STA 4210 Regression Analysis

Deborah Burr

August 17, 2018

OVERVIEW: REGRESSION ANALYSIS

What is STA 4210 about?

Two key things to tune into:

- Relationships
- Models

What can you do with regression?

- Explore data
- Make predictions
- Do inference

Example: The Association Between Number of Beers Consumed and Blood Alcohol Content

Sixteen individuals at Ohio State University took part in an experiment to determine how the number of 12 oz. beers consumed affects Blood Alcohol Content as measured by the breathalyzer machine. Subjects selected from a hat a piece of paper that had a number on it (1–9), and they had to consume that number of beers.

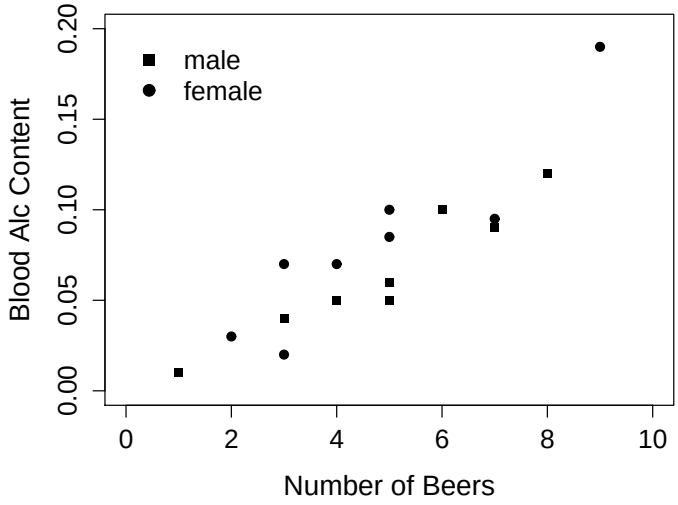
```
BAC  No.of.Beers  Weight  Sex
.10  5            132     female
```

General questions:

- How strongly related are the variables BAC and No.of.Beers?
- How can you predict BAC for a person if you know their No.of.Beers?
- How can you assess the accuracy of your prediction?
- Are there other ways you can use this data set to form better predictions?

Full Data for Blood Alcohol Content Example

BAC	NOB	W	S
.10	5	132	f
.03	2	128	f
.19	9	110	f
.12	8	192	m
.04	3	172	m
.095	7	250	f
.07	3	125	f
.06	5	175	m
.02	3	175	f
.05	5	275	m
.07	4	130	f
.10	6	168	m
.085	5	128	f
.09	7	246	m
.01	1	164	m
.05	4	175	m



In the BAC example, the two variables $X =$ No.of.Beers and $Y =$ BAC are both quantitative/continuous.

The overall trend of the relationship in the scatterplot looks linear.

Our first model assumes that a straight line is an appropriate summary of the overall trend.

In this example, our first goal is prediction. What about the goals of exploration and inference?

Discuss.

Example. What Variables are Associated with House Selling Price?

We study the *general linear model* in this course, and the model allows more than one predictor variable. The multiple regression model is introduced in Chapters 6 and 7.

The data set "house selling prices" contains observations on 200 home sales in Oregon. There are ten variables in the dataset; below are the first three rows:

First three rows:

```
  price size acres     lot beds baths age garage condition agec
1 232.5 1679  0.23 10018.8    3   1.5  35      1         0    M
2 470.0 4494  0.52 22651.2    5   4.0  38      1         0    M
3 150.0 2542  0.11  4791.6    4   0.0   5      1         0    N
```

Question: How can we use this data to discover which variables as a group, best explain `price`?

This is not an inference question. However, there are some statistical tools which enable us to approach this data exploration in a systematic fashion. (Ch. 9)

The General Linear Model

The general linear model includes simple linear regression but goes far beyond it. (There is always just one response variable Y.)

For example, the model allows:

- Two or more quantitative predictors $(X_1, X_2, \ldots, X_k)$ (multiple regression, Chs. 6 and 7).
- Categorical predictors (e.g. analysis of variance models, STA 4211)
- Curvature in the relationship between Xs and Y (quadratic regression, Section 8.1).

We go into other ways the model can be adapted to real-world situations in Chapter 8.

Tools

- Inferential methods are mainly based on the t and F methods, in Chs. 6 and 7.
- We will need linear algebra (matrix arithmetic) to handle the multiple regression model. The necessary background will be covered in the course (Ch. 5).
- The statistical programming language R for computation and graphics.

Outline of topics

- Some review of Statistics I (normal distribution, t procedures for comparing two means, correlation coefficient r)
- Simple linear regression model, and fitting the model (Ch. 1). Includes least-squares and maximum likelihood.
- Introduction to R
- Inference in simple linear regression (2.1, 2.5, 2.7, 2.8, 2.9)
- Some useful diagnostic tools, from Ch. 3
- Matrix algebra for simple regression, Ch. 5
- Multiple regression, Ch. 6
- Special issues in multiple regression, Ch. 7
- Quadratic regression, Ch. 8
- Variable selection, Ch. 9

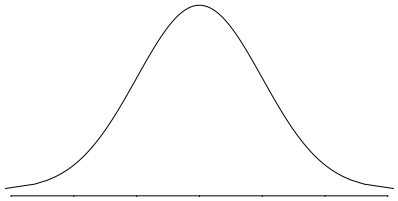
REVIEW of STATISTICS I

Topics:

- The histogram
- The normal distribution
- The two-sample pooled t procedures for inference
- The t distribution
- Measuring association between two continuous variables (Pearson's r)

The Normal Distribution

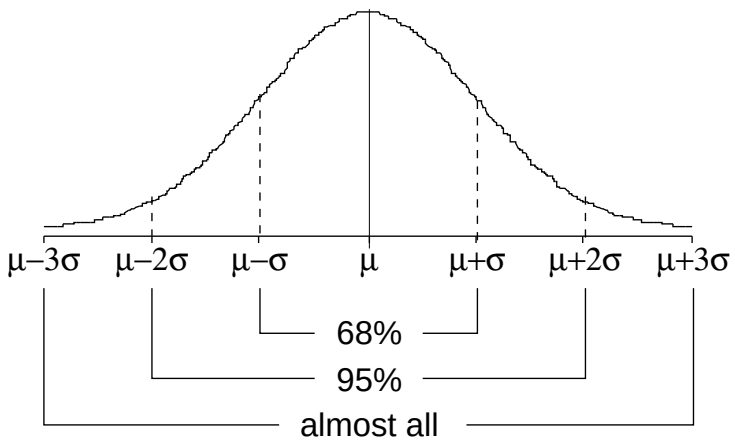
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



μ and σ are parameters of this distribution. Effect on curve if μ increases? Effect on curve if σ decreases?

Basic Properties of the Normal Distribution

1. Curve is symmetric, centered at the mean μ.
2. 50% of area lies to right of μ.
3. The SD σ measures the spread: The bigger the value of σ, the more spread out and flatter the curve.
4. Area under the curve is always 1.
5. (a) 68% of area is within σ of μ,
   (b) 95% of area is within 2σ of μ,
   (c) $\sim 99.7\%$ of area is within 3σ of μ.



Comparison of means from two normal samples (any sample sizes)

Basic Framework

Assume:

1. The Y's are a random sample of size n_Y from a normal population with mean μ_Y and SD σ_Y, both unknown; $\bar{Y}$ is the sample mean and S_Y is the sample SD.
2. The Z's are a random sample of size n_Z from a normal population with mean μ_Z and SD σ_Z, both unknown; $\bar{Z}$ is the sample mean and S_Z is the sample SD.
3. The two samples are independent of one another.
4. The two standard deviations are equal; that is, $\sigma_Y = \sigma_Z$.

Want:

1. A confidence interval for $\mu_Y - \mu_Z$.
2. Hypothesis tests concerning μ_Y and μ_Z.

The standard two-sample t-test:

Form

$$T = \frac{\bar{Y} - \bar{Z}}{\sqrt{\left(\frac{1}{n_Y} + \frac{1}{n_Z}\right)\frac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{(n_Y + n_Z - 2)}}}.$$

If the null hypothesis is true, and the two population variances are equal, then the distribution of this is t with $n_Y + n_Z - 2$ df.

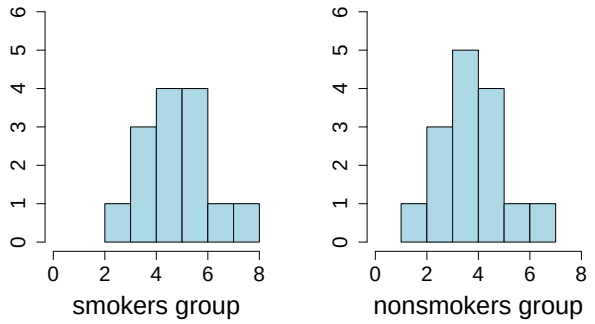
Example: A physician named A.W. Andrews developed way of scoring a photograph of a face for wrinkles. The scores range from 1 to 10, with 1 indicating no wrinkles, 10 indicating a severe case.

He photographed 29 women aged 45–55, of whom 14 smoked a pack a day and 15 did not smoke. He reported that the 14 who smoked can be considered a random sample from the population of women who smoke a pack a day, and the 15 who didn't smoke can be considered a random sample from the population of women who don't smoke.[1]

Scores:

S: $2, 6, 4, 6, 4, 5, 5, 6, 5, 6, 7, 5, 4, 8$ $n_s = 14$ $\bar{Y}_s = 5.2$ $S_s = 1.48$

NS: $1, 5, 3, 5, 3, 4, 4, 5, 4, 5, 6, 4, 3, 7, 4$ $n_{ns} = 15$ $\bar{Y}_{ns} = 4.2$ $S_{ns} = 1.42$



[1] Obviously, this can't really be true.


```
Welch Modified Two Sample t-test

data: smokers and nonsmokers
t=1.88, df=26.69, p-value=0.0711
alternative hyp.: true difference in means is not 0
95 percent confidence interval: (-0.09,  2.12)
sample estimates:  mean of y: 5.21  mean of z: 4.20
```

```
Standard Two Sample t-test

data: smokers and nonsmokers
t=1.88, df=27, p-value=0.0706
alternative hyp.: true difference in means is not 0
95 percent confidence interval: (-0.09, 2.12)
sample estimates:  mean of y: 5.21  mean of z: 4.20
```

Comments on the Analysis

- Conclusion:

- The CI is $(-0.09, 2.12)$, which means that we are "95% confident" that $-0.09 < \mu_{\text{s}} - \mu_{\text{ns}} < 2.12$. This interval contains 0.

- Need to do an informal check of normality. You do this by looking at the histograms. The assumption of normality is not really critical, and unless there is something glaring, you usually don't worry about it.

- Side note on Welch's method. The number of df is not necessarily an integer. Actually, there are t-distributions with any number of df. They are not tabulated, but this is not a problem, because you will always do this using a software package anyway.

The standard two-sample t procedures: Theory

Basic Facts

1. $E(\bar{Y} - \bar{Z}) = \mu_Y - \mu_Z$
2. $\text{SD}(\bar{Y} - \bar{Z}) = \sigma\sqrt{\dfrac{1}{n_Y} + \dfrac{1}{n_Z}}$
3. $\bar{Y} - \bar{Z}$ has a normal distribution.
4. $\dfrac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{\sigma\sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}} \sim \mathcal{N}(0, 1).$
5. $\dfrac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{\hat{\sigma}\sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}} \sim t_{n_Y + n_Z - 2}$

,

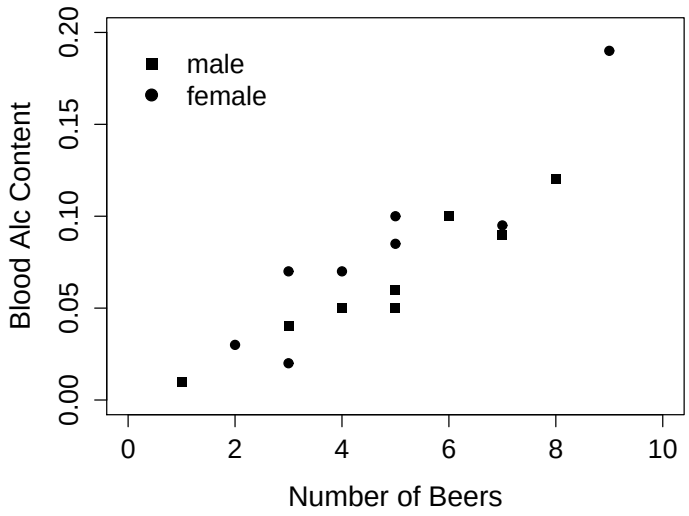
where $\hat{\sigma} = \dfrac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{(n_Y + n_Z - 2)}$

Correlation

We will discuss:

- What correlation is, and what it does.
- Caveats about the use of correlation.

Recall the BAC data, p. 3. Below is the scatterplot of $Y =$ BAC vs. $X =$ No.of.Beers again.



Question: How strongly related are the variables BAC and No.of.Beers?

Correlation

Recall: If X and Y are two random variables, the correlation between them is

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X \sigma_Y}.$$

Suppose we are considering n individuals, and for each we measure two variables X and Y, obtaining pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The *Pearson correlation coefficient* (or just "correlation coefficient") is

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

Correlation is a measure of association between two variables X and Y.

The correlation coefficient measures *linear* association, i.e. the extent of clustering around a *straight* line.

The correlation coefficient is always between -1 and $+1$.

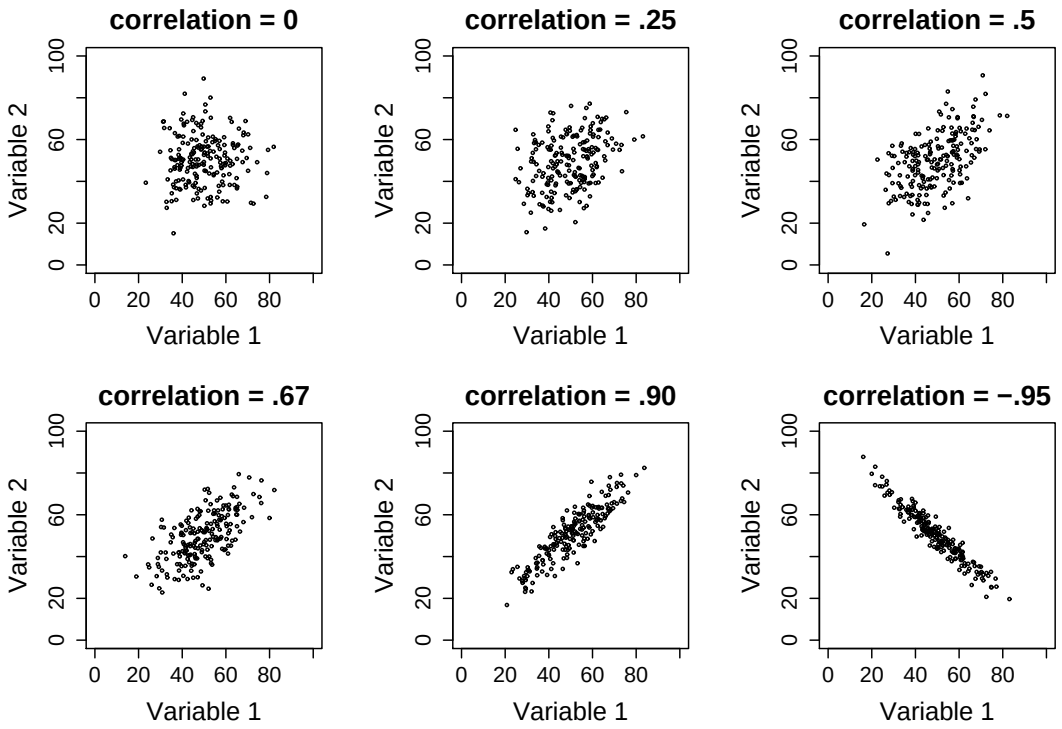
The closer the correlation is to 1 (or -1), the stronger the association. If the correlation is 0 then there is no (linear) association.

When a scatterplot shows an increasing trend (i.e. as x increases, y generally increases), the correlation is positive.

When a scatterplot shows a decreasing trend (i.e. as x increases, y generally decreases), the correlation is negative.

We will see later that r^2 (called the "coefficient of determination") plays an important role in regression analysis.

The actual numerical value of r is difficult to interpret; i.e. does correlation $= .7$ indicate a strong association?

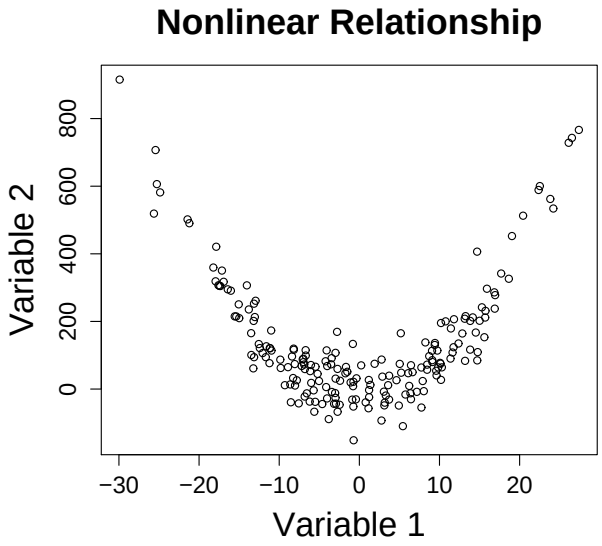


In R:

```
> cor(BAC, No.of.Beers)
[1] 0.8943381
```

Cautions:

1. **Beware of overinterpretation of correlations.** For example, .5 sounds impressive, but is actually very weak.
2. **Correlation is a measure of *linear* association.** Linear refers to a *straight* line. Two variables may be strongly associated, but if association is not linear, then correlation may not be high.

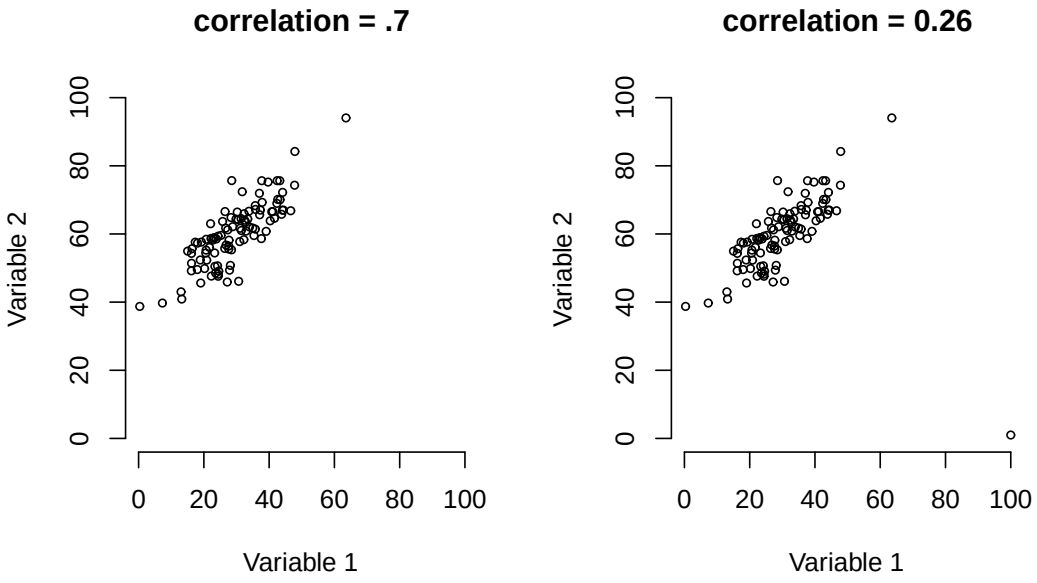


Therefore:

- If you want to see whether there is a relationship between two variables, should always look at the scatterplot.
- Use correlation coefficient to summarize association only if cloud of points looks like an ellipsoid.

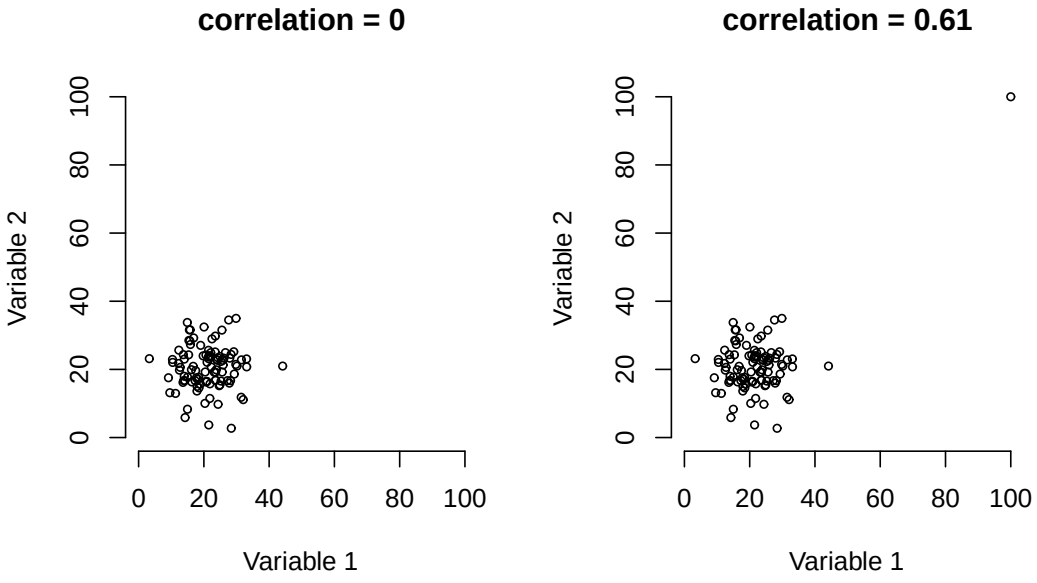
3 Correlation is very sensitive to outliers.

(a) A single observation can bring a high correlation way down.



3 Correlation is very sensitive to outliers.

(b) A single observation can make correlation close to 1, when in fact there really isn't any association.



ASSOCIATION AND CAUSATION

Example: Two surveys of housing prices were done, one in Gainesville, other in Boston. In each case two variables were recorded:

Variable 1: temperature outside at time of sale.

Variable 2: selling price.

Noted that $r = -.8$

??

4 **Association is not causation.**
Sometimes there is a strong correlation, but this is due to some third ("confounding") variable that is affecting variables 1 and 2.

A variable is a confounding variable if it has an effect on both the response and the explanatory variables, but is not one of the explanatory variables listed in the data summary.

Selling Price / Temperature example:

Recall:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Correlation is not affected by any of the following:

1. Adding a fixed number (e.g. 8) to variable 1.
2. Same for variable 2.
3. Multiplying by a fixed positive number (e.g. 2) variable 1.
4. Same for variable 2.
5. Interchanging roles of variables 1 and 2.

Can simultaneously do any combination of 1–5 and the correlation will still not change.

CHAPTERS 1 and 2 Linear Regression with One Predictor Variable

Overview

- The regression model, with and without the normality assumption, 1.1, 1.3 and 1.8
- Use of R for regression, R tutorial and lecture notes
- Least squares and the regression line, 1.6
- Examples of uses of the regression line, 1.2, 1.4 and 1.5
- Variation about the regression line, 1.7
- Informal prediction
- Maximum likelihood for simple linear regression, 1.8
- Confidence intervals and hypothesis tests for the slope (and other) parameters, 2.1, 2.4
- Using regression methods for prediction (prediction intervals), 2.5
- Sums of squares and the coefficient of determination, 1.7 and 2.7