

# Linear regression and correlation

Ariel Alonso Abad

Catholic University of Leuven

## Association and correlation, their scientific relevance

- Discovering associations is fundamental in science
- Many scientific hypotheses are stated in terms of correlation or lack of correlation
- Although correlation does not imply causation, causation does imply correlation. That is, although a correlational study cannot definitely prove a causal hypothesis, it may rule one out
- Some variables simply cannot be manipulated for ethical reasons. Other variables, such as birth order, sex, and age are inherently correlational because they cannot be manipulated and, therefore, the scientific knowledge concerning them must be based on correlation evidence

## Association and correlation, their scientific relevance

- Once correlation is known it can be used to make predictions
- When we know a score on one measure we can make a more accurate prediction of another measure that is highly related to it. The stronger the relationship between/among variables the more accurate the prediction
- Practical evidence from correlation studies can lead to testing that evidence under controlled experimental conditions
- Complex correlational statistics like multiple regression and partial correlation allow the correlation between two variables to be recalculated after the influence of other variables is removed

## Kalama study

### Kalama study

As part of an investigation into the physical development of children a health scientist measured the age (in months) and the height (in cm) of 12 children in the Kalama province in Egypt.

**Research question:** Is there a relationship between length and age?

**Data**

age	18	19	20	21	22	23	24	25	26	27	28	29
height	76.1	77.0	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5

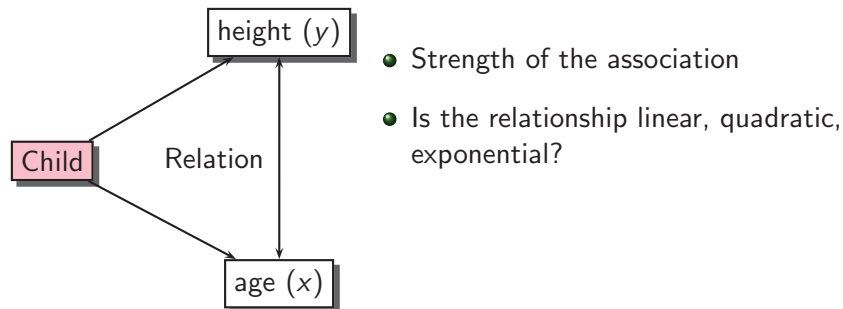
## Kalama study

### Kalama study

As part of an investigation into the physical development of children a health scientist measured the age (months) and the height (cm) of 12 children in the Kalama province in Egypt.

**Research question:** Is there a relationship between length and age?

Two variables measured for every child in the sample



## Pearson correlation coefficient

Correlation coefficient

$$r_{xy}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}}$$

- Range:  $-1 \leq r_{xy} \leq 1$
- Perfect positive correlation between  $x$  and  $y$ :  $r_{xy} = 1$
- No correlation between  $x$  and  $y$ :  $r_{xy} = 0$
- Perfect negative correlation between  $x$  and  $y$ :  $r_{xy} = -1$

## Reading the data in R

```
> # Defining working directory
> setwd("C:\\R-code-data")
>
> ## Reading the data
>
> kalama=read.table("kalama.txt", header=T)
> kalama
>
  age height
1  18   76.1
2  19   77.0
3  20   78.1
4  21   78.2
5  22   78.8
6  23   79.7
7  24   79.9
8  25   81.1
9  26   81.2
10 27   81.8
11 28   82.8
12 29   83.5
>
```

## Descriptive statistics in R

```
> ## Descriptive Statistics
>
> options(digits=2)
> descrip.kalama<-stat.desc(kalama[,c("age","height")],basic=TRUE, desc=TRUE)
> descrip.kalama
>
      age  height
nbr.val  12.00 12.000
min      18.00 76.100
max      29.00 83.500
range    11.00  7.400
sum      282.00 958.200
median   23.50 79.800
mean     23.50 79.850
SE.mean   1.04  0.665
CI.mean.0.95 2.29  1.463
var       13.00  5.301
std.dev   3.61  2.302
coef.var   0.15  0.029
>
```

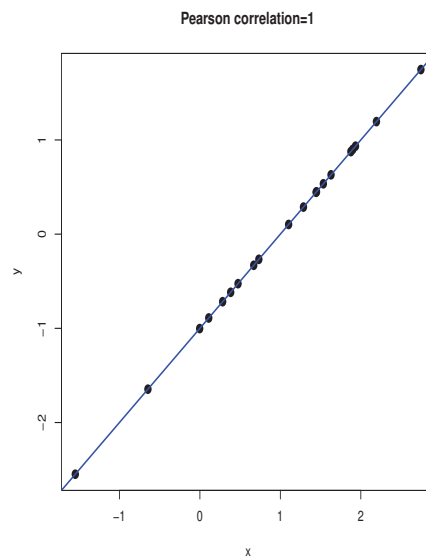
## Estimating correlations in R

```
> ## Calculating the covariance and correlatio
> cov.age.height<-cov(kalama$age,kalama$height)
> corr.age.height<-cor(kalama$age,kalama$height)
> cov.age.height
[1] 8.3
> corr.age.height
[1] 0.99
> ## Testing if the population correlation is zero
> corr.age.height.test= cor.test(kalama$age, kalama$height,
+                               alternative="two.sided", method = "pearson")
> corr.age.height.test
```

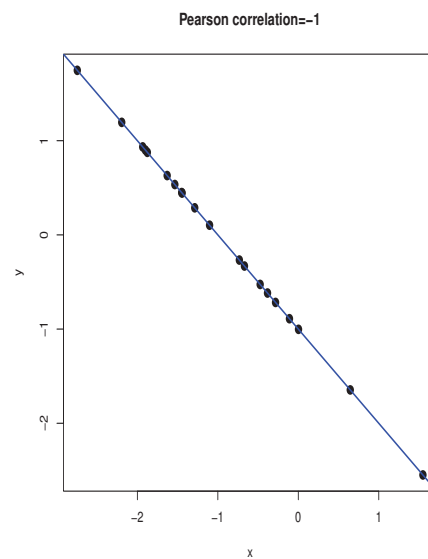
Pearson's product-moment correlation

```
data: kalama$age and kalama$height
t = 30, df = 10, p-value = 4.428e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.98 1.00
sample estimates:
 cor
0.99
>
```

## Pearson correlation coefficient



## Pearson correlation coefficient

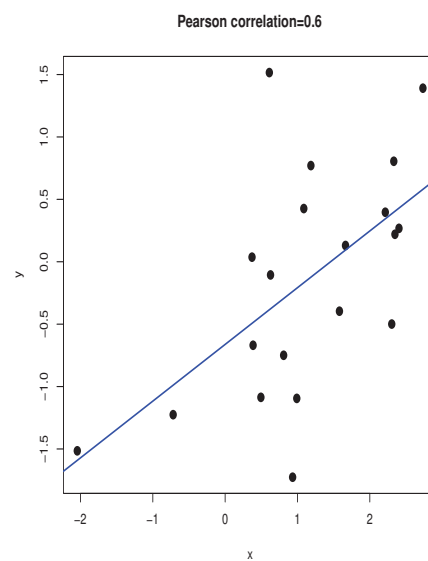


Alonso, A.

Linear regression

10 / 77

## Pearson correlation coefficient

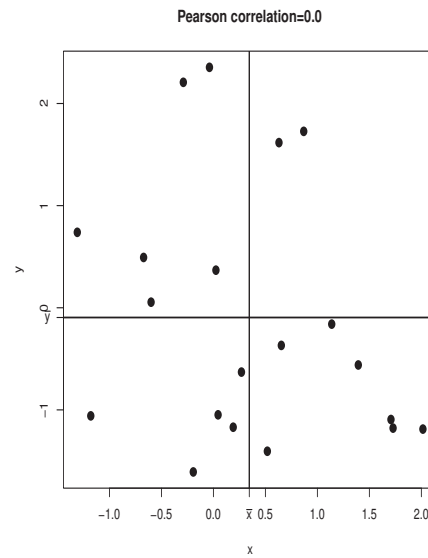


Alonso, A.

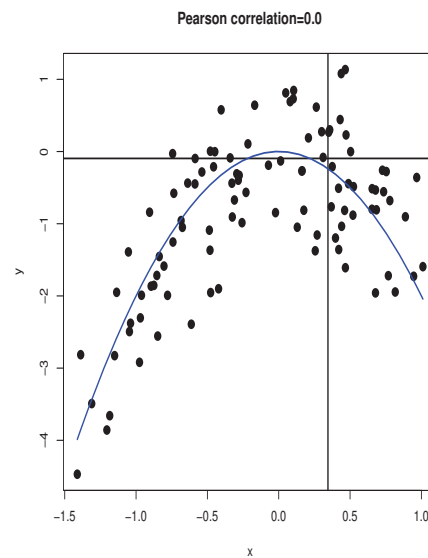
Linear regression

10 / 77

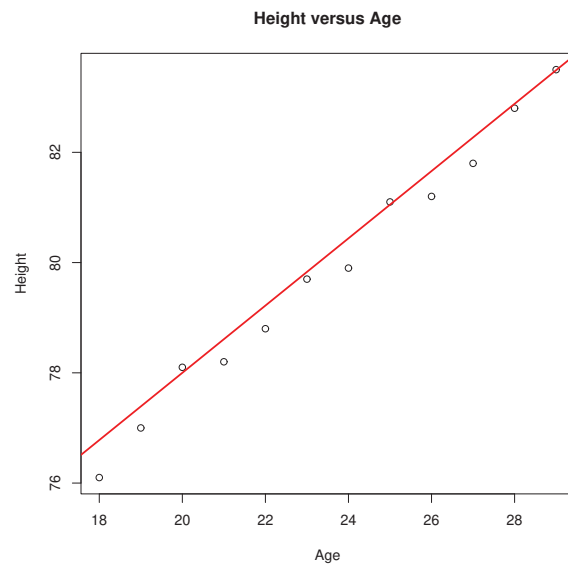
## Pearson correlation coefficient



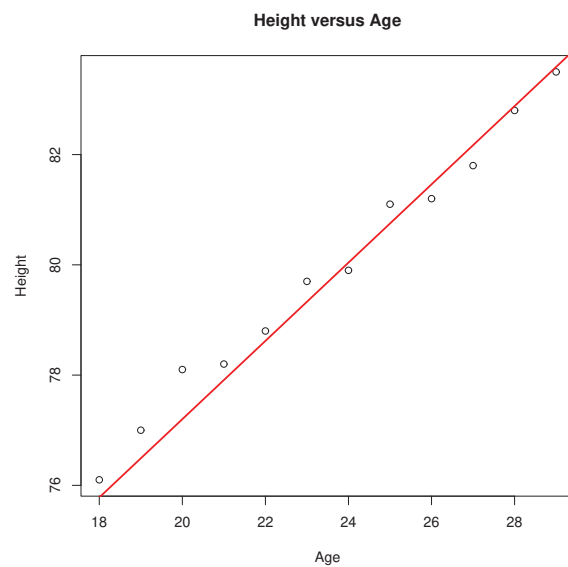
## Pearson correlation coefficient



## Kalama study ( $r_K = 0.994$ ): Best line

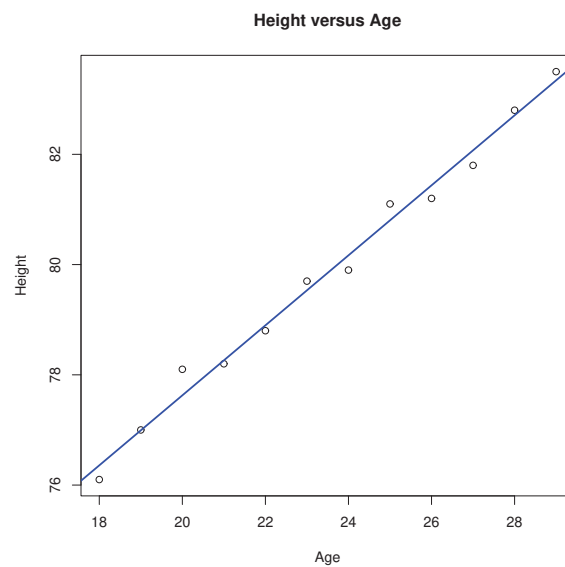


## Kalama study ( $r_K = 0.994$ ): Best line





## Kalama study ( $r_K = 0.994$ ): Best line



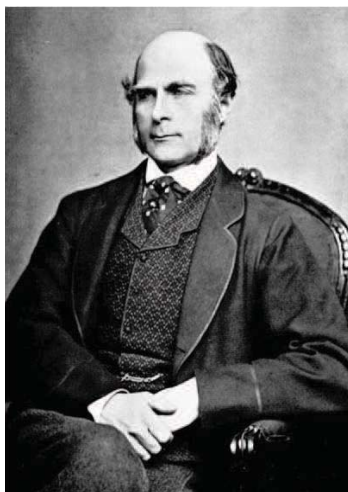
## When to use Regression Analysis

- Regression analysis is used for explaining or modeling the relationship between a single variable  $Y$ , called the response output or dependent variable, and one or more predictor or explanatory variables,  $\mathbf{X}' = (X_1, \dots, X_p)$
- When  $p = 1$  it is called **simple** regression but when  $p > 1$  it is called **multiple** regression
- When there is more than one  $Y$ , then it is called multivariate multiple regression which we won't be covering here
- The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical

## Regression Analysis: Possible objectives

- Prediction of future observations
- Assessment of the effect of, or relationship between, explanatory variables on the response
- A general description of data structure
- Extensions exist to handle multivariate responses, binary responses (logistic regression analysis) and count responses (Poisson regression)

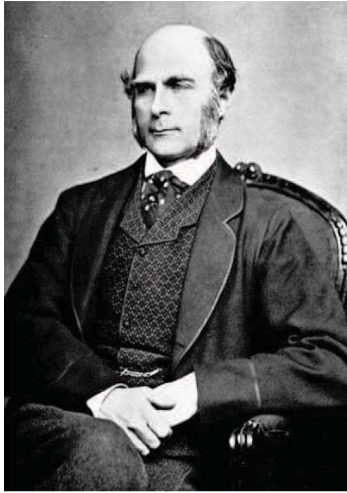
## Francis Galton



- Cousin of Charles Darwin
- Regression and correlation
- The phenomenon of regression towards the mean

"Regression towards mediocrity in hereditary stature". Journal of the Anthropological Institute 15 (1886), 246-263.

## Francis Galton

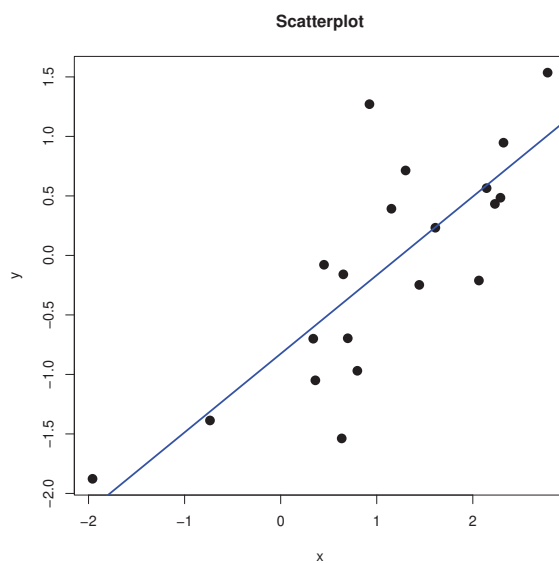


Francis Galton coined the term regression to mediocrity in 1875 in reference to the simple regression equation in the form

$$\frac{y - \bar{y}}{s_y} = r \left( \frac{x - \bar{x}}{s_x} \right).$$

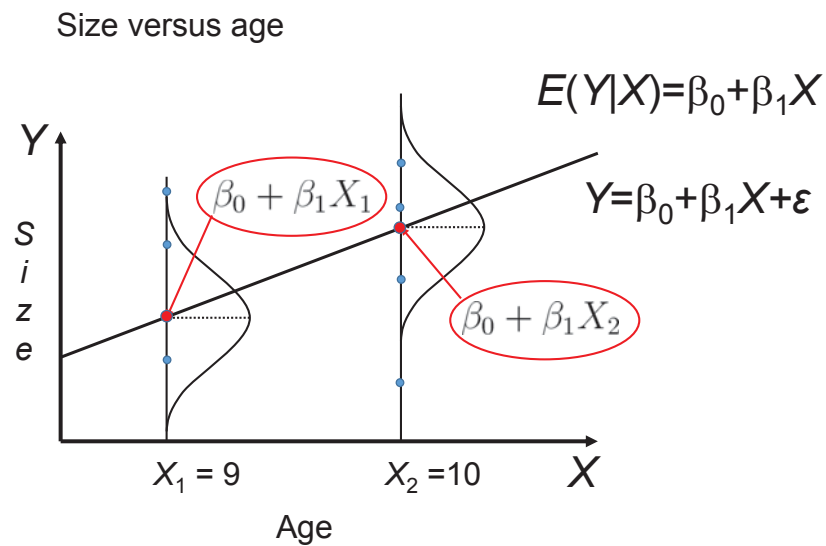
Sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers. The **regression** effect.

## Linear regression



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

## Linear regression



## Formal Statement of the Model

For each unit  $i = 1, \dots, n$ , the value of explanatory variable  $X_i$  and the response  $Y_i$  are recorded. *Simple Linear Regression* model.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

### Assumptions

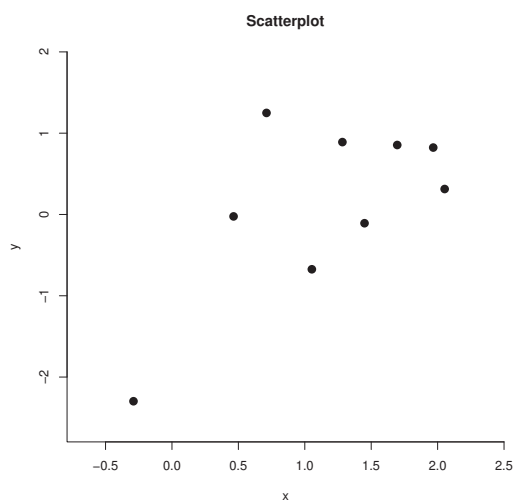
- 1 The value of  $X_i$  is precisely known.
- 2  $Y_i$  is a continuous random variable.
- 3  $\beta_0$  and  $\beta_1$  are parameters. That is, they are: unknown, constant and do not depend on the research unit.
- 4  $\varepsilon_i$  is a random error term. It is not observable.

## Formal Statement of the Model

### Additional assumptions

- 5 For two different units,  $i$  and  $j$ ,  $\varepsilon_i$  and  $\varepsilon_j$  are independent.
- 6  $X_i$  and  $\varepsilon_i$  are independent.
- 7  $\varepsilon_i \sim N(0, \sigma^2)$  for all  $i$ , i.e.,  $\varepsilon_i$  is normally distributed with  $E(\varepsilon_i) = 0$ , and  $Var(\varepsilon_i) = \sigma^2$  for all  $i$

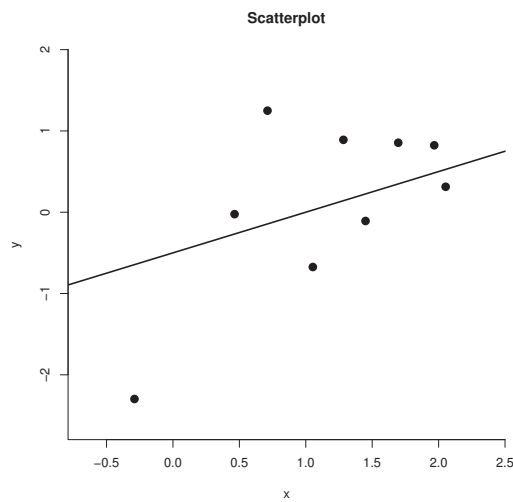
## Least squares method



Which line fits the data best?

•  $Y = \beta_0 + \beta_1 X + \varepsilon$

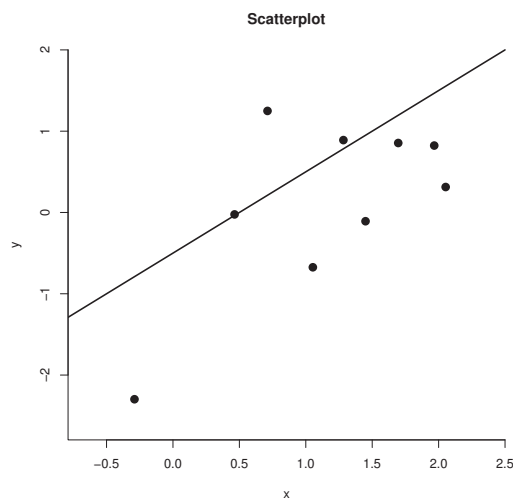
## Least squares method



Which line fits the data best?

•  $\hat{Y} = -0.5 + 0.5X$

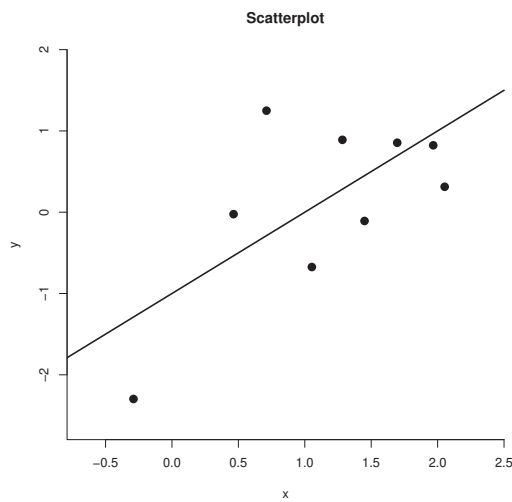
## Least squares method



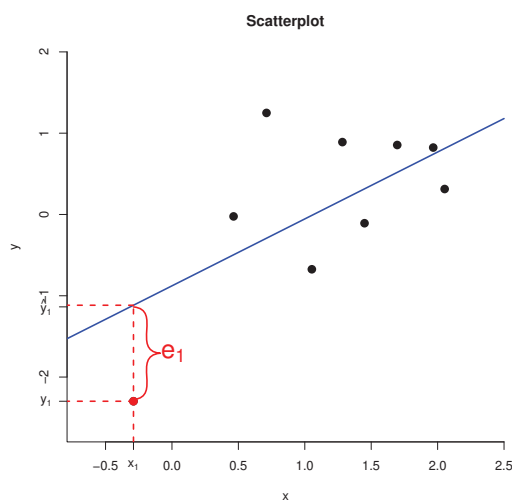
Which line fits the data best?

•  $\hat{Y} = -0.5 + X$

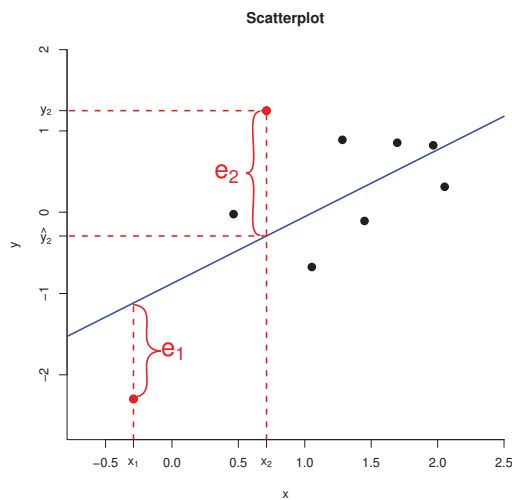
## Least squares method



## Least squares method



## Least squares method



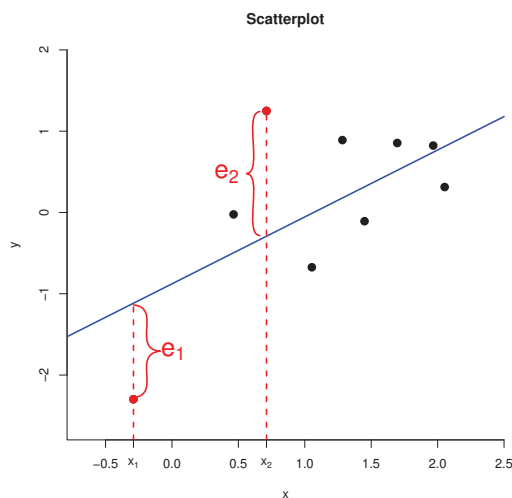
Which line fits the data best?

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$$

$$y_2 = \hat{y}_2 + e_2$$

$$e_2 = y_2 - \hat{y}_2$$

## Least squares method



Find the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that minimize

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_i e_i^2,$$

where

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$



## Estimation of the Regression Parameters

One needs to minimize

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Taking partial derivatives of  $Q(\hat{\beta}_0, \hat{\beta}_1)$  w.r.t.  $\hat{\beta}_0$  and  $\hat{\beta}_1$  and setting the resulting expressions equal to zero leads to the so-called *Normal Equations*

$$\begin{aligned}\sum Y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \\ \sum X_i Y_i &= \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2\end{aligned}$$

## Estimation of the Regression Parameters

A little algebra yields the ordinary least squares estimators for the parameters (OLS)

$$\hat{\beta}_1 = b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{SS_{XY}}{SS_{XX}}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

where

$$SS_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) \text{ and } SS_{XX} = \sum (X_i - \bar{X})^2$$

## Estimation of the Regression Parameters

A little algebra yields the ordinary least squares estimators for the parameters (OLS)

### Estimated model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{\beta}_1 = b_1 = r_{xy} \frac{S_y}{S_x}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

where

$$S_x^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 \text{ and } S_y^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

## Fitted Regression Line

With  $\hat{\beta}_0$  and  $\hat{\beta}_1$  one can compute the fitted model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

The estimated model should be *close* to the true linear regression model. Thus, one can think of  $\hat{Y}_i$  as the estimated mean response at  $X = X_i$

What about  $\sigma^2$ ?

## Estimation of $\sigma^2$

The minimum value of  $Q(\hat{\beta}_0, \hat{\beta}_1)$  is denoted as  $SSE$

- It is the sum of squares deviations between the observations and the fitted line.
- It is a measure of how well the fitted line fits the data.

$$\begin{aligned} SSE = Q(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2 \end{aligned}$$

where  $e_i$  is called the *residual* for observation  $i$ .

## Estimation of $\sigma^2$

### Note

- 1 The residual  $e_i = Y_i - \hat{Y}_i$  is the difference between observed and predicted values at  $X_i$ .
- 2 We can think of  $e_i$  as an estimator of the error  $\varepsilon_i$ .
- 3 The residuals are a fundamental tool to check the adequacy of the model.

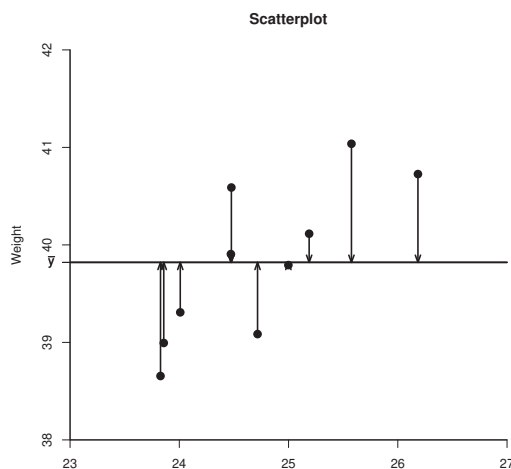
## Estimation of $\sigma^2$

- Recall that  $\sigma^2$  is the common variance for  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ .
- Because  $e_1, e_2, \dots, e_n$  estimate the  $\varepsilon$ 's,  $SSE$  should provide some information about the true variance  $\sigma^2$ .
- In fact,

$$s^2 = MSE = \frac{SSE}{n - 2}$$

is an *unbiased* estimator of  $\sigma^2$ .

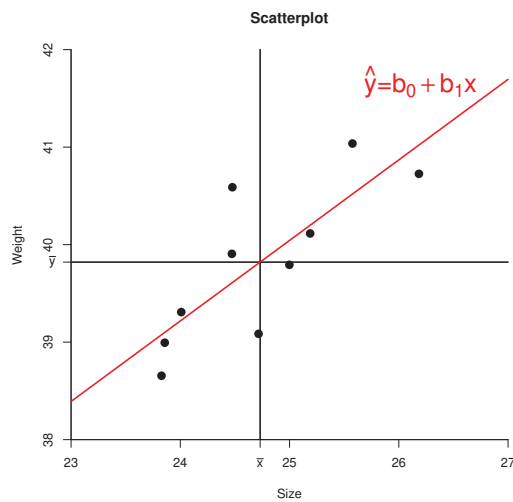
## Sources of variation



### Variation in $Y$

$$SSTO = \sum_i (y_i - \bar{y})^2$$

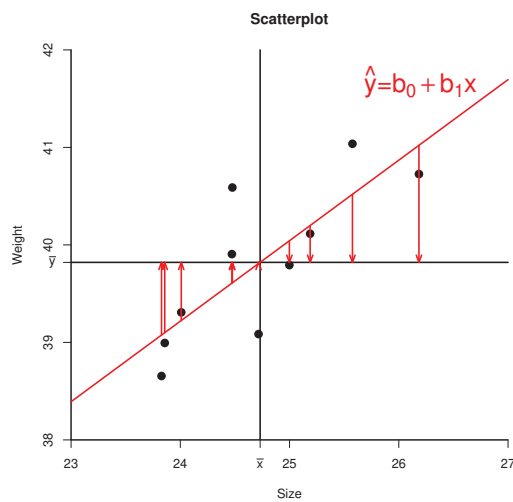
## Sources of variation



### Variation in $Y$

$$SSTO = \sum_i (y_i - \bar{y})^2$$

## Sources of variation

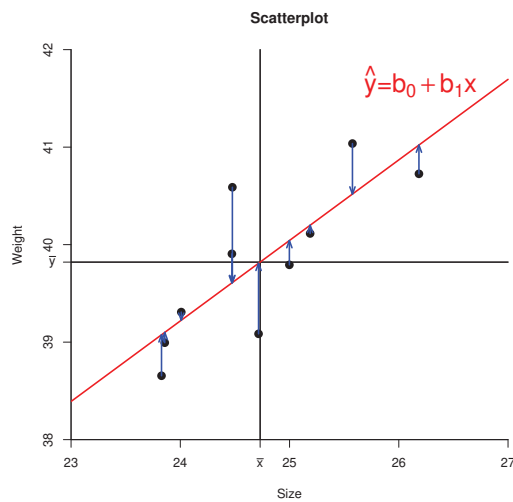


### Variation in $Y$

$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

## Sources of variation



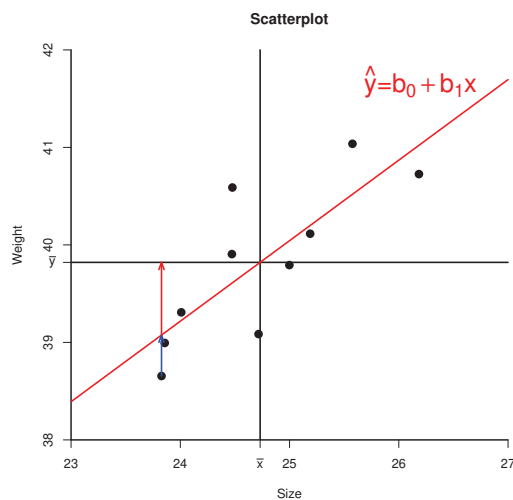
### Variation in Y

$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

## Sources of variation



### Variation in Y

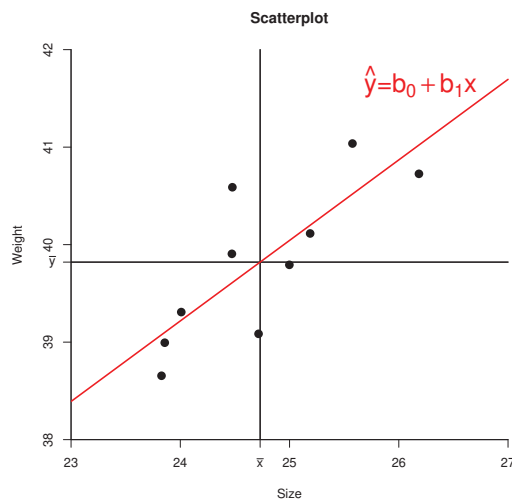
$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SSTO = SSR + SSE$$

## Sources of variation



### Variation in Y

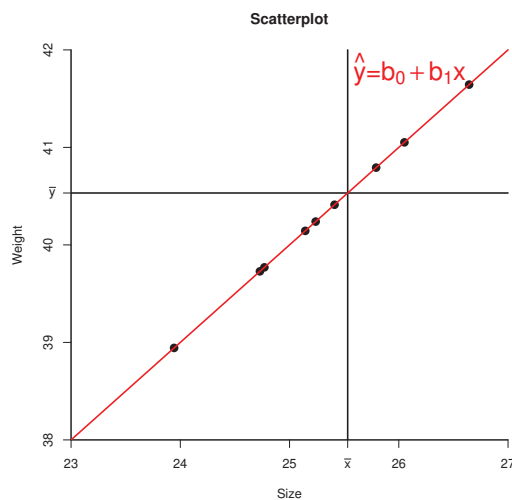
$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SSTO = SSR + SSE$$

## Sources of variation



### Variation in Y

$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = 0$$

$$SSTO = SSR$$

## The sum of the squares

$$SSTO = SSR + SSE$$

*SSTO*: Total variation in the response  $Y$

*SSE*: The variation in  $Y$  not explained by the model

*SSR*: The variation in  $Y$  explained by the model

### The sum of the squares

We can decompose the total sum of squares in two different sums of squares: the residual and regression sum of squares.

## Coefficient of determination

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

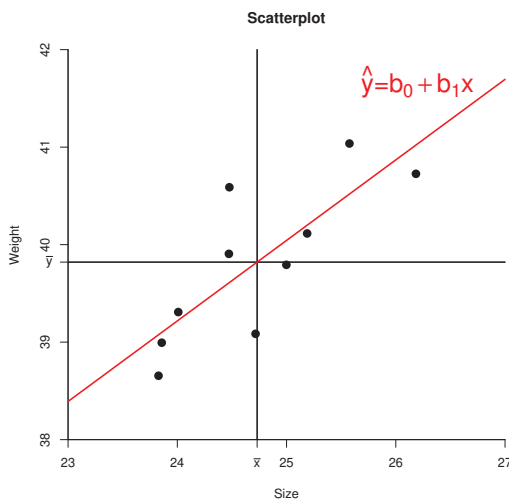
### Coefficient of determination

The coefficient of determination is a measure of the proportion of the total variation in the response that can be explained by the linear regression model.

- The coefficient of determination is always between 0 and 1



## Sources of variation

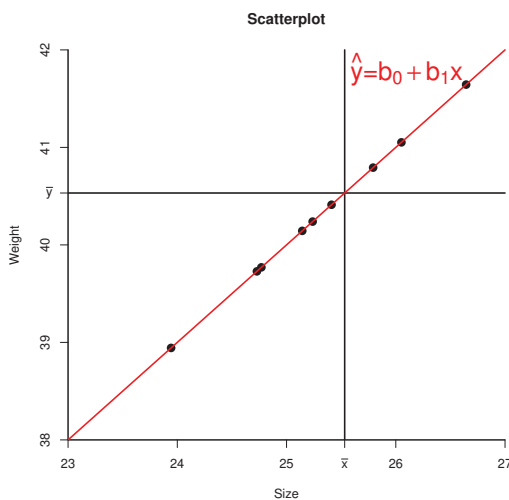


### Variation van $Y$

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$
- The larger the better

## Sources of variation



### Variation van $Y$

$$R^2 = \frac{SSR}{SSTO} = 1$$

## Linear regression: R code and output

```
> ## Fitting the model
>
> res<-lm(height~age, data=kalama)
> kalama.anova<-anova(res)
> kalama.summary<-summary(res)
> kalama.anova
>
Analysis of Variance Table

Response: height
      Df Sum Sq Mean Sq F value    Pr(>F)
age      1  57.655   57.655   879.99 4.428e-11 ***
Residuals 10   0.655    0.066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

## Linear regression: R code and output

```
> kalama.summary
>
Call:
lm(formula = height ~ age, data = kalama)

Residuals:
    Min       1Q   Median       3Q      Max
-0.27238 -0.24248 -0.02762  0.16014  0.47238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283    0.5084   127.71 < 2e-16 ***
age          0.6350     0.0214    29.66 4.43e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9876
F-statistic: 880 on 1 and 10 DF, p-value: 4.428e-11
>
```

## Kalama Study: R Output

Anova					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	57.65	57.65	879.99	0.0000
Residuals	10	0.66	0.07		
Total	11	58.31			

Coefficients				
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	64.9283	0.5084	127.71	0.0000
age	0.6350	0.0214	29.66	0.0000

## Kalama Study: R Output

Coefficients				
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	64.9283	0.5084	127.71	0.0000
age	0.6350	0.0214	29.66	0.0000

- $y = \beta_0 + \beta_1 \cdot x + \epsilon$
- $b_0 = \bar{y} - b_1 \bar{x} = 64.928$
- $b_1 = r_{xy} \frac{s_y}{s_x} = 0.635$
- What does this p-value give?

## Inference

$$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$$

Coefficients

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	64.9283	0.5084	127.71	0.0000
age	0.6350	0.0214	29.66	0.0000

## Kalama Study

Anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	57.65	57.65	879.99	0.0000
Residuals	10	0.66	0.07		
Total	11	58.31			

- $r_{xy}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}} = 0.994$

- $R^2 = \frac{SSR}{SST} = \frac{57.65}{58.31} = 0.989$

$$r_{xy} = \sqrt{R^2}$$

## Kalama Study

Anova					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	57.65	57.65	879.99	0.0000
Residuals	10	0.66	0.07		
Total	11	58.31			

- $\hat{\sigma}^2 = MSE = \frac{SS_{Error}}{12 - 2} = \frac{0.66}{10} = 0.07$
- $R^2 = \frac{SSR}{SST} = \frac{57.65}{58.31} = 0.989$
- A substantial proportion of the variation in the outcome, 98.9%, is explained by the linear regression model.

## Multiple linear regression

A multiple regression model is used to explain a dependent variable  $Y$  in terms of one or more independent variables  $\mathbf{X}' = (1, X_1, \dots, X_{p-1})$ .

If  $Y$  is a quantitative random variable and the elements in  $\mathbf{X}$  can take both quantitative and qualitative values, then one can consider a *regression model*

$$Y = f(\mathbf{X}) + \epsilon,$$

with  $\mathbf{X}$  and  $\epsilon$  independent and  $E(\epsilon) = 0$ ,  $\text{Var}(\epsilon) = \sigma^2$ . Often, it is also assumed that  $\epsilon$  is normally distributed.

The previous model essentially describes the average behavior of  $Y$  as a function  $f(\cdot)$  of  $\mathbf{X}$ , i.e.,  $E(Y) = f(\mathbf{X})$ .

## The regression model

Taylor's theorem states that if  $f$  is differentiable at certain point  $\mathbf{a} \in \mathbb{R}^{p-1}$  then

$$f(\mathbf{X}) = f(\mathbf{a}) + (\mathbf{X} - \mathbf{a})' \boldsymbol{\beta}_* + |\mathbf{X} - \mathbf{a}| h(\mathbf{X}), \quad \lim_{\mathbf{X} \rightarrow \mathbf{a}} h(\mathbf{X}) = 0.$$

Therefore, at least locally (close to  $\mathbf{a}$ ),  $f(\cdot)$  can often be approximated by a *linear* model, i.e.,  $f(\mathbf{X}) = \mathbf{X}' \boldsymbol{\beta} = \beta_0 + \sum \beta_j X_j$ .

$$\begin{aligned} Y &\approx \mathbf{X}' \boldsymbol{\beta} + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon \end{aligned}$$

The previous regression model is linear in the parameters and, hence, it is called a linear regression model.

**Non-linear Regression Model:**  $Y = \beta_0 + \beta_1 X_1^{\beta_2} + \epsilon$

## General linear regression model

**Model:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

where

- $X_1, \dots, X_{p-1}$  are known predictor variables.
- $\beta_0, \beta_1, \dots, \beta_{p-1}$  are unknown parameters.
- $\epsilon$  is an error term. It is often assumed that  $\epsilon \sim N(0, \sigma^2)$ .

## Interpretation of the parameters

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

- This response function is a hyperplane, which is a plane in more than two dimensions.
- The parameter  $\beta_k$  indicates the change in the mean response  $E(Y|\mathbf{X})$  with a unit increase in the predictor variable  $X_k$ , when all other predictor variables in the regression model are held constant.
- $E(Y|\mathbf{X} = 0) = \beta_0$ . The intercept gives the average response when all covariates are zero. It may not be interpretable unless the covariates are centered.

## Categorical covariates: Dummy variables

### Example

- $Y$  length in hospital stay
- $X_1$ : patient's age
- $X_2$ : gender coded as female (1) - male (0)
- Main effects model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Males

$$E(Y) = \beta_0 + \beta_1 X_1$$

Females

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2$$

## Categorical covariates: Dummy variables

### Example

- $Y$  length in hospital stay
- $X_1$ : patient's age
- $X_2$ : gender coded as female (1) - male (0)
- Main effects model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Males

$$E(Y) = \beta_0 + \beta_1 X_1$$

Females

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$$

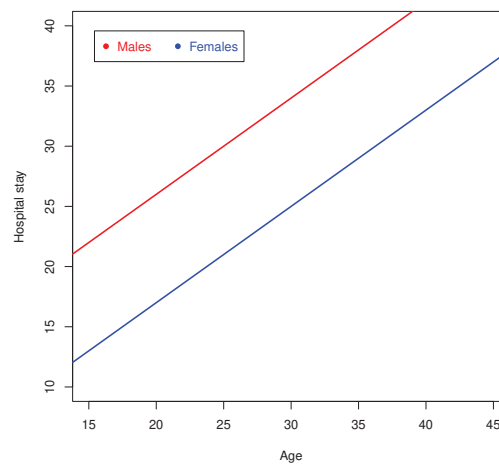
## Main effects model: Parallel lines

Males

$$E(Y) = \beta_0 + \beta_1 X_1$$

Females

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$$





## Categorical covariates: Dummy variables

### Example

- $Y$  length in hospital stay
- $X_1$ : patient's age
- $X_2$ : gender coded as female (1) - male (0)
- Interaction model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

Males

$$E(Y) = \beta_0 + \beta_1 X_1$$

Females

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1$$

## Categorical covariates: Dummy variables

### Example

- $Y$  length in hospital stay
- $X_1$ : patient's age
- $X_2$ : gender coded as female (1) - male (0)
- Interaction model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

Males

$$E(Y) = \beta_0 + \beta_1 X_1$$

Females

$$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

## Categorical covariates: Dummy variables

### Example

- $Y$  length in hospital stay
- $X_1$ : patient's age
- $X_2$ : gender coded as female (1) - male (0)
- Interaction model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Males

$$E(Y) = \beta_0 + \beta_1 X_1$$

Females

$$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1$$

- It is still a linear model:  $X_3 = X_1 X_2$

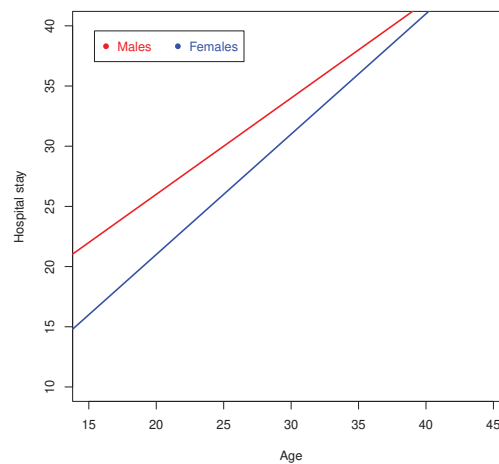
## Interaction model: Non-parallel lines

Males

$$E(Y) = \beta_0 + \beta_1 X_1$$

Females

$$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1$$



## Categorical covariates: Dummy variables

### Example

- $Y$  length in hospital stay
- $X_1$ : patient's age
- $X_2$ : female (1) - male (0)
- $X$  disability status: 3 levels
  - ① Not disabled
  - ② Partially disabled
  - ③ Fully disabled

## Categorical covariates: Dummy variables

For a factor with  $r = 3$  levels, one needs to consider  $(r - 1) = 2$  indicator (dummy) variables as predictors:

$$x_3 = \begin{cases} 1 & \text{Not disabled} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{Partially disabled} \\ 0 & \text{otherwise} \end{cases}$$

Main effects model

$$Y = \beta_0 + \beta_1 X_1 + \overbrace{\beta_2 X_2}^{\text{gender}} + \underbrace{\beta_3 X_3 + \beta_4 X_4}_{\text{disability status}} + \varepsilon$$

## Categorical covariates: Dummy variables

For a factor with  $r = 3$  levels, one needs to consider  $(r - 1) = 2$  indicator (dummy) variables as predictors:

$$x_3 = \begin{cases} 1 & \text{Not disabled} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{Partially disabled} \\ 0 & \text{otherwise} \end{cases}$$

Interaction model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_3 + \beta_4 X_4}_{\text{disability status}} + \underbrace{\beta_5 X_1 X_3 + \beta_6 X_1 X_4}_{\text{interaction: disability-age}} + \varepsilon$$

## Great flexibility

- Polynomial regression:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$

## Great flexibility

- Polynomial regression:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$ , with  $X_3 = X_1^2$

## Great flexibility

- Polynomial regression:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- Transformed variables:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

## Great flexibility

- Polynomial regression:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- Transformed variables:

$$Y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i}$$

## Great flexibility

- Polynomial regression:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- Transformed variables:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Transformed variables:

$$\frac{1}{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

## Great flexibility

- Polynomial regression:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$

- Transformed variables:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Transformed variables:

$$Y = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$$

## Matrix formulation

Let us consider the following multiple regression model for the  $i$ th subject in the study, with  $i = 1, 2, \dots, n$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_{p-1} X_{p-1i} + \varepsilon_i$$

Collecting all the information on all subjects into vectors and matrices, the previous model can be written as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{p-11} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & \dots & x_{p-1n} \end{pmatrix}}_{\mathbf{X}} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

## Matrix formulation

General linear regression model

$$\begin{matrix} \mathbf{Y} & = & \mathbf{X} & \cdot & \boldsymbol{\beta} & + & \boldsymbol{\varepsilon} \\ (n \times 1) & & (n \times p) & & (p \times 1) & & (n \times 1) \end{matrix}$$

- $\mathbf{Y}$  is the response vector.
- $\boldsymbol{\beta}$  is a vector of parameters.
- $\mathbf{X}$  is a matrix of known covariates with no measurement error.
- $\boldsymbol{\varepsilon}$  is a vector of errors with  $E(\boldsymbol{\varepsilon}) = 0$  and  $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \mathbf{I}$ , where  $\mathbf{I} = \text{diag}(1)$  is the so-called identity matrix.
- $\boldsymbol{\varepsilon}$  and  $\mathbf{X}$  are assumed to be independent of each other.

## Interpreting the the model: $\mathbf{Y}$

- $\mathbf{Y}$  is called the regressand, response variable, criterion variable, or dependent variable.
- The decision as to which variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables.
- Alternatively, there may be an operational reason, in which case there need be no presumption of causality.



## Interpreting the the model: $X$

- $X$ : Its elements are called regressors, explanatory variables, covariates, input variables, predictor variables, or independent variables.
- Usually a constant is included as one of the regressors. The corresponding element of  $\beta$  is called the intercept.
- Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero.

## Interpreting the the model: $X$

- Sometimes one of the regressors can be a non-linear function of another regressor, as in polynomial regression. The model remains linear as long as it is linear in the parameter vector  $\beta$ .
- The regressors may be viewed either as random variables, which one simply observes, or they can be considered as predetermined fixed values which one can choose.
- Both interpretations may be appropriate in different cases, and they generally lead to the same estimation procedures; however different approaches to asymptotic analysis are used in these two situations.

## Interpreting the the model: $\beta$

- $\beta$  is a  $p$ -dimensional parameter vector. Its elements are called effects, or regression coefficients.
- Statistical estimation and inference in linear regression focuses on  $\beta$ .
- The elements of this parameter vector are interpreted as the partial derivatives of the dependent variable with respect to the various independent variables.

## Interpreting the the model: $\varepsilon$

- $\varepsilon$  is called the error term, disturbance term, or noise.
- This variable captures all other factors which influence the dependent variable  $\mathbf{Y}$  other than the regressors  $\mathbf{X}$ .
- The relationship between the error term and the regressors, for example whether they are correlated, is a crucial step in formulating a linear regression model, as it will determine the method to use for estimation.
- Typically, one assumes that  $\mathbf{X}$  and  $\varepsilon$  are independent.

## Estimating the model

Like before, the parameters can be estimated based on the ordinary least squares criterion (OLS)

$$Q = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_p X_{p-1i})^2$$

i.e., finding the values  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$  that minimize  $Q$ .

The solution to this optimization problem is given by the solution  $\hat{\boldsymbol{\beta}}$  of the system of normal equations

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

## Fitted values and residuals

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_{p-1} X_{p-1i}$
- Residuals  $e_i = Y_i - \hat{Y}_i$ 
  - $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$
  - $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ .
  - $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$
  - $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$  with  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  (hat matrix)
  - $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
  - $\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$

## The sum of the squares

$$SST = \sum_i (Y_i - \bar{Y})^2, SSR = \sum_i (\hat{Y}_i - \bar{Y})^2, SSE = \sum_i (Y_i - \hat{Y}_i)^2$$

$$SST = SSR + SSE$$

*SST*: Total variation in the response

*SSR*: The variation explained by the model (covariates)

*SSE*: The variation not explained by the model (covariates)

- Coefficient of determination:  $R^2 = \frac{SSR}{SST}$ , interpretation idem
- $\hat{\sigma}^2 = MSE = \frac{SSE}{n - p}$

## Inferences

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0 \quad H_A : \text{not all } \beta_k \text{ equal zero}$$

Anova Table

Source of variation	SS	df	MS
Regression	$SSR$	$p - 1$	$MSR = \frac{SSR}{p - 1}$
Error	$SSE$	$n - p$	$MSE = \frac{SSE}{n - p}$
Total	$SSTO$	$n - 1$	

- Under the null  $F = \frac{MSR}{MSE} \sim F(p - 1, n - p)$

## Inferences

$$H_0 : E(Y|\mathbf{X}) = \beta_0 \quad H_A : E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

Anova Table

Source of variation	SS	df	MS
Regression	$SSR$	$p - 1$	$MSR = \frac{SSR}{p - 1}$
Error	$SSE$	$n - p$	$MSE = \frac{SSE}{n - p}$
Total	$SSTO$	$n - 1$	

- Under the null  $F = \frac{MSR}{MSE} \sim F(p - 1, n - p)$

## Inferences: $\beta_k$

$$H_0 : \beta_k = 0 \quad H_A : \beta_k \neq 0$$

- Test statistics:

$$t = \frac{\hat{\beta}_k}{s\{\hat{\beta}_k\}} \sim t(1 - \alpha/2; n - p)$$

- Confidence interval:

$$\hat{\beta}_k \pm t(1 - \alpha/2; n - p)s\{\hat{\beta}_k\}$$

## Comparing nested models

### Likelihood ratio tests

- Null hypothesis of interest equals  $H_0 : \beta \in \Theta_{\beta,0}$ , for some subspace  $\Theta_{\beta,0}$  of the parameter space  $\Theta_\beta$

## Comparing nested models

### Likelihood ratio tests

- Null hypothesis of interest equals  $H_0 : \beta \in \Theta_{\beta,0}$ , for some subspace  $\Theta_{\beta,0}$  of the parameter space  $\Theta_\beta$
- For instance,  
$$H_0 : E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 \quad H_A : E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$$

## Comparing nested models

### Likelihood ratio tests

- Null hypothesis of interest equals  $H_0 : \beta \in \Theta_{\beta,0}$ , for some subspace  $\Theta_{\beta,0}$  of the parameter space  $\Theta_\beta$
- For instance,
$$H_0 : \beta_3 = \beta_4 = 0 \quad H_A : \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$
- Notation:
  - $L_{ML}$ : ML likelihood function
  - $\hat{\beta}_{ML,0}$ : MLE under  $H_0$
  - $\hat{\beta}_{ML}$ : MLE under general model

## Likelihood ratio tests

- Test statistic:

$$-2 \ln \lambda_N = -2 \ln \left[ \frac{L_{ML}(\hat{\beta}_{ML,0})}{L_{ML}(\hat{\beta}_{ML})} \right]$$

- Asymptotic null distribution:  $\chi^2$  with d.f. equal to the difference in dimension of  $\Theta_\beta$  and  $\Theta_{\beta,0}$ .
- An equivalent F-test can also be used.

## Patient satisfaction

### Case study

A hospital administrator wanted to study the relation between patient satisfaction ( $Y$ ) and patient's age ( $X_1$ , in years), severity of illness ( $X_2$ , an index), and anxiety level ( $X_3$ , an index).

The administrator randomly selected 46 patients and collected data on the previous variables. Larger values of  $Y$ ,  $X_2$ , and  $X_3$  are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

## R code: Patient satisfaction

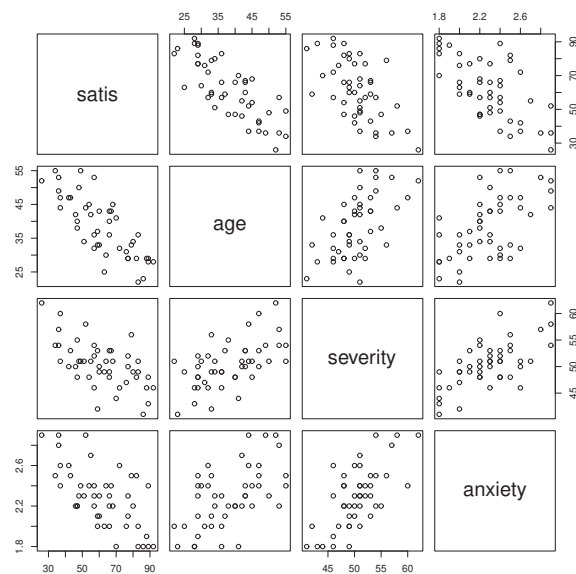
```
> ## Reading the data
>
> satisfaction=read.table("satisfaction.txt", header=T)
> head(satisfaction,10)
>
  satis age severity anxiety
1    48  50      51     2.3
2    57  36      46     2.3
3    66  40      48     2.2
4    70  41      44     1.8
5    89  28      43     1.8
6    36  49      54     2.9
7    46  42      50     2.2
8    54  45      48     2.4
9    26  52      62     2.9
10   77  29      50     2.1
>
```



## R code: Patient satisfaction

```
> ## Exploring the data
>
> cor(satisfaction)
>
      satis      age      severity      anxiety
satis  1.0000000 -0.7867555 -0.6029417 -0.6445910
age    -0.7867555  1.0000000  0.5679505  0.5696775
severity -0.6029417  0.5679505  1.0000000  0.6705287
anxiety -0.6445910  0.5696775  0.6705287  1.0000000
>
> options(digits=2)
> descrip.satisfaction<-stat.desc(satisfaction,basic=TRUE, desc=TRUE)
> descrip.satisfaction
>
      satis      age      severity      anxiety
nbr.val  46.00  46.00  4.6e+01  46.000
min      26.00  22.00  4.1e+01  1.800
max      92.00  55.00  6.2e+01  2.900
range    66.00  33.00  2.1e+01  1.100
median   60.00  37.50  5.0e+01  2.300
mean     61.57  38.39  5.0e+01  2.287
SE.mean   2.54   1.31  6.4e-01  0.044
var      297.10  79.53  1.9e+01  0.090
std.dev   17.24   8.92  4.3e+00  0.299
coef.var   0.28   0.23  8.6e-02  0.131
>
> plot(satisfaction)
>
```

## R code: Patient satisfaction



## R code: Patient satisfaction

```
> ## Fitting the model
>
>
> satisfaction.lm<-lm(satis~age+severity+anxiety, data=satisfaction)
> satisfaction.summary<-summary(satisfaction.lm)
> satisfaction.summary

Call:
lm(formula = satis ~ age + severity + anxiety, data = satisfaction)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  158.491     18.126    8.74 5.3e-11 ***
age          -1.142      0.215   -5.31 3.8e-06 ***
severity     -0.442      0.492   -0.90  0.374
anxiety     -13.470      7.100   -1.90  0.065 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10 on 42 degrees of freedom
Multiple R-squared:  0.682,    Adjusted R-squared:  0.659
F-statistic: 30.1 on 3 and 42 DF,  p-value: 1.54e-10
>
```

## R code: Patient satisfaction

```
> ## Likelihood ratio test null model versus full model
>
>
> satisfaction.lm.int<-lm(satis~1, data=satisfaction) # Null model
> anova(satisfaction.lm.int,satisfaction.lm)          # Null versus full
>
Analysis of Variance Table

Model 1: satis ~ 1
Model 2: satis ~ age + severity + anxiety
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     45 13369
2     42  4249   3      9120 30.1 1.5e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

## R code: Patient satisfaction

```
> ## Likelihood ratio test null model versus full model
>
>  $satis = \beta_0 + \epsilon \Leftrightarrow satis = \beta_0 + \beta_1age + \beta_2severity + \beta_3anxiety + \epsilon$ 
> satisfaction.lm.int<-lm(satis~1, data=satisfaction) # Null model
> anova(satisfaction.lm.int,satisfaction.lm)          # Null versus full
>
Analysis of Variance Table

Model 1: satis ~ 1
Model 2: satis ~ age + severity + anxiety
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1      45 13369
2      42 4249  3      9120 30.1 1.5e-10 ***  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> ## Previous analysis with summary function

Multiple R-squared:  0.682,    Adjusted R-squared:  0.659
F-statistic: 30.1 on 3 and 42 DF,  p-value: 1.54e-10  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ 
>
```

## R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.anova<-anova(satisfaction.lm)
> satisfaction.anova
>
Analysis of Variance Table

Response: satis
      Df Sum Sq Mean Sq F value Pr(>F)
age      1  8275      8275  81.80 2.1e-11 ***
severity  1   481       481   4.75  0.035 *
anxiety   1   364       364   3.60  0.065 .
Residuals 42  4249       101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

## R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.anova<-anova(satisfaction.lm)
> satisfaction.anova
>
Analysis of Variance Table

Response: satis
      Df Sum Sq Mean Sq F value Pr(>F)
age      1  8275    8275   81.80 2.1e-11 ***
severity  1   481     481    4.75  0.035 *
anxiety   1   364     364    3.60  0.065 .
Residuals 42  4249     101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> ## Previous analysis with summary function
>
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  158.491    18.126     8.74 5.3e-11 ***
age          -1.142     0.215    -5.31 3.8e-06 ***
severity     -0.442     0.492    -0.90  0.374
anxiety      -13.470     7.100    -1.90  0.065 .
>
```

## R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.lm2<-lm(satis~age+anxiety+severity, data=satisfaction)
> satisfaction.anova2<-anova(satisfaction.lm2)
> satisfaction.anova2
>
Analysis of Variance Table

Response: satis
      Df Sum Sq Mean Sq F value Pr(>F)
age      1  8275    8275   81.80 2.1e-11 ***
anxiety   1   763     763    7.55  0.0088 **
severity   1    82      82    0.81  0.3741
Residuals 42  4249     101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>
> ## Previous analysis with summary function
>
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  158.491    18.126     8.74 5.3e-11 ***
age          -1.142     0.215    -5.31 3.8e-06 ***
severity     -0.442     0.492    -0.90  0.374
anxiety      -13.470     7.100    -1.90  0.065 .
>
```

## R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.anova<-anova(satisfaction.lm)
> satisfaction.anova
>
Analysis of Variance Table

Response: satis
          Df Sum Sq Mean Sq F value    Pr(>F)
age         1   8275     8275   81.80 2.1e-11 ***
severity    1    481      481    4.75  0.035  *
anxiety     1    364      364    3.60  0.065  .
Residuals  42   4249      101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

## Final model

### Final model:

$$Y_i = 145.941 - 1.2X_{1i} - 16.742X_{3i} + \varepsilon_i$$

```
> ## Final model
>
> satisfaction.lm.final<-lm(satis~age+anxiety, data=satisfaction)
> satisfaction.final.summary<-summary(satisfaction.lm.final)
> satisfaction.final.summary
>
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept)  145.941     11.525   12.66  4.2e-16 ***
age          -1.200       0.204   -5.88  5.4e-07 ***
anxiety     -16.742       6.081   -2.75  0.0086 **
---
>
```

## Inference for mean response

- In many applications one wants to estimate and/or make inferences about the mean of the response  $Y$  for a given value of the predictors  $\mathbf{X}_h$

$$Y_h = E(Y|\mathbf{X}_h) = \mathbf{X}_h' \boldsymbol{\beta}$$

- Point estimate:  $\hat{Y}_h = \mathbf{X}_h' \hat{\boldsymbol{\beta}}$ .
- Confidence intervals are used for inferences.

## Inference for mean response

- $E(Y_h) = \mathbf{X}_h' \boldsymbol{\beta}$
- $\hat{Y}_h = \mathbf{X}_h' \hat{\boldsymbol{\beta}}$
- $s^2(\hat{Y}_h) = \mathbf{X}_h' \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{X}_h$
- $s^2(\hat{Y}_h) = \text{MSE}(\mathbf{X}_h' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_h)$
- Hence  $1 - \alpha$  confidence limits are

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p) \cdot s(\hat{Y}_h)$$

## Prediction of a new observation

- Let now assume that one wants to predict a new observation  $Y_{h,new}$  corresponding to a given level of the predictor variable  $\mathbf{X} = \mathbf{X}_h$ .
- This is typically done by constructing a  $1 - \alpha$  prediction interval, i.e., by finding values  $Y_{h,low}$  and  $Y_{h,up}$  so that

$$P(Y_{h,low} \leq Y_{h,new} \leq Y_{h,up} | \mathbf{X} = \mathbf{X}_h) = 1 - \alpha$$

- The prediction interval is given by

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p) s\{pred\}$$

$$\text{where } s^2\{pred\} = MSE + s^2(\hat{Y}_h)$$

## R code: Predicting a new observation

```
> ## Predicting a new observation
>
> newdata = data.frame(age=43, anxiety=2.7)
> pred.w.plim <- predict(satisfaction.lm.final, newdata, interval="predict")
> pred.w.clim <- predict(satisfaction.lm.final, newdata, interval = "confidence")
> pred.w.plim
>
> fit lwr upr
1 49 28 70
>
> pred.w.clim
>
> fit lwr upr
1 49 44 54
>
```