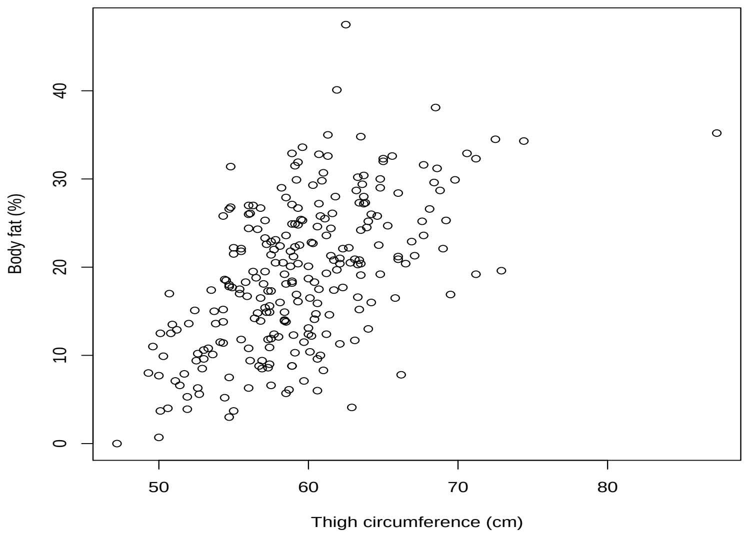
University of California, Los Angeles
Department of Statistics

Nicolas Christou
Ivo Dinov

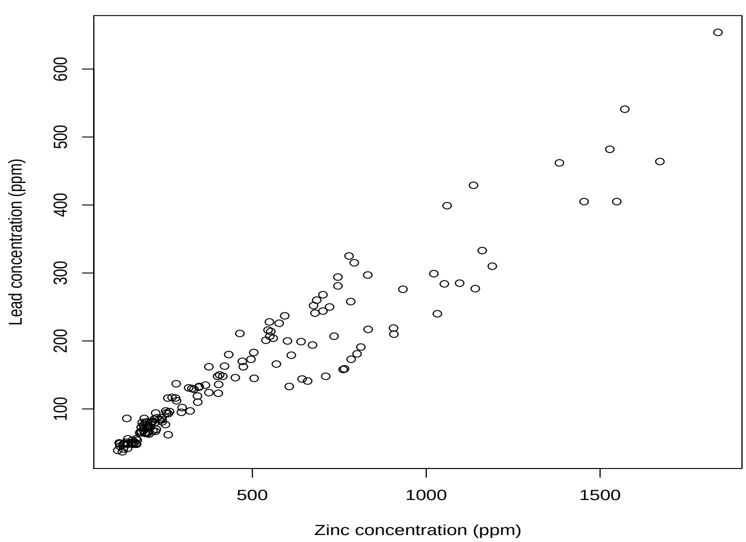
Simple regression analysis

Introduction:

Regression analysis is a statistical method aiming at discovering how one variable is related to another variable. It is useful in predicting one variable from another variable. Consider the following "scatterplot" of the percentage of body fat against thigh circumference (cm). This data set is described in detail in the handout on R.



And another one:



What do you observe?

Is there an equation that can model the picture above?

- Regression model equation:

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

  where

  - y dependent variable (random)
  - x independent variable (non-random)
  - β_0 intercept (non-random)
  - β_1 slope (non-random)
  - ϵ random error term, $\epsilon \sim N(0, \sigma)$

- Using the method of least squares we estimate β_0 and β_1:

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

  or easier for calculations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$$

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

- The fitted line is:

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

- The difference between the observed and the fitted y_i is the residual. It is computed as

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Covariance between y and x:

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

- Coefficient of correlation (r):

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\texttt{cov}(x, y)}{\texttt{sd}(x)\texttt{sd}(y)}$$

  Or easier for calculations:

$$r = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}\sqrt{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}}}$$

  Always $-1 \leq r \leq 1$.

- Useful things to know:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n} \quad \text{and} \quad \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}$$

Simple regression analysis - example

The data below give the mileage per gallon (y) obtained by a test automobile when using gasoline of varying octane (x):

y	x	xy	y^2	x^2
13.0	89	1157.0	169.00	7921
13.5	93	1255.5	182.25	8649
13.0	87	1131.0	169.00	7569
13.2	90	1188.0	174.24	8100
13.3	89	1183.7	176.89	7921
13.8	95	1311.0	190.44	9025
14.3	100	1430.0	204.49	10000
14.0	98	1372.0	196.00	9604
$\sum_{i=1}^{8}y_i=108.1$	$\sum_{i=1}^{8}x_i=741$	$\sum_{i=1}^{8}x_iy_i=10028.2$	$\sum_{i=1}^{8}y_i^2=1462.31$	$\sum_{i=1}^{8}x_i^2=68789$

a. **Find the least squares estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.**

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}=\frac{10028.2-\frac{1}{8}(741)(108.1)}{68789-\frac{741^2}{8}}=0.100325.$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=\frac{108.1}{8}-0.100325\frac{741}{8}=4.2199.$$

Therefore the fitted line is: $\hat{y}_i=4.2199+0.100325x_i$.

b. **Compute the fitted values and residuals.**

Using the fitted line $\hat{y}_i=4.2199+0.100325x_i$ we can find the fitted values and residuals. For example, the first fitted value is: $\hat{y}_1=4.2199+0.100325(89)=13.1488$, and the first residual is $e_1=y_1-\hat{y}_1=13.0-13.1488=-0.14888$, etc. The table below shows all the fitted values and residuals.

$\hat{y}_i$	e_i	e_i^2
13.14883	-0.14882	0.02215
13.55013	-0.05013	0.00251
12.94818	0.05183	0.00269
13.24915	-0.04915	0.00242
13.14883	0.15118	0.02285
13.75078	0.04922	0.00242
14.25240	0.04760	0.00227
14.05175	-0.05175	0.00268
	$\sum_{i=1}^{n}e_i=0$	$\sum_{i=1}^{n}e_i^2=0.05998$

c. Compute the covariance between y and x.

$$\texttt{cov}(y,x) = \frac{1}{8-1}\left[10028.2 - \frac{1}{8}(741)(108.1)\right] = 2.21.$$

d. verify that $\texttt{sd}(x) = 4.689$ and $\texttt{sd}(y) = 0.479$ and then calculate the correlation coefficient.

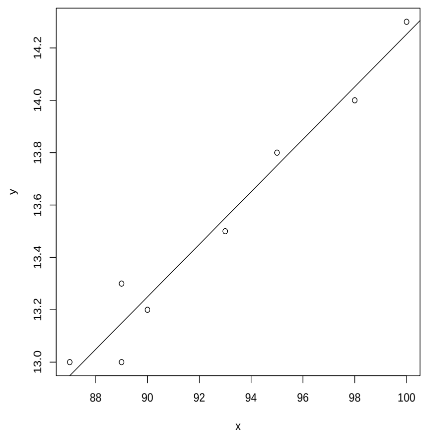
$$r = \frac{2.21}{(0.479)(4.689)} = 0.984.$$

The same example can be done with few simple commands in R:

```
#Enter the data:
> x <- c(89,93,87,90,89,95,100,98)
> y <- c(13,13.5,13,13.2,13.3,13.8,14.3,14)

#Run the regression of y on x:
> ex <- lm(y ~x)

#Display the results:
> summary(ex)

Call:
lm(formula = y ~ x)

Residuals:
       Min         1Q     Median         3Q        Max
-0.1488221 -0.0505280 -0.0007717  0.0498781  0.1511779

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.21990    0.74743   5.646  0.00132 **
x            0.10032    0.00806  12.447 1.64e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09999 on 6 degrees of freedom
Multiple R-squared: 0.9627, Adjusted R-squared: 0.9565
F-statistic: 154.9 on 1 and 6 DF,  p-value: 1.643e-05


#Compute the covariance, standard deviations, and correlation coefficient:
> cov(y,x)
> sd(y)
> sd(x)
> cor(y,x)
```

Plot y on x and add the regression fitted line on the plot:

```
> ex <- lm(y ~ x)
> plot(x, y)
> abline(ex)
```



Here is the plot:
The object `ex` contains the following:

```
names(ex)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"
```

We can list the fitted values or the residuals using

```
ex$fitted.values
ex$residuals
```

Predict a new value of y using the function `predict`:

```
pred_new <- predict(ex, se.fit=TRUE, data.frame(x=96))
pred_new$fit
       1
13.85110
```

The value above was computed by:

$$\hat{y} = 4.2199 + 0.100325(96) = 13.8511.$$

For SOCR demos see here:

http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_AnalysisActivities_SLR

http://www.socr.ucla.edu/htmls/ana/SimpleRegression_Analysis.html

Interactive Regression: http://www.socr.ucla.edu/htmls/game/Bivariate_Game.html

Simple regression in R - examples

Example 1:
We will use the following data:

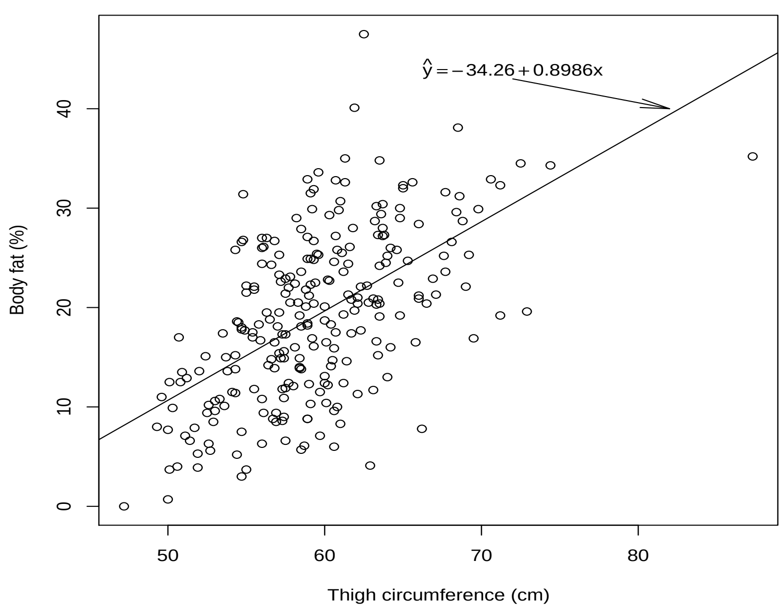
```
data1 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/body_fat.txt", header=TRUE)
```

This file contains data on percentage of body fat determined by underwater weighing and various body circumference measurements for 251 men. Here is the variable description:

Variable	Description
x_1	Density determined from underwater weighing
x_2	Percent body fat from Siri's (1956) equation
x_3	Age (years)
x_4	Weight (lbs)
x_5	Height (inches)
x_6	Neck circumference (cm)
x_7	Chest circumference (cm)
x_8	Abdomen 2 circumference (cm)
x_9	Hip circumference (cm)
x_{10}	Thigh circumference (cm)
x_{11}	Knee circumference (cm)
x_{12}	Ankle circumference (cm)
x_{13}	Biceps (extended) circumference (cm)
x_{14}	Forearm circumference (cm)
x_{15}	Wrist circumference (cm)

We want to run the regression of x_2 (percentage body fat) on x_{10} (thigh circumference). Here is the regression output:

```
ex1 <- lm(data1$x2 ~data1$x10)
summary(ex1)

Call:
lm(formula = data1$x2~ data1$x10)

Residuals:
     Min       1Q   Median       3Q      Max
-18.1601  -4.7707  -0.1076   4.5219  25.5994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -34.26252    4.99529  -6.859 5.46e-11 ***
data$x10      0.89861    0.08373  10.732  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.947 on 249 degrees of freedom
Multiple R-squared: 0.3163, Adjusted R-squared: 0.3135
F-statistic: 115.2 on 1 and 249 DF,  p-value: < 2.2e-16
```



Example 2:
Access the data:

```
data2 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/soil_complete.txt", header=TRUE)
```

This data set consists of 6 variables. The first two columns are the x and y coordinates, and the last four columns are the concentration of cadmium, copper, lead and zinc in *ppm* at 155 locations. We will run the regression of lead against zinc. Our goal is to build a regression model to predict the lead concentration from the zinc concentration. Here is the regression output.

```
ex2 <- lm(data2$lead ~data2$zinc)
summary(ex2)

Call:
lm(formula = data2$lead ~ data2$zinc)

Residuals:
    Min      1Q  Median      3Q     Max
-79.853 -12.945  -1.646  15.339 104.200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.367688   4.344268   3.998 9.92e-05 ***
data2$zinc   0.289523   0.007296  39.681  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 33.24 on 153 degrees of freedom
Multiple R-squared: 0.9114, Adjusted R-squared: 0.9109
F-statistic:  1575 on 1 and 153 DF,  p-value: < 2.2e-16
```

For more data sets on the SOCR website:

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data