Linear regression and correlation

Ariel Alonso Abad

Catholic University of Leuven

General Information

- This course will be divided in **two** parts or modules
  - **Part I**: Prof. Ariel Alonso Abad
  - **Part II**: Prof. Rob Jelier
- **Teaching Plan (Part I)**:
  - Lecture 1: Linear regression and correlation
  - Lecture 2: Generalized linear models: Logistic Regression
  - Lecture 3: Multilevel Models: Longitudinal data
  - Lecture 4: Multilevel Models: Cluster data
  - Lecture 5: Missing data
  - Project-day

Evaluation

- Project and an exam=20 points. **Project**: 4 points. **Exam**: 16 points
- **Project and project-day:**:
  - After first lecture students organize themselves in 6 tutorial groups within 5 days
  - Email Prof. Alonso list with the members of each group (names and student numbers)
  - Toledo → Course Documents → Part I → Projects
  - Report with a detailed discussion of the analysis written copy the project-day
  - Email electronic copy and the R code at least two days before the project-day
  - Project-day

Project-day

- Each group has to present the results of the analysis
- Presenter will be chosen by Prof. Ariel Alonso Abad
- Time for discussion, members of other tutorial groups expected to ask questions
- Evaluation
  - Report
  - Presentation
  - Defense of the analysis
  - Questions
- Exam=methodological and practical part
- **Use Toledo**

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

Association and correlation

Association and correlation

Two random variables are dependent, if the probability of an outcome for one variable dependents on the outcome of the other.

Relationship between height and weight

If we measure length and weight in a group of children we will find that relatively tall children, as an average, also have a larger body weight and vice-versa.

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

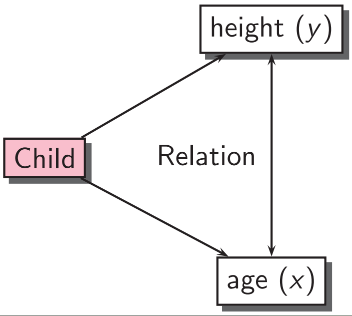
Kalama study

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



- Strength of the association
- Is the relationship linear, quadratic, exponential?

Measuring association

Several measures have been developed for continuous variables

- Covariance
- Pearson correlation coefficient
- Mutual information
- Informational coefficient of correlation

Covariance

Common variability of 2 variables x en y. Sample of outcomes $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$\mathrm{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Covariance

Drawbacks of the covariance

$$\text{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Sensitive to the spread/variability of x and y
- Sensitive to differences in scale

Covariance

Kalama study

	Scale of height	
	Meters	Centimeters
cov(*age*, *height*)	0.082	8.254

Pearson correlation coefficient

Solution: Standardize the variables

- z-score for x_i: $\dfrac{(x_i - \bar{x})}{s_x}$
- z-score for y_i: $\dfrac{(y_i - \bar{y})}{s_y}$

Pearson correlation coefficient

$$\mathrm{r}_{xy}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

$$\mathrm{r}_{xy}(x, y) = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

Pearson correlation coefficient

- A correlation coefficient indicates the direction and strength of the association between two variables
- Pearson correlation coefficient takes values between -1 (perfect negative correlation) and +1 (a perfect positive correlation)
- A value near zero indicates that the variables do not show any **linear** relation
- A positive correlation means that a large value of one variable is often associated with a large value of the other one
- For negative correlation the reverse is true: large values on one variable are often associated with small values on the other one

Kalama study: Pearson correlation coefficient

Correlation between age and length

$$r_{xy}(x,y) = \frac{1}{n-1}\sum_i \frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$$

$$= \frac{1}{11}(1.53 \cdot 1.63 + 1.25 \cdot 1.24 + \ldots + 1.53 \cdot 1.59)$$

$$r_K(x,y) = 0.994$$

Length and age

Correlation is strong and positive

Pearson correlation: Scale invariant

Kalama studie

Measure	Scale of height: Meters	Centimeters
corr(age, $height$)	0.994	0.994
cov(age, $height$)	0.082	8.254

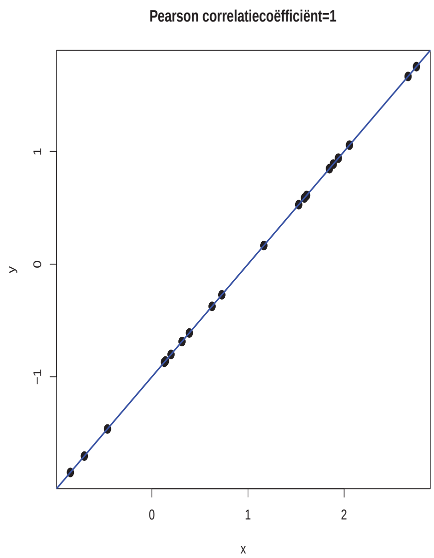
Pearson correlation coefficient

Correlation coefficient

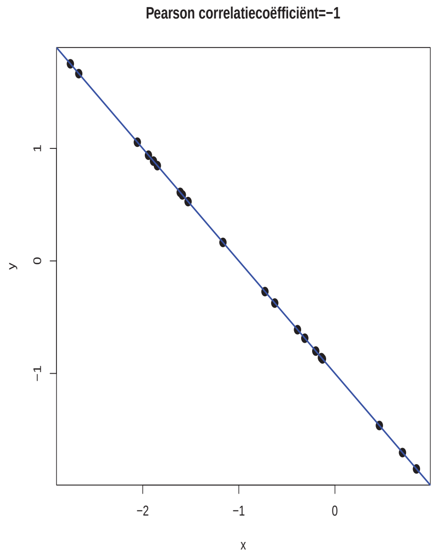
$$r_{xy}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{s_x^2 s_y^2}}$$

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

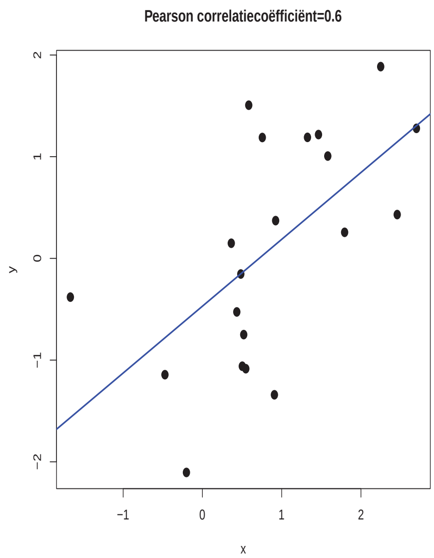
Pearson correlation coefficient



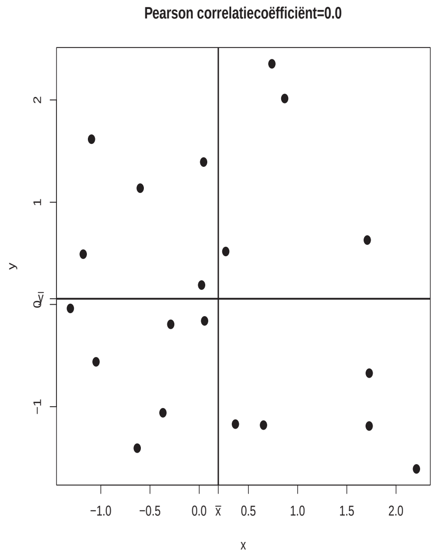
Pearson correlation coefficient



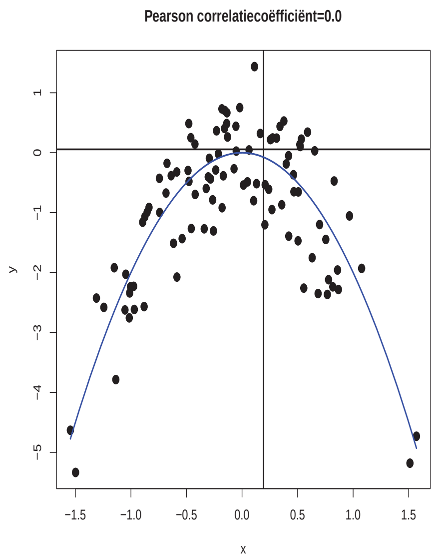
Pearson correlation coefficient



Pearson correlation coefficient



Pearson correlation coefficient



Estimating correlations 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
>
```

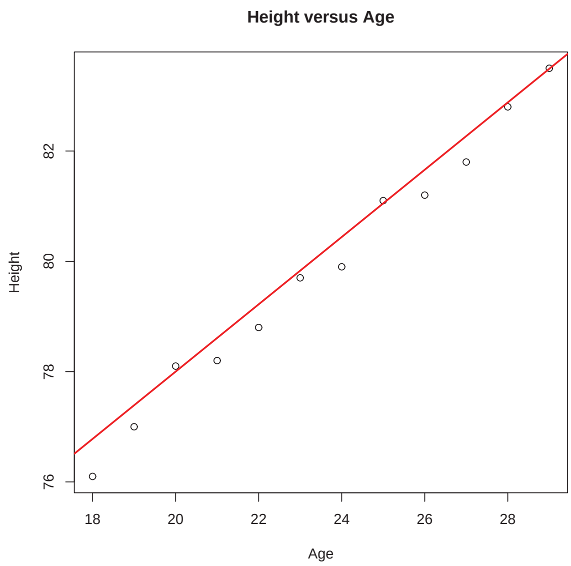

Estimating correlations in R

```
> ## 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
>
```

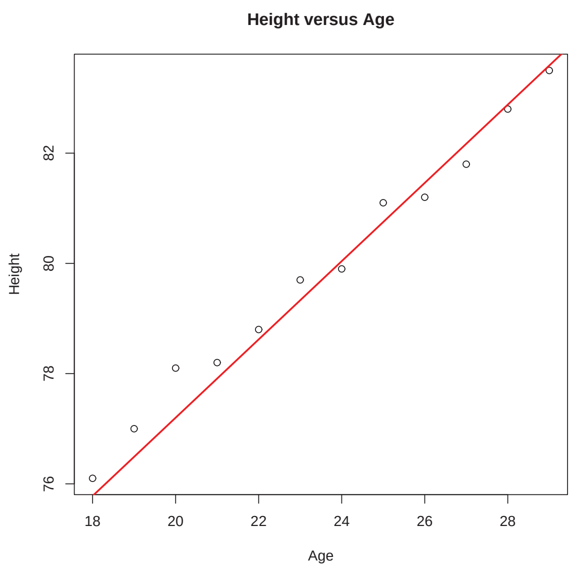
Estimating correlations in R

```
> ## 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
>
```

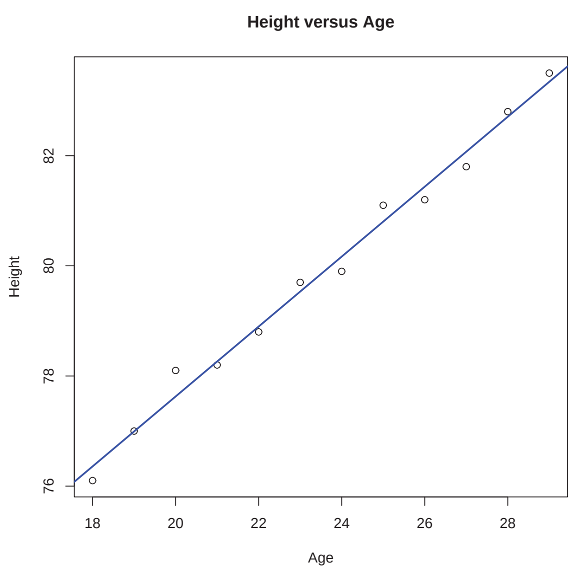

Kalama study ($r_K = 0.994$): Best line



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



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



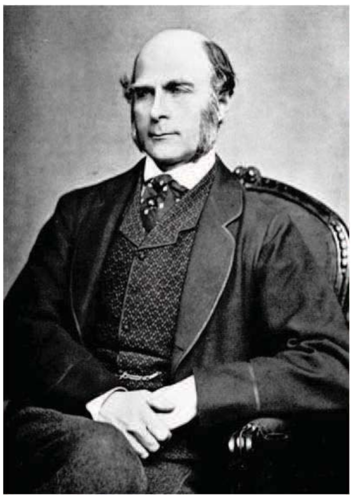
When to use Regression Analysis

- 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, $\boldsymbol{X}' = (X_1, \ldots, X_p)$
- 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 wont 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)

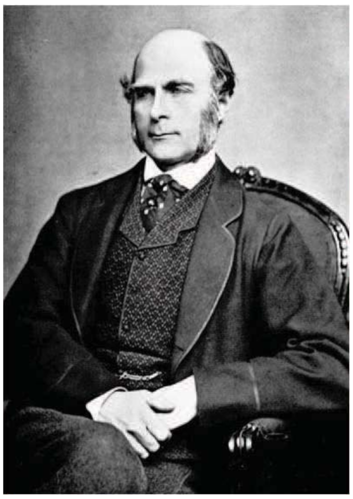
Francis Galton



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

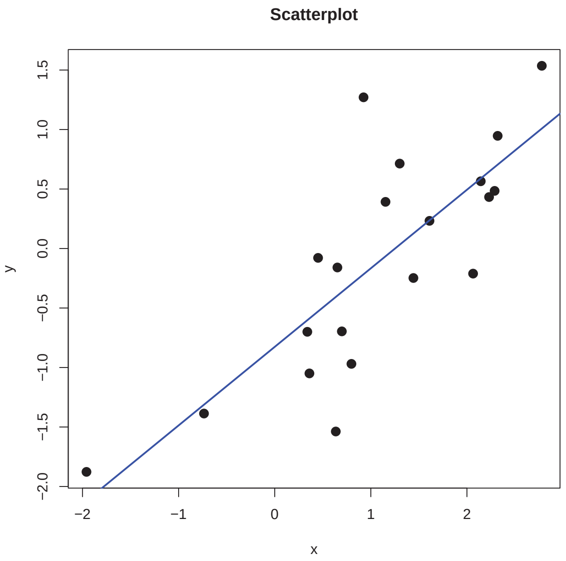
"Regression towards mediocrity in hereditary stature". Journal of the Anthropological Institute 15 (1886), 246-263.

Francis Galton



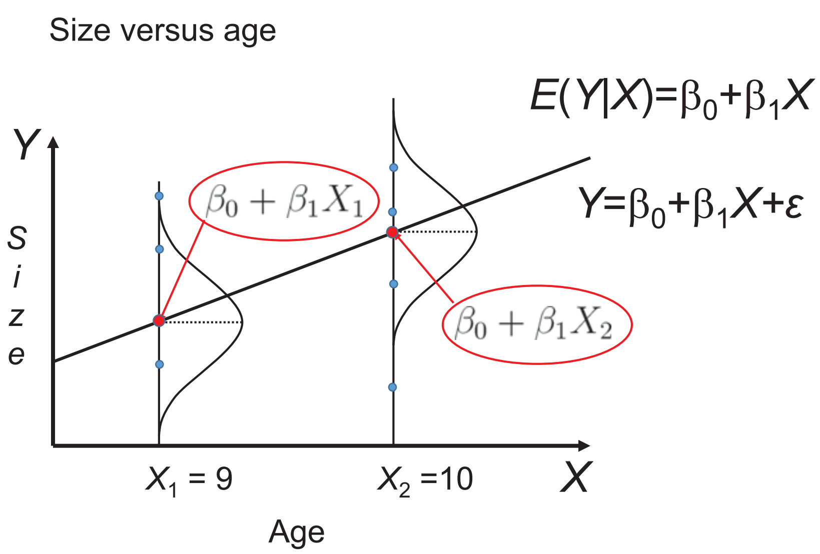
Galton noticed that sons of tall parents tend to be tall but not as tall as their parents while sons of short fathers tend to be short but not as short as their fathers.
He considered this tendency to be a **regression** to "mediocrity"

Linear regression



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

Linear regression



Formal Statement of the Model

For each unit $i = 1, \ldots, n$, 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

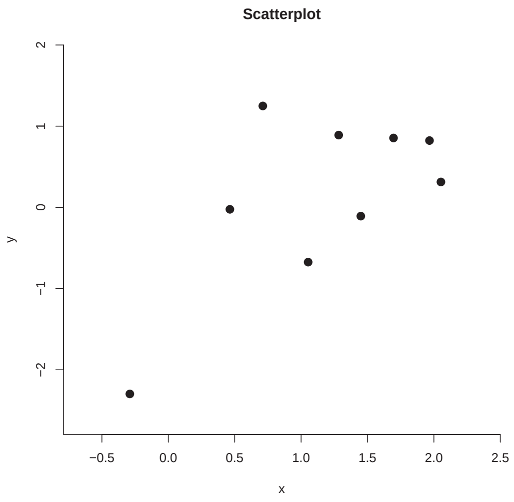
1. The value of X_i is precisely known.
2. Y_i is a continuous random variable.
3. β_0 and β_1 are parameters. That is, they are: unknown, constant and do not depend on the research unit.
4. ε_i is a random error term. It is not observable.

Formal Statement of the Model

Additional assumptions

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

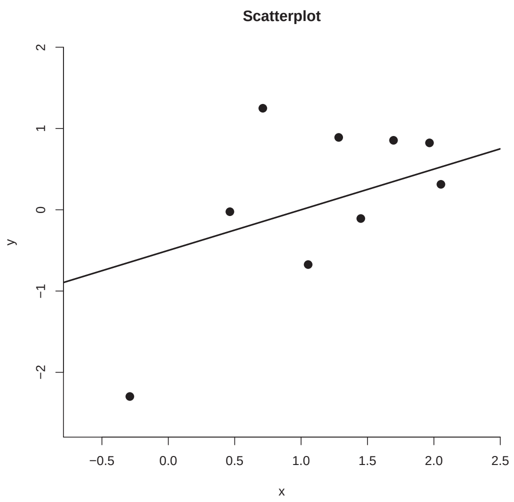
Least squares method



Which line fits the data best?

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

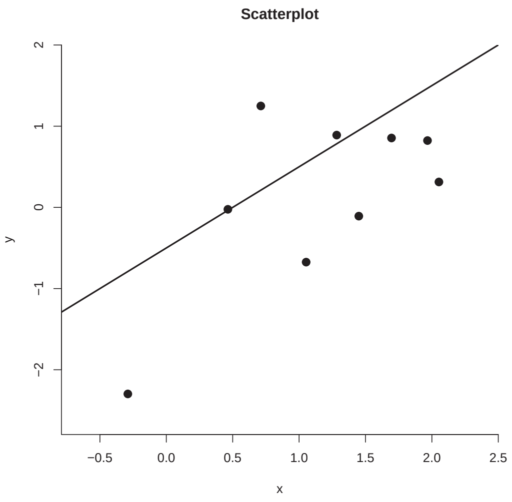
Least squares method



Which line fits the data best?

- $Y = -0.5 + 0.5X + \varepsilon$

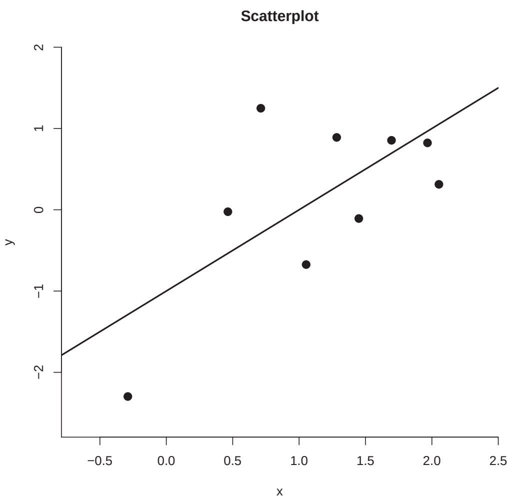
Least squares method



Which line fits the data best?

- $Y = -0.5 + X + \varepsilon$

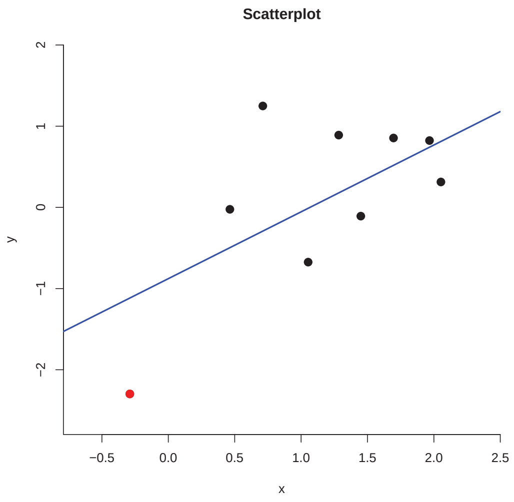
Least squares method



Which line fits the data best?

- $Y = -0.26 + 0.45X + \varepsilon$

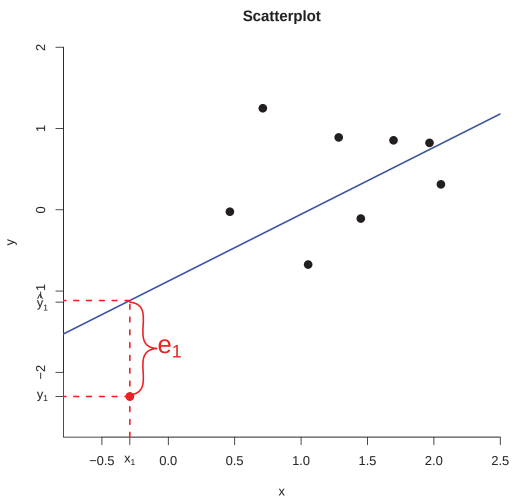
Least squares method



Which line fits the data best?

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

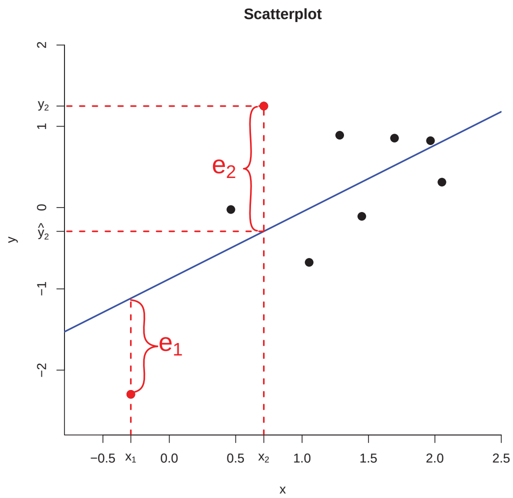
Least squares method



Which line fits the data best?

$$\hat{y}_1 = \beta_0 + \beta_1 x_1$$
$$y_1 = \hat{y}_1 + e_1$$
$$e_1 = y_1 - \hat{y}_1$$

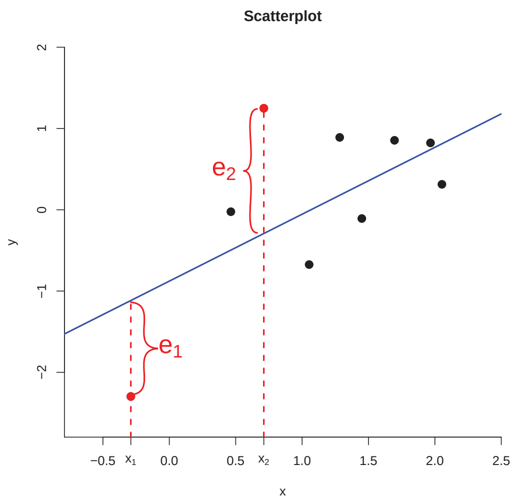
Least squares method



Which line fits the data best?

$$\hat{y}_2 = \beta_0 + \beta_1 x_2$$
$$y_2 = \hat{y}_2 + e_2$$
$$e_2 = y_2 - \hat{y}_2$$

Least squares method



Find the values of β_0 and β_1 that minimize $SSE = \sum_i e_i^2$, where

$$\sum_i e_i^2 = \sum_i (y_i - \beta_0 + \beta_1 x_i)^2$$

Estimated model

$$\hat{y} = b_0 + b_1 x$$
$$\hat{\beta}_1 = b_1 = r_{xy} \frac{s_y}{s_x}$$
$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$$

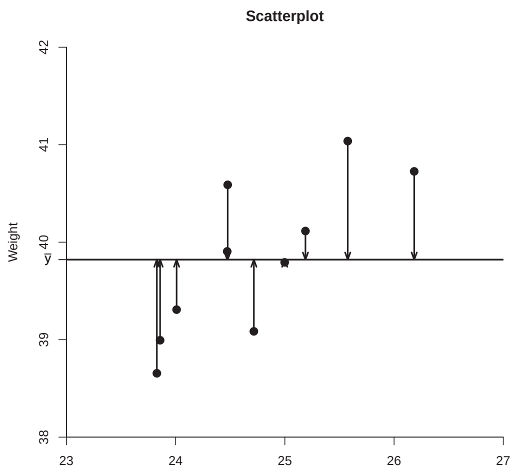
What about σ^2?

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

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

is an *unbiased* estimator of σ^2.

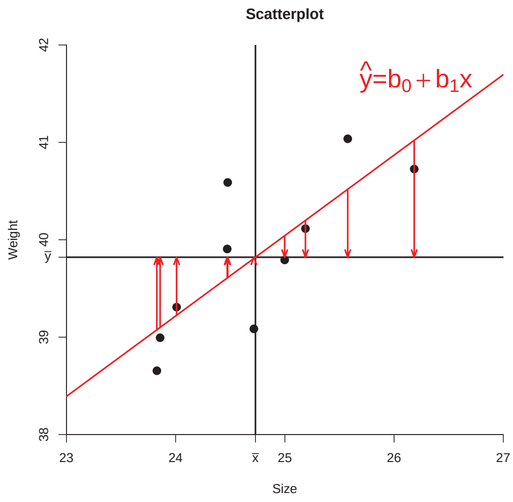
Sources of variation



Variation in Y

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

Sources of variation

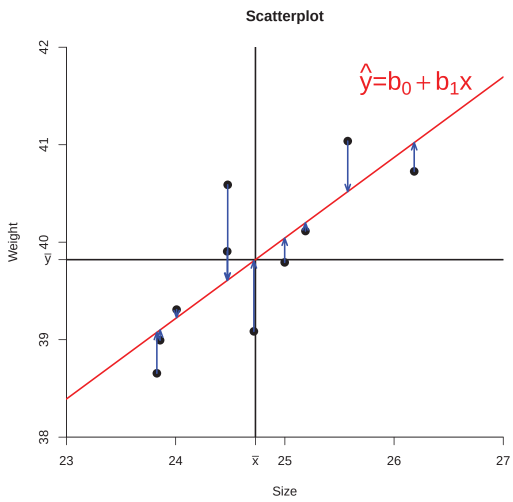


Variation in Y

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

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

Sources of variation



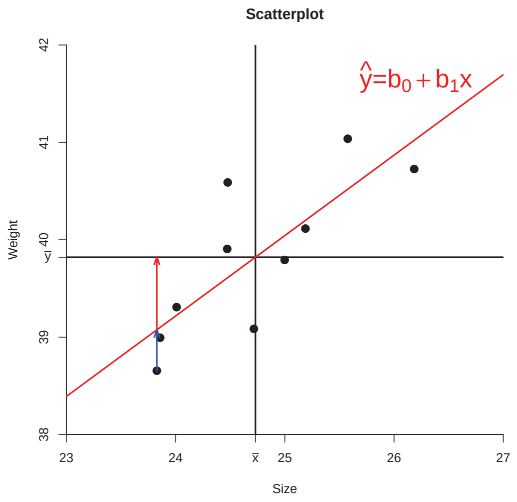
Variation in Y

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

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

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

Sources of variation



Variation in Y

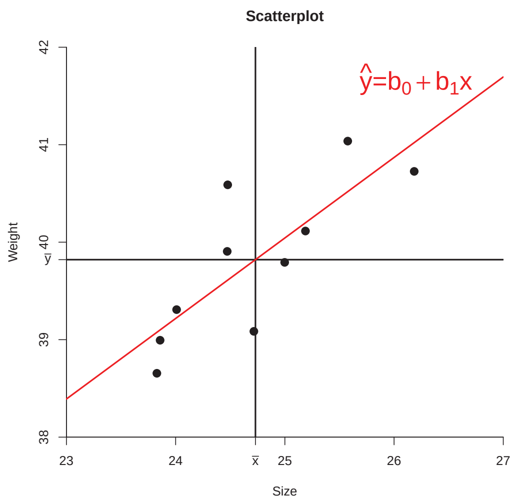
$$SS_{Total} = \sum_i (y_i - \bar{y})^2$$

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

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

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

Sources of variation



Variation in Y

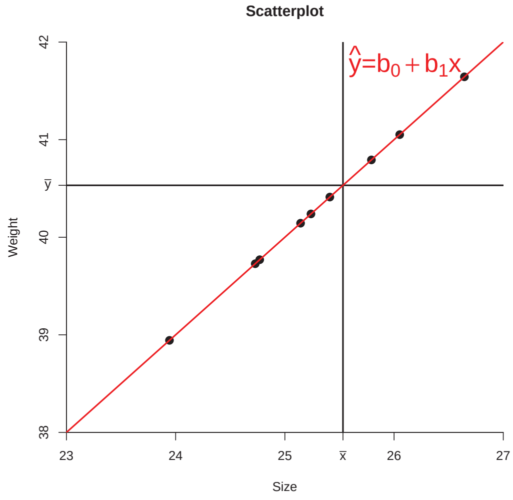
$$SS_{Total} = \sum_i (y_i - \bar{y})^2$$

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

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

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

Sources of variation



Variation in Y

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

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

$$SS_{Error} = 0$$

$$SS_{Total} = SS_{Regression}$$

The sum of the squares

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

SS_{Total}: Total variation in the observations

SS_{Error}: The variation not explained by the model

$SS_{Regression}$: The variation 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

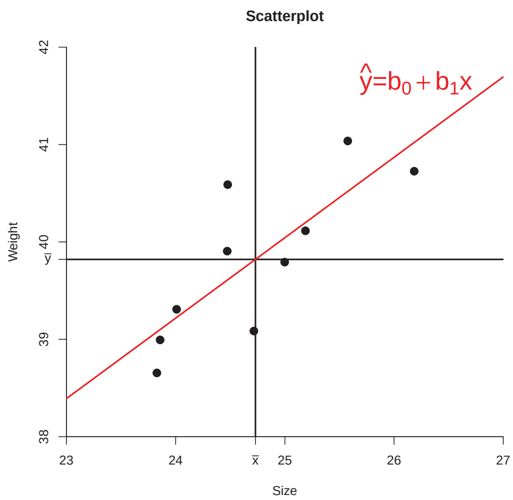
$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Coefficient of determination

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

- The coefficient of determination is always between 0 en 1

Sources of variation

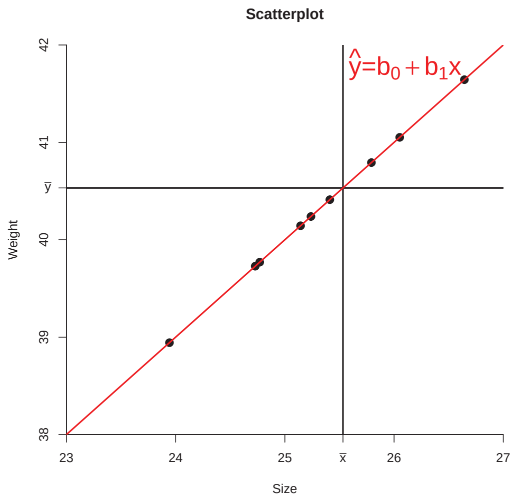


Variation van Y

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

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

Sources of variation



Variation van Y

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1$$

Simple linear regression

Kalama study: Descriptive Statistics

Variable	Mean	Standard deviation
Age (x)	23.50	3.6
Height (y)	79.85	2.3

$r_{xy} = 0.994$

Kalama study: Estimated model

Model: $\hat{y} = b_0 + b_1 x$.

$$b_1 = r_{xy}\frac{s_y}{s_x} = 0.994 \cdot \frac{2.30}{3.60} = 0.635$$

$$b_0 = \bar{y} - b_1\bar{x} = 79.85 - 0.635 \cdot 23.50 = 64.93$$

Linear regression: R code

```
> ## 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

```
> 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

- $y = \beta_0 + \beta_1 \cdot x + \epsilon$
- $b_0 = \bar{y} - b_1\bar{x} = 64.928$

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

- $y = \beta_0 + \beta_1 \cdot x + \epsilon$
- $b_0 = \bar{y} - b_1\bar{x} = 64.928$
- $b_1 = \mathrm{r}_{xy}\frac{s_y}{s_x} = 0.635$

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

- $y = \beta_0 + \beta_1 \cdot x + \epsilon$
- $b_0 = \bar{y} - b_1\bar{x} = 64.928$
- $b_1 = \mathrm{r}_{xy}\frac{s_y}{s_x} = 0.635$
- 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			

- $r_{xy}(x,y) = \dfrac{\text{cov}(x,y)}{\sqrt{s_x^2 s_y^2}} = 0.994$
- $R^2 = \dfrac{SS_{Regression}}{SS_{Total}} = \dfrac{57.65}{58.31} = 0.989$

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			

- $r_{xy}(x, y) = \dfrac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}} = 0.994$
- $R^2 = \dfrac{SS_{Regression}}{SS_{Total}} = \dfrac{57.65}{58.31} = 0.989$

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

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			

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

Multiple linear regression

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

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

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

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

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

The regression model

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

$$f(\boldsymbol{X}) = f(\boldsymbol{a}) + \beta_*'(\boldsymbol{X} - \boldsymbol{a}) + h(\boldsymbol{X})|\mathbf{X} - \mathbf{a}|, \qquad \lim_{\boldsymbol{X} \to \boldsymbol{a}} h(\boldsymbol{X}) = 0.$$

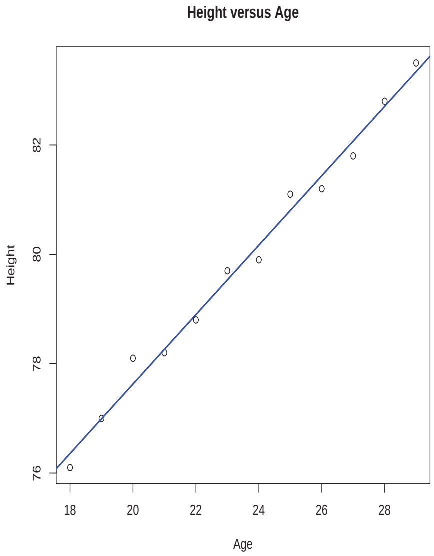
Therefore, at least locally (close to $\boldsymbol{a}$), $f(\cdot)$ can often be approximated by a *linear* model, i.e., $f(\boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{X} = \sum \beta_j X_j$.

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

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

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

Danger of extrapolation



Within the range of the data

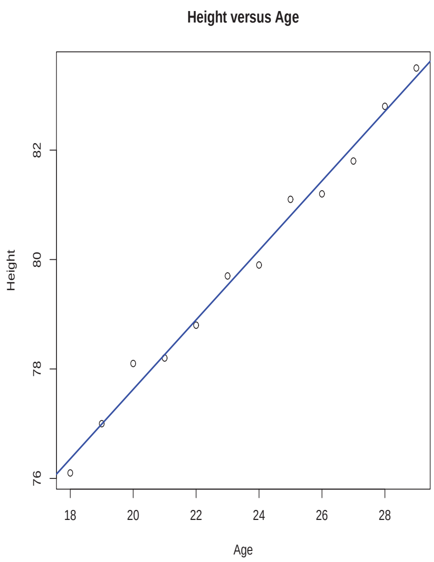
age= 27.5 months (2.29 years)
Average height=?

$\hat{y} = 64.92 + 0.635x$

$\hat{y} = 64.92 + 0.635 \cdot 27.5 = 82.38$

Average height= 0.82 m

Danger of extrapolation



Outside the range of the data

Age= 480 months (40 years)
Average height=?

$\hat{y} = 64.92 + 0.635x$

$\hat{y} = 64.92 + 0.635 \cdot 480 = 369.7$

Average height= **3.7** m

Interpretation of the parameters

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

- This response function is a hyperplane, which is a plane in more than two dimensions.
- The parameter β_k indicates the change in the mean response $E(Y|\boldsymbol{X})$ with a unit increase in the predictor variable X_k, when all other predictor variables in the regression model are held constant.
- $E(Y|\boldsymbol{X} = \boldsymbol{0}) = \beta_0$. 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	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2$

Categorical covariates: Dummy variables

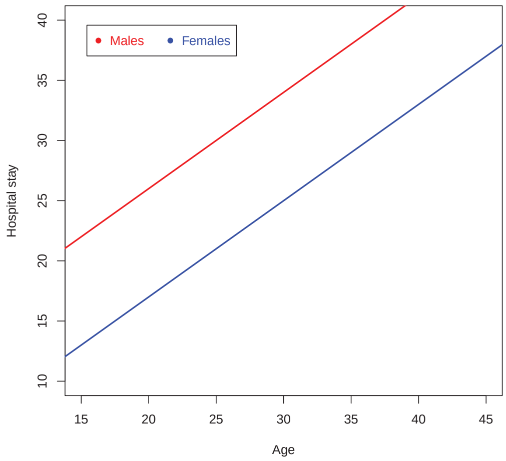
Example

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

Main effects model: Parallel lines

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



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

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

Categorical covariates: Dummy variables

Example

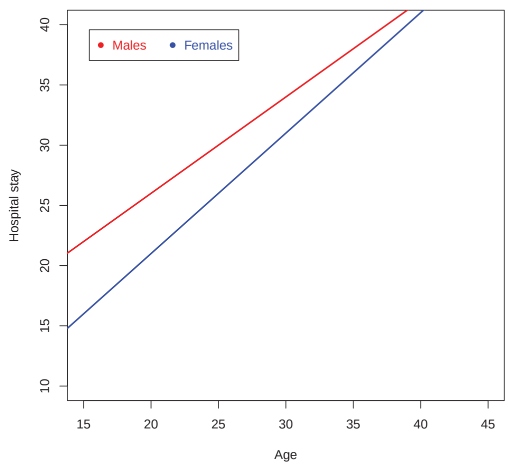
- Y length in hospital stay
- X_1: patient's age
- X_2: gender coded as female (1) - male (0)
- Interaction model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

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

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

Interaction model: Non-parallel lines

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



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: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$, 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

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

It can be written as

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

Matrix Formulation

General linear regression model

$$\underset{(n \times 1)}{\boldsymbol{Y}} = \underset{(n \times p)}{\boldsymbol{X}} \cdot \underset{(p \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}$$

- $\boldsymbol{Y}$ response vector.
- $\boldsymbol{\beta}$ parameters vector.
- $\boldsymbol{X}$ matrix of known constants.
- $\varepsilon_i \sim N(0, \sigma^2)$ independent and identically distributed.
- $\boldsymbol{\varepsilon}$ vector, $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \boldsymbol{I}$

Estimating the model

Least squares criterion

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_p X_{p-1\,i})^2$$

find the values $\beta_0, \beta_1, \ldots, \beta_{p-1}$ that minimize Q.

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

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

and $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

Fitted values and residuals

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

The sum of the squares

$$SS_{Total} = \sum_i (y_i - \bar{y})^2, \quad SS_{Regression} = \sum_i (\hat{y}_i - \bar{y})^2,$$
$$SS_{Error} = \sum_i (y_i - \hat{y}_i)^2$$

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

SS_{Total}: Total variation in the observations

SS_{Error}: The variation not explained by the model

$SS_{Regression}$: The variation explained by the model

- Coefficient of determination: $R^2 = \dfrac{SS_{Regression}}{SS_{Total}}$, interpretation idem
- $\hat{\sigma}^2 = MSE = \dfrac{SSE}{n-p}$

Inferences

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

Anova Table

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

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

Inferences

$$\mathrm{H}_0 : E(Y|\boldsymbol{X}) = \beta_0 \quad \mathrm{H}_A : E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

Anova Table

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

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

Inferences: β_k

$$\mathrm{H}_0 : \beta_k = 0 \quad \mathrm{H}_A : \beta_k \neq 0$$

- 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 $H_0 : \boldsymbol{\beta} \in \Theta_{\boldsymbol{\beta},0}$, for some subspace $\Theta_{\boldsymbol{\beta},0}$ of the parameter space $\Theta_{\boldsymbol{\beta}}$
- For instance,

$$\mathrm{H}_0 : E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 \quad \mathrm{H}_A : E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$$

Comparing nested models
Likelihood ratio tests

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

$$\mathrm{H}_0 : \beta_3 = \beta_4 = 0 \quad \mathrm{H}_A : \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$

- Notation:
  - L_{ML}: ML likelihood function
  - $\widehat{\boldsymbol{\beta}}_{ML,0}$: MLE under H_0
  - $\widehat{\boldsymbol{\beta}}_{ML}$: MLE under general model

Likelihood ratio tests

- Test statistic:

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

- Asymptotic null distribution: χ^2 with d.f. equal to the difference in dimension of $\Theta_{\boldsymbol{\beta}}$ and $\Theta_{\boldsymbol{\beta},0}$.

- An equivalent F-test can also be used.

Patient satisfaction

Case study

A hospital administrator wished 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.

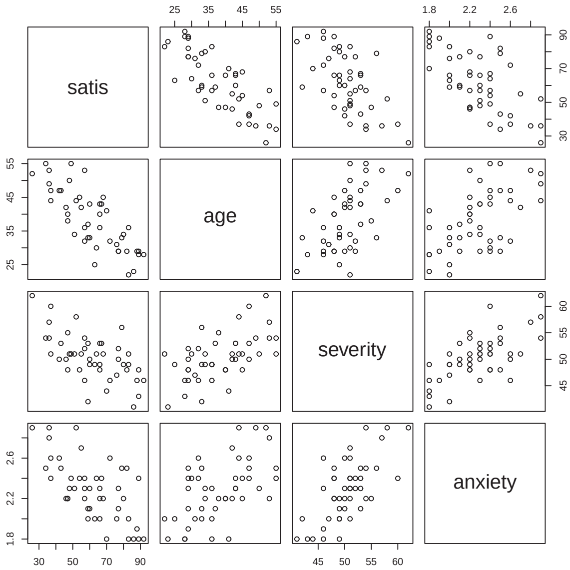
R code: Patient satisfaction

```
> ## 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
>
> ## 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
>
```


R code: Patient satisfaction

```
> ## Sequential building og 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 .   # Same p-value as before (-1.9)^2=3.6
```

R code: Patient satisfaction

```
> ## Sequential building og 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     # Same p-value as before (-0.9)^2=0.81
anxiety      -13.470      7.100   -1.90    0.065 .
>
```


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 **
---
>
```

Predicting the outcome

The $1-\alpha$ prediction limits for a new observation Y_{new} corresponding to the covariate vector $\boldsymbol{X}_{new}$ is given by

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

where $s^2\{pred\} = MSE + s^2\{\hat{Y}_{new}\} = MSE[1 + \boldsymbol{X}'_{new}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_{new}]$ and $\hat{Y}_{new} = \hat{E}(Y|\boldsymbol{X} = \boldsymbol{X}_{new}) = \boldsymbol{X}_{new}\hat{\boldsymbol{\beta}}$

- We predict the new observation Y_{new} using the average of Y when $\boldsymbol{X} = \boldsymbol{X}_{new}$
- The variance of $\hat{Y}_{new}$ as an estimator of the conditional expectation is $s^2\{\hat{Y}_{new}\}$
- The variance of $\hat{Y}_{new}$ as an predictor of Y_{new} is larger, namely, $MSE + s^2\{\hat{Y}_{new}\}$

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
>
```