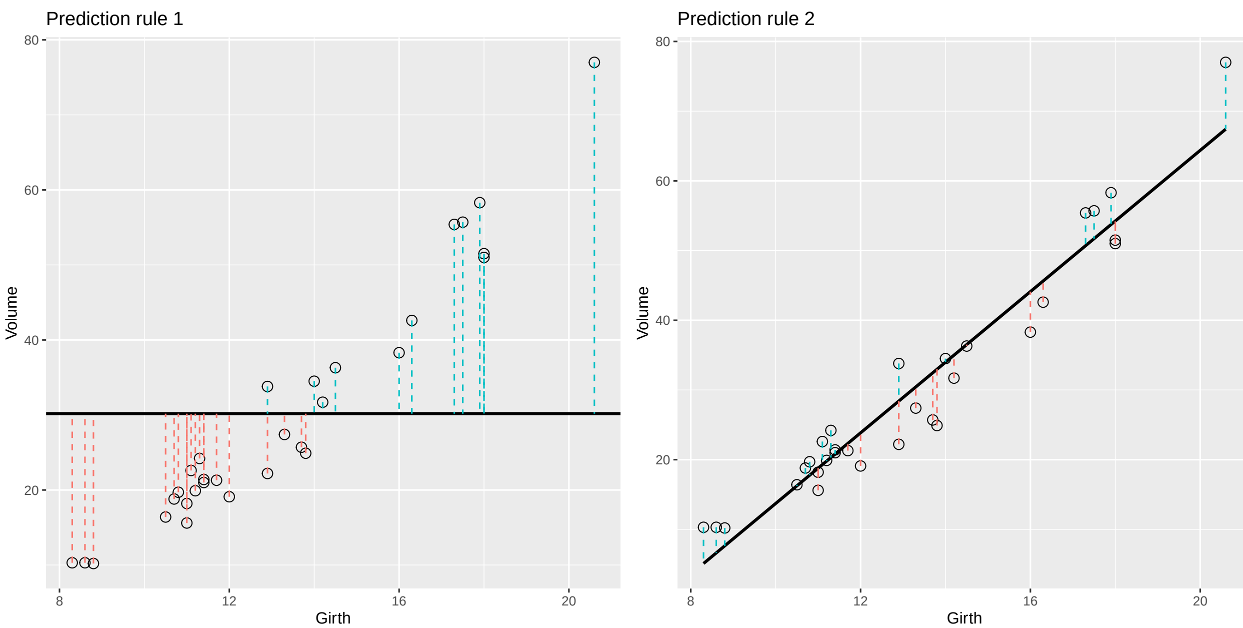
$$r^2 = \frac{TSS - SSE}{TSS}$$

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

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once per session.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.

## `geom_smooth()` using formula = 'y ~ x'
```



- 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.

```r
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 Introduction to multiple regression model

3.1 Multiple regression model

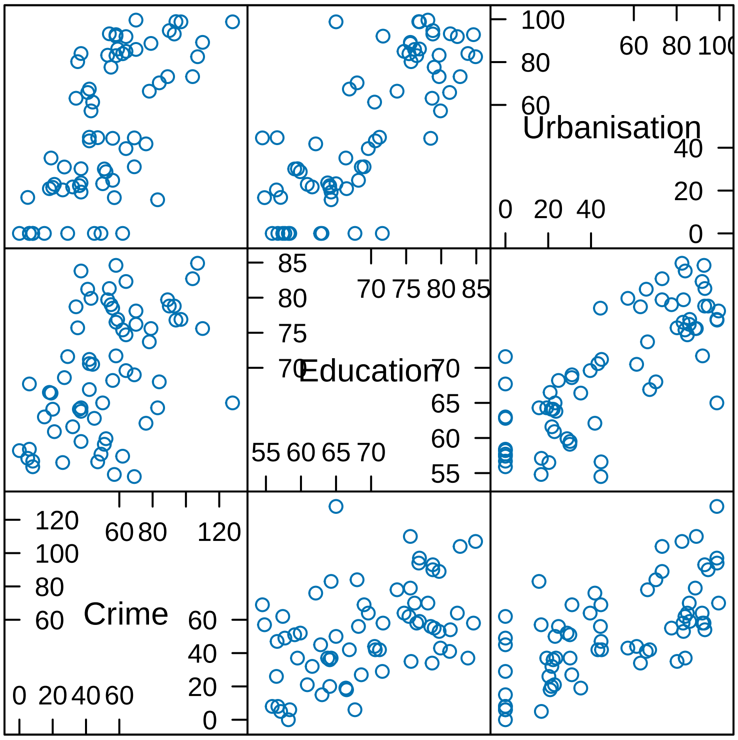
- We look at data set containing measurements from the 67 counties of Florida.
- Our focus is on
  - The response y: `Crime` which is the crime rate
  - The predictor x_1: `Education` which is proportion of the population with high school exam
  - The predictor x_2: `Urbanisation` which is proportion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

3.2 Example

```r
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
```

```r
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 ϵ are random noise with mean zero and standard deviation σ.
- 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
```

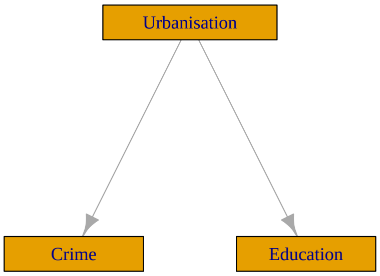
- From the output we find the prediction equation

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

- 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 Example from last lecture

4.1 Crime data set

- The data are measurements from 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

```
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
```

4.2 Multiple regression model for crime data

- Both `Education` (x_1) and `Urbanisation` (x_2) were correlated with `Crime` (y).
- We consider the model

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

- The errors ϵ are random terms with a $\texttt{norm}(0, \sigma)$ distribution.
- The graph for the mean response is a 2-dimensional plane in a 3-dimensional space.

```
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

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

5 The general model

5.1 Regression model

- In a multiple regression we have
  - a response variable Y.
  - k predictor variables $x_1, x_2, \dots, x_k$.
- Multiple regression model:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

- The **systematic** part of the model says that **the mean response** is a linear function of the predictors:

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

  - Here $E(Y|x_1, x_2, \dots, x_k)$ is used to denote the mean value of the response when we know the value of the predictors $x_1, \dots, x_k$.
- The **error** ϵ is a random variable having a distribution with
  - a normal distribution
  - mean 0
  - standard deviation σ

5.2 Interpretation of parameters

- In the multiple linear regression model

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

  - The parameter α is the `Intercept`, corresponding to the mean response, when all predictors are equal to zero.
  - The parameters $(\beta_1, \beta_2, \cdots, \beta_k)$ are called **partial regression coefficients**.
- Imagine that all predictors but x_1 are held fixed. Then

$$E(y|x_1, x_2, \dots, x_k) = \tilde{\alpha} + \beta_1 x_1,$$

  where

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

- So the mean response depends linearly on x_1 when all other variables are kept fixed.
  - The line has slope β_1, which describes the change in the mean response, when x_1 is changed one unit.
  - The rate of change β_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.

6 Estimation

6.1 Estimation of model parameters

- Suppose we have a sample of size n.
- Based on the sample, we estimate $(\alpha, \beta_1, \beta_2, \dots, \beta_k)$ by the values $(a, b_1, b_2, \dots, b_k)$.
- Based on this estimate we obtain the **prediction equation** (or **regression equation**) as

$$\widehat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

- The difference between our observations and the predictions made by the prediction equation are called the **residuals** $e_i = y_i - \widehat{y}_i$.
- The estimates $(a, b_1, b_2, \dots, b_k)$ are chosen by the **least squares method**, which seeks to minimize the sum of squared residuals

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

6.2 Estimation of error variance

- Recall that σ^2 is the variance of the error terms in the model. Using the residuals as approximations of the errors in our sample, we estimate σ^2 by

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n - k - 1}.$$

- Rather than n we divide 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 n minus the number of parameters in the regression equation.
  - Currently we have $k + 1$ parameters: 1 intercept and k slopes.

7 Multiple R-squared

7.1 Multiple R^2

- We can compare two models to predict the response y.
- Model 1: We do not use the predictors, and use $\bar{y}$ to predict any y-measurement. The corresponding sum of squared prediction errors 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 with predictors $x_1, \dots, x_k$ to predict y. The sum of squared prediction errors is now
  - $SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ 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 squared prediction errors, when we use $x_1, x_2, \dots, x_k$ as explanatory variables.
  - We say that R^2 is the proportion of the total variation in the data that can be explained by $x_1, \dots, x_k$.
- Properties:
  - $0 \leq R^2 \leq 1$
  - $R^2 = 0$ means $TSS = SSE$, so the model does not improve when we include $x_1, \dots, x_k$ in the model.
  - $R^2 = 1$ means $SSE = 0$, so the observations are predicted perfectly by $x_1, \dots, x_k$.

7.2 Example

```r
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 $\widehat{y} = 59 - 0.58x_1 + 0.68x_2$
- The estimate for σ is $s = 20.82$ (`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 output provides a test of the hypothesis $H_0 : \beta_1 = 0$ corresponding to no effect of the predictor x_1.
  - The estimate $b_1 = -0.5834$ has standard error (`Std. Error`) $se = 0.4725$ with corresponding observed t-value (`t value`) $t_{\text{obs}} = \frac{-0.5834}{0.4725} = -1.235$.
  - This means that b_1 isn't significantly different from zero, since the p-value (`Pr(>|t|)`) is 22%. That means that we can exclude `Education` as a predictor.

7.3 Example

- Our final model is then a simple linear regression:

```r
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 as we will see next.

7.4 F-test for comparing two models

- We can compare two models, where one is obtained from the other by setting m parameters to zero, by an F-test.
- We can compare R^2 for the two models via the following F-statistic:

$$F_{obs} = \frac{(R_2^2 - R_1^2)/\mathrm{df}_1}{(1 - R_2^2)/\mathrm{df}_2}$$

  - R_2^2 corresponds to the larger model and R_1^2 corresponds to the smaller model.
  - Large values of F_{obs} means that R_2^2 is large compared to R_1^2, pointing towards the alternative hypothesis.
  - $\mathrm{df}_1 = m$ is the number of parameters set to zero in the null-hypothesis.
  - $\mathrm{df}_2 = n - k - 1$ where n is sample size and $k + 1$ is the number of unknown parameters in the larger model.
- In R the calculations are done using `anova`:

```r
anova(model2, model)
```

```
## Analysis of Variance Table
##
## Model 1: Crime ~ Urbanisation
## Model 2: Crime ~ Education + Urbanisation
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     65 28391
## 2     64 27730  1    660.61 1.5247 0.2214
```

8 Overall F-test for effect of predictors

8.1 F-test

- We consider the hypothesis

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

  against the alternative that at least one β_i are non-zero.
  - This is the hypothesis that there is no effect of any of the predictors.
- As test statistic we use

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

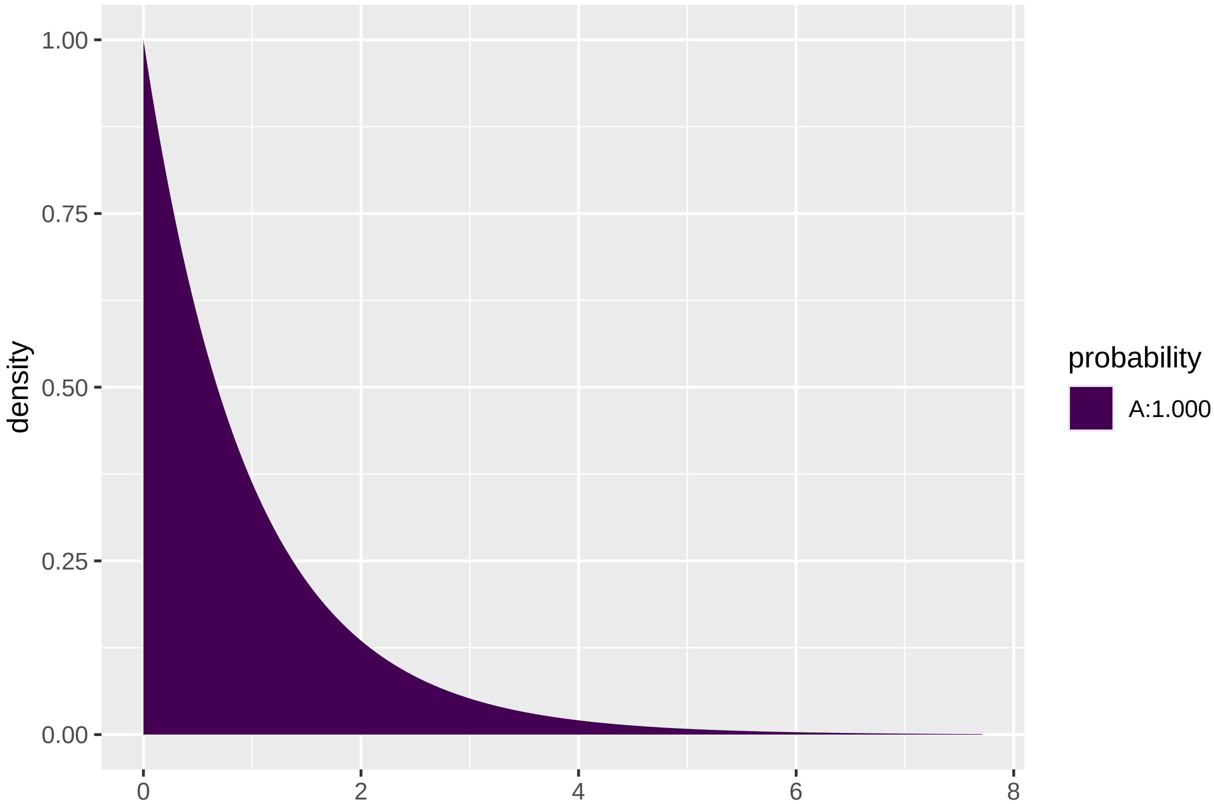
- Large values of R^2 implies large values of F_{obs}, which points to the alternative hypothesis.
- Thus, the p-value is the probability of observing something larger than the computed F_{obs}.
- The distribution of F_{obs} under the null-hypothesis is an F-distribution with degrees of freedom
  - $df_1 = k$ (the number of parameters set to zero in the null-hypothesis).

  - $df_2 = n - k - 1$ (number of observations minus number of unknown parameters in the model).

8.2 Example

- We return to `Crime` and the prediction equation $\widehat{y} = 59 - 0.58x_1 + 0.68x_2$, where $n = 67$ and $R^2 = 0.4714$.
- We test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$. We have
  - $df_1 = k = 2$ since 2 parameters are set to zero under H_0.
  - $df_2 = n - k - 1 = 67 - 2 - 1 = 64$.
  - Then we can calculate $F_{obs} = \frac{R^2/k}{(1-R^2)/(n-k-1)} = 28.54$
- 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-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
```

9 Interaction model

9.1 Interaction between effects of predictors

- Could it be possible that a combination of `Education` and `Urbanisation` is good for prediction? We investigate this using the **interaction 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 $x_1 x_2$.
- If we fix x_2 in this model, the mean response is linear in x_1 with intercept $\alpha + \beta_2 x_2$ and slope $\beta_1 + \beta_3 x_2$, since

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

  - The slope for x_1 now depends on the value of x_2!
  - Interaction means that the effect of x_1 on the response depends on the value of x_2.
  - Interaction **does not** mean that x_1 and x_2 affect each other.

9.2 Example - interaction model

- We fit the model for the `Crime` data set:

```r
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_1x_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**. It happens because the predictors are highly correlated.
- It also illustrates that we can have different models with equally good predictive properties.
  - In the case of an interaction model we always choose the model without interaction because 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?

10 Multiple linear regression with categorical predictors

10.1 Dummy variables

- Suppose we want to predict the response variable using a categorical predictor variable x with k categories.
- We choose one group, say group k, as the reference category.
- For the remaining groups $1, \dots, k-1$, we define dummy variables

$$z_i = \begin{cases} 0, & \text{if } x \neq i, \\ 1, & \text{if } x = i. \end{cases}$$

  for $i = 1, \dots, k-1$.

  - The dummy variable z_i is 1 if an observation is in group i and 0 otherwise.
  - When all dummy variables $z_i = 0$, $i = 1, \dots, k-1$, it means that the observation belongs to the reference group k.
- We can use the variables $z_1, \dots, z_{k-1}$ in a multiple regression along with other predictor variables.

10.2 Example

- Consider the dataset `mtcars`. We are interested in how engine type `vs` (categorical) and weight of the car `wt` (quantitative, x_1) are associated with fuel consumption `mpg`.
- The variable `vs` is already coded as a dummy variable z in R, taking the value 1 if the engine is v-shaped and 0 otherwise.
- The multiple regression model becomes

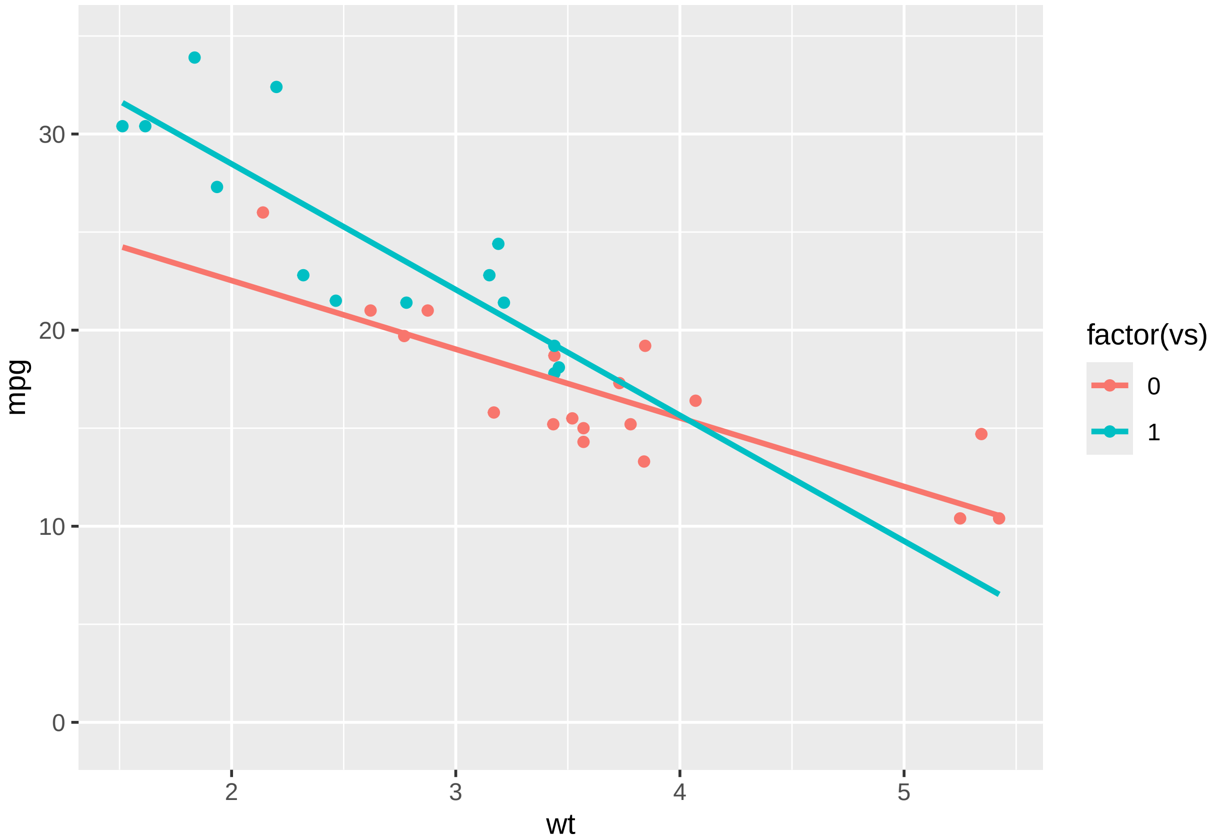
$$E(Y|x_1, z) = \alpha + \beta_1 x_1 + \beta_2 z.$$

  - For $z = 0$:
$$E(Y|x_1, z) = \alpha + \beta_1 x_1.$$
  - For $z = 1$:
$$E(Y|x_1, z) = \alpha + \beta_2 + \beta_1 x_1.$$
- So we get two different regression lines for the two groups.
  - The lines have a common slope β_1 (parallel lines).
  - The lines have different intercepts. The difference in intercepts is β_2.

10.3 Example

- 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(mpg ~ wt, color = ~factor(vs), group=~factor(vs), data = mtcars) %>% gf_lm()
```



- An unclear picture, but a tendency to decreasing number of miles per gallon with increasing weight for both engine types.
- The slope of the lines for the two engine types look different. But is the difference significant? Or can the difference be explained by sampling variation?

10.4 Example

- We fit a multiple regression model without interaction in R:

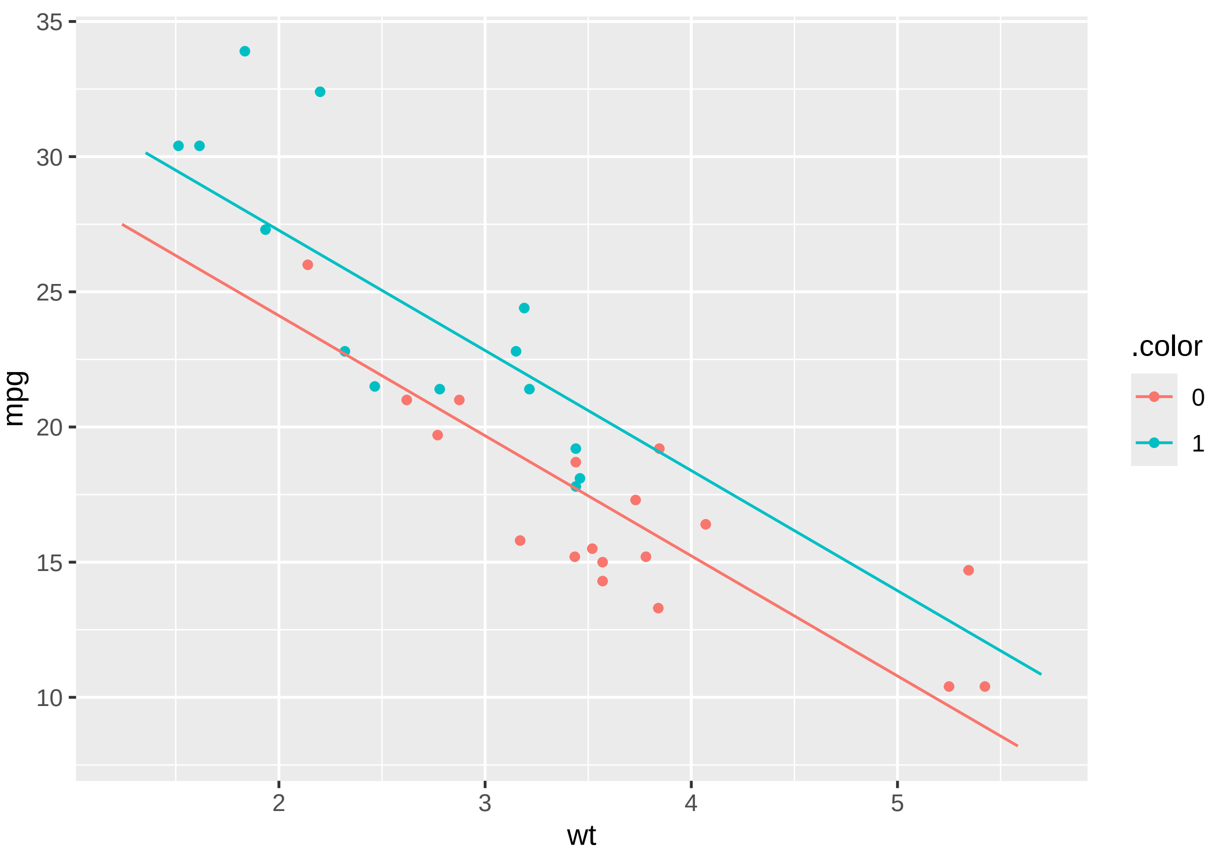
```
model1 <- lm(mpg ~ wt + factor(vs) , data = mtcars)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(vs), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7071 -2.4415 -0.3129  1.4319  6.0156
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.0042     2.3554  14.012 1.92e-14 ***
## wt           -4.4428     0.6134  -7.243 5.63e-08 ***
## factor(vs)1   3.1544     1.1907   2.649   0.0129 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.78 on 29 degrees of freedom
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.7873
## F-statistic: 58.36 on 2 and 29 DF,  p-value: 6.818e-11
```

- The common slope to `wt` is estimated to be $\widehat{\beta}_1 = -4.44$, with corresponding p-value $5.63 \cdot 10^{-8}$, so the effect of `wt` is significantly different from zero.
  - The estimate is negative, so increasing weight decreases the number of miles per gallon.
- The estimate for intercept in the reference group ("not v-shaped") is $\widehat{\alpha} = 33.0$, which is significantly different from zero if we test at level 5% (this test is not really of interest).
- The difference between intercepts for the two engine types is $\widehat{\beta}_1 = 3.15$, which is significant with p-value=1%.
  - This suggests that the regression lines are not the same for the two engine types.
  - The value 3.15 is the vertical distance between the two regression lines.

```r
plotModel(model1)
```



10.5 Example: Prediction equations

```r
summary(model1)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ wt + factor(vs), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7071 -2.4415 -0.3129  1.4319  6.0156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.0042     2.3554  14.012 1.92e-14 ***
## wt           -4.4428     0.6134  -7.243 5.63e-08 ***
## factor(vs)1   3.1544     1.1907   2.649   0.0129 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.78 on 29 degrees of freedom
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.7873
## F-statistic: 58.36 on 2 and 29 DF,  p-value: 6.818e-11
```

- Reference/baseline group (not v-shaped):

$$\widehat{y} = 33.0 - 4.44x$$

- V-shaped:

$$\widehat{y} = 33.0 + 3.15 - 4.44x = 36.15 - 4.44x.$$

10.6 Interaction model

- We can expand the regression model by including an interaction between x and z:

$$E(y|x,z) = \alpha + \beta_1 x + \beta_2 z + \beta_3 z \cdot x.$$

- This yields a regression line for engine type:
- Not v-shaped ($z = 0$): $E(y|x,z) = \alpha + \beta_1 x$
- V-shaped ($z = 1$): $E(y|x,z) = \alpha + \beta_2 + (\beta_1 + \beta_3)x$.
- β_2 is *the difference* in `Intercept` between the two groups, while β_3 is *the difference* in `slope` between the two groups.

10.7 Example: Prediction equations

- When we use `*` in the model formula we include interaction between `vs` and `wt`:

```
model2 <- lm(mpg ~ wt * factor(vs), data = mtcars)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ wt * factor(vs), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9950 -1.7881 -0.3423  1.2935  5.2061
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     29.5314     2.6221  11.263 6.55e-12 ***
## wt              -3.5013     0.6915  -5.063 2.33e-05 ***
```

```
## factor(vs)1     11.7667     3.7638   3.126   0.0041 **
## wt:factor(vs)1  -2.9097     1.2157  -2.393   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.578 on 28 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8171
## F-statistic: 47.16 on 3 and 28 DF,  p-value: 4.497e-11
```

- We use the output to write the prediction equations:
  - Reference/baseline group (not v-shaped):
$$\widehat{y} = 29.5 - 3.5x$$
  - V-shaped:
$$\widehat{y} = (29.5 + 11.8) + (-3.5 - 2.9)x = 41.3 - 6.4x.$$

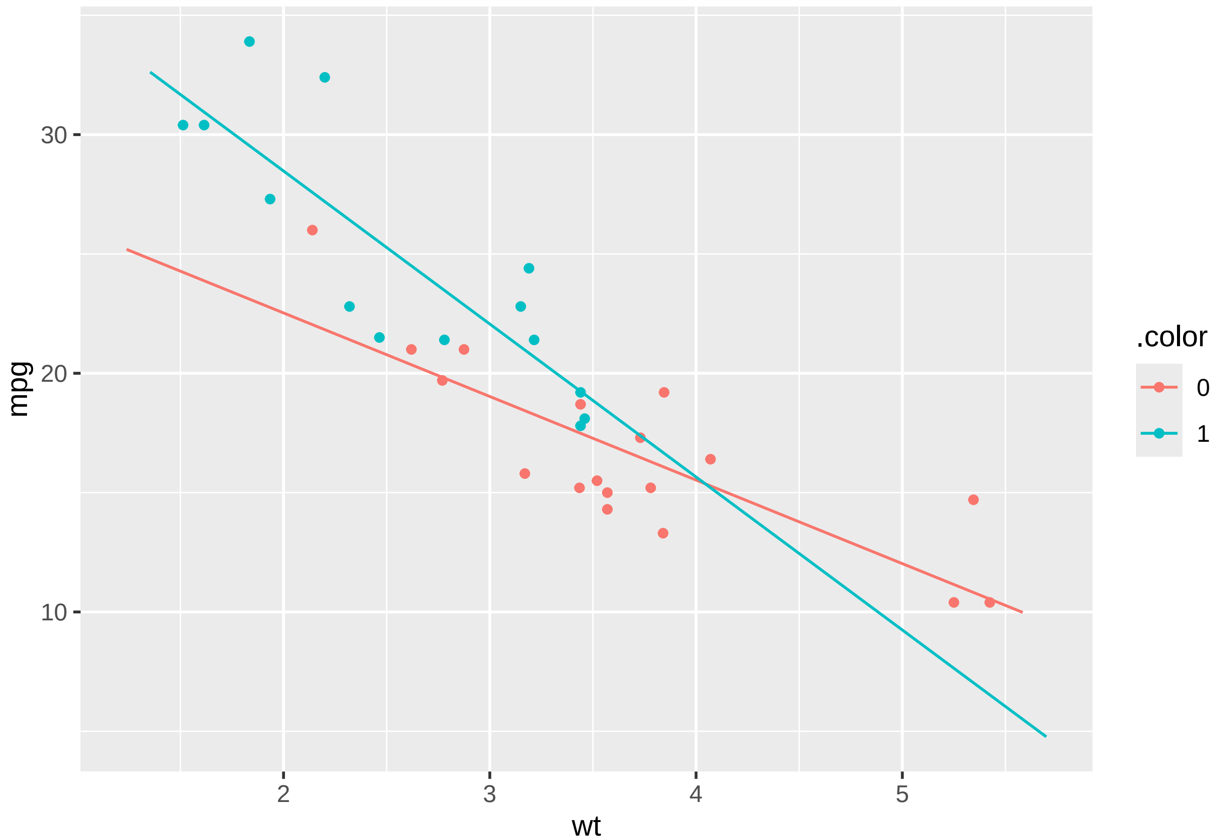
10.8 Example: Individual tests

```
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ wt * factor(vs), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9950 -1.7881 -0.3423  1.2935  5.2061
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     29.5314     2.6221  11.263 6.55e-12 ***
## wt              -3.5013     0.6915  -5.063 2.33e-05 ***
## factor(vs)1     11.7667     3.7638   3.126   0.0041 **
## wt:factor(vs)1  -2.9097     1.2157  -2.393   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.578 on 28 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8171
## F-statistic: 47.16 on 3 and 28 DF,  p-value: 4.497e-11
```

- *The difference* in slope between the two engine types is estimated to $\widehat{\beta}_3 = -2.9$ which is significant with p-value=0.0236, so the slopes are significantly different.

```
plotModel(model2)
```



10.9 Hierarchy of models

- Always test for no interaction ($\beta_3 = 0$) before making tests for main effects ($\beta_1 = 0$ or $\beta_2 = 0$).

$$Model: E(Y|x, z) = \alpha + \beta_1 x + \beta_2 z + \beta_3 xz$$

Interaction

$\downarrow$ H_0: $\beta_3 = 0$

Main effects only

H_0: $\beta_2 = 0$ $\swarrow$ $\searrow$ H_0: $\beta_1 = 0$

Only effect of x   Only effect of z

H_0: $\beta_1 = 0$ $\searrow$ $\swarrow$ H_0: $\beta_2 = 0$

No effect of x or z

10.10 F-test

- We can compare the two models in the `mtcars` example, namely the model with and without interaction via

```r
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + factor(vs)
## Model 2: mpg ~ wt * factor(vs)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 224.09
## 2     28 186.03  1    38.062 5.7287 0.02363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```