Marek Gągolewski
Konstancja Bobecka-Wesołowska
Przemysław Grzegorzewski

Computer Statistics with R

9. Regression Analysis

Faculty of Mathematics and Information Science
Warsaw University of Technology
[Last update: January 10, 2013]

Contents

Info

These tutorials are likely to contain bugs and typos. In case you find any don't hesitate to *contact us*! Thanks in advance!

9.1. Preliminaries

9.1.1. Coefficients of correlation

The most popular sample-based correlation coefficients may be calculated in R by calling the `cor()` function. By default, Pearson's r is used. This coefficient indicates the degree of linear dependence between two tied variables (two vectors of equal length) and serves as a good estimator of the theoretical correlation coefficient $\varrho = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y)$ if X, Y are normally distributed. The closer the coefficient is to either -1 or 1, the stronger the **linear** relationship between the variables.

The r coefficient is not robust when applied to data that do not come from normal distributions. In such cases we should use non-parametric (rank-based) measures of association. They may be calculated by passing an additional argument to the `cor()` function: `method="spearman"` for the Spearman ρ coefficient and `"kendall"` for Kendall's τ.

To test whether the (theoretical) correlation coefficient ϱ is significant, i.e.

$$\begin{aligned} H_0 &: \varrho = 0, \text{non-correlated variables} \\ K &: \varrho \neq 0, \end{aligned}$$

we may apply the `cor.test()` function. By default it uses Pearson's r, but this behavior may be altered with the `method` argument.

9.1.2. Regression analysis

Regression analysis is a statistical tool used to investigate potential functional relationships between two or more variables. It tries to express a variable Y (the *dependent / response / effect variable*, or regressand) as a function $f_{\mathbf{a}}(\mathbf{X})$ of a set of variables $\mathbf{X} = (X_1, \ldots, X_m)$, called the *independent / explanatory variables*, regressors, or predictors. The function $f_{\mathbf{a}}$ is parametrized by a vector $\mathbf{a} = (a_1, \ldots a_k)$, and the main goal of regression analysis is to find its estimate, $\widehat{\mathbf{a}}$.

9.1.2.1. Linear regression

Among the simplest regression tasks we have the *linear regression*. Here, the class of functions used to explain Y is of the form

$$f_{\mathbf{a}}(\mathbf{X}) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_mX_m + \varepsilon, \tag{9.1}$$

where $\mathbf{a} = (a_0, \ldots, a_m)$. In particular, the $m = 1$ case is called the *simple linear regression*. This gives the statistical model of the form

$$Y = aX + b + \varepsilon, \tag{9.2}$$

where ε is an *error* (*residual, disturbance*) *term*. Note that b is called an *intercept*.

To build a linear model, i.e. to find „the best" estimator of $\mathbf{a}$, the `lm()` function may be used. The function needs a *formula* as its first argument – it is used to define the set of variables to be considered. Its syntax is as follows:

$$\mathtt{Y} \sim \mathtt{X_1} + \mathtt{X_2} + \cdots + \mathtt{X_m}. \tag{9.3}$$

To build a model without the intercept, we write:

$$\mathtt{Y} \sim \mathtt{X_1} + \mathtt{X_2} + \cdots + \mathtt{X_n} - \mathtt{1}. \tag{9.4}$$

More information may be found in the manual: `?lm`.

The returned fitted model is a `list` object. It may then be used to apply additional computations, such as `summary()` (model diagnostics and details), `predict()` (prediction), or `plot()` (graphical diagnostics).

9.2. Examples

Ex. 9.1. Generate samples of size $n = 200$ from the following bivariate $\mathrm{NN}(\mu_X, \sigma_X, \mu_Y, \sigma_Y, \varrho)$ distributions.

1. The bivariate normal distribution $\mathrm{NN}(0, 1, 0, 1, 0)$.
2. The bivariate normal distribution $\mathrm{NN}(0, 1, 2, 1, 0.6)$.

Estimate the correlation coefficient ϱ and verify its significance.

Solution.
(a) Random deviates from the $\mathrm{NN}(0, 1, 0, 1, 0)$ distribution may be created by generating two independent (this implies $\varrho = 0$) vectors from the $\mathrm{N}(0, 1)$ distribution.

```
n <- 200
x <- rnorm(n)
y <- rnorm(n)
```

Let us calculate different sample correlation coefficients:

```
cor(x, y)
## [1] 0.05848
cor(x, y, method = "spearman")  # for comparison - non-parametric estimator
## [1] 0.05617
cor(x, y, method = "kendall")  # for comparison - non-parametric estimator
## [1] 0.03588
```

We may test for the significance of the theoretical correlation coefficient ($H_0 : \varrho = 0$, $K : \varrho \neq 0$):

```
cor.test(x, y)
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 0.8243, df = 198, p-value = 0.4108
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08092  0.19563
## sample estimates:
##     cor
## 0.05848
```

Please, draw conclusions.

(b) Let U, V i.i.d. $\mathrm{N}(0, 1)$, and set $X := \mu_1 + \sigma_1 U$, $Y := \mu_2 + \sigma_2 \left(\varrho U + \sqrt{1 - \varrho^2} V\right)$. It may be shown that in such case (X, Y) follows the $\mathrm{NN}(\mu_1, \sigma_1, \mu_2, \sigma_2, \varrho)$ distribution (ϱ is the theoretical coefficient of correlation).

Let us generate n random deviates from $\mathrm{NN}(0, 1, 2, 1, 0.6)$.

```
u <- rnorm(n)
v <- rnorm(n)
x2 <- 0 + 1 * u
y2 <- 2 + 1 * (0.6 * u + sqrt(1 - 0.6^2) * v)
```

Details

The covariance matrix of the NN$(\mu_1, \sigma_1, \mu_2, \sigma_2, \varrho)$ distribution is given by:

$$\mathbf{C} = \begin{bmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{9.5}$$

which in our case is equal to:

```
(C <- matrix(c(1, 0.6, 0.6, 1), nrow = 2))
```

```
##      [,1] [,2]
## [1,]  1.0  0.6
## [2,]  0.6  1.0
```

To generate a random sample, also the `mvrnorm()` function from the `MASS()` package may be used:

```
library("MASS")  # load the library
sample <- mvrnorm(n, c(0, 2), C)
head(sample)  # print some observations
```

```
##         [,1]    [,2]
## [1,] -2.3932 -0.2273
## [2,] -0.7434  2.6390
## [3,] -1.4237  1.8627
## [4,]  1.9282  2.5280
## [5,] -0.1147  1.3211
## [6,]  2.5933  3.3002
```

Here are the correlation coefficient estimates, $\hat{\varrho}$:

```
cor(x2, y2)
```

```
## [1] 0.6476
```

```
cor(x2, y2, method = "spearman")  # for comparison
```

```
## [1] 0.628
```

```
cor(x2, y2, method = "kendall")  # for comparison
```

```
## [1] 0.4511
```

Test for significance (Pearson's r-based):

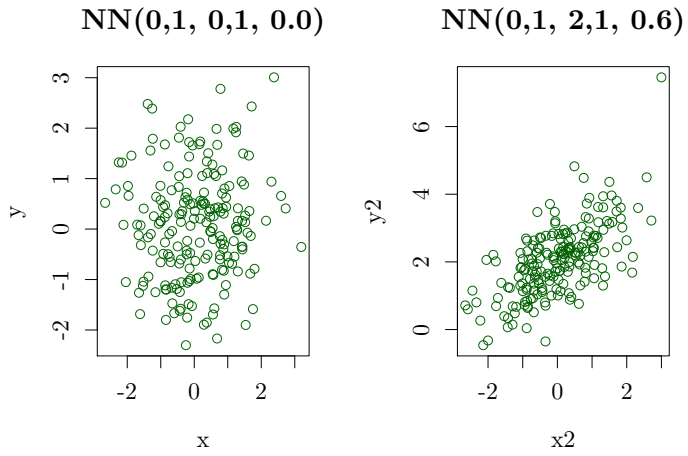
```
cor.test(x2, y2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x2 and y2
## t = 11.96, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5590 0.7215
## sample estimates:
##    cor
## 0.6476
```

Draw conclusions.

Additionally, please compare the scatter plots for the two samples.

```
par(mfrow = c(1, 2))
plot(x, y, main = "NN(0,1, 0,1, 0.0)", col = "darkgreen")
plot(x2, y2, main = "NN(0,1, 2,1, 0.6)", col = "darkgreen")
```



⊡

Ex. 9.2. Two professors ranked their favorite students (labeled A–K) according to their skills. Is there any dependence between the professors' opinions? Test an appropriate hypothesis. Use significance level of 0.05.

Student	A	B	C	D	E	F	G	H	I	J	K
Prof. X	1,	7,	8,	3,	6,	10,	9,	2,	11	4,	5
Prof. Y	4,	8,	10,	1,	5,	9,	11,	3,	7,	2,	6

Solution.

```
x <- c(1, 7, 8, 3, 6, 10, 9, 2, 11, 4, 5)
y <- c(4, 8, 10, 1, 5, 9, 11, 3, 7, 2, 6)
```

The two vectors store the ranks given by each lecturer. We are interested whether the ranks are correlated.

Of course, here Pearson's r shouldn't be used. Let us determine the value of a rank correlation coefficient:

```
cor(x, y, method = "spearman")
```

```
## [1] 0.7909
```

Let's apply a test for the significance of the correlation coefficient ϱ. We verify $H_0 : \varrho = 0$ against $K : \varrho \neq 0$.

```
cor.test(x, y, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 46, p-value = 0.006061
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##    rho
## 0.7909
```

At a significance level of $\alpha = 0.05$ we reject the null hypothesis stating that the data are not linearly dependent (Test for significance of Spearman's ρ, $S = 46$, p-value $= 0.0061$).

Thus, we may assume that there is a statistically significant (positive) relationship between the professors' opinions.

⊡

Ex. 9.3. 10 farmers from the same country applied varying amounts of fertilizer and obtained varying yields of rye. Data below are given in hundredweight for fertilizer and tons for crop yield.

Fertilizer	8.3,	9.2,	7.7,	8.4,	8.8,	9.6,	10.3,	8.7,	9.1,	9.4
Yield	14.1,	15.1,	13.6,	13.9,	14.6,	15.8,	16.6,	14.1,	14.9,	15.6

1. Find a linear regression equation between the amount of fertilizer applied and crop yield.
2. How well does the independent variable explain the dependent variable in this regression model?
3. Predict average yield for 9, 10 and 11 hundredweight of fertilizer employed.

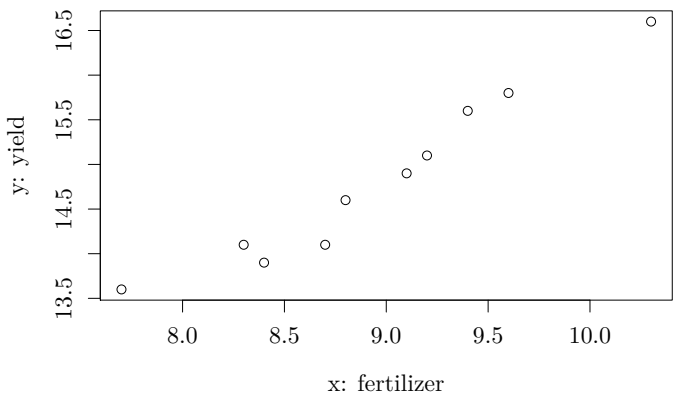
Solution.

First we input our data into R.

```
# input data
x <- c(8.3, 9.2, 7.7, 8.4, 8.8, 9.6, 10.3, 8.7, 9.1, 9.4)  # fertilizer
y <- c(14.1, 15.1, 13.6, 13.9, 14.6, 15.8, 16.6, 14.1, 14.9, 15.6)  # yield
```

Let us draw a scatter plot for y as a function of x.

```
plot(x, y, xlab = "x: fertilizer", ylab = "y: yield")
```



Intuitively, it might be a good idea to try to express the amount of yield as a linear function of fertilizer applied. The intuition is supported by the results of linear correlation coefficient's estimation:

```
cor(x, y)
## [1] 0.9728
cor.test(x, y)
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 11.88, df = 8, p-value = 2.312e-06
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.8857 0.9938
## sample estimates:
##    cor
## 0.9728
```

Therefore we will fit a regression line to the data set, so that for the i-th observation we will have

$$y_i = ax_i + b + \varepsilon_i, \tag{9.6}$$

where ε_i is an error term, or *residual*. A (really good) method of the coefficients a and b estimation is called the *least squares*, in which we try to find the estimates $\widehat{a}, \widehat{b}$ which minimize the sum of squared residuals, i.e.

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - (\widehat{a}x_i + \widehat{b})\right)^2. \tag{9.7}$$

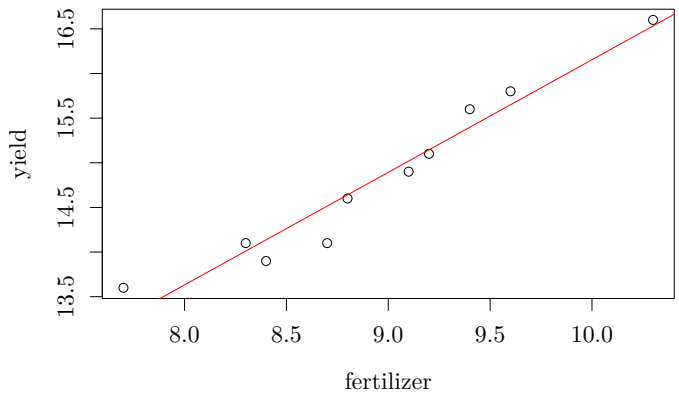
Note that some requirements have to be fulfilled here. The error term (a) should be a sample of i.i.d. normally distributed random variables with expectation equal to 0, and (b) it should not statistically depend on the values of the independent variables.

An R function for fitting linear models by the least squares method is called `lm()`.

```
# fit a linear model
m <- lm(y ~ x)
# print the estimated coefficients
m$coefficients
```

```
## (Intercept)           x
##       3.544       1.261
```

```
# now, draw a scatter plot with a regression line
plot(y ~ x, xlab = "fertilizer", ylab = "yield")
abline(m, col = "red")
```



R tells us that

$$Y = 1.261X + 3.5444 \tag{9.8}$$

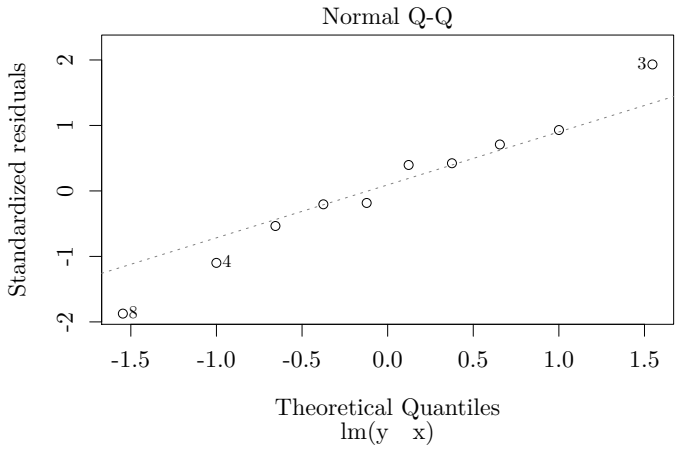
is a best-fit result (using least squares) in this case. This regression line is drawn in the figure above.

We should check whether the results we obtained are sound. We pose ourselves a question: How well does the independent variable (X, the amount of fertilizer applied) explain the dependent variable (Y, the amount of yield)? Is yield really a linear function of fertilizer, and the resulting "imprecision" is only due to an error in measurements or due to some "not really important statistical factors"?

First, let us look at the distribution of residuals. We need to verify some assumptions.

1. Are they normally distributed (good) or not (bad)? Look at the figure below and answer the question.

```
# Normal Q-Q plot for residuals: qqnorm(m$residuals); qqline(m$residuals)
# or, similarly (a simpler way):
plot(m, which = 2)
```



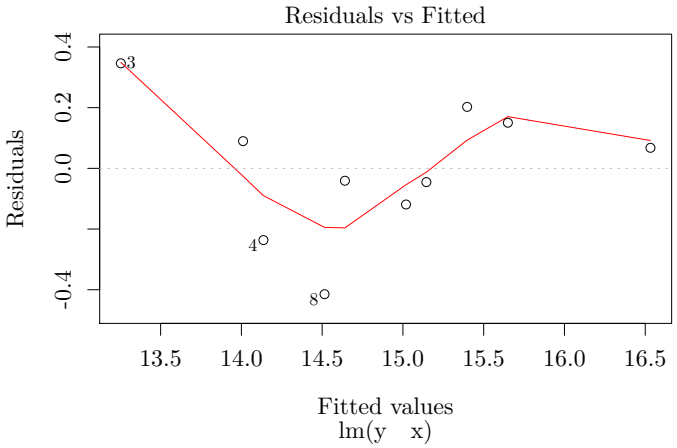
Shapiro-Wilk test for normality will give us a ‘more formal‘ answer:

```
shapiro.test(m$residuals)


##
##  Shapiro-Wilk normality test
##
## data:  m$residuals
## W = 0.9846, p-value = 0.9852
```

2. Do the residuals seem to be homoskedastic, i.e. do they have the same variance (good), or rather the error term depends on the values of the independent variables (bad)? Do they lie chaotically near the $y = 0$ line (good), or is there some kind of regularity in their dispersion (bad)? Look at the figure below and answer the question.

```
# Plot of residuals:
plot(m, which = 1)
```



Now let us perform a more thorough diagnosis of the fitted model.

```
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4148 -0.1007  0.0134  0.1352  0.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.544      0.953    3.72   0.0059 **
## x              1.261      0.106   11.88  2.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.235 on 8 degrees of freedom
## Multiple R-squared:  0.946,Adjusted R-squared:  0.94
## F-statistic:  141 on 1 and 8 DF,  p-value: 2.31e-06
```

The function `summary()` gives us some general information on the fitted linear model's appropriateness. Among the most important items are:

1. `Coefficients`: t-tests for the significance of the estimated coefficients. For each of the model parameters, this test verifies the null hypothesis H_0 : the coefficient equals 0, against K : the coefficient significantly differs from 0. If the resulting p-values, denoted `Pr(>|t|)`, are small, we may suspect that the estimates are significant, i.e. they are carrying important information rather than being only a statistical "noise" (note that e.g. if $Y = 0X + b + \varepsilon$, then X does not give any information about Y).

   In our example both $\widehat{a}$ and $\widehat{b}$ are significant ($\alpha = 0.05$).

   This test is very important in multiple regression, that is when we have many independent variables ($m > 1$). It can indicate redundant variables (not carrying much information on the dependent variable).

2. `Multiple R-squared`: The coefficient of determination, R^2, is defined as a fraction of variance of the dependent variable explained by the model. In our case, about 94.6% of the the yield amount variability can be explained by a linear function of

fertilizer applied. The rest is a great unknown: maybe these are the effects of the measurement error, the weather, vandalism etc.? Anyway, that is a good result. Of course, $R^2 = 1$ is the ideal, but extremely rare (and thus suspicious), outcome.

3. `F-statistic`: F-test (analysis of variance) for the *whole* model. It tests a null hypothesis H_0 : all the coefficients are equal to 0, against $K : \neg H_0$. Simply, if H_0 held, the fitted linear model would be inappropriate.

   In our case the p-value suggests that on any rational significance level we should reject H_0, i.e. it is not absurd to express the dependent variable in terms of a linear function of the independent variable.

With such a good fit we may try to make a prediction of the average yield for different amounts of fertilizer applied. For example:

```
# We pass the x-values for which we want to make predictions in a
# one-column data frame. The column name should be the same as the name of
# independent variable:
xpred <- data.frame(x = c(9, 10, 11))
predict(m, xpred)
```

```
##     1     2     3
## 14.89 16.15 17.41
```

Note that the prediction for $x = 11$ (column `3`) is an extrapolation and yet assumes the same yield amount trend for unobserved values of fertilizer. That might be misleading!

⊡

Ex. 9.4. The following table lists yearly income and residential value of 9 randomly chosen families living in some city district.

Income (10^3 PLN)	360,	640,	490,	210,	280,	470,	580,	190,	320
Res. value (10^6 PLN)	1.49,	3.10,	2.60,	0.92,	1.26,	2.42,	2.88,	0.81,	1.34

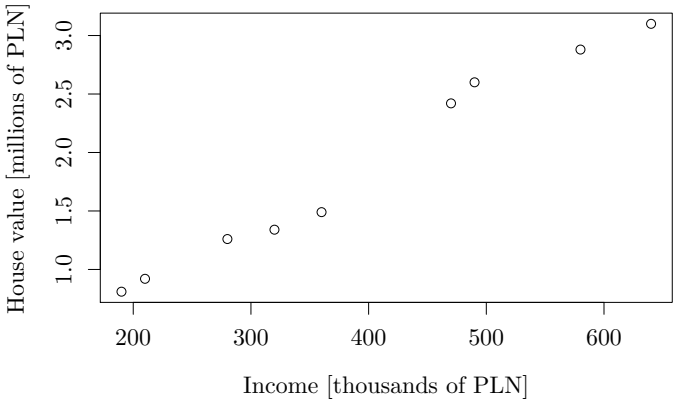
1. Find a linear regression line.
2. How well does the independent variable explain the dependent variable in this regression model?
3. Predict residential value of a family with income of 400 000 PLN.
4. Construct a 95% confidence interval for the predicted value.

Solution.
First we input the data in R:

```
income <- c(360, 640, 490, 210, 280, 470, 580, 190, 320)
value <- c(1.49, 3.1, 2.6, 0.92, 1.26, 2.42, 2.88, 0.81, 1.34)
```

Let us check whether the relationship between the independent variable X (income) and the dependent Y (residential value), is of linear type:

```
plot(income, value, xlab="Income [thousands of PLN]",
     ylab="House value [millions of PLN]")
```



Let us fit a linear regression model:

```
mdist <- lm(value ~ income)
(desc <- summary(mdist))
```

```
##
## Call:
## lm(formula = value ~ income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19776 -0.10925  0.00695  0.04732  0.20584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.268439   0.128168   -2.09    0.074 .
## income       0.005434   0.000304   17.86  4.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.138 on 7 degrees of freedom
## Multiple R-squared:  0.979,Adjusted R-squared:  0.975
## F-statistic:  319 on 1 and 7 DF,  p-value: 4.25e-07
```

The parameter estimates may be read from the `Coefficients` table (column: `Estimate`) or displayed using:

```
mdist$coefficients
```

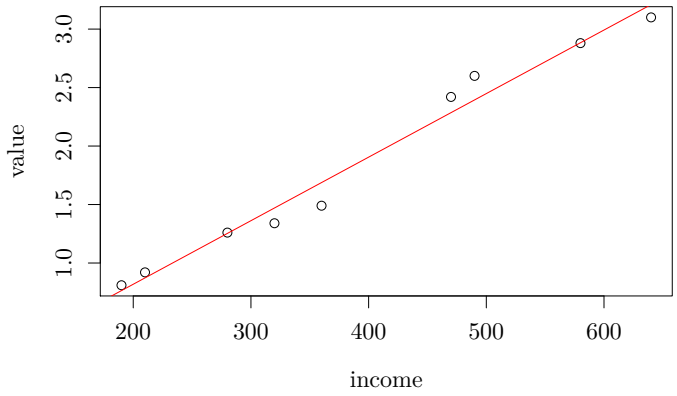
```
## (Intercept)      income
##   -0.268439    0.005434
```

`(Intercept)` means $\hat{b}$, and `income` — $\hat{a}$, the coefficient standing near X in the model equation. Therefore, the calculated regression line may be written as:

$$Y = -0.2684 + 0.0054X + \varepsilon.$$

Let us add the regression line to the scatter plot.

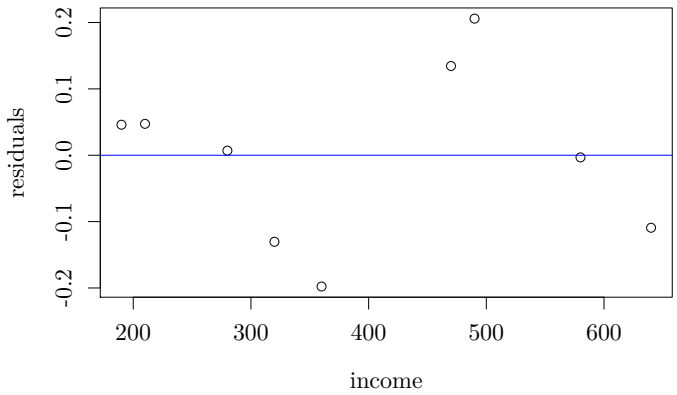
```
plot(income, value)
abline(mdist, col = "red")
```



Now let us check how well does the independent variable explain the dependent variable in this model.

Residual analysis. The scatter plot of residuals as a function of X.

```
plot(income, mdist$residuals, ylab = "residuals", xlab = "income")
abline(h = 0, col = "blue")
```



That's OK.

Shapiro-Wilk's test for normality of the residuals.

```
shapiro.test(mdist$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mdist$residuals
## W = 0.9706, p-value = 0.9002
```

That's OK.

Coefficient of determination. R^2 is equal to:

```
desc$r.squared
```

```
## [1] 0.9785
```

That's OK.

Additionally, we may test for significance of the Pearson correlation coefficient.

```
cor.test(income, value)
```

```
##
##  Pearson's product-moment correlation
##
## data:  income and value
## t = 17.86, df = 7, p-value = 4.254e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9476 0.9978
## sample estimates:
##    cor
## 0.9892
```

That's OK.

F-test. Null hypothesis $H : b = 0$ and $a = 0$ (there is no linear dependency) *vs* $K : \neg H_0$.

```
desc$fstatistic
```

```
## value numdf dendf
## 319.1   1.0   7.0
```

```
anova(mdist)
```

```
## Analysis of Variance Table
##
## Response: value
##           Df Sum Sq Mean Sq F value  Pr(>F)
## income     1   6.06    6.06     319 4.3e-07 ***
## Residuals  7   0.13    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mdist)$"Pr(>F)"[1]  # p-value
```

```
## [1] 4.254e-07
```

That's OK.

t-test. (test for significance of the coefficients). $H_1 : a = 0, H_2 : b = 0$ against $K_1 : a \neq 0, K_2 : b \neq 0$.

```
desc$coefficients
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.268439  0.1281678  -2.094 7.448e-02
## income       0.005434  0.0003042  17.863 4.254e-07
```

Info

Note that the t-test for the significance of coefficient b suggests that we could try to remove the intercept from the model.

```
mdist2 <- lm(value ~ income - 1)
(desc2 <- summary(mdist2))
```

```
##
## Call:
## lm(formula = value ~ income - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2521 -0.1094 -0.0949  0.0733  0.2289
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## income  0.00484    0.00013    37.2    3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.164 on 8 degrees of freedom
## Multiple R-squared:  0.994,Adjusted R-squared:  0.994
## F-statistic: 1.38e+03 on 1 and 8 DF,  p-value: 2.99e-10
```

The obtained model is given by:

$$Y = -0.2684X + \varepsilon.$$

Draw conclusions.

Prediction:

```
(prdctd <- data.frame(income = 400))
```

```
##   income
## 1    400
```

```
predict(mdist, prdctd)
```

```
##     1
## 1.905
```

95% confidence interval for the predicted value:

```
predict(mdist, prdctd, interval = "prediction", level = 0.95)  # c.i. for prediction
```

```
##     fit   lwr   upr
## 1 1.905 1.562 2.249
```

```
predict(mdist, prdctd, interval = "confidence", level = 0.95)  # c.i. for the mean
```

```
##     fit   lwr   upr
## 1 1.905 1.796 2.014
```

The bounds of the confidence intervals may be found in the `lwr` (*lower*) and `upr` (*upper*) columns.

⊡

Ex. 9.5. The table below shows the size of the population in some country (in thousands). Find and verify an exponential regression model for the data. Assuming the same trend in the future, predict the size of the population in 2020.

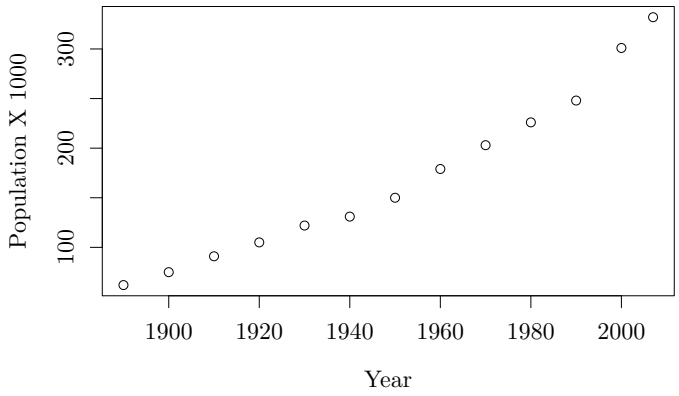
Year	1890,	1900,	1910,	1920,	1930,	1940,	1950,
Population	62,	75,	80,	93,	102,	121,	140,

Year	1960,	1970,	1980,	1990,	2000,	2010
Population	179,	203,	226,	260,	301,	332

Solution.
Data input:

```
yr <- seq(1890, 2010, by = 10)
pop <- c(62, 75, 80, 93, 102, 121, 140, 179, 203, 226, 260, 301, 332)
```

```
plot(yr, pop, xlab="Year", ylab="Population X 1000")
```



Fitting a linear model possibly would make no sense (please, check that). Let us try a different model (as indicated in the task) that can be easily transformed to a linear one so that the `lm()` function may be used.

The *exponential model* is given by: $Y = \exp(a + bX)$

This non-linear model may be linearized by substituting: $z := \ln(y)$ and then $Z = a + bX$.

```
logpop <- log(pop)
popmod <- lm(logpop ~ yr)
summary(popmod)
```

```
##
## Call:
## lm(formula = logpop ~ yr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06566 -0.03550  0.00755  0.01151  0.06550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.31e+01   6.07e-01   -38.0  5.1e-13 ***
## yr           1.44e-02   3.11e-04    46.2  6.0e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.042 on 11 degrees of freedom
## Multiple R-squared:  0.995,Adjusted R-squared:  0.994
## F-statistic: 2.13e+03 on 1 and 11 DF,  p-value: 5.99e-14
```

or equivalently:

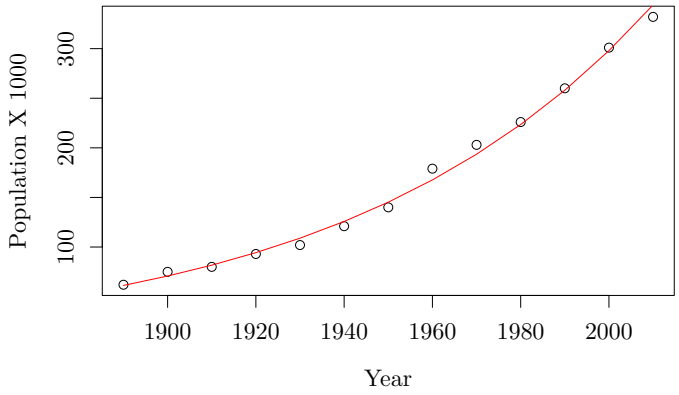
```
lm(log(pop) ~ yr)
```

```
##
## Call:
```

```
## lm(formula = log(pop) ~ yr)
##
## Coefficients:
## (Intercept)           yr
##    -23.0532       0.0144
```

Please, analyze the above results by yourself.

```
plot(yr, pop, xlab="Year", ylab="Population X 1000")
yfit <- exp(popmod$coef[1]+popmod$coef[2]*yr)
lines(yr, yfit, col=2)
```



Prediction:

```
newyr <- data.frame(yr = 2020)
(newlogpop <- predict(popmod, newyr, interval = "prediction"))
```

```
##     fit   lwr   upr
## 1 5.984 5.877 6.092
```

```
exp(newlogpop)  # the inverse of log()
```

```
##     fit   lwr   upr
## 1 397.2 356.8 442.1
```

⊡

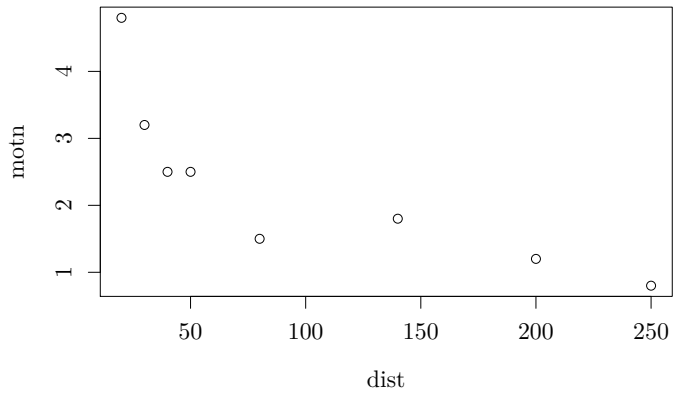
Ex. 9.6. 8 peak ground motion measures (Y) [mm] were made during an earthquake at different distances from the epicenter (X) [km]. The results are summarized in the table below.

Distance	20,	30,	40,	50,	80,	140,	200,	250
Vibration	4.8,	3.2,	2.5,	2.5,	1.5,	1.8,	1.2,	0.8

1. Find a regression model that explains the variability of Y well.
2. Estimate the ground motion at 100 km from the epicenter.

Solution.

```
dist <- c(20, 30, 40, 50, 80, 140, 200, 250)
motn <- c(4.8, 3.2, 2.5, 2.5, 1.5, 1.8, 1.2, 0.8)
plot(dist, motn)
```



Let us find the best non-linear model.

(a) Exponential model: $Y = \exp(a+bX)$. Linearization: $Z := \ln(Y)$ and then $Z = a+bX$.

```
yp <- log(motn)
t1 <- lm(yp ~ dist)
summary(t1)
```

```
##
## Call:
## lm(formula = yp ~ dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4129 -0.0990  0.0152  0.1016  0.3866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.30317    0.14213    9.17  9.5e-05 ***
## dist        -0.00606    0.00110   -5.52   0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.25 on 6 degrees of freedom
## Multiple R-squared:  0.835,Adjusted R-squared:  0.808
## F-statistic: 30.4 on 1 and 6 DF,  p-value: 0.00149
```

(b) Multiplicative model: $Y = aX^b$. Linearization: $Z := \ln(Y)$, $U := \ln(X)$, $a' := \ln(a)$. Then $Z = a' + bU$.

```
xp <- log(dist)
t2 <- lm(yp ~ xp)
summary(t2)
```

```
##
## Call:
## lm(formula = yp ~ xp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2165 -0.1295 -0.0113  0.1084  0.2969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   3.2139     0.3312    9.70  6.9e-05 ***
## xp           -0.5915     0.0761   -7.78  0.00024 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.185 on 6 degrees of freedom
## Multiple R-squared:  0.91,Adjusted R-squared:  0.895
## F-statistic: 60.5 on 1 and 6 DF,  p-value: 0.000238
```

(c) Another model: $Y = \frac{1}{a+bX}$. Linearization: $V := 1/Y$, then $V = a + bX$.

```
yb <- 1/motn
t3 <- lm(yb ~ dist)
summary(t3)
```

```
##
## Call:
## lm(formula = yb ~ dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16568 -0.08000  0.00388  0.06648  0.16675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.204817   0.069423    2.95  0.02560 *
## dist        0.003689   0.000537    6.87  0.00047 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.122 on 6 degrees of freedom
## Multiple R-squared:  0.887,Adjusted R-squared:  0.869
## F-statistic: 47.3 on 1 and 6 DF,  p-value: 0.000467
```

(d) Yet another model: $Y = a + \frac{b}{X}$ Linearization: $W := 1/X$ and then $Y = a + bW$.

```
xb <- 1/dist
t4 <- lm(motn ~ xb)
summary(t4)
```

```
##
## Call:
## lm(formula = motn ~ xb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2676 -0.2134 -0.0519  0.1509  0.4870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.755      0.166    4.56   0.0039 **
## xb            78.091      6.704   11.65  2.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.285 on 6 degrees of freedom
## Multiple R-squared:  0.958,Adjusted R-squared:  0.951
## F-statistic:  136 on 1 and 6 DF,  p-value: 2.41e-05
```

Conclusion: Among the proposed models, the best fit is obtained for $Y = a + \frac{b}{X}$. Let us

predict the vibrations at 100 km.

```
pw <- data.frame(xb = 1/100)
predict(t4, pw, interval = "confidence")
```

```
##     fit   lwr   upr
## 1 1.536 1.244 1.829
```

⊡