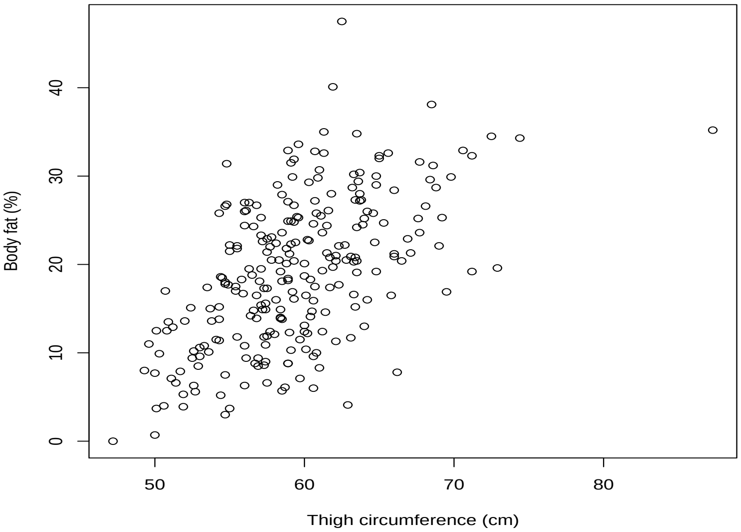
University of California, Los Angeles
Department of Statistics

Statistics 100C **Instructor: Nicolas Christou**

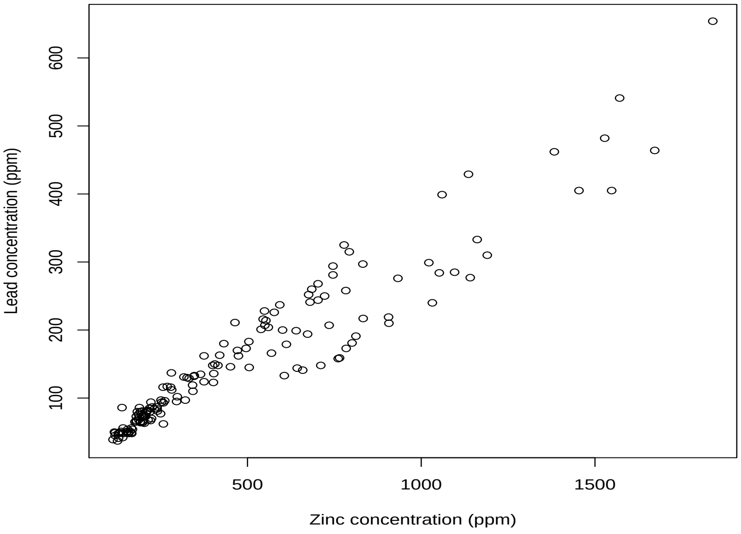
Simple regression analysis

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: This is the concentration of lead against the concentration of zinc (see handout on `R` for more details on this data set).



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 response variable (random)
  - x predictor 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:

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

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

- The fitted line is:

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

- Distribution of $\hat{\beta}_1$ and $\hat{\beta}_0$:

  $$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right), \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

- The standard deviation σ is unknown and it is estimated with the "residual standard error" which measures the variability around the fitted line. It is computed as follows:

  $$s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$$

  where

  $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is called the residual (the difference between the observed y_i value and the fitted value $\hat{y}_i$.

- Coefficient of determination:
  The total variation in y (total sum of squares $SST = \sum_{i=1}^n (y_i - \bar{y})^2$) is equal to the regression sum of squares ($SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$) plus the error sum of squares ($SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$):

  $$SST = SSR + SSE$$

  The percentage of the variation in y that can be explained by x is called coefficient of determination (R^2):

  $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad \text{Always} \;\; 0 \leq R^2 \leq 1$$

- Useful:

  $$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \Rightarrow SST = (n-1)s_y^2 \quad \text{where } s_y^2 \text{ is the variance of } y.$$

- Coefficient of correlation (r):

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

  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 $-1 \leq r \leq 1$ and $R^2 = r^2$.

- Another formula for r:

  $$r = \hat{\beta}_1 \frac{s_x}{s_y}$$

  where s_x, s_y are the standard deviations of x and y.

- Sample covariance between y and x:

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

  Therefore

  $$r = \frac{\text{cov}(x, y)}{s_x s_y} \Rightarrow \text{cov(x,y)} = r s_x s_y \;\; \text{and} \;\; \hat{\beta}_1 = r \frac{s_y}{s_x}$$

- Standard error of $\hat{\beta}_1$ and $\hat{\beta}_0$:

  $$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{s_e}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}}$$

  and

  $$s_{\hat{\beta}_0} = s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}}$$

- Testing for linear relationship between y and x:

  $$\begin{aligned} H_0 &: \quad \beta_1 = 0 \\ H_a &: \quad \beta_1 \neq 0 \end{aligned}$$

  Test statistic:

  $$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

  Reject H_0 (i.e. there is linear relationship) if $t > t_{\frac{\alpha}{2};n-2}$ or $t < -t_{\frac{\alpha}{2};n-2}$

- Confidence interval for β_1:

  $$\hat{\beta}_1 - t_{\frac{\alpha}{2};n-2}\ s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2};n-2}\ s_{\hat{\beta}_1}$$

  Or β_1 falls in:

  $$\hat{\beta}_1 \pm t_{\frac{\alpha}{2};n-2}\ s_{\hat{\beta}_1}$$

- Prediction interval for y for a given x (when $x_i = x_g$):

  $$\hat{y}\text{g} \pm t_{\frac{\alpha}{2};n-2}\ s_e\sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \text{where } \hat{y}_g = \hat{\beta}_0 + \hat{\beta}_1 x_g.$$

- Confidence interval for the mean value of y for a given x (when $x_i = x_g$):

  $$\hat{y}\text{g} \pm t_{\frac{\alpha}{2};n-2}\ s_e\sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \text{where } \hat{y}_g = \hat{\beta}_0 + \hat{\beta}_1 x_g.$$

- Useful things to know:

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

Simple regression analysis - A simple 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
$\sum_{i=1}^{8} y_i = 108.1$	$\sum_{i=1}^{8} x_i = 741$	$\sum_{i=1}^{8} x_i y_i = 10028.2$	$\sum_{i=1}^{8} y_i^2 = 1462.31$	$\sum_{i=1}^{8} x_i^2 = 68789$

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

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

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

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. **Find the estimate of σ^2.**

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{0.05998}{8-2} = 0.009997.$$

Therefore, $s_e = \sqrt{0.009997} = 0.09999$.

d. **Compute the standard error of $\hat{\beta}_1$.**

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}} = \frac{0.09999}{\sqrt{68789 - \frac{741^2}{8}}} = 0.00806.$$

e. **Construct a 95% confidence interval for $\hat{\beta}_1$.**
The parameter β_1 falls in:

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2};n-2}\ s_{\hat{\beta}_1} \quad \text{or} \quad 0.100325 \pm 2.447(0.00806)$$

Therefore we are 95% confident that β_1 falls in the interval: $0.0806 \leq \beta_1 \leq 0.12$.

f. **Estimate the miles per gallon for an octane gasoline level of 94.**

$$\hat{y} = 4.2199 + 0.100325(94) = 13.65.$$

g. **Compute the coefficient of determination, R^2.**

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n e_i^2}{(n-1)s_y^2} = 1 - \frac{0.05998}{7(0.2298)} = 0.9627.$$

Therefore, 96.27% of the variation in Y can be explained by x.

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
```

Simple regression in R - examples

Example 1:
We will use the following data:

```
data1 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics100C/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
y	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 Y (percentage body fat) on x_2 (thigh circumference). Here is the regression output:

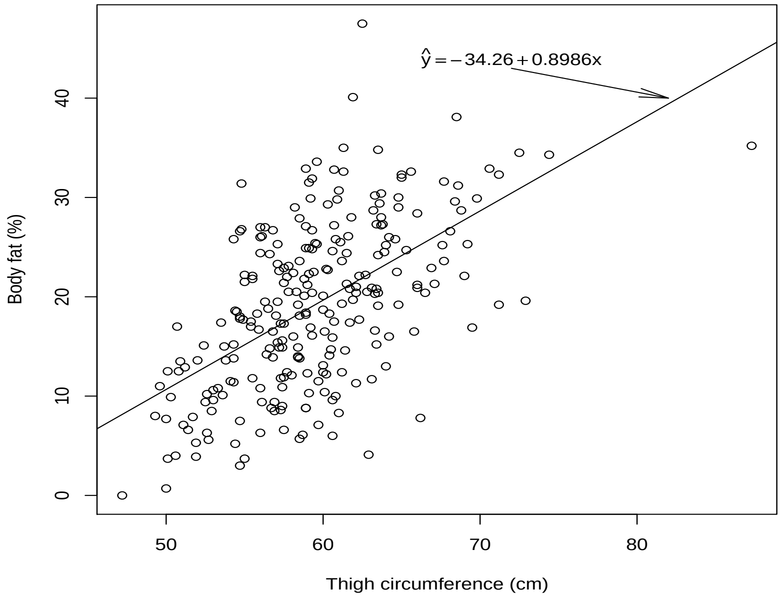
```
ex1 <- lm(data1$y ~data1$x10)
summary(ex1)

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:
Here are the data:

```
data2 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics100C/soil.txt", header=TRUE)
```

This data set consists of 4 variables. The first two columns are the x and y coordinates, and the last two columns are the concentration of 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 
```

Exercise:

a. Construct the histogram of `lead` and `zinc` and comment.

b. Transform the data to get a bell-shaped histogram.

c. Plot the transform data of lead on the transform data of zinc and compare this scatterplot with the scatterplot of the original data.

d. Run the regression of the transform data of lead on the transform data of zinc and compare the R^2 of this regression to the R^2 using the original data.