

# 02323 Introduction to Statistics

## Lecture 8: Simple linear regression

DTU Compute  
Technical University of Denmark  
2800 Lyngby – Denmark

# Overview

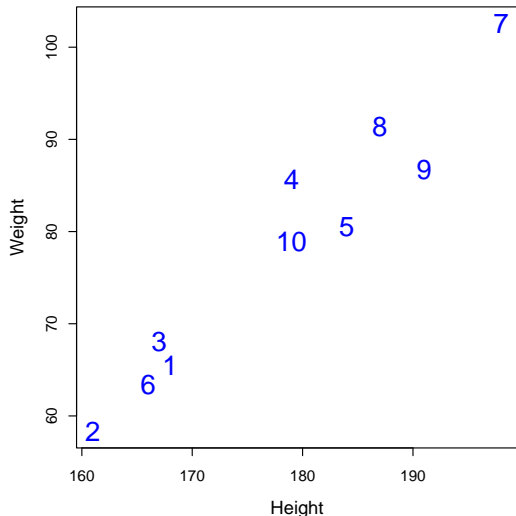
- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

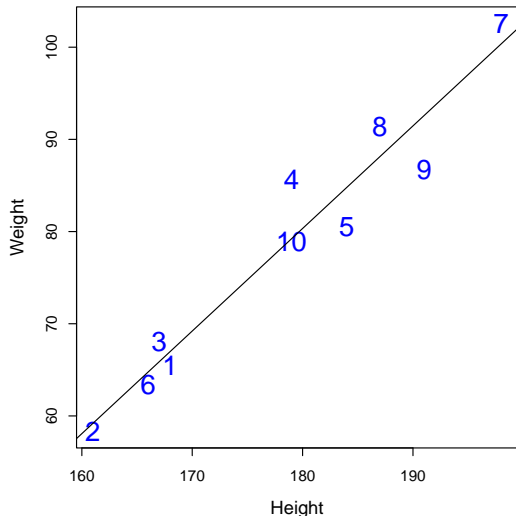
# Example: Height-Weight

Heights ( $x_i$ )	168	161	167	179	184	166	198	187	191	179
Weights ( $y_i$ )	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



# Example: Height-Weight

Heights ( $x_i$ )	168	161	167	179	184	166	198	187	191	179
Weights ( $y_i$ )	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

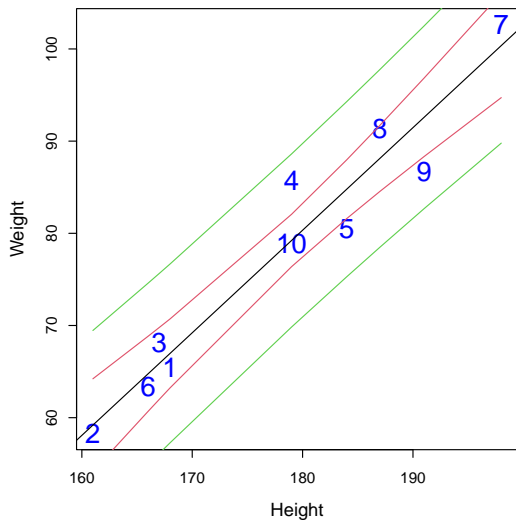


Heights ( $x_i$ )	168	161	167	179	184	166	198	187	191	179
Weights ( $y_i$ )	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
summary(lm(y ~ x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.876 -1.451 -0.608  2.234  6.477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -119.958     18.897   -6.35  0.00022 ***
## x              1.113       0.106   10.50  5.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.88 on 8 degrees of freedom
## Multiple R-squared:  0.932, Adjusted R-squared:  0.924
## F-statistic: 110 on 1 and 8 DF, p-value: 5.87e-06
```

Heights ( $x_i$ )	168	161	167	179	184	166	198	187	191	179
Weights ( $y_i$ )	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



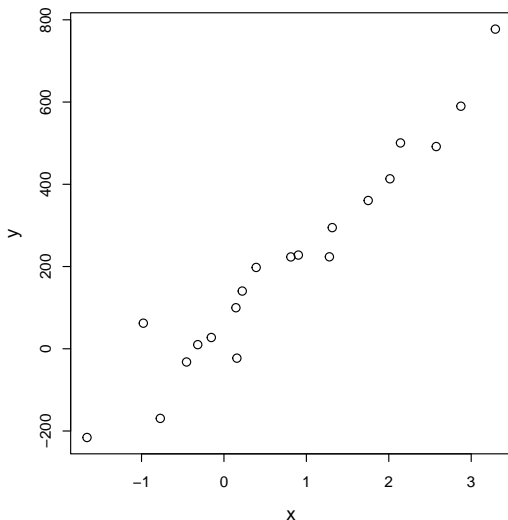
# Overview

- 1 Example: Height-Weight
- 2 Linear regression model**
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation



# A scatter plot of some data

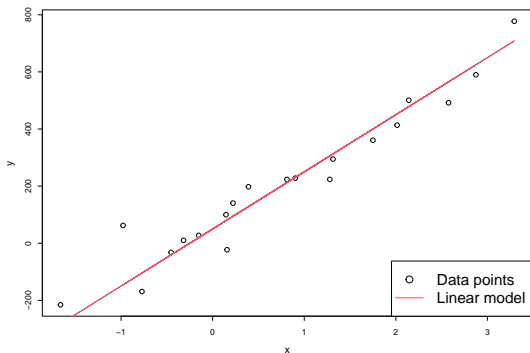
- We have  $n$  pairs of data points  $(x_i, y_i)$ .



# Express a linear model

- Express a linear model:

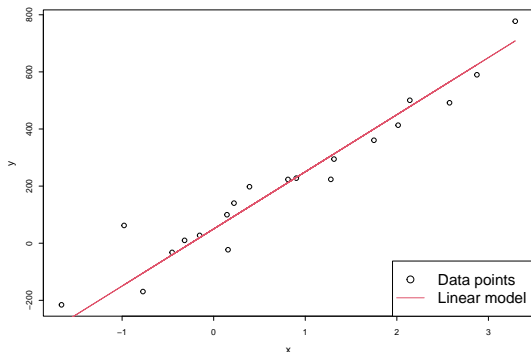
$$y_i = \beta_0 + \beta_1 x_i + ?$$



# Express a linear model

- Express a linear model:

$$y_i = \beta_0 + \beta_1 x_i + ?$$



- Something is missing: Description of the *random variation*.

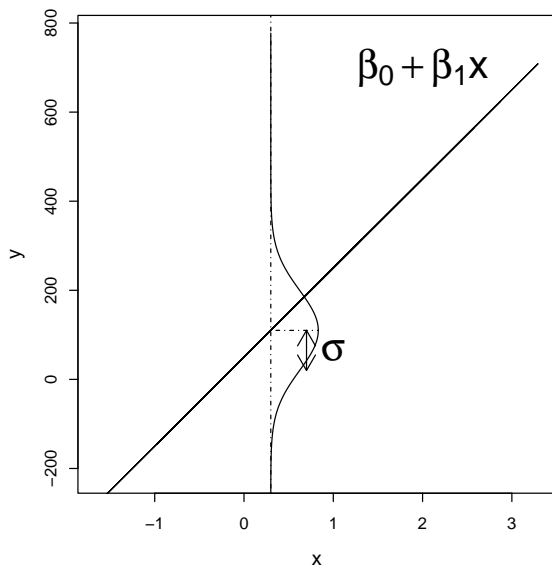
# Express a linear regression model

- Express the *linear regression model*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

- $Y_i$  is the *dependent/outcome variable*. A random variable.
- $x_i$  is an *independent/explanatory variable*. Deterministic numbers.
- $\varepsilon_i$  is the deviation/error. A random variable.
- We assume that the  $\varepsilon_i$ ,  $i = 1, \dots, n$ , are *independent and identically distributed (i.i.d.)*, with  $\varepsilon_i \sim N(0, \sigma^2)$ .

# Illustration of statistical model



# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method**
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

# Least squares method

- How can we estimate the parameters  $\beta_0$  and  $\beta_1$ ?

# Least squares method

- How can we estimate the parameters  $\beta_0$  and  $\beta_1$ ?
- Good idea: Minimize the variance  $\sigma^2$  of the residuals.



# Least squares method

- How can we estimate the parameters  $\beta_0$  and  $\beta_1$ ?
- Good idea: Minimize the variance  $\sigma^2$  of the residuals.
- But how?

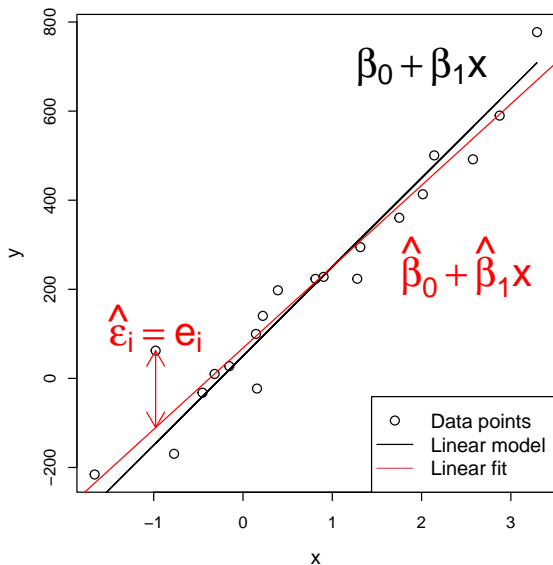
# Least squares method

- How can we estimate the parameters  $\beta_0$  and  $\beta_1$ ?
- Good idea: Minimize the variance  $\sigma^2$  of the residuals.
- But how?
- Minimize the Residual Sum of Squares (*RSS*),

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

$\hat{\beta}_0$  and  $\hat{\beta}_1$  minimize the RSS.

## Illustration of model, data and fit



## Least squares estimator

Theorem 5.4 (here as estimators, as in the book)

The least squares estimators of  $\beta_0$  and  $\beta_1$  are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

where  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ .

# Least squares estimates

Theorem 5.4 (here as *estimates*)

The least squares estimates of  $\beta_0$  and  $\beta_1$  are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ .

# R example

```

set.seed(100)

# Generate x
x <- runif(n = 20, min = -2, max = 4)

# Simulate y
beta0 <- 50; beta1 <- 200; sigma <- 90
y <- beta0 + beta1 * x + rnorm(n = length(x), mean = 0, sd = sigma)

# From here: like for the analysis of 'real data', we have data in x and y:

# Scatter plot of y against x
plot(x, y)

# Find the least squares estimates, use Theorem 5.4
(beta1hat <- sum( (y - mean(y))*(x-mean(x)) ) / sum( (x-mean(x))^2 ))
(beta0hat <- mean(y) - beta1hat*mean(x))

# Use lm() to find the estimates
lm(y ~ x)

# Plot the fitted line
abline(lm(y ~ x), col="red")

```

# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?**
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

The parameter estimates are random variables

What if we took a new sample?

Would the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  be the same?



## The parameter estimates are random variables

What if we took a new sample?

Would the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  be the same?

No, they are random variables!

If we took a new sample, we would get another realisation.

## The parameter estimates are random variables

What if we took a new sample?

Would the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  be the same?

No, they are random variables!

If we took a new sample, we would get another realisation.

What are the (sampling) distributions of the parameter estimates ...

... in a linear regression model w. normal distributed errors?

## The parameter estimates are random variables

What if we took a new sample?

Would the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  be the same?

No, they are random variables!

If we took a new sample, we would get another realisation.

What are the (sampling) distributions of the parameter estimates ...

... in a linear regression model w. normal distributed errors?

This may be investigated using simulation ...

Let's go to R!

## The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

- $\hat{\beta}_0$  and  $\hat{\beta}_1$  are normal distributed and their variance can be estimated:

### Theorem 5.8 (first part)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$$

$$\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x} \sigma^2}{S_{xx}}$$

- We won't use the covariance  $\text{Cov}[\hat{\beta}_0, \hat{\beta}_1]$  for now.

# Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$

## Theorem 5.8 (second part)

$\sigma^2$  is usually replaced by its estimate,  $\hat{\sigma}^2$ , the *central estimator of  $\sigma^2$* :

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of  $\sigma^2$  is used, the variances also become estimates. We'll refer to them as  $\hat{\sigma}_{\beta_0}^2$  and  $\hat{\sigma}_{\beta_1}^2$ .

## Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$

### Theorem 5.8 (second part)

$\sigma^2$  is usually replaced by its estimate,  $\hat{\sigma}^2$ , the *central estimator of  $\sigma^2$* :

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of  $\sigma^2$  is used, the variances also become estimates. We'll refer to them as  $\hat{\sigma}_{\beta_0}^2$  and  $\hat{\sigma}_{\beta_1}^2$ .

Estimates of standard deviations of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  (equations 5-43 and 5-44):

$$\hat{\sigma}_{\beta_0} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}; \quad \hat{\sigma}_{\beta_1} = \hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$**
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

## Hypothesis tests for $\beta_0$ and $\beta_1$

We can carry out hypothesis tests for the parameters in a linear regression model:

$$H_{0,i} : \beta_i = \beta_{0,i}$$

$$H_{1,i} : \beta_i \neq \beta_{1,i}$$



## Hypothesis tests for $\beta_0$ and $\beta_1$

We can carry out hypothesis tests for the parameters in a linear regression model:

$$H_{0,i} : \beta_i = \beta_{0,i}$$

$$H_{1,i} : \beta_i \neq \beta_{1,i}$$

### Theorem 5.12

Under the null-hypotheses ( $\beta_0 = \beta_{0,0}$  and  $\beta_1 = \beta_{0,1}$ ) the statistics

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}; \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

are  $t$ -distributed with  $n - 2$  degrees of freedom, and inference should be based on this distribution.

## Hypothesis tests for $\beta_0$ and $\beta_1$

- See Example 5.13 for an example of a hypothesis test.
- Test if the parameters are significantly different from 0:

$$H_{0,i} : \beta_i = 0, \quad H_{1,i} : \beta_i \neq 0$$

```
# Read data into R
```

```
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
```

```
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)
```

```
# Fit model to data
```

```
fit <- lm(y ~ x)
```

```
# Look at model summary to find Tobs-values and p-values
```

```
summary(fit)
```

## Confidence intervals for $\beta_0$ and $\beta_1$

### Method 5.15

$(1 - \alpha)$  confidence intervals for  $\beta_0$  and  $\beta_1$  are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_0}$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_1}$$

where  $t_{1-\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of a  $t$ -distribution with  $n - 2$  degrees of freedom.

- Remember that  $\hat{\sigma}_{\beta_0}$  and  $\hat{\sigma}_{\beta_1}$  may be found using equations 5-43 and 5-44.
- In R, we can find  $\hat{\sigma}_{\beta_0}$  and  $\hat{\sigma}_{\beta_1}$  under "Std. Error" from `summary(fit)`.

# Illustration of CIs by simulation

```

# Number of repetitions (here: CIs)
nRepeat <- 1000

# Empty logical vector of length nRepeat
TrueValInCI <- logical(nRepeat)

# Repeat the simulation and estimation nRepeat times:
for(i in 1:nRepeat){
  # Generate x
  x <- runif(n = 20, min = -2, max = 4)
  # Simulate y
  beta0 = 50; beta1 = 200; sigma = 90
  y <- beta0 + beta1 * x + rnorm(n = length(x), mean = 0, sd = sigma)
  # Use lm() to fit model
  fit <- lm(y ~ x)
  # Use confint() to compute 95% CI for intercept
  ci <- confint(fit, "(Intercept)", level=0.95)
  # Was the 'true' intercept included in the interval? (covered)
  (TrueValInCI[i] <- ci[1] < beta0 & beta0 < ci[2])
}

# How often was the true intercept included in the CI?
sum(TrueValInCI) / nRepeat

```

# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line**
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

## Method 5.18 Confidence interval for $\beta_0 + \beta_1 x_0$

- The confidence interval for  $\beta_0 + \beta_1 x_0$  corresponds to a confidence interval for the line at the point  $x_0$ .
- The  $100(1 - \alpha)\%$  CI is computed by

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

## Method 5.18 Prediction interval for $\beta_0 + \beta_1 x_0 + \varepsilon_0$

- The prediction interval for  $Y_0$  is found using a value  $x_0$ .
- This is done *before*  $Y_0$  is observed, using

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

- In  $100(1 - \alpha)\%$  of cases, the prediction interval will contain the observed  $y_0$ .
- For a given  $\alpha$ , a prediction interval is wider than a confidence interval.

# Example of confidence intervals for the line

```

# Generate x
x <- runif(n = 20, min = -2, max = 4)

# Simulate y
beta0 = 50; beta1 = 200; sigma = 90
y <- beta0 + beta1 * x + rnorm(n = length(x), sd = sigma)

# Use lm() to fit model
fit <- lm(y ~ x)

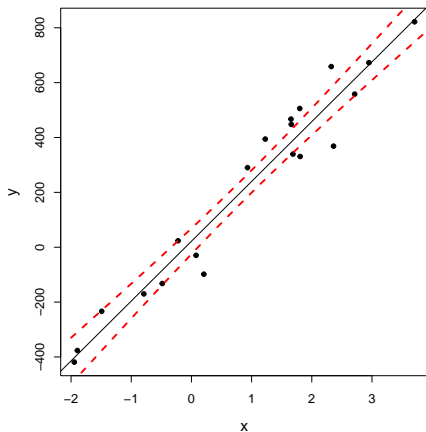
# Make a sequence of 100 x-values
xval <- seq(from = -2, to = 6, length.out = 100)

# Use the predict function
CI <- predict(fit, newdata = data.frame(x = xval),
              interval = "confidence",
              level = 0.95)

# Check what we got
head(CI)

# Plot the data, model fit and intervals
plot(x, y, pch = 20)
abline(fit)
lines(xval, CI[, "lwr"], lty=2, col = "red", lwd = 2)
lines(xval, CI[, "upr"], lty=2, col = "red", lwd = 2)

```





# Example of prediction intervals for the line

```

# Generate x
x <- runif(n = 20, min = -2, max = 4)

# Simulate y
beta0 = 50; beta1 = 200; sigma = 90
y <- beta0 + beta1 * x + rnorm(n = length(x), sd = sigma)

# Use lm() to fit model
fit <- lm(y ~ x)

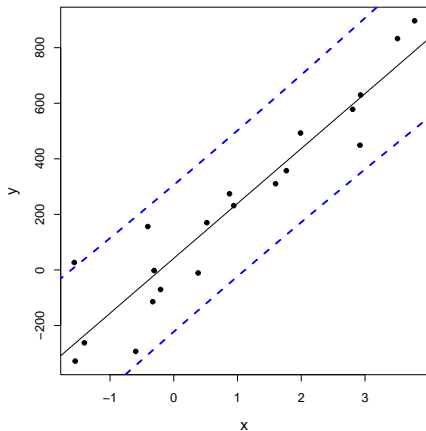
# Make a sequence of 100 x-values
xval <- seq(from = -2, to = 6, length.out = 100)

# Use the predict function
PI <- predict(fit, newdata = data.frame(x = xval),
             interval = "prediction",
             level = 0.95)

# Check what we got
head(PI)

# Plot the data, model fit and intervals
plot(x, y, pch = 20)
abline(fit)
lines(xval, PI[, "lwr"], lty = 2, col = "blue", lwd = 2)
lines(xval, PI[, "upr"], lty = 2, col = "blue", lwd = 2)

```



# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'**
- 8 Correlation
- 9 Residual Analysis: Model validation

# What more do we get from summary()?

```
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -216.86  -66.09   -7.16   58.48  293.37
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      41.8         30.9    1.35    0.19
## x                197.6         16.4   12.05  4.7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122 on 18 degrees of freedom
## Multiple R-squared:  0.89, Adjusted R-squared:  0.884
## F-statistic: 145 on 1 and 18 DF, p-value: 4.73e-10
```

# summary(lm(y~x))

• Residuals:            Min            1Q            Median            3Q            Max

# summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

# summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
     The residuals': minimum, 1st quartile, median, 3rd quartile, maximum
- Coefficients:  
     Estimate Std. Error t value Pr(>|t|) "stars"

## summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:  
Estimate Std. Error t value Pr(>|t|) "stars"

The coefficients':

$\hat{\beta}_i$             $\hat{\sigma}_{\beta_i}$             $t_{\text{obs}}$             $p$ -value

- The test is  $H_{0,i} : \beta_i = 0$  vs.  $H_{1,i} : \beta_i \neq 0$
- The stars indicate which size category the  $p$ -value belongs to.

## summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:  
                  Estimate Std. Error t value Pr(>|t|) "stars"

The coefficients':

$\hat{\beta}_i$             $\hat{\sigma}_{\beta_i}$             $t_{\text{obs}}$             $p\text{-value}$

- The test is  $H_{0,i} : \beta_i = 0$  vs.  $H_{1,i} : \beta_i \neq 0$
  - The stars indicate which size category the  $p$ -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom



## summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:  
Estimate Std. Error t value Pr(>|t|) "stars"

The coefficients':

$\hat{\beta}_i$             $\hat{\sigma}_{\beta_i}$             $t_{\text{obs}}$             $p$ -value

- The test is  $H_{0,i} : \beta_i = 0$  vs.  $H_{1,i} : \beta_i \neq 0$
  - The stars indicate which size category the  $p$ -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom  
 $\varepsilon_i \sim N(0, \sigma^2)$ , the output shows  $\hat{\sigma}$  and  $\nu$  degrees of freedom (used for hypothesis tests, CIs, PIs etc.)

## summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:  
                  Estimate Std. Error t value Pr(>|t|) "stars"

The coefficients':

$\hat{\beta}_i$             $\hat{\sigma}_{\beta_i}$             $t_{\text{obs}}$             $p\text{-value}$

- The test is  $H_{0,i} : \beta_i = 0$  vs.  $H_{1,i} : \beta_i \neq 0$
  - The stars indicate which size category the  $p$ -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom  
 $\varepsilon_i \sim N(0, \sigma^2)$ , the output shows  $\hat{\sigma}$  and  $\nu$  degrees of freedom (used for hypothesis tests, CIs, PIs etc.)
- Multiple R-squared:   XXX

## summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:  
Estimate Std. Error t value Pr(>|t|) "stars"

The coefficients':

$\hat{\beta}_i$             $\hat{\sigma}_{\beta_i}$             $t_{\text{obs}}$             $p\text{-value}$

- The test is  $H_{0,i} : \beta_i = 0$  vs.  $H_{1,i} : \beta_i \neq 0$
  - The stars indicate which size category the  $p$ -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom  
 $\varepsilon_i \sim N(0, \sigma^2)$ , the output shows  $\hat{\sigma}$  and  $\nu$  degrees of freedom (used for hypothesis tests, CIs, PIs etc.)
- Multiple R-squared:   XXX  
Explained variation  $r^2$ .

## summary(lm(y~x))

- Residuals:           Min           1Q    Median           3Q           Max  
The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:  
Estimate Std. Error t value Pr(>|t|) "stars"

The coefficients':

$\hat{\beta}_i$             $\hat{\sigma}_{\beta_i}$             $t_{\text{obs}}$             $p\text{-value}$

- The test is  $H_{0,i} : \beta_i = 0$  vs.  $H_{1,i} : \beta_i \neq 0$
  - The stars indicate which size category the  $p$ -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom  
 $\varepsilon_i \sim N(0, \sigma^2)$ , the output shows  $\hat{\sigma}$  and  $\nu$  degrees of freedom (used for hypothesis tests, CIs, PIs etc.)
- Multiple R-squared:   XXX  
Explained variation  $r^2$ .
- The rest we don't use in this course.

# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation**
- 9 Residual Analysis: Model validation

# Explained variation and correlation

- Explained variation in a model is  $r^2$ , in summary "Multiple R-squared".
- Found as

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ .

- The proportion of the total variability explained by the model.

# Explained variation and correlation

- The correlation  $\rho$  is a measure of *linear relation* between two random variables.
- Estimated (i.e. empirical) correlation satisfies that

$$\hat{\rho} = r = \sqrt{r^2} \operatorname{sgn}(\hat{\beta}_1)$$

where  $\operatorname{sgn}(\hat{\beta}_1)$  is:  $-1$  for  $\hat{\beta}_1 \leq 0$  and  $1$  for  $\hat{\beta}_1 > 0$

# Explained variation and correlation

- The correlation  $\rho$  is a measure of *linear relation* between two random variables.
- Estimated (i.e. empirical) correlation satisfies that

$$\hat{\rho} = r = \sqrt{r^2} \operatorname{sgn}(\hat{\beta}_1)$$

where  $\operatorname{sgn}(\hat{\beta}_1)$  is:  $-1$  for  $\hat{\beta}_1 \leq 0$  and  $1$  for  $\hat{\beta}_1 > 0$

- Hence:
  - Positive correlation when positive slope.
  - Negative correlation when negative slope.



# Test for significance of correlation

- Test for significance of correlation (linear relation) between two variables

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

is equivalent to

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

where  $\hat{\beta}_1$  is the estimated slope in a simple linear regression model

## Example: Correlation and $R^2$ for height-weight data

```
# Read data into R

x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Fit model to data
fit <- lm(y ~ x)

# Scatter plot of data with fitted line
plot(x,y, xlab = "Height", ylab = "Weight")
abline(fit, col="red")

# See summary
summary(fit)

# Correlation between x and y
cor(x,y)

# Squared correlation is the "Multiple R-squared" from summary(fit)
cor(x,y)^2
```

# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 **Residual Analysis: Model validation**

# Residual Analysis

## Method 5.28

- Check normality assumptions with a qq-plot.
- Check (non-)systematic behavior by plotting the residuals,  $e_i$ , as a function of the fitted values  $\hat{y}_i$ .

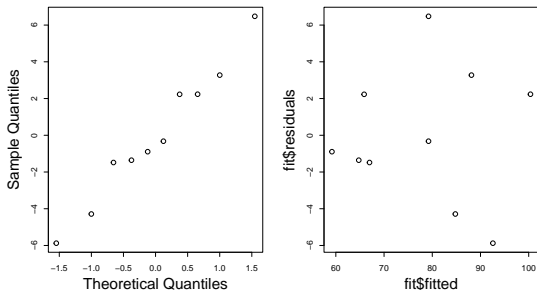
## (Method 5.29)

- Is the independence assumption reasonable?

# Residual analysis in R

```
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)
fit <- lm(y ~ x)
```

```
par(mfrow = c(1, 2))
qqnorm(fit$residuals, main = "", cex.lab = 1.5)
plot(fit$fitted, fit$residuals, cex.lab = 1.5)
```



# Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for  $\beta_0$  and  $\beta_1$
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation