Linear regression

6

Outline

1. Introduction
2. Simple linear regression
   - Model and parameter estimation
   - Inference in simple linear regression
   - Adequacy of the regression model
3. Multiple linear regression
   - Model and parameter estimation
   - Inference in multiple linear regression
   - Multicollinearity
   - Dummy variables

Introduction

- Joint study of two variables
- Dependence between two variables
- Regression

$$y = f(x) + u$$

Outline

1. Introduction
2. Simple linear regression
   - Model and parameter estimation
   - Inference in simple linear regression
   - Adequacy of the regression model
3. Multiple linear regression
   - Model and parameter estimation
   - Inference in multiple linear regression
   - Multicollinearity
   - Dummy variables

Simple linear regression model

- Linear regression
- History of linear regression

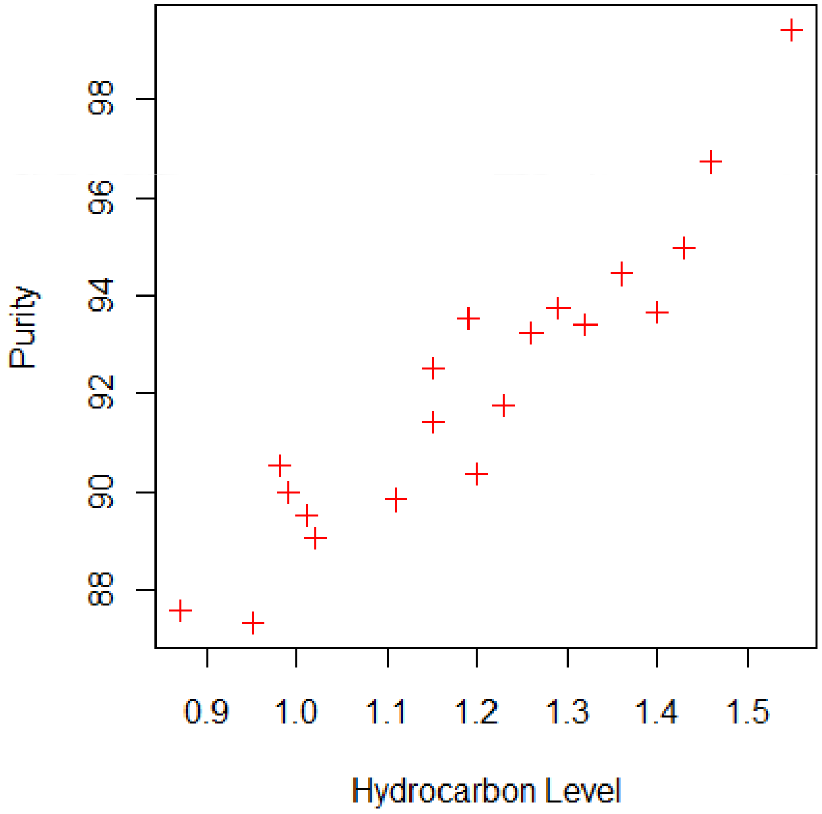
$$y = \beta_0 + \beta_1 x + u$$

Example:
Oxygen purity in a distillation process

Table 6-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Example:
Oxygen purity in a distillation process



Simple linear regression model

- n pairs (x_i, y_i)
- Aim: predict Y with information from X
- X: independent variable (explanatory or covariate)
- Y: dependent variable (to be fitted)

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

β_0 and β_1 regression coefficients

β_0 intercept

β_1 slope

Simple linear regression model

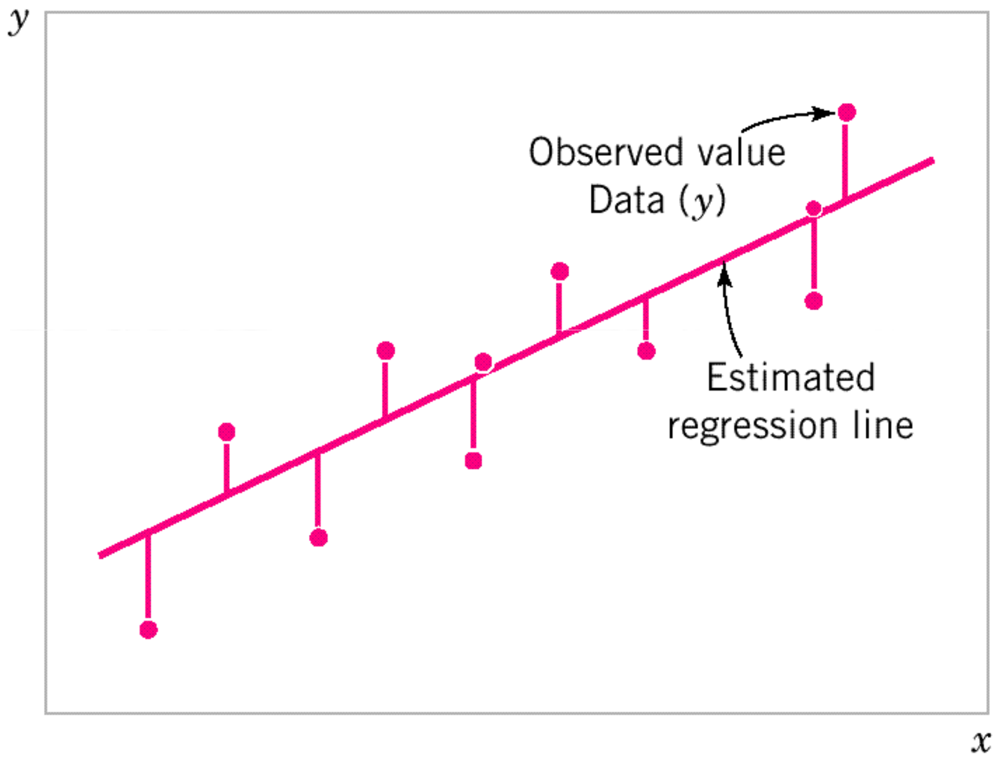


Figure 6-6 Deviations of the data from the estimated regression model.

Model assumptions

1. Linearity
2. Homogeneity, $E[u_i] = 0$
3. Homocedasticity, $\mathrm{Var}[u_i] = \sigma^2$
4. Independence, u_i indep. u_j, in particular $E[u_i u_j] = 0$
5. Normality, $u_i \sim N(0, \sigma)$

Model assumptions

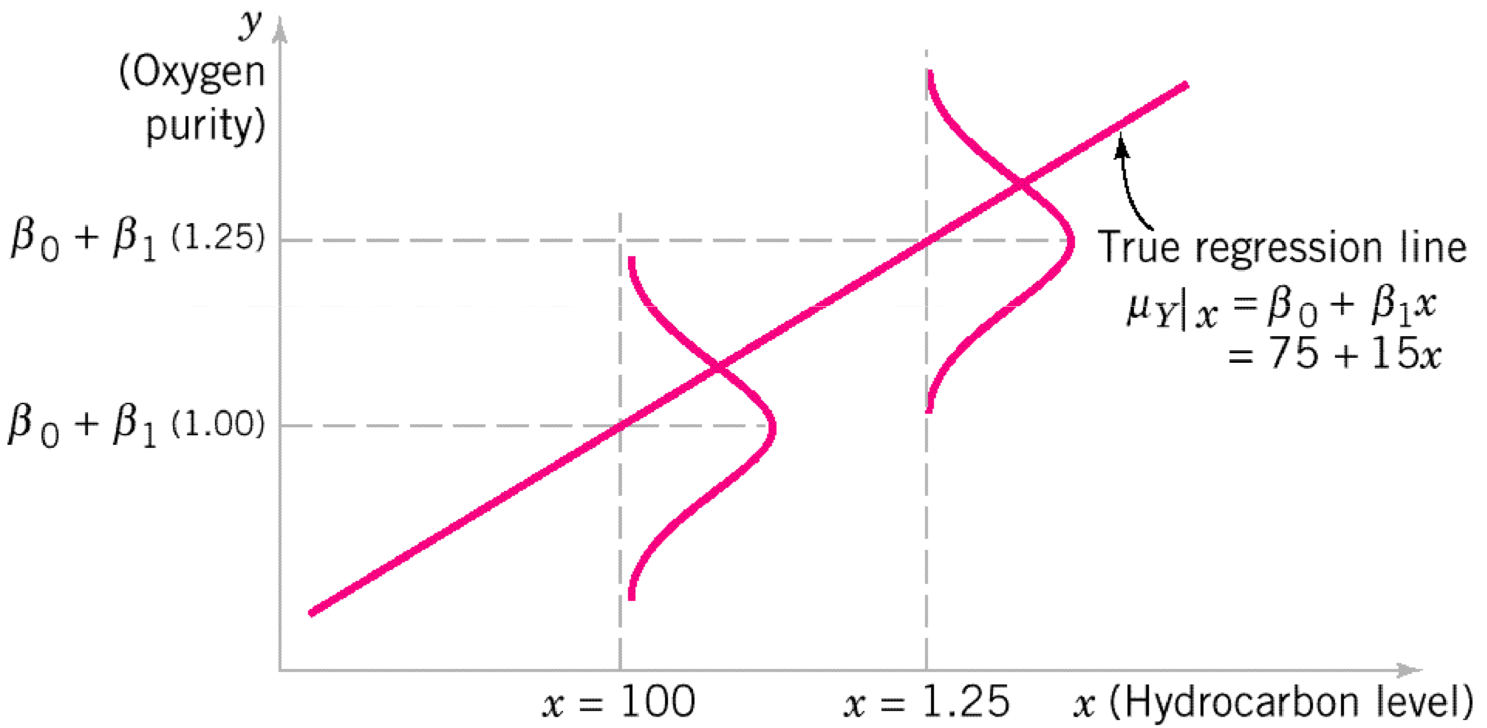
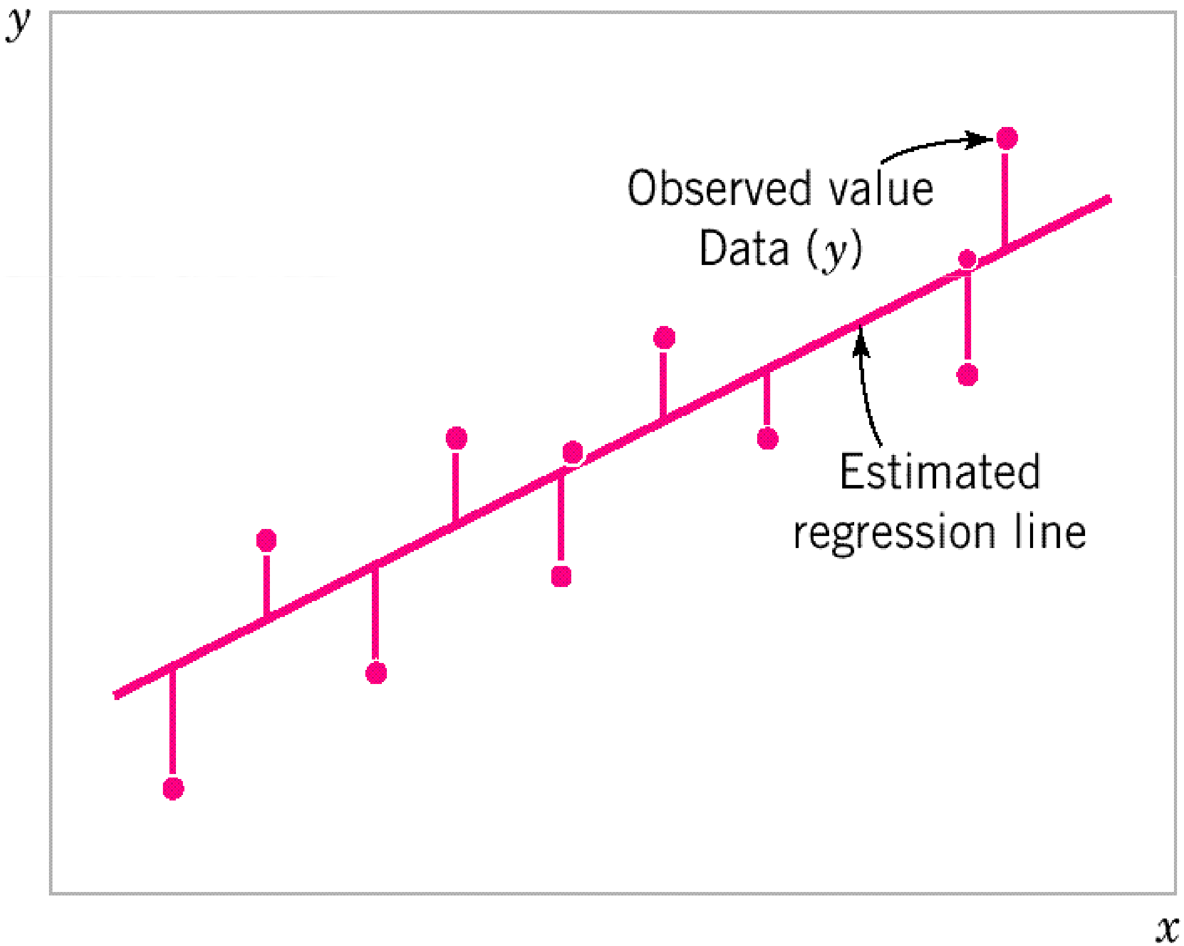


Figure 6-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

Least squares



Least squares (Gauss, 1809)

- Aim: Find β_0 and β_1 that best fit our data.
- Equation:

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

- Residual errors:

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

- Minimize:

$$\sum_{i=1}^{n} e_i^2$$

Least squares (Gauss, 1809)

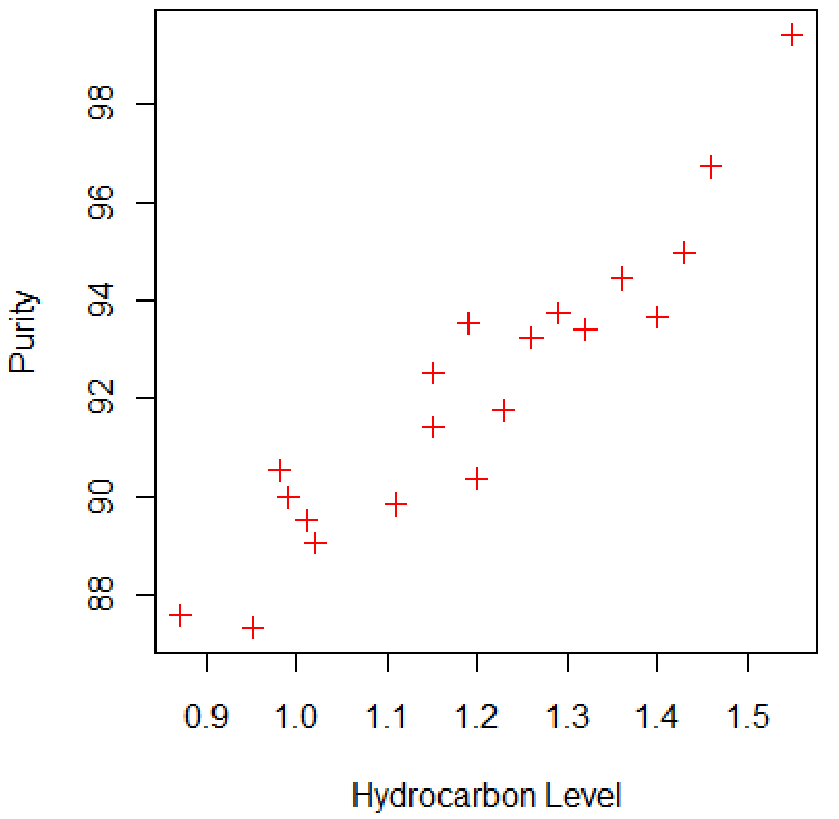
- Estimators:

$$\hat{\beta}_1 = \frac{S_{X,Y}}{S_X^2}$$

$$\hat{\beta}_0 = \bar{y} - \frac{S_{X,Y}}{S_X^2}\bar{x}$$

$$\hat{y}_i = \bar{y} + \hat{\beta}_1\left(x_i - \bar{x}\right)$$

Example:
Oxygen purity in a distillation process



Example:
Oxygen purity in a distillation process

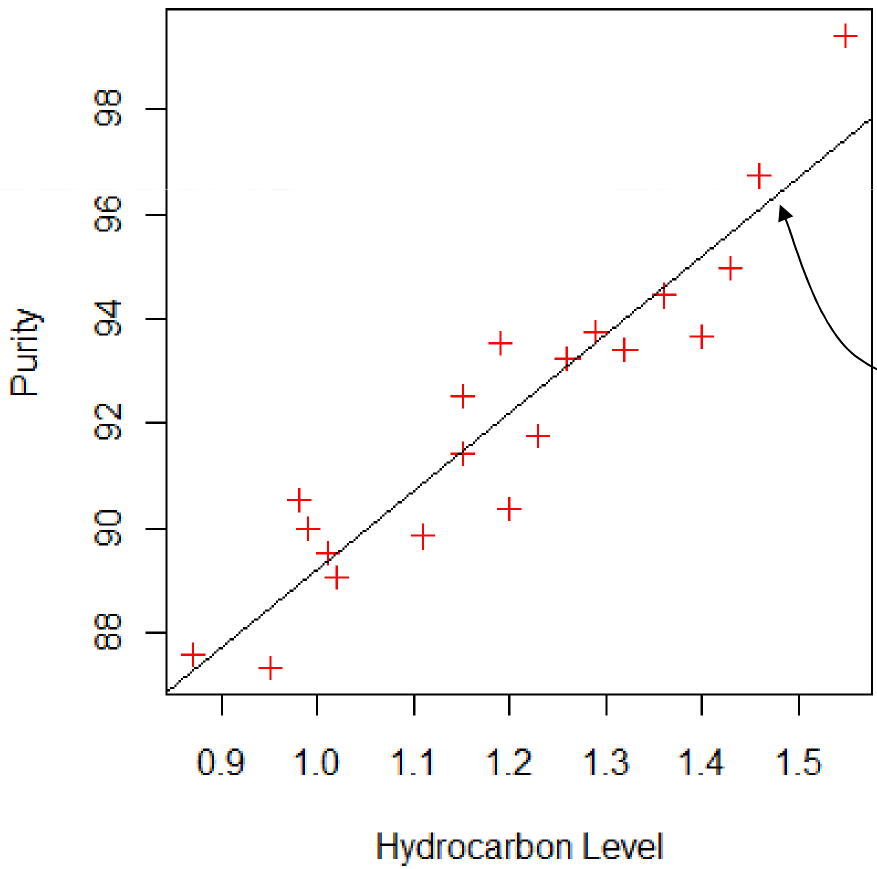
$$n = 20 \qquad \bar{x} = 1.196 \qquad \bar{y} = 92.16$$

$$S_x^2 = 0.681 \qquad S_{xy} = 10.177$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{10.177}{0.681} = 14.95 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.16 - (14.95)1.196 = 74.28$$

$$\hat{y} = 74.28 + 14.95x$$

Example:
Oxygen purity in a distillation process



$$\hat{y} = 74.28 + 14.95x$$

Example:
Oxygen purity in a distillation process

- Statistical packages (R) solve it with a click

```
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      74.28        14.95
```

Estimating the variance

Residual variance $S^2(e) = \dfrac{\sum e_i^2}{n-2}$ UNBIASED

Residual standard error $S(e) = \sqrt{\dfrac{\sum e_i^2}{n-2}}$

Inference in simple linear regression

$$\hat{\beta}_0 \sim N\left(\beta_0, \sqrt{\frac{\sigma^2}{n}\left(1+\frac{\overline{x}^2}{S_x^2}\right)}\right) \qquad S(\hat{\beta}_0) = \sqrt{\frac{S^2(e)}{n}\left(1+\frac{\overline{x}^2}{S_x^2}\right)}$$

$$\frac{\hat{\beta}_0 - \beta_0}{S(\hat{\beta}_0)} \sim t_{n-2}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \sqrt{\frac{\sigma^2}{nS_x^2}}\right) \qquad S(\hat{\beta}_1) = \sqrt{\frac{S^2(e)}{nS_x^2}}$$

$$\frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} \sim t_{n-2}$$

Inference in simple linear regression

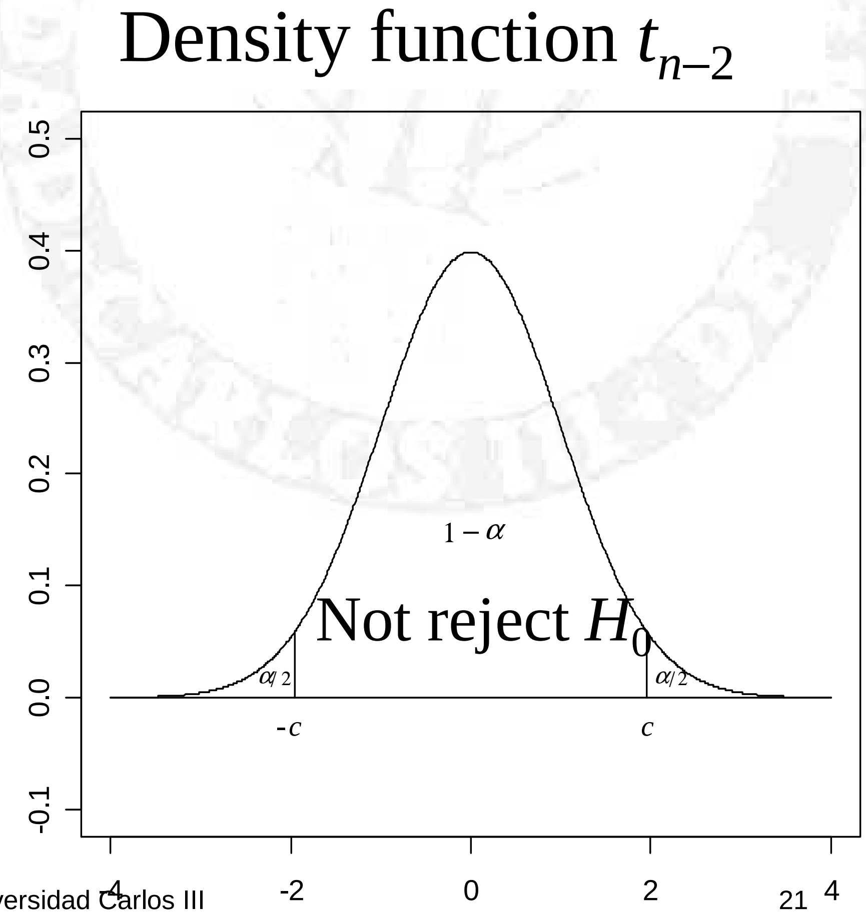
$H_0\colon \beta_i = 0$

$H_1\colon \beta_i \neq 0$

If $\quad t = \left| \dfrac{\hat{\beta}_i}{S(\hat{\beta}_i)} \right| > t_{n-2,\alpha/2}$

we reject H_0.

$c = t_{n-2,\alpha/2}$



Example:
Oxygen purity in a distillation process

```
>summary(lm(y~x))
Call:
lm(formula = y ~ x)
Coefficients:
             Estimate  Std. Error   t     value Pr(>|t|)
(Intercept)   74.283     1.593    46.62   < 2e-16 ***
x             14.947     1.317    11.35  1.23e-09 ***

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared: 0.8774,     Adjusted R-squared: 0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

Adequacy of the regression model
Sum of squares identity

- Sum of Squares identity: $SS_T = SS_R + SS_E$

$$SS_T = \text{Total Sum of Squares} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_R = \text{Regression Sum of Squares} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_E = \text{Error Sum of Squares} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- ANOVA test

$$\text{If } \beta_1 = 0, \text{ then } \frac{SS_R}{SS_E/(n-2)} \sim F_{1,n-2}$$

Example:
Oxygen purity in a distillation process

```
>summary(lm(y~x))
Call:
lm(formula = y ~ x)
Coefficients:
             Estimate  Std. Error   t     value Pr(>|t|)
(Intercept)  74.283     1.593    46.62   < 2e-16 ***
x            14.947     1.317    11.35  1.23e-09 ***

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared: 0.8774,     Adjusted R-squared: 0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

R^2 coefficient

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{nS_Y^2} = \frac{S_{X,Y}^2}{S_X^2 S_Y^2}$$

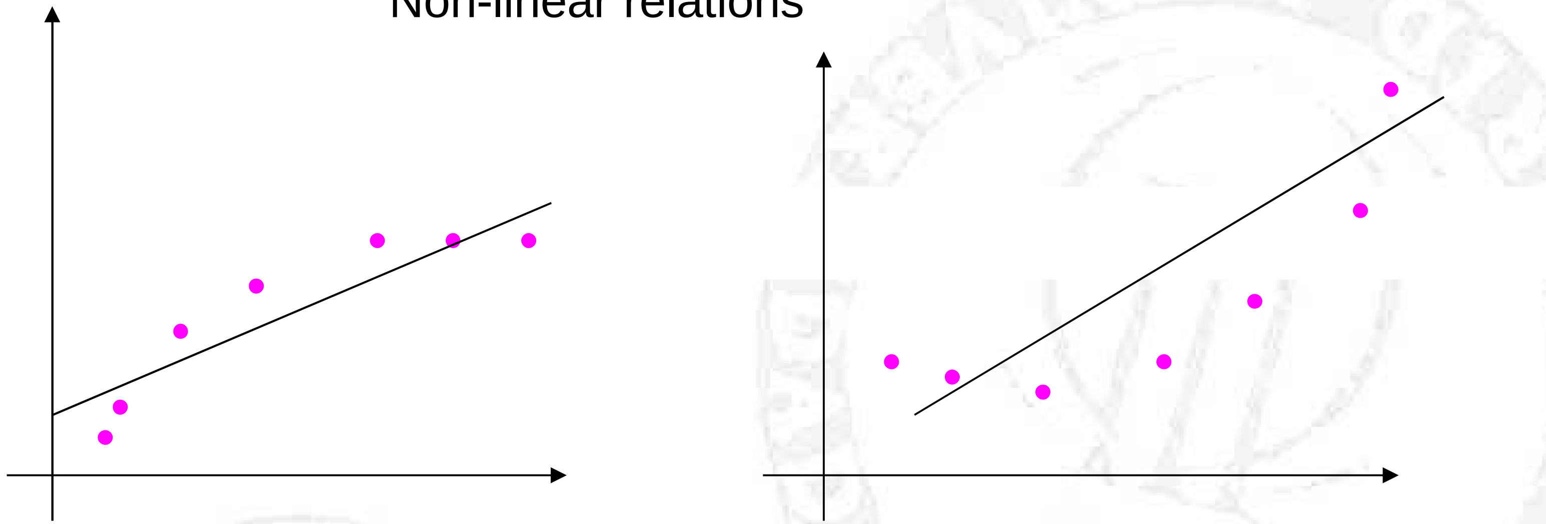
- Commonly given as a percentage.
- Represents the percentage of variability explained by the regression model.

Adequacy of the regression model
Diagnostic graphs

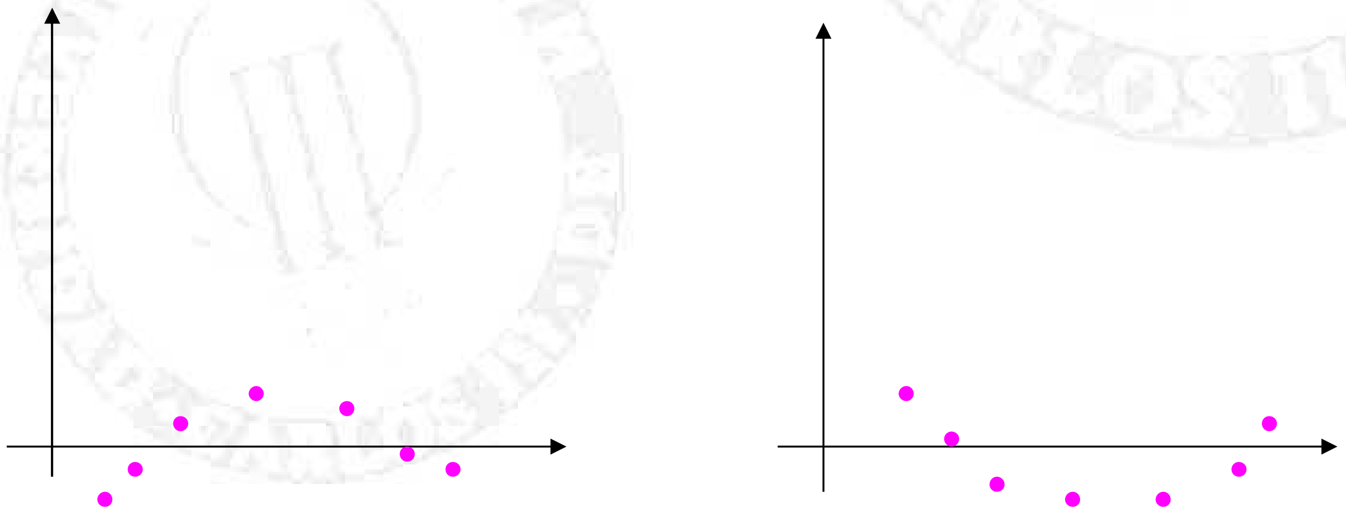
Once the regression model has been fitted:

- Study the residual errors to check that the model assumptions are fullfilled.
- If the model assumptions are not fullfilled, the variables must be transformed.

Non-linear relations

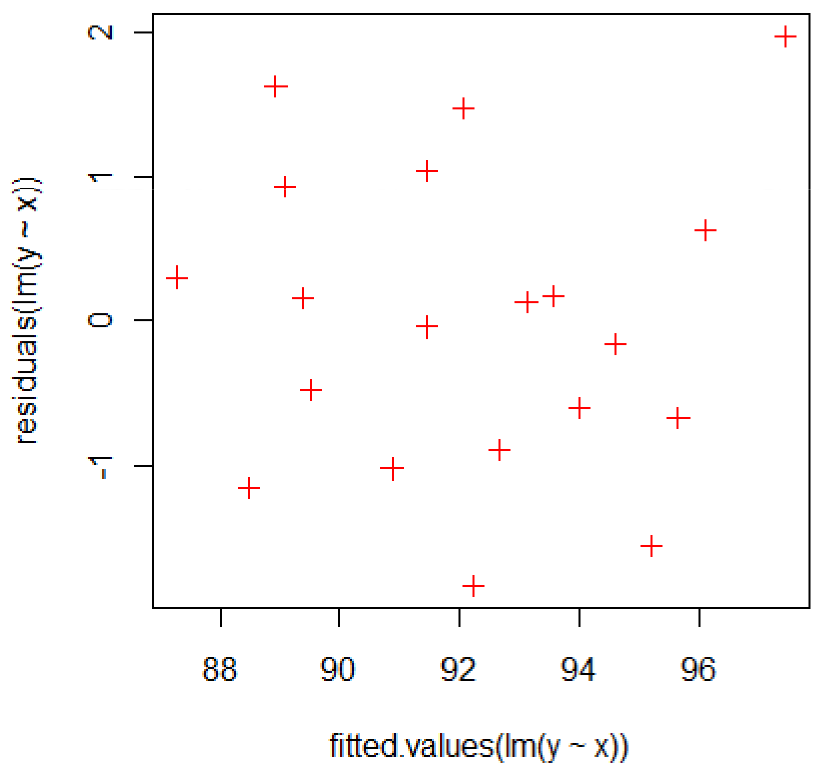


Residual graphs



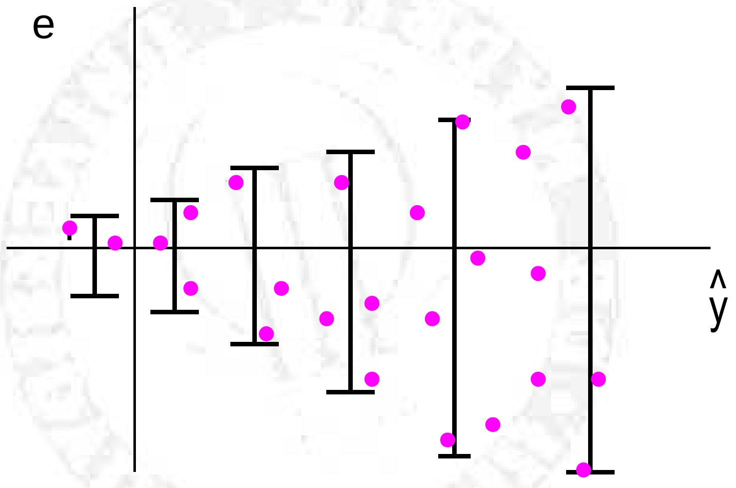
Linearity

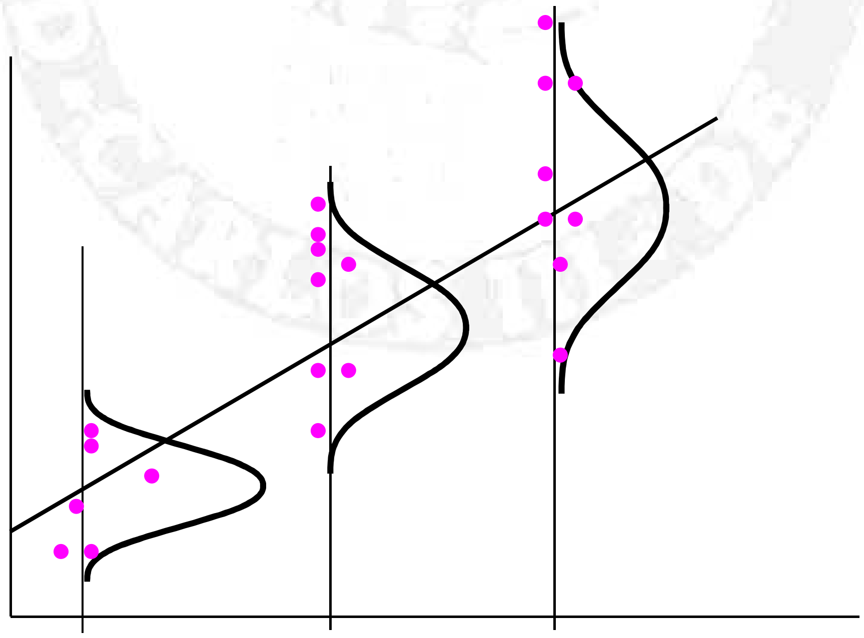
- The relationship between *x* and *y* should be linear.
- Check that there is no structure in the fitted vs. residuals graph.



Homocedasticity

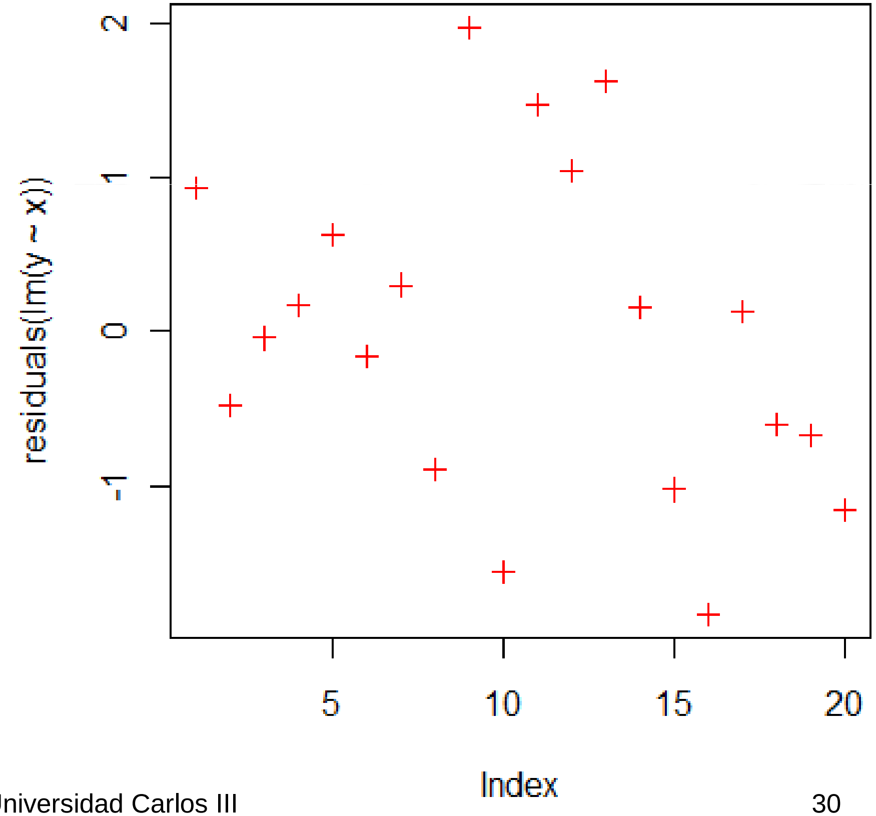
The variance of residual error must be approximately the same for all levels of the explanatory variable.





Independence

- Data should not be correlated with time.
- Check that there is no (time) tendency in residual errors.



Normality

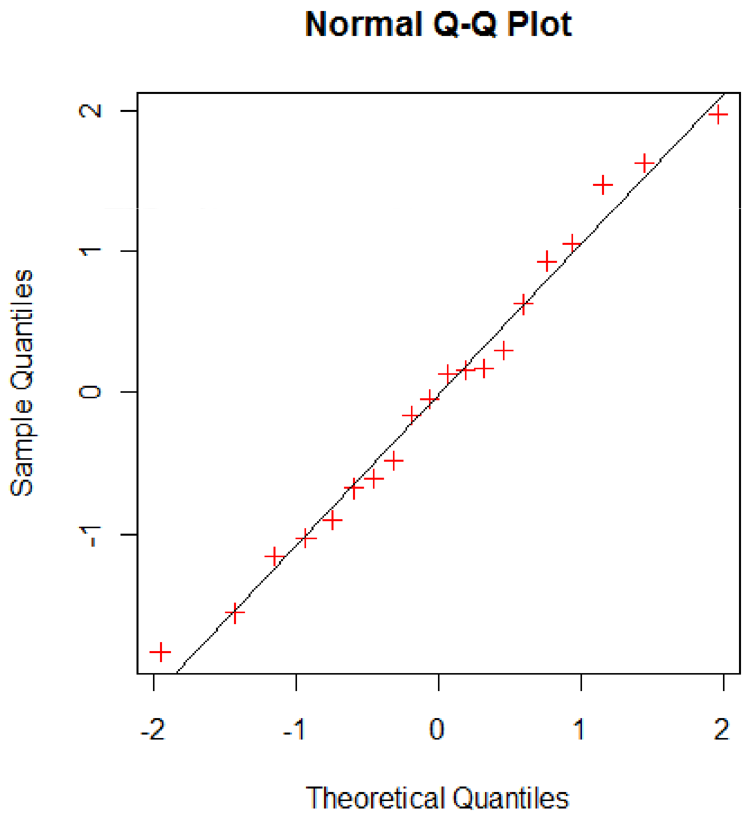
Check residual errors are normaly distributed.

> shapiro.test(residuals(lm(y~x)))

Shapiro-Wilk normality test

data: residuals(lm(y ~ x))

W = 0.9796, p-value = 0.9293



Numerical interpretation of the coefficients

Once we have determined the regression coefficients:

- $y=a+bx$

  When x is enlarged 1 unit, y enlarges b units .

- $\ln(y)=a+bx$

  When x is enlarged 1 unit, y enlarges by $100b$% .

- $\ln(y)=a+b\ln(x)$

  When x is enlarged by 1%, y enlarges by b% .

- $y=a+b\ln(x)$

  When x is enlarged by 1%, y enlarges $b/100$ units .

Outline

1. Introduction
2. Simple linear regression
   - Model and parameter estimation
   - Inference in simple linear regression
   - Adequacy of the regression model
3. Multiple linear regression
   - Model and parameter estimation
   - Inference in multiple linear regression
   - Multicollinearity
   - Dummy variables

Multiple linear regression model

- Joint study of several variables (more than two).
- Several independent variables x_i are used (jointly) to predict a depedent variable y
- Useage of all available information.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

Multiple linear regression model

- n observations $(x_{i1},\ldots, x_{ik},y_i)$
- Aim: predict y with information from $x_1,\ldots,x_k$
- $x_1,\ldots,x_k$: independent variables (regressors)
- y: dependent (response) variable (to be predicted)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

$\beta_0, \beta_1, \ldots, \beta_k$ regression coefficients

Example: Ice cream consumption

4-week periods from March 18, 1951 to July 11, 1953	Consumption of ice cream Y (pints)	Price of ice cream X_1 (dollars per pint)	Weekly family income X_2 (dollars)	Mean temperature X_3 (degrees Fahrenheit)
1	.386	.270	78	41
2	.374	.282	79	56
3	.393	.277	81	63
4	.425	.280	80	68
5	.406	.272	76	69
6	.344	.262	78	65
7	.327	.275	82	61
8	.288	.267	79	47
9	.269	.265	76	32
10	.256	.277	79	24
11	.286	.282	82	28
12	.298	.270	85	26
13	.329	.272	86	32
14	.318	.287	83	40
15	.381	.277	84	55
16	.381	.287	82	63
17	.470	.280	80	72
18	.443	.277	78	72
19	.386	.277	84	67
20	.342	.277	86	60
21	.319	.292	85	44
22	.307	.287	87	40
23	.284	.277	94	32
24	.326	.285	92	27
25	.309	.282	95	28
26	.359	.265	96	33
27	.376	.265	94	41
28	.416	.265	96	52
29	.437	.268	91	64
30	.548	.260	90	71

Y ice cream consumption

X_1 price

X_2 family income

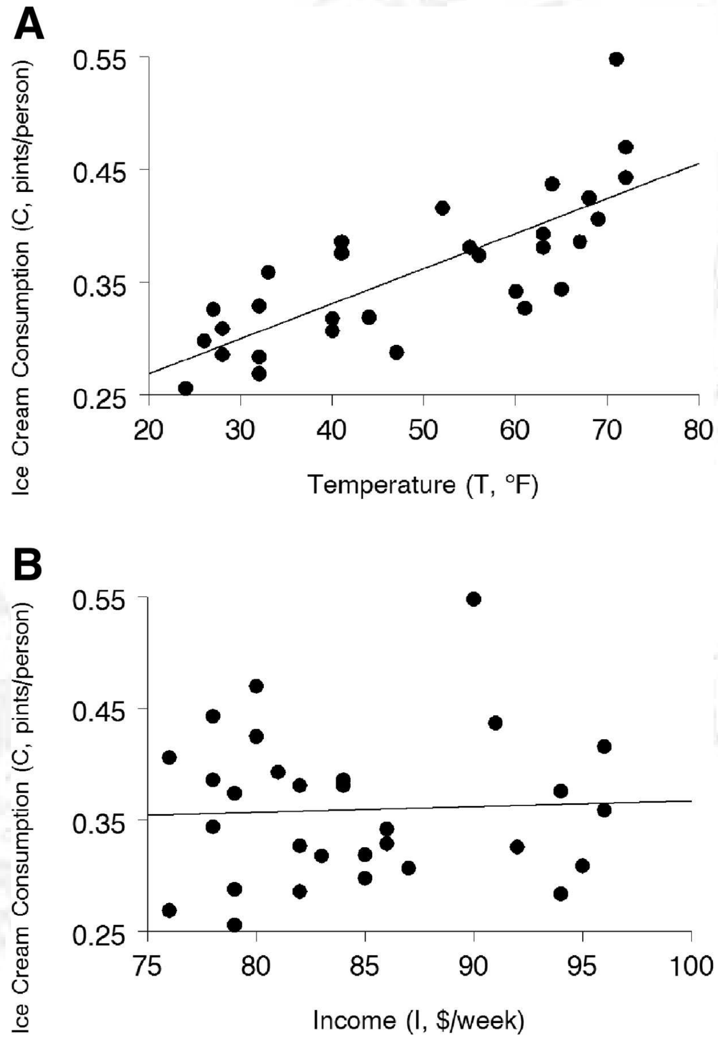
X_3 temperature

Figure 1. A, Scatterplot of ice cream consumption (C) vs temperature (T) showing the best-fit simple regression line as described by Equation 6 in the text.

Y ice cream consumption

X_2 family income

X_3 temperature

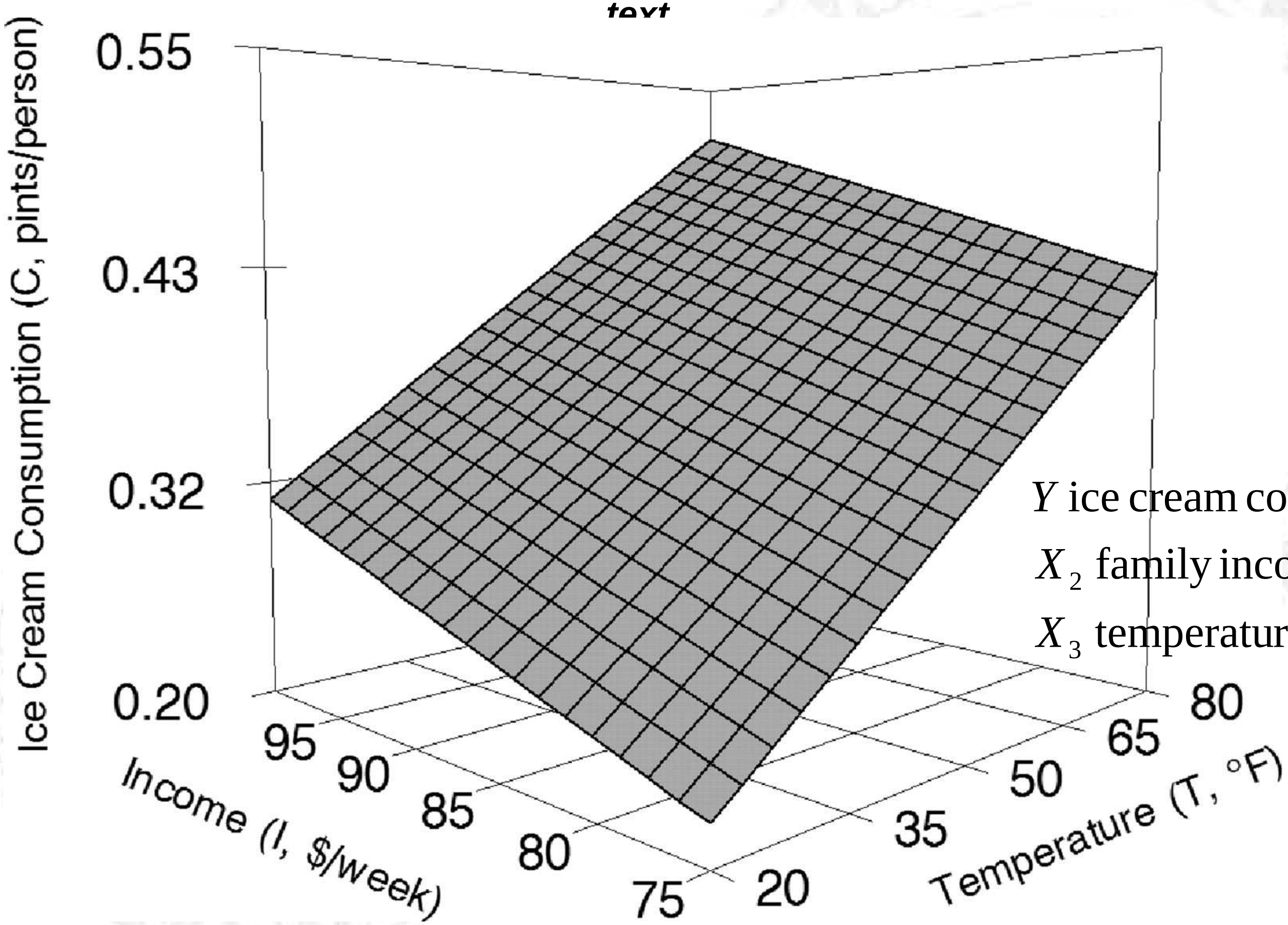


Slinker B K , Glantz S A Circulation 2008;117:1732-1737

American Heart Association

Learn and Live

Figure 2. Three-dimensional plot of the best-fit multiple regression plane relating ice cream consumption (C) to both temperature (T) and income (I), as described by Equation 9 in the text



Y ice cream consumption

X_2 family income

X_3 temperature

Slinker B K , Glantz S A Circulation 2008;117:1732-1737

American Heart Association

Learn and Live

Model assumptions

1. Linearity, data approx. belong to a hyperplane
2. Homogeneity, $E[u_i]= 0$
3. Homocedasticity, $\mathrm{Var}[u_i]=\sigma^2$
4. Independence, u_i indep. u_j, in particular $E[u_i u_j]= 0$
5. Normality, $u_i \sim N(0, \sigma)$

More about the assumptions

- Summarizing:

$$y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \sigma\right)$$

Extra hypothesis

- The sample size (n) is greater than k+1
- The explanatory variables are linearly independent.

Matrix approach to linear regression

- We can write the multiple linear regression model as

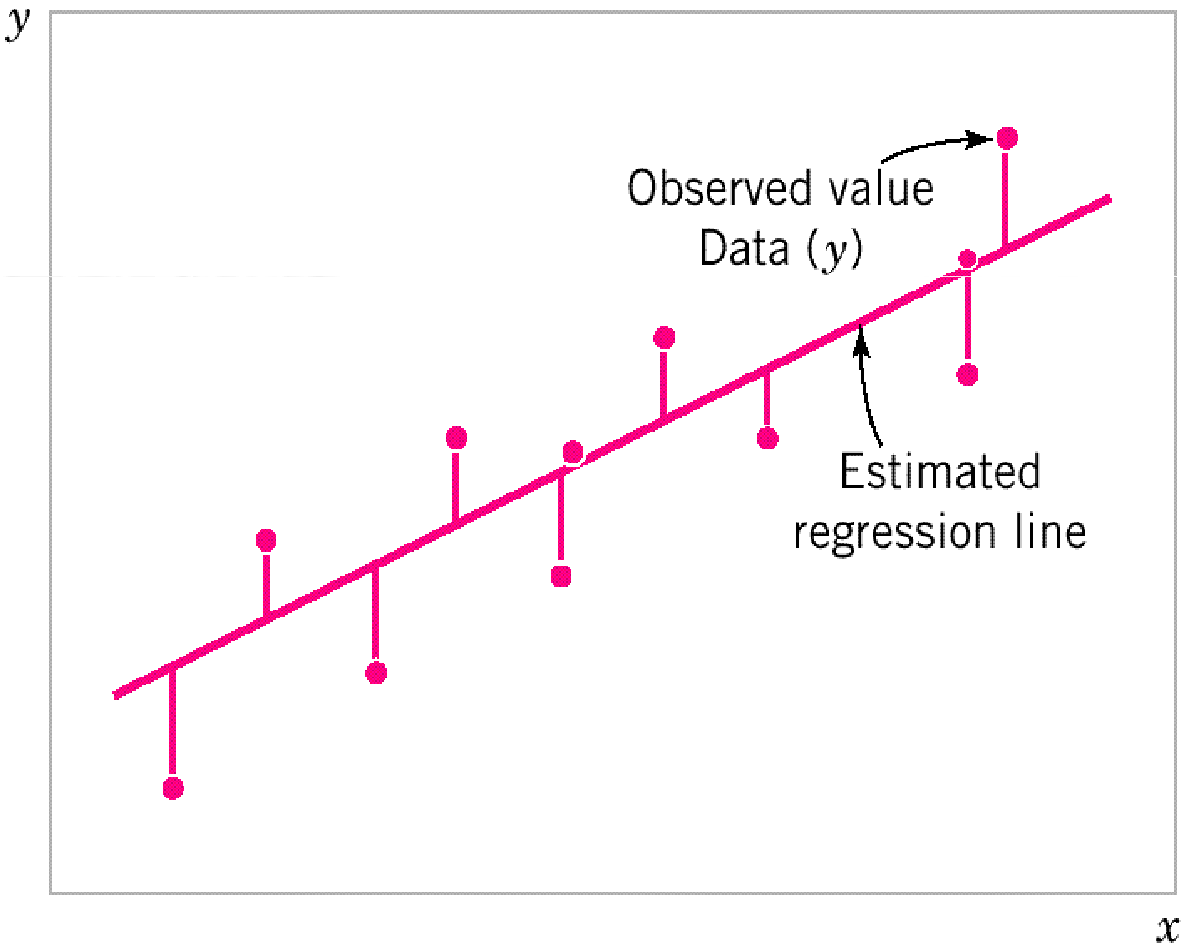
$$Y = X\beta + U$$

with:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} ; \; X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} ; \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} ; U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$U \sim N\left(0_n, \sigma^2 I_n\right) \; ; \; Y \sim N\left(X\beta, \sigma^2 I_n\right)$$

Least squares



Least squares

- Aim: Find $\beta_0, \beta_1, \ldots, \beta_k$ that best fit our data.
- Equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$$

- Residual errors:

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}\right)$$

- Minimize:

$$\sum_{i=1}^{n} e_i^2$$

Least squares estimators

- Estimator of the regression coefficients:

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y$$

Example: Ice cream comsumption

> lm(y~x2+x3)

Call:

lm(formula = y ~ x2 + x3)

Coefficients:

(Intercept)	x2	x3
-0.113195	0.003530	0.003543

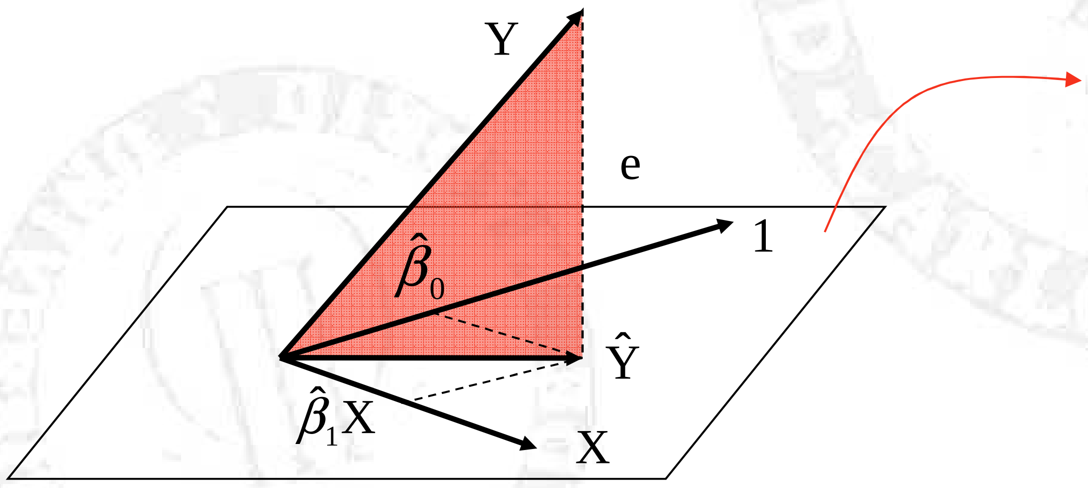
Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

Geometrical interpretation



Vector subspace spanned by the columns of *X*

Variance estimation

- The variance σ^2 is estimated by means of the residual variance

$$S^2(e) = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

- It is an unbiased estimator of σ^2 and further

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{n-k-1}$$

Inference in multiple linear regression

$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y$ is normally distributed, thus

$$\hat{\beta} \sim N\left(\beta, \sqrt{\sigma^2 \left(X^t X\right)^{-1}}\right) \quad ; \quad \hat{\beta}_{i-1} \sim N\left(\beta_{i-1}, \sqrt{\sigma^2 \left(X^t X\right)^{-1}_{ii}}\right)$$

$$Var\left[\hat{\beta}_{i-1}\right] = \sigma^2 \left(X^t X\right)^{-1}_{ii}$$

Variance σ^2 is usually unknown and estimated by the residual variance

$$S\left(\hat{\beta}_{i-1}\right) = \sqrt{\left(X^t X\right)^{-1}_{ii} S^2(e)}$$

Inference in multiple linear regression

In order to determine whether x_i contributes significantly to the multiple linear regression model, we must test

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0.$$

The null hypothesis is rejected if:

$$\left| \frac{\hat{\beta}_i}{S(\hat{\beta}_i)} \right| > t_{n-k-1,\alpha/2}$$

Example: Ice cream comsumption

```
> summary(lm(y~x2+x3))
Call:
lm(formula = y ~ x2 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.113195   0.108280  -1.045  0.30511
x2           0.003530   0.001170   3.017  0.00551 **
x3           0.003543   0.000445   7.963 1.47e-08 ***

Residual standard error: 0.03722 on 27 degrees of freedom
Multiple R-squared: 0.7021,     Adjusted R-squared:  0.68
F-statistic: 31.81 on 2 and 27 DF,  p-value: 7.957e-08
```

Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

Sum of squares identity

- Sum of Squares identity: $SS_T = SS_R + SS_E$

$$SS_T = \text{Total Sum of Squares} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_R = \text{Regression Sum of Squares} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_E = \text{Error Sum of Squares} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

R^2 coefficient

- The R^2 coefficient is given by:

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SS_E}{SS_T}$$

- The adjusted R^2 coefficient takes into account the number of model parameters, and only increases if the residual variance decreases

$$\bar{R}^2 = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 1 - \frac{S^2(e)}{SS_T/(n-1)}$$

Example: Ice cream comsumption

```
> summary(lm(y~x2+x3))
Call:
lm(formula = y ~ x2 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.113195   0.108280  -1.045  0.30511
x2           0.003530   0.001170   3.017  0.00551 **
x3           0.003543   0.000445   7.963 1.47e-08 ***

Residual standard error: 0.03722 on 27 degrees of freedom
Multiple R-squared: 0.7021,     Adjusted R-squared:  0.68
F-statistic: 31.81 on 2 and 27 DF,  p-value: 7.957e-08
```

Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

ANOVA test

If $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, then $\dfrac{\mathrm{SS_R}/k}{\mathrm{SS_E}/(n-k-1)} \sim F_{k,n-k-1}$

We can check wheter there exist some linear relation between the response variable and the regressors testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \text{some } \beta_i \neq 0.$$

The null hypothesis is rejected if:

$$\frac{\mathrm{SS_R}/k}{\mathrm{SS_E}/(n-k-1)} > F_{k,n-k-1,\alpha}$$

Example: Ice cream comsumption

```
> summary(lm(y~x2+x3))
Call:
lm(formula = y ~ x2 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.113195   0.108280  -1.045  0.30511
x2           0.003530   0.001170   3.017  0.00551 **
x3           0.003543   0.000445   7.963 1.47e-08 ***

Residual standard error: 0.03722 on 27 degrees of freedom
Multiple R-squared: 0.7021,     Adjusted R-squared:  0.68
F-statistic: 31.81 on 2 and 27 DF,  p-value: 7.957e-08
```

Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

Multicollinearity

- Multicollinearity is a common problem in multiple linear regression. It appears when there exist strong dependencies among the regressor variables x_i
- In the presence of multicollinearity $det(X^tX) \cong 0$
- It often happens that all indenpendent variables contribute significantly to their simple models, but not to the multiple model.

Selection of variables & diagnostic graphs

- The best model is selected among the ones all whose independent variables contribute significantly to it.
- To keep the number of variables reasonably low, we choose the model with the highest adjusted R^2 coefficient.
- Diagnositic graphs: As for simple regression.

Example: Ice cream comsumption

```
> summary(lm(y~x1))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9230     0.3964   2.329   0.0273 *
x1           -2.0472     1.4393  -1.422   0.1660

> summary(lm(y~x2))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.316715   0.168665   1.878   0.0709 .
x2          0.000505   0.001988   0.254   0.8014

> summary(lm(y~x3))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2068621  0.0247002   8.375 4.13e-09 ***
x3          0.0031074  0.0004779   6.502 4.79e-07 ***
```

Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

Example: Ice cream comsumption

```
> summary(lm(y~x2+x3))
Call:
lm(formula = y ~ x2 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.113195   0.108280  -1.045  0.30511
x2           0.003530   0.001170   3.017  0.00551 **
x3           0.003543   0.000445   7.963 1.47e-08 ***

Residual standard error: 0.03722 on 27 degrees of freedom
Multiple R-squared: 0.7021,     Adjusted R-squared:  0.68
F-statistic: 31.81 on 2 and 27 DF,  p-value: 7.957e-08
```

Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

Example: Ice cream comsumption

```
> summary(lm(y~x1+x2))
Call:
lm(formula = y ~ x1 + x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9002400  0.4550344   1.978   0.0582 .
x1          -2.0300382  1.4738940  -1.377   0.1797
x2           0.0002135  0.0019687   0.108   0.9144

Residual standard error: 0.06583 on 27 degrees of freedom
Multiple R-squared: 0.0678,     Adjusted R-squared: -0.001257
F-statistic: 0.9818 on 2 and 27 DF,  p-value: 0.3876
```

Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

Example: Ice cream comsumption

```
> summary(lm(y~x1+x3))
Call:
lm(formula = y ~ x1 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.59655    0.25831   2.309   0.0288 *
x1          -1.40176    0.92509  -1.515   0.1413
x3           0.00303    0.00047   6.448 6.56e-07 ***

Residual standard error: 0.04132 on 27 degrees of freedom
Multiple R-squared: 0.6328,     Adjusted R-squared: 0.6056
F-statistic: 23.27 on 2 and 27 DF,  p-value: 1.336e-06
```

Y ice cream consumption

X_1 price

X_2 family income

X_3 temperature

Example: Ice cream comsumption

```
> summary(lm(y~x1+x2+x3))
Call:
lm(formula = y ~ x1 + x2 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
x1          -1.0444140  0.8343573  -1.252  0.22180
x2           0.0033078  0.0011714   2.824  0.00899 **
x3           0.0034584  0.0004455   7.762  3.1e-08 ***

Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-squared: 0.719,      Adjusted R-squared: 0.6866
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

Y ice cream consumption

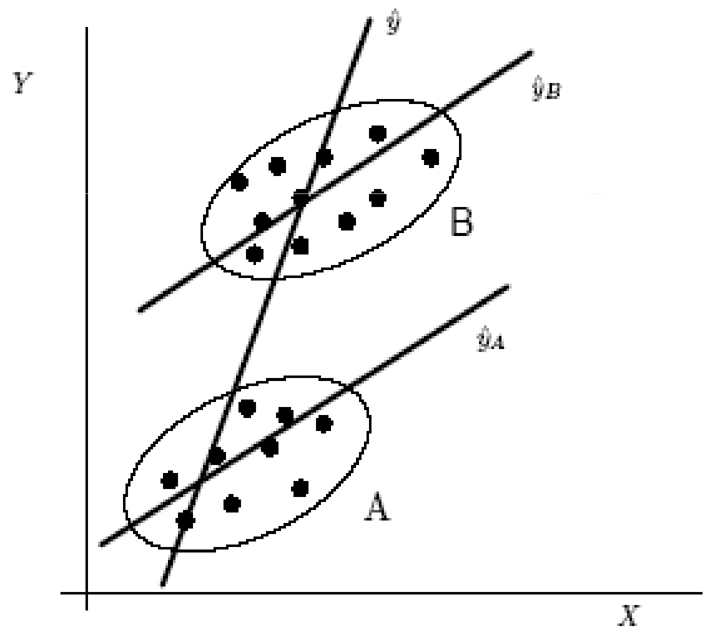
X_1 price

X_2 family income

X_3 temperature

Dummy variables

In a sample we might have observations from two different groups.



Example: Female and male individuals.

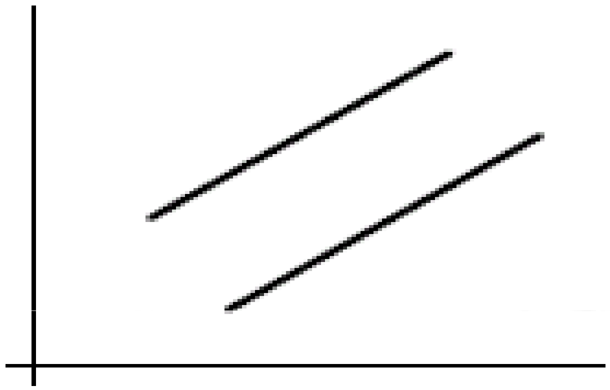
Dummy variables

- A dummy (or indicator) variable represents the group:

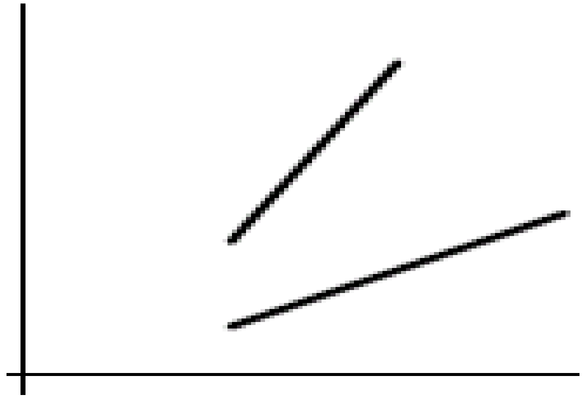
$$d_i = \begin{cases} 0 \text{ if the } i\text{-th observation belongs to group A} \\ 1 \text{ if the } i\text{-th observation belongs to group B} \end{cases}$$

Dummy variables

$$y = \beta_0 + \beta_1 x + \beta_2 d + u$$



$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + u$$



Dummy variables

- It may happen that there are more than two groups.
- In case we have s groups, we must introduce $s-1$ dummy variables d_t , $1 \leq t \leq s-1$

$$d_{it} = \begin{cases} 1 & \text{if the } i\text{-th observation belongs to group } t \\ 0 & \text{otherwise} \end{cases}$$

$$y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_3 + u$$