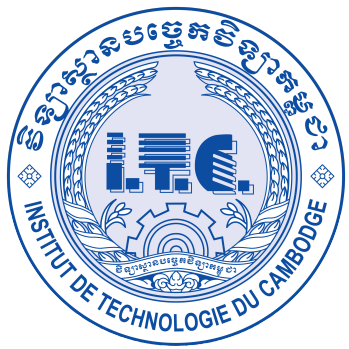
Regression Analysis

Chapter 02
Simple Linear Regression

Phauk Sokkhey
phauk.sokkhey@itc.edu.kh
Nhim Malai
nhim.malai@itc.edu.kh

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia



Contents

Standard Model of Simple Linear Regression

Simple Linear Regression Model

- Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.
- *The key point in Simple Linear Regression is that the dependent variable must be a continuous/real value. However, the independent variable can be measured on continuous, discrete or categorical values.*

Standard Model of Simple Linear Regression

Simple Linear regression algorithm has mainly two objectives:

- **Model the relationship between the two variables:** Such as the relationship between Income and expenditure, experience and Salary, etc.
- **Forecasting new observations:** Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.

Example: Munich Rent Index

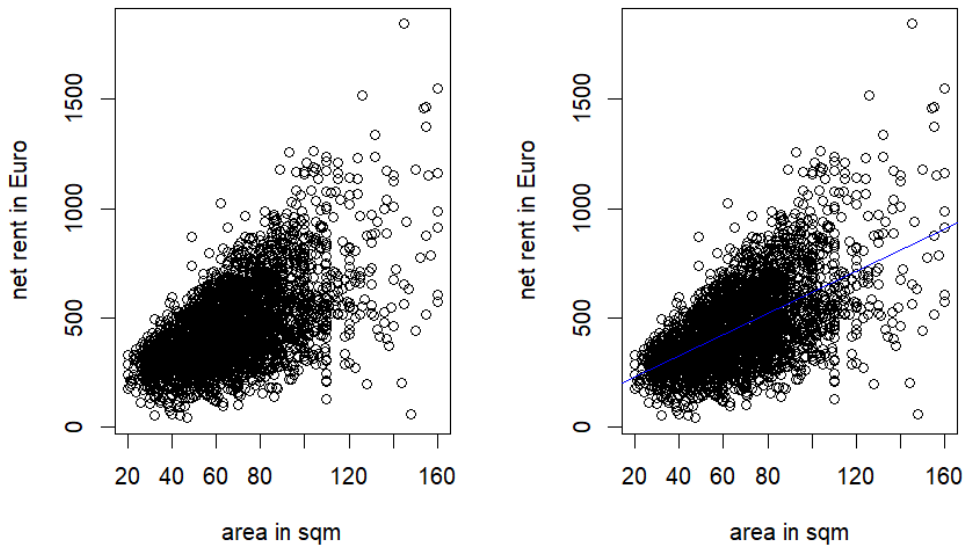


Figure: Munich rent index: scatter plot between net rent and area for apartments in 1999 (left). In the right panel, a regression line is additionally included.

- The scatter plot displays an approximate linear relationship between *rent* and *area*, i.e.,

$$rent_i = \beta_0 + \beta_1 \cdot area_i + \varepsilon_i$$

Standard Model of Simple Linear Regression

Simple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, ..., n$$

where ε_i i.i.d and $\varepsilon_i \sim N(0, \sigma^2)$

- n: total number of observations (subjects, elements) in the dataset
- x_i: the value of regressor of observation $i = 1, ..., n$
- y_i: the outcome of observation $i = 1, ..., n$

Unknown parameters

- β_0 (Intercept): point in which the line intercepts the y-axis;
- β_1 (Slope): increase in Y per unit change in X.

Assumptions of (Simple) Linear Regressions

Assumptions for (Simple) Linear Regressions:

- **Linearity**: the relationship between x and y is linear.
- **Independence of errors**: there is no relationship between residuals and response y; in other words, y is independent of each other.
- **Normality of errors**: the residuals must be approximately normally distributed.
- **Equal variance (homoscedasticity)**: the variance of the residuals is the same for all values of x.

Interpretation of Parameters

From

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

Due to the assumption of $\varepsilon_i \sim N(0, \sigma^2)$ or $E(\varepsilon_i) = E(\varepsilon_i | x_i) = 0$, we obtain a useful property of conditional mean:

$$E(y|x) = \beta_0 + \beta_1 x$$

Interpretation of Parameters

- The interpretation of the parameter β_1 follows from the identity

$$E(y|x+1) - E(y|x) = [\beta_0 + \beta_1(x+1)] - [\beta_0 + \beta_1 x] = \beta_1$$

  The parameter β_1 is thus the average increase in the outcome when the regressor increases with one unit. This parameter is also the slope of the regression line. The parameter is often referred to as the regression coefficient.

- The interpretation of the parameter β_0 follows from the identity

$$E(y|x=0) = \beta_0 + \beta_1 \cdot 0 = \beta_0$$

  The parameter β_0 is thus the average outcome when the regressor takes the value zero. It is the intercept of the regression line.

Example: Munich Rent Index

Model

$$rent_i = \beta_0 + \beta_1 \cdot area_i + \varepsilon_i, \qquad i = 1, ..., 3082$$

where ε_i is i.i.d and $\varepsilon_i \sim N(0, \sigma^2)$

- $rent_i$: the monthly net rent per month (in Euro) for observation $i = 1, ..., n$
- $area_i$: living area in square meters for observation $i = 1, ..., n$

Contents

Least Square Estimator (LSE)

- LSE minimizes *sum of square errors*.
- The *fitted value* for observation i is given by,

$$\hat{y} = \hat{E}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sum of square errors is denoted by,

$$SSE(\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

- LSE can be written as:

$$\hat{\beta} = \mathbf{ArgMin}_{\beta} SSE(\beta)$$

Matrix Notation

- Parameter vector: $\boldsymbol{\beta}^t = (\beta_0, \beta_1)$
- Estimate: $\hat{\boldsymbol{\beta}}^t = (\hat{\beta}_0, \hat{\beta}_1)$
- Outcome vector: $\boldsymbol{Y}^t = (y_1, ..., y_n)$
- Design matrix ($n \times 2$ matrix):

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

  The ith row of X is represented by $\boldsymbol{x}_i^t = (1, x_i)$
- Error vector: $\boldsymbol{\varepsilon}^t = (\varepsilon_i, ..., \varepsilon_n)$

Matrix Notation

- Simple regression model can be written as:

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i$$

or as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with ε_i i.i.d and $E(\varepsilon_i) = E(\varepsilon_i|x_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$

- Sum of square errors becomes:

$$SSE(\beta) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

- LSE becomes:

$$\hat{\beta} = \mathbf{ArgMin}_{\beta} RSS(\beta) = \mathbf{ArgMin}_{\beta} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

LSE for Simple Linear Regression

Theorem 1 (LSE for (Simple) Linear Regression).

Assume that the model is correct and the $n \times 2$ design matrix has rank 2. Then, the LSE of β is given by:

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

and this solution is unique.

After working out the matrix formulation:

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Proof

We have

$$\begin{aligned}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 &= (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{Y}^t\boldsymbol{Y}-\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{Y}-\boldsymbol{Y}^t\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}^t\end{aligned}$$

Applying vector differentiation,

$$\frac{d}{d\boldsymbol{\beta}}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = -2\boldsymbol{X}^t\boldsymbol{Y}+2\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}$$

The LSE of $\boldsymbol{\beta}$ satisfies

$$\frac{d}{d\boldsymbol{\beta}}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = 0$$

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{Y}$$

The solution ($\boldsymbol{X}$ has full rank and hence $\boldsymbol{X}^t\boldsymbol{X}$ is invertible) is thus given by

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

Proof (continue)

Show that the matrix partial derivative of second order (Hessian matrix) is positive definite

$$\frac{d^2}{d\beta d\beta^t}\|\boldsymbol{Y}-\boldsymbol{X}\beta\|^2=\frac{d}{d\beta}(-2\boldsymbol{X}^t\boldsymbol{Y}+2\boldsymbol{X}^t\boldsymbol{X}\beta)=2\boldsymbol{X}^t\boldsymbol{X}$$

The 2×2 matrix $\boldsymbol{X}^t\boldsymbol{X}$ is positive definite because $\boldsymbol{X}$ is full of rank.

Show that the solution is unique Suppose that there are two different solutions (say $\hat{\beta}_1$ and $\hat{\beta}_2$ such that $\hat{\beta}_1\neq\hat{\beta}_2$), then it holds that

$$\boldsymbol{X}^t\boldsymbol{Y}=\boldsymbol{X}^t\boldsymbol{X}\hat{\beta}_1=\boldsymbol{X}^t\boldsymbol{X}\hat{\beta}_2$$

Hence, $\boldsymbol{X}^t\boldsymbol{X}(\hat{\beta}_1-\hat{\beta}_2)=0$. Because $\boldsymbol{X}^t\boldsymbol{X}$ is of full rank, the unique solution of the equation given by $\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\nu}$ is provided by the null solution $\boldsymbol{\nu}=\boldsymbol{0}$. Therefore the following equality must hold true: $\hat{\beta}_1-\hat{\beta}_2=\boldsymbol{0}$. Hence, the supposition $\hat{\beta}_1\neq\hat{\beta}_2$ gives a contradiction and hence $\hat{\beta}_1=\hat{\beta}_2$ must be true, i.e. there is only one unique solution.

Proof (continue)

This solution could also be found by computing the partial derivatives of *SEE* with respect to the two parameters, setting these derivatives to zero and solving the system of equations for the two parameters (our solution applies to multiple linear regression).

$$\frac{\partial}{\partial \beta_0} SSE(\boldsymbol{\beta}) = 0 \qquad \frac{\partial}{\partial \beta_1} SSE(\boldsymbol{\beta}) = 0$$

This gives

$$\frac{\partial}{\partial \beta_0} SSE(\boldsymbol{\beta}) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial}{\partial \beta_1} SSE(\boldsymbol{\beta}) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Example: Munich Rent Index

R-code

```
library(gamlss.data)
data(rent99)

slm <- lm(rent~area, data = rent99)
slm
```

Output

```
##Call:
##lm(formula = rent ~ area, data = rent99)
##
##Coefficients:
##(Intercept)         area
##    134.592        4.821
```

The estimated (or fitted) regression line is given by

$$\hat{y}_i = 134.60 + 4.82 x_i$$

- The slope parameter $\hat{\beta}_1 = 4.82$ can be interpreted as if the living area increase by $1m^2$, the rent increase by about 4.82 Euro on average.
- The intercept parameter $\beta_0 = 134.60$ has no direct physical interpretation because there are no apartments with an area of $0m^2$. This issue can be resolved by first centering the regressor which is illustrated next.

R-code

```
library(dplyr)
rent99 <- rent99 %>%
  mutate(area.centered = area - mean(area))
mean(rent99$area)
slm2 <- lm(rent~area.centered, data = rent99)
slm2
```

Output

```
##[1] 67.37476
##
##Call:
##lm(formula = rent ~ area.centered, data = rent99)
##
##Coefficients:
##  (Intercept)  area.centered
##      459.437          4.821
```

Now the intercept is estimated by $\hat{\beta}_0 = 459.44$. When the centered regressor area is equal to 0 meaning that the average area is $67.37m^2$ with the net rent on average estimated to be 459.44 Euro.

Mean and Variance of LSE

Theorem 2 (Mean and Variance of the LSE).

Assume that the model is correct and $rank(\boldsymbol{X}) = 2$ $(2 \leq n)$, *then the following holds:*

1. $E(\hat{\boldsymbol{\beta}}) = \beta$ *(the LSE is an* **unbiased** *estimator of* β*)*
2. $Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2$

After working out the matrix multiplication and inversion:

$$Var(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Proof

Part 1: The unbiasedness of $\hat{\boldsymbol{\beta}}$

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((\boldsymbol{X}^t\boldsymbol{X})^{-1})\boldsymbol{X}^t\boldsymbol{Y} \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t E(\boldsymbol{Y}) \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Proof (continue)

Part 2: For the covariance matrix of $\hat{\beta}$ we will need $Var\,\boldsymbol{Y}$. On the diagonal of this matrix we find $Var(y_i) = Var(\varepsilon_i) - \sigma^2$ and on the off-diagonal positions we need the covariances $Cov(y_i, y_j)$ where $i \neq j$. All these covariances are equal to zero because the independence between outcomes is assumed. Hence, the covariance matrix of $\hat{\beta}$ becomes

$$\begin{aligned} Var(\hat{\beta}) &= Var((\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}) \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}\,Var(\boldsymbol{Y})[(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t]^t \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2 \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2 \end{aligned}$$

An Estimator of Variance σ^2

Theorem 3 (An unbiased estimator of σ^2).

Assume that the model is correct and let $\boldsymbol{x}_i^t$ denotes the ith row of $\boldsymbol{X}(i = 1, ..., n)$. Then

$$MSE = \frac{SEE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\hat{\boldsymbol{\beta}})^2}{n-2} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-2}$$

is an unbiased estimator of σ^2

Theorem 4 (Sampling distribution of MSE).

Assume that the model is correct. Then

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$$

Contents

Distribution of the Standardised and the Studentised LSE

- $\boldsymbol{\sigma}^2_{\beta_j} = Var(\hat{\beta}_j)$ is the variance of $\hat{\beta}_j$ where $j = 0, 1$. This is thus that appropriate diagonal element of $Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2$ which is also denoted by $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$.

Corollary 5 (Distribution of the standardised LSE).

Assume that the model is correct. Then, the standardised parameter estimator of β_j is

$$\frac{\hat{\beta}_j - \beta_j}{\boldsymbol{\sigma}_{\beta_j}} \sim N(0, 1)$$

As $n \to \infty$,

$$\frac{\hat{\beta}_j - \beta_j}{\boldsymbol{\sigma}_{\beta_j}} \xrightarrow{d} N(0, 1)$$

- The variance σ^2 is unknown and replaced by its estimator MSE.
- The estimator of $Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2$ is denoted by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}MSE$. The estimator of $\boldsymbol{\sigma}^2_{\boldsymbol{\beta}_j}$ is denoted by $\hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}_j}$ or $\boldsymbol{S}^2_{\boldsymbol{\beta}_j}$.
- $\boldsymbol{S}^2_{\boldsymbol{\beta}_j}$ is the variance of parameter estimator $\hat{\beta}_j$.
- $\boldsymbol{S}_{\boldsymbol{\beta}_j}$ is the **standard error** (SE or se) of parameter estimator $\hat{\beta}_j$.

Theorem 6 (Distribution of the studentised LSE).

Assume that the model is correct, Then the studentised estimator of β_j is

$$\frac{\hat{\beta}_j - \beta_j}{\boldsymbol{S}_{\beta_j}} \sim t_{n-2}$$

As $n \to \infty$

$$\frac{\hat{\beta}_j - \beta_j}{\boldsymbol{S}_{\beta_j}} \sim N(0,1)$$

Confidence Intervals

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}$$

For a t-distribution with $n-2$ degree of freedom, say $T \sim t_{n-2}$, it follows by the definition that

$$P(-t_{n-2,1-\frac{\alpha}{2}} < T < t_{n-2,1-\frac{\alpha}{2}}) = 1 - \alpha$$

Hence, with $T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}$,

$$P(-t_{n-2,1-\frac{\alpha}{2}} < \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} < t_{n-2,1-\frac{\alpha}{2}}) = 1 - \alpha$$

implies that

$$P(\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}_{\beta_1} < \beta_1 < \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}_{\beta_1}) = 1 - \alpha$$

From this equality, the $100(1-\alpha)\%$ confidence interval (CI) of β_1 is given by

$$[\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}_{\beta_1}, \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}_{\beta_1}]$$

Hypothesis Testing

Hypothesis

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$

Test statistic

$$T = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}$$

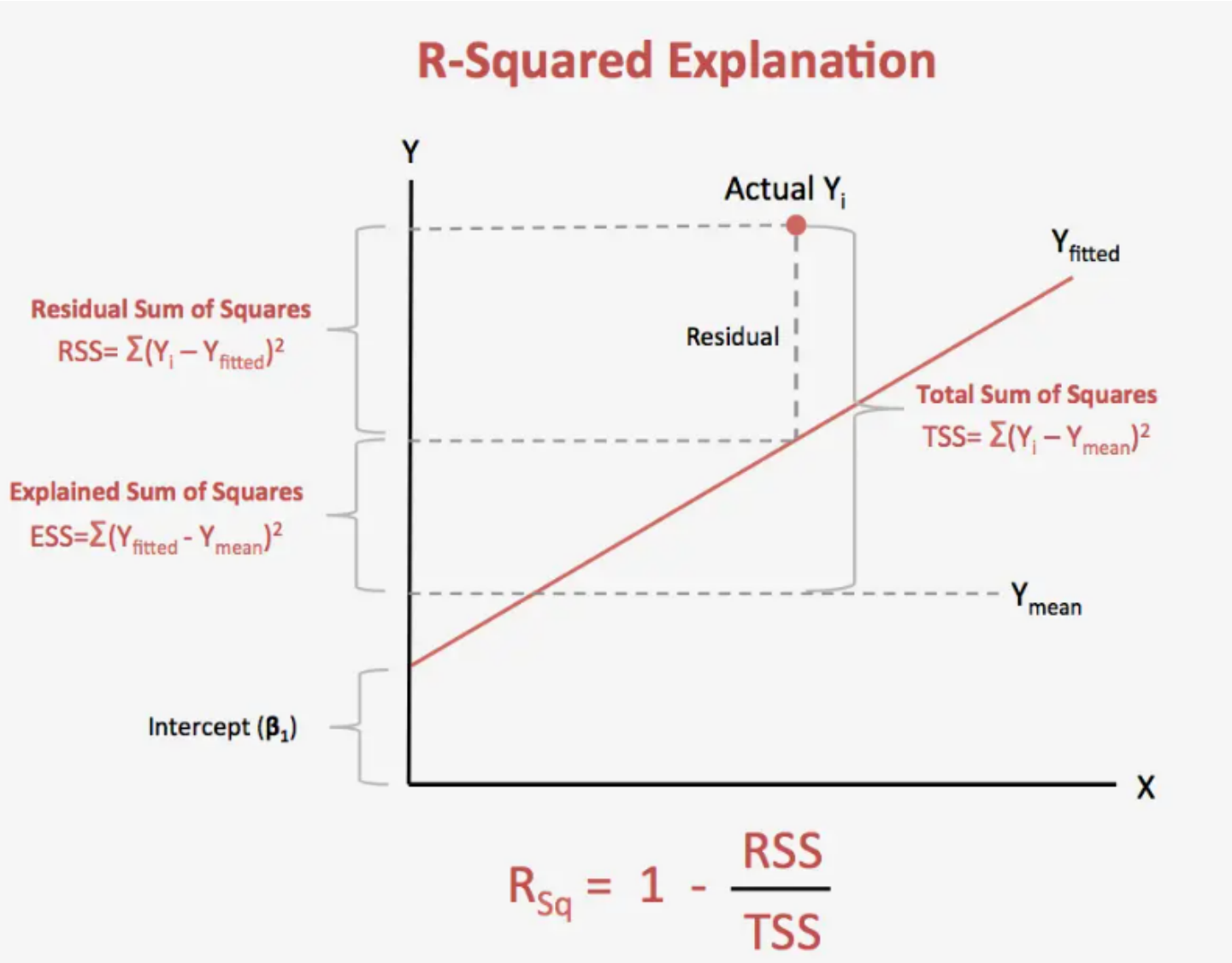
Rejection region

$$[-t_{n-2,1-\frac{\alpha}{2}}, t_{n-2,1-\frac{\alpha}{2}}]$$

p-value

$$p - value = P(|T| \geq |t|) = 2P(T \geq |t|)$$

Measuring Goodness of Fit



Measuring Goodness of Fit

Coefficient of Determination

- **R-squared:** This measures the variation of a regression model. R-squared either increases or remains the same when new predictors are added to the model.

$$R^2 = 1 - \frac{SSE}{SST} = \frac{S_{xy}}{S_{xx} \times S_{yy}} = r_{xy}^2$$

- **Adjusted R-squared:** This measures the variation for a multiple regression model, and helps you determine goodness of fit.

$$Adjusted - R^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1}$$

where n is the size of datasets, k is the number of independent variable and R^2 is the r-square value.

Example: Munich Rent Index

R-code

```
slm <- lm(rent~area, data = rent99)
summary(slm)
```

Output

```
#Call:
#lm(formula = rent ~ area, data = rent99)
#
#Residuals:
#    Min      1Q  Median      3Q     Max
#-786.63 -104.88   -5.69   95.93 1009.68
#
#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept) 134.5922     8.6135   15.63   <2e-16 ***
#area          4.8215     0.1206   39.98   <2e-16 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 158.8 on 3080 degrees of freedom
#Multiple R-squared:  0.3417,Adjusted R-squared:  0.3415
#F-statistic:  1599 on 1 and 3080 DF,  p-value: < 2.2e-16
```

R-code

```
confint(slm)
```

Output

```
#                 2.5 %     97.5 %
#(Intercept) 117.703417 151.480972
#area          4.585017   5.057912
```

Contents

Linearity of the regression models

The conditional mean of the outcome must satisfy

$$E(y|x) = \beta_0 + \beta_1 x$$

If the parameters are known, then this is equivalent to the condition

$$0 = E(\varepsilon|x) = E(y - \beta_0 - \beta_1 x|x)$$

If there are replicated outcomes available for a given x, then $E(y - \beta_0 - \beta_1 x|x)$ can be (unbiasedly) estimated as the sample mean of the residuals $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ for which $x_i = x$. These average residuals can be computed for all $x \in \{x_1, ..., x_n\}$.

Example: Munich Rent Index

R-code

```
e<-slm$residuals
x.all<-unique(rent99$area)
ave.e<-c()
for (x in x.all){
  ave.e<-c(ave.e, mean(e[rent99$area==x]))
}
plot(rent99$area, e, cex.lab=1.5, cex.axis=1.5,
     xlab="area in sqm", ylab="residuals")
points(x.all, ave.e, col=2, pch=8)
abline(h=0, lty=2)
```

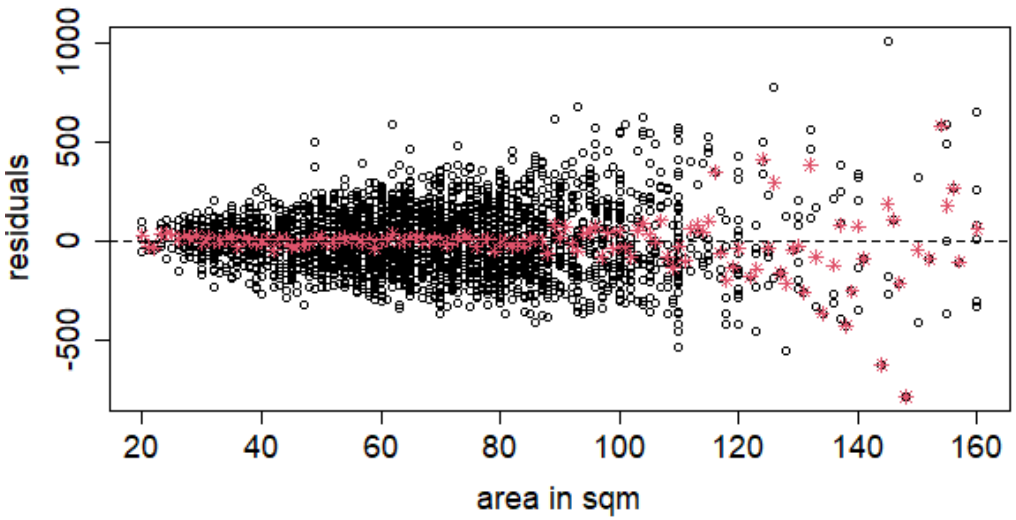


Figure: Scatter plot of the residuals against the area in sqm (Munich Rent Index Example). The red stars represent the sample means of the residuals for a given area in sqm

This figure shows no systematic pattern between average residuals and the regressor (area in sqm), therefore, we can conclude that the linearity assumption is satisfied.

Independence of errors

There is no relationship between residuals and response y, if the parameters are known, then this equivalence to condition

$$0 = E(\varepsilon|y) = E(y - \beta_0 - \beta_1 x|y)$$

R-code

```
e<-slm$residuals
y<-slm$fitted.values
plot(y, e, cex.lab=1.5, cex.axis=1.5,
     xlab="fitted values", ylab="residuals", main="")
abline(h=0, lty=2)
```

Example: Munich Rent Index

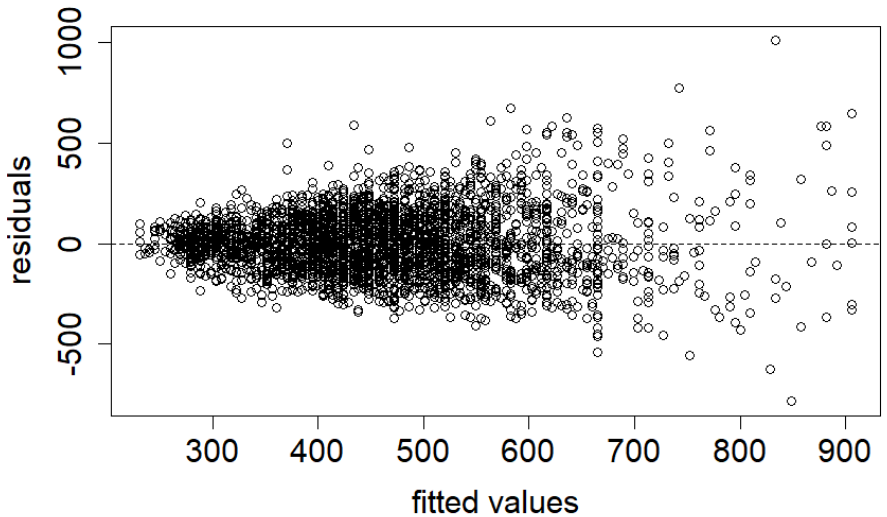


Figure: Scatter plot of the residuals against the fitted values of net rent.

This figure shows no systematic pattern between residuals and the fitted values of net rent, therefore, we can conclude that the independence of errors assumption holds.

Normality of the errors

The model implies that

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 x_i | x_i \sim N(0, \sigma^2)$$

The normal QQ-plots can be used to assess this assumption.

R-code

```
qqnorm(slm$residuals, cex.lab=1.5, cex.axis=1.5,
       xlab = "expexted quantiles",
       ylab = "residuals", main = "")
qqline(slm$residuals)
```

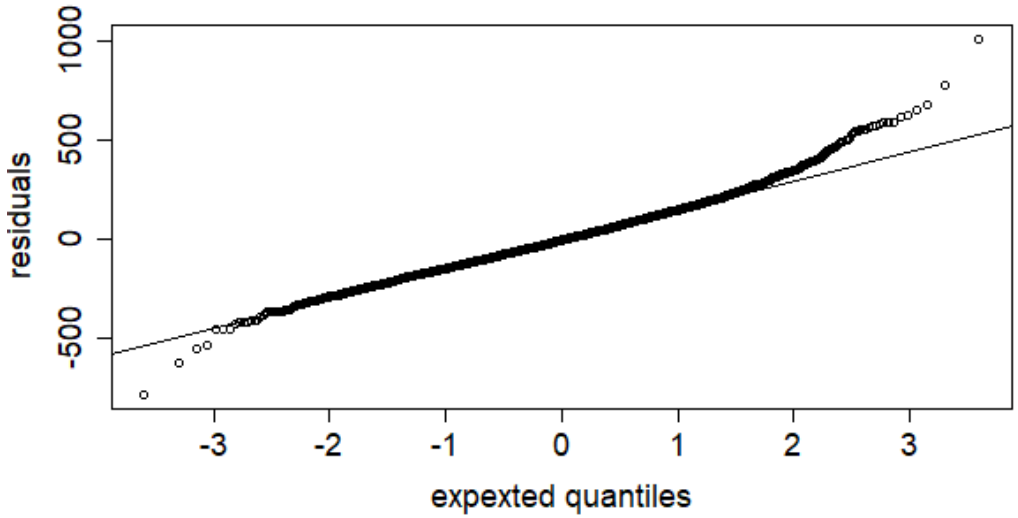


Figure: Normal QQ-plot of the residuals of the Munich Rent Index Example

The normal QQ-plot of residuals shows that most of the points are close to the straight line with no systematic pattern and there are some larger and a few smaller deviations at both ends of the tails however this is relatively small compared to all observations (3082). Thus, the normality assumption is hold. Note that the normality assumption is not problematic with a large sample size.

Homoscedasticity

The model implies that

$$Var(\varepsilon_i) = Var(\varepsilon_i|x_i) = Var(y_i|x_i) = \sigma^2$$

meaning that the variance of the outcomes (and of the error terms) is constant and does not depend on the values of the regressor.

Example: Munich Rent Index

R-code

```
par(mfrow=c(1,2))
area <- unique(rent99$area)
var.y<-c()
for(x in area){
  var.y<-c(var.y, sum(slm$residuals[rent99$area==x]^2))
}
plot(area, var.y, cex.lab=1.5, cex.axis=1.5,
     xlab="area in sqm", ylab="sample variance", col=2, pch=8)
abline(h=sum(slm$residuals^2)/3082, lty=2)
# note that the reference line is at MSE (n-2)/n

plot(area, sqrt(var.y), cex.lab=1.5, cex.axis=1.5,
     xlab="area in sqm", ylab="simple standard deviation",
     col=2, pch=8)
abline(h=sqrt(sum(slm$residuals^2)/3082), lty=2)
# note that the reference line is at sqrt MSE (n-2)/n
par(mfrow=c(1,1))
```

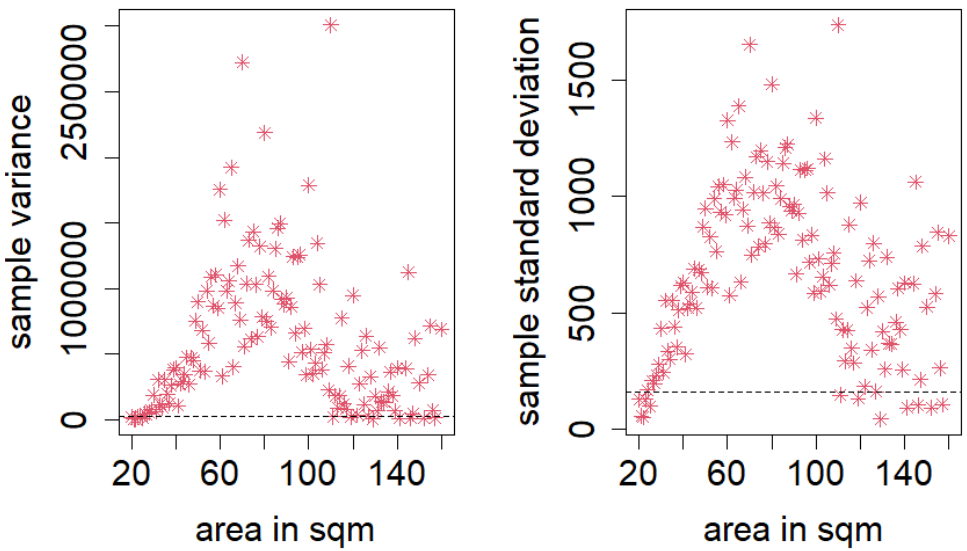


Figure: Sample variance (left) and standard deviation (right) against the regressor (area in sqm). The horizontal reference line corresponds to the MSE (left) and $\sqrt{MSE}$ (right).

A parabolic relationship is observed between sample variance and the regressor (left) and between sample standard deviation and the regressor (right). The assumption of homoscedasticity might **not hold**.

Munich Rent Index

R-code

```
par(mfrow=c(1,2))
e<-slm$residuals
plot(rent99$area, e^2, cex.lab=1.5, cex.axis=1.5,
     xlab="area in sqm",
     ylab="squared residuals")
abline(h=sum(slm$residuals^2)/(3082-2), lty=2)

e<=slm$residuals
plot(rent99$area, abs(e),cex.lab=1.5, cex.axis=1.5,
     xlab="area in sqm",
     ylab="absolute value of the residuals")
abline(h=sum(abs(slm$residuals))/3082, lty=2)
par(mfrow=c(1,1))
```

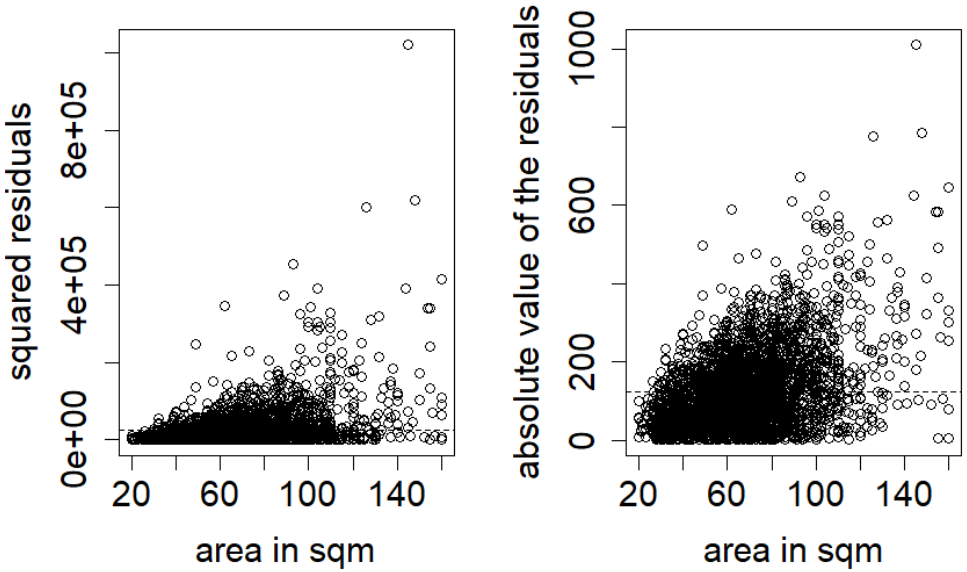


Figure: Scatter plots of ε_i^2 against x_i and of $|\varepsilon_i|$ against x_i. The horizontal reference lines correspond to *MSE* (left) and $\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|$ (right).

It can be observed that larger areas have larger squared residuals (left) and absolute values of the residuals. Linear relationships can be observed. Thus, the assumption of homoscedasticity might **not hold**.

Contents

Predictions

- x^*: a new case, future values, not involved in parameters estimation
- y^*: the response is not yet observed
- $m(x) = \beta_0 + \beta_1 x$: the conditional mean

Assume the outcome, y^*, behaves according to the same statistical model as the observed sample. Then, the model becomes,

$$y^* = m(x) + \varepsilon^* = \beta_0 + \beta_1 x^* + \varepsilon^*$$

with $\varepsilon^* \sim N(0, \sigma^2)$.

- A point prediction of y^*, say $\tilde{y}^*$, is given by

$$\tilde{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

- The prediction error is given by

$$\varepsilon^* = \hat{y}(x) - y^*$$

Prediction Interval

- The distribution of the prediction error is given by

$$\hat{y}(x) - y^* \sim N(0, \sigma_m^2(x) + \sigma^2)$$

Theorem 7 (Prediction interval for a given x and with normal error terms).

Assume that the model is correct. Then, for a given x and a confidence level $1 - \alpha$

$$P\left(\hat{y}(x) - t_{n-2,1-\frac{\alpha}{2}}\sqrt{\sigma_m^2(x) + MSE} \leq y^* \leq \hat{y}(x) + t_{n-2,1-\frac{\alpha}{2}}\sqrt{\sigma_m^2(x) + MSE}\right) = 1 - \alpha$$

Proof

For the standardised prediction error, we have

$$\frac{\hat{y}(x) - y^*}{\sqrt{\sigma_m^2(x) + \sigma^2}} \sim N(0, 1)$$

- $\sigma_m^2(x)$ (variance of LSE) is proportional to σ^2, then $\sigma_m^2(x) = \sigma_m'^2(x)\sigma^2$
- σ^2 is estimated by MSE
- $\sigma_m^2(x) + \sigma^2$ is estimated by $MSE(\sigma_m'^2(x) + 1)$.
- $(n-2)MSE/\sigma^2 \sim \chi_{n-2}^2$

Hence,

$$\frac{\hat{y}(x) - y^*}{\sqrt{MSE(\sigma_m'^2(x) + 1)}} \sim t_{n-2}$$

Thus,

$$P\left(-t_{n-2,1-\alpha/2} \leq \frac{\hat{y}(x) - y^*}{\sqrt{MSE(\sigma_m'^2(x) + 1)}} \leq t_{n-2,1-\alpha/2}\right) = 1 - \alpha$$

Example: Munich Rent Index

- For the sake of illustration, even if the homoscedasticity assumption might not hold, we will still do a prediction on this example.
- We want to predict the net rent of an apartment with a living area of 75 sqm.

R-code

```
predict(slm, newdata = data.frame(area = 75),
        interval = "prediction", level = 0.99)
```

Output

```
##      fit      lwr      upr
##1 496.202 86.89523 905.5088
```

Interpretation: it can be predicted that an apartment with a living area of 75 sqm has net rent of 496.20 Euro and within a probability of 95% this price is ranging from 86.90 Euro to 905.51 Euro.

Contents

Binary Dummy Regressors

- In the two-sample problem, we are interested in comparing two means.
- Dummy regressor x_i is defined as

$$\begin{aligned} x_i &= 1 \text{ if observation } i \text{ belongs to group B} \\ &= 0 \text{ if observation } i \text{ belongs to group A} \end{aligned}$$

- The regression model

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

with ε_i i.i.d. $N(0, \sigma^2)$

- The model is equivalent to

$$y_i | x_i = 0 \sim N(\beta_0, \sigma^2) \quad \text{and} \quad y_i | x_i = 1 \sim N(\beta_0 + \beta_1, \sigma^2)$$

- Outcomes for group A and B

$$\mu_A = E(y|x = 0) = \beta_0 \quad \text{and} \quad \mu_B = E(y|x = 1) = \beta_0 + \beta_1$$

- Effect size

$$\delta = \mu_B - \mu_A = \beta_1$$

Example: Munich Rent Index

Research Question: does the quality of the kitchen affect the net rent of the apartment?

- Model

$$rent_i = \beta_0 + \beta_1 \cdot kitchen_i + \varepsilon_i \qquad i = 1, ..., 3082$$

where ε_i i.i.d. $N(0, \sigma^2)$

  - $rent_i$: the monthly net rent per month (in Euro)
  - $kitchen_i$: quality of kitchen: 0 - standard, and 1 - premium

R-code

```
slm_kitchen <- lm(rent~kitchen, data=rent99)
summary(slm_kitchen)
```

Output

```
#Call:
#lm(formula = rent ~ kitchen, data = rent99)
#
#Residuals:
#    Min      1Q  Median      3Q     Max
#-481.36 -134.07  -30.13   98.07 1390.84
#
#Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept)  452.545      3.552 127.421   <2e-16 ***
#kitchen1     162.145     17.227   9.412   <2e-16 **
#---
#Signif. codes:
#0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 192.9 on 3080 degrees of freedom
#Multiple R-squared:  0.02796,Adjusted R-squared:  0.02764
#F-statistic: 88.59 on 1 and 3080 DF,  p-value: < 2.2e-16
```

For the same problem, let's do the same analysis but now with the two-sample t-test method

R-code

```
t.test(rent~kitchen, data=rent99)
```

Output

```
# Welch Two Sample t-test
#
#data:  rent by kitchen
#t = -7.7057, df = 137.55, p-value = 2.329e-12
#alternative hypothesis: true difference in means between group
#0 and group 1 is not equal to 0
#95 percent confidence interval:
# -203.7537 -120.5370
#sample estimates:
#mean in group 0 mean in group 1
#       452.5452        614.6905
```

The agreement between the results can be seen directly.

Contents

Association versus Causation

Form Munich Rent Index example:

- There is a positive significant effect of living area (*area*) on net rent (*rent*) with $p < 0.05$
- There is a positive significant effect of quality of kitchen (*kitchen*) on net rent (*rent*) with $p - value < 0.05$.

All these conclusions refer to **associations** between a regressor and (mean) outcome, but they do not necessarily imply a **causation**.
In case of a **causation** interpretation, think of the following questions:

- Can we conclude that increasing a living area causes the average net rent to become larger?
- Can we conclude that a premium kitchen causes the average net rent to become larger?

Causal Inference

- **Causal inference** is a discipline in statistics that aims to develop methods that can be used to assess causal relationships.
- For simplicity, a two-sample problem will be used for illustration.
- Define two counterfactual outcomes:
  - $y_i(1)$ is the outcome of subject i is subject i would belong to group B ($x_i = 1$)
  - $y_i(0)$ is the outcome of subject i is subject i would belong to group A ($x_i = 0$)
- The observed y_i can be written as

$$y_i = x_i y_i(1) + (1 - x_i) y_i(0)$$

- The **causal effect** for subject i is $y_i(1) - y_i(0)$

- We are interested in the **average causal effect**, defined in terms of population averages:

$$E(y(1) - y(0)) = E(y(1)) - E(y(0))$$

- If we have an unbiased estimator for $E(y(1))$ and $E((0))$, then we also have an unbiased estimator of the causal effect. This holds true only for studies that involve **complete randomisation**, but it does not hold in general.

Observational Study and Experimental Study

- **Observational study** is when the regressor x_i and the outcome y_i variables are sampled together, without the researcher having control over the decision of which subject i is assigned to which group.
- **Experimental study** is when the subjects in the study were randomized over the grou