SF 2930 Regression Analysis

Lecture 1

Simple Linear Regression

Timo Koski

KTH Royal Institute of Technology

2023-01-18

Seal, Hilary L: Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model, Biometrika, 1967

Regression analysis is regarded as one of the oldest and most time-tested topics in mathematical statistics. The earliest form of the linear regression was the least squares method, which was published by Adrien-Marie Legendre, in 1805, and by Georg Friedrich Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun.

Nowadays, linear regression plays an important role, e.g., in machine learning. The linear regression algorithm is one of the fundamental supervised machine-learning algorithms due to its relative simplicity and well-known properties.

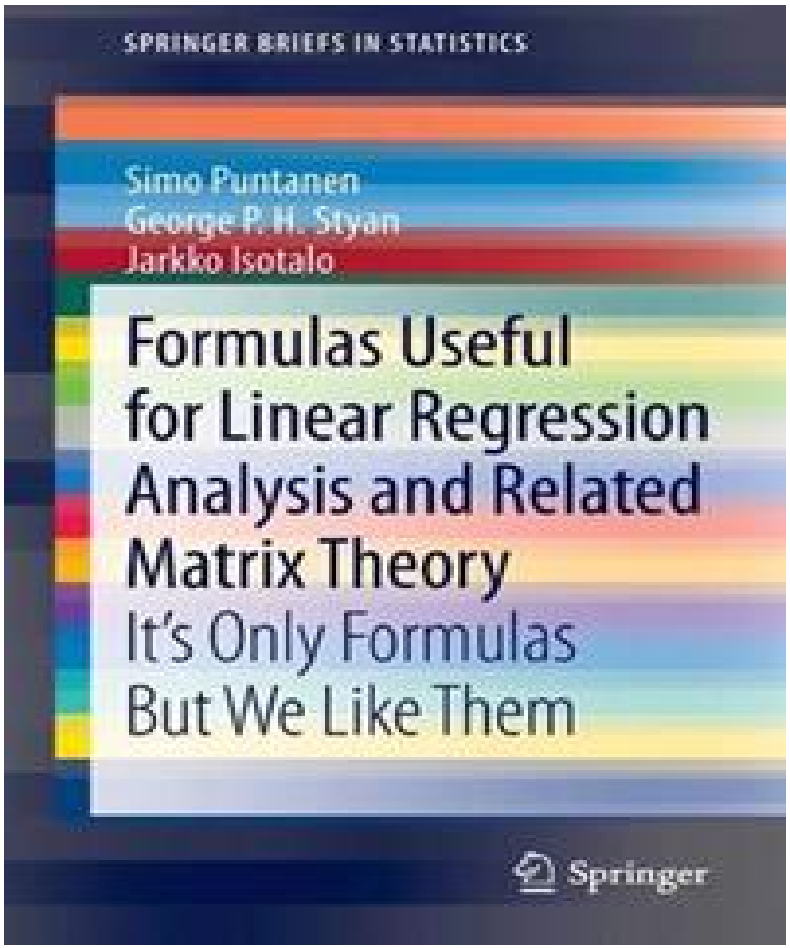
- Multiple Linear Regression is a special instance of Feedforward Neural Networks with a single layer, to be seen later.
- Graph Regression
- Control Engineering (System Identification), Signal Processing, Time Series Analysis
- Genetic Epidemiology
- Econometrics

The Textbook

D. Montgomery, E. Peck, G. Vining: Introduction to Linear Regression Analysis. WileyInterscience, 5th Edition (2012).
ISBN-10: 978-0-470-54281-1. 645 pages.
Acronym in the sequel: MPV. MPV is digitally available via KTHB.

There is a complete solutions manual : D. Montgomery, E. Peck, G. Vining: Solutions Manual to accompany Introduction to Linear Regression Analysis. 5th Edition, also digitally available via KTHB.

Matrix Calculus Formuas for Linear Regression Analysis (digitally available at KTHB)



Simple linear regression

The theory and practice of regression deals with the following situation: There are n pairs of values (real numbers)

$$\mathcal{D}_{tr} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

- y_i: values of a dependent variable y a.k.a. called the 'outcome' or 'response' variable, or a 'label' in machine learning jargon.

Simple linear regression

The theory and practice of regression deals with the following situation: There are n pairs of values (real numbers)

$$\mathcal{D}_{tr} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

- y_i: values of a dependent variable y a.k.a. called the 'outcome' or 'response' variable, or a 'label' in machine learning jargon.
- x_i: values of an independent variable x a.k.a 'predictor', 'covariate', 'explanatory variable' or 'feature'.

Simple linear regression

The theory and practice of regression deals with the following situation: There are n pairs of values (real numbers)

$$\mathcal{D}_{tr} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

- y_i: values of a dependent variable y a.k.a. called the 'outcome' or 'response' variable, or a 'label' in machine learning jargon.
- x_i: values of an independent variable x a.k.a 'predictor', 'covariate', 'explanatory variable' or 'feature'.
- $$y = \beta_0 + \beta_1 x \tag{1}$$

  is called the *theoretic line of regression*. This line is not likely to hold exactly in $\mathcal{D}_{tr}$. The idea is to **fit** a line of this kind to the data in $\mathcal{D}_{tr}$ in an approximate sense. We see next how this is done.

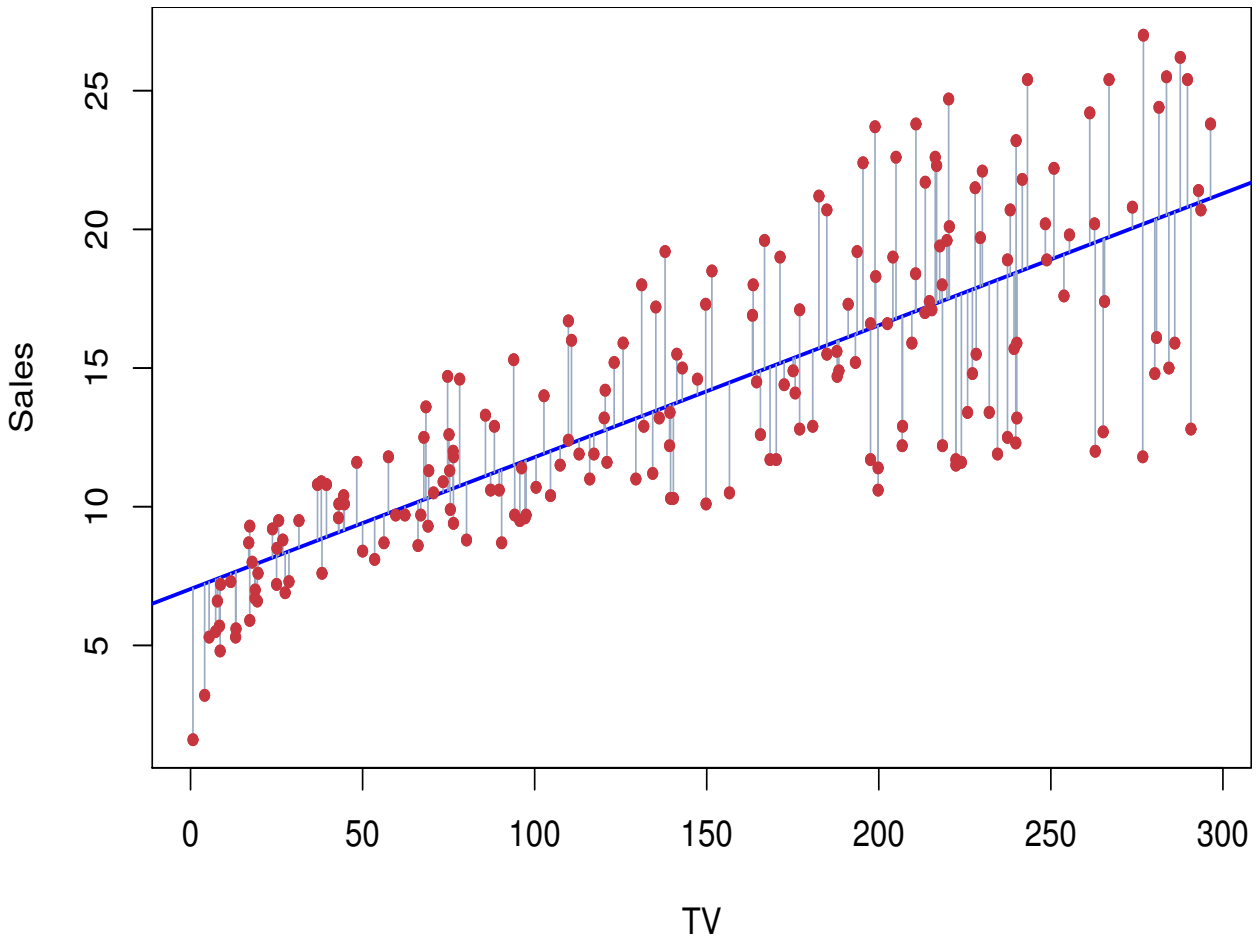
Simple linear regression & Least Squares 1.

We fit a line like the theoretic line of regression by estimating the parameters β_0 and β_1 by the Method of Least Squares, i.e. we minimize the sum of the squared **vertical** distances between the theoretic line and values of the response y, i.e., we minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

This is also known (as a special case of) Ordinary Least Squares (OLS) regression

LSE: Vertical Distances[1]



[1] by Courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert *An introduction to statistical learning*, Chapter 3

Fitting: Simple linear regression & Least Squares 2.

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

β_0 in is called the intercept, β_1 is the slope. The values $\hat{\beta}_0$ and $\hat{\beta}_1$ attaining the minimum are known as the *Least Squares-Estimates (LSE)* of β_0 and β_1, respectively.

Fitting: Why not?

$$\bar{Q}(\beta_0, \beta_1) = \sum_{i=1}^{n} | \, y_i - \beta_0 - \beta_1 x_i \, |$$

Fitting: Simple linear regression & Least Squares 3.

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

By a straightforward partial differentiation w.r.t β_0 och β_1 followed by setting the partial derivatives equal to 0, and by solving the resulting system of two linear equations, we find (c.f. Appendix D) that the LSE are $\hat{\beta}_1$ and $\hat{\beta}_0$ given by

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

where $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

4.

The straight line

$$\hat{y} := \hat{\beta}_0 + \hat{\beta}_1 x.$$

is called the *estimated line of regression* or the *predictor*. The vertical distances e_i from y_i to the estimated line of regression at x_i,

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

are called *observed least squares residuals*. Q_0 is defined as

$$Q_0 := Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{e}_i^2.$$

and is called the **residual sum of squares** (RSS).

Simple linear regression & Least Squares

5.

Observed least squares residuals

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

$$Q_0 := Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{e}_i^2.$$

We can, of course, compute the observed residuals for any pair (β_0, β_1)

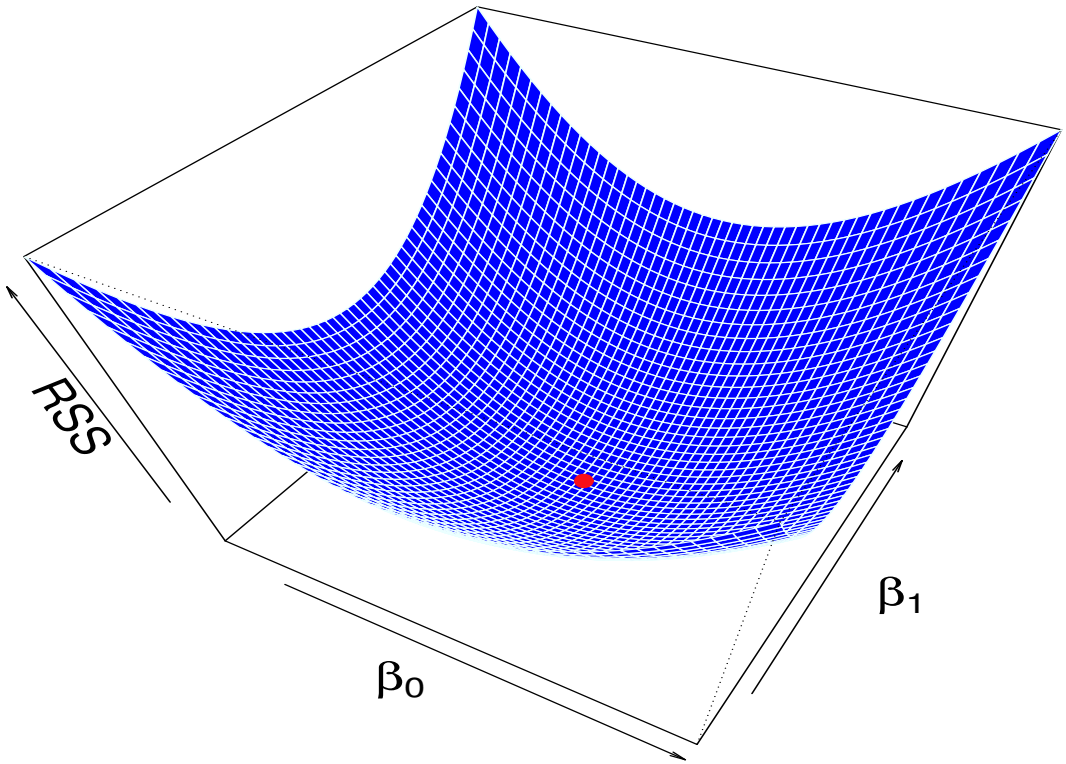
$$e_i = \beta_0 - \beta_1 x_i$$

and

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2.$$

The function $Q(\beta_0, \beta_1)$ is plotted for a data set in the book of James et.al. in the next slide.

LSE: RSS [2] $Q(\beta_0, \beta_1)$ PLOTTED, THE RED POINT IS $Q_0 := Q(\hat{\beta}_0, \hat{\beta}_1)$



[2] by Courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert *An introduction to statistical learning*, Chapter 3

Supervised Learning

In the parlance of machine learning, we have now used the training set

$$\mathcal{D}_{tr} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

to learn the regression model. The next step of learning in practice is to use a test set, i.e., data pairs from the same source not used in LSE

$$\mathcal{D}_{test} = \{(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})\}$$

and compute using the test residuals using the learned model predictor $e_i^{(t)} := y_{n+i} - \hat{\beta}_0 - \hat{\beta}_1 x_{n+i}$ for $i = 1, \ldots, m$, and the residual sum of squares $\sum_{i=1}^{m} \left(e_i^{(t)}\right)^2$ in order to compare the coefficients of determination (see below) for both the training set and the test set.

Semisupervised Learning

The training set is

$$\mathcal{D}_{tr} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

We want to learn the labels $\hat{y}$ for

$$\{x_{n+1}, \ldots, x_{n+m}\}$$

and do this by

$$\hat{y}_{n+i} = \hat{\beta}_0 - \hat{\beta}_1 x_{n+i} \quad \text{for } i = 1, \ldots, m.$$

Simple linear regression: A toy example

This can be done by hand (or by a hand-held calculator). The training set is

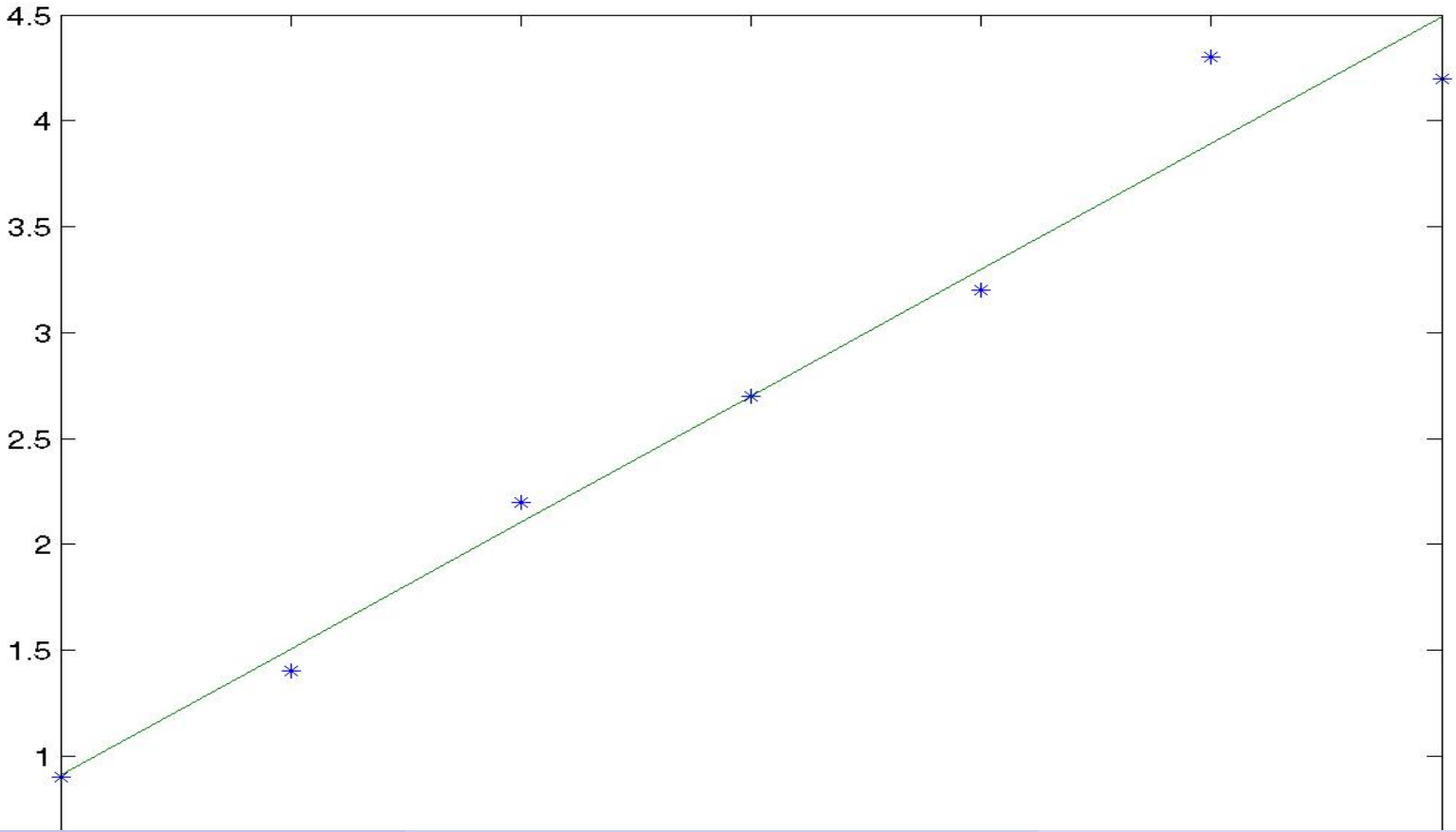
$$\mathcal{D}_{tr} = \{(1, 0.9), (2, 1.4), (3, 2.2), (4, 2.7), (5, 3.2), (6, 4.3), (7, 4.2)\}$$

LSE yields the predictor

$$\hat{y} = 0.3143 + 0.5964x, \quad Q_0 = 0.2796, \quad s^2 := \frac{\sum_{i=1}^{n} Q_0}{7-2} = 0.0559$$

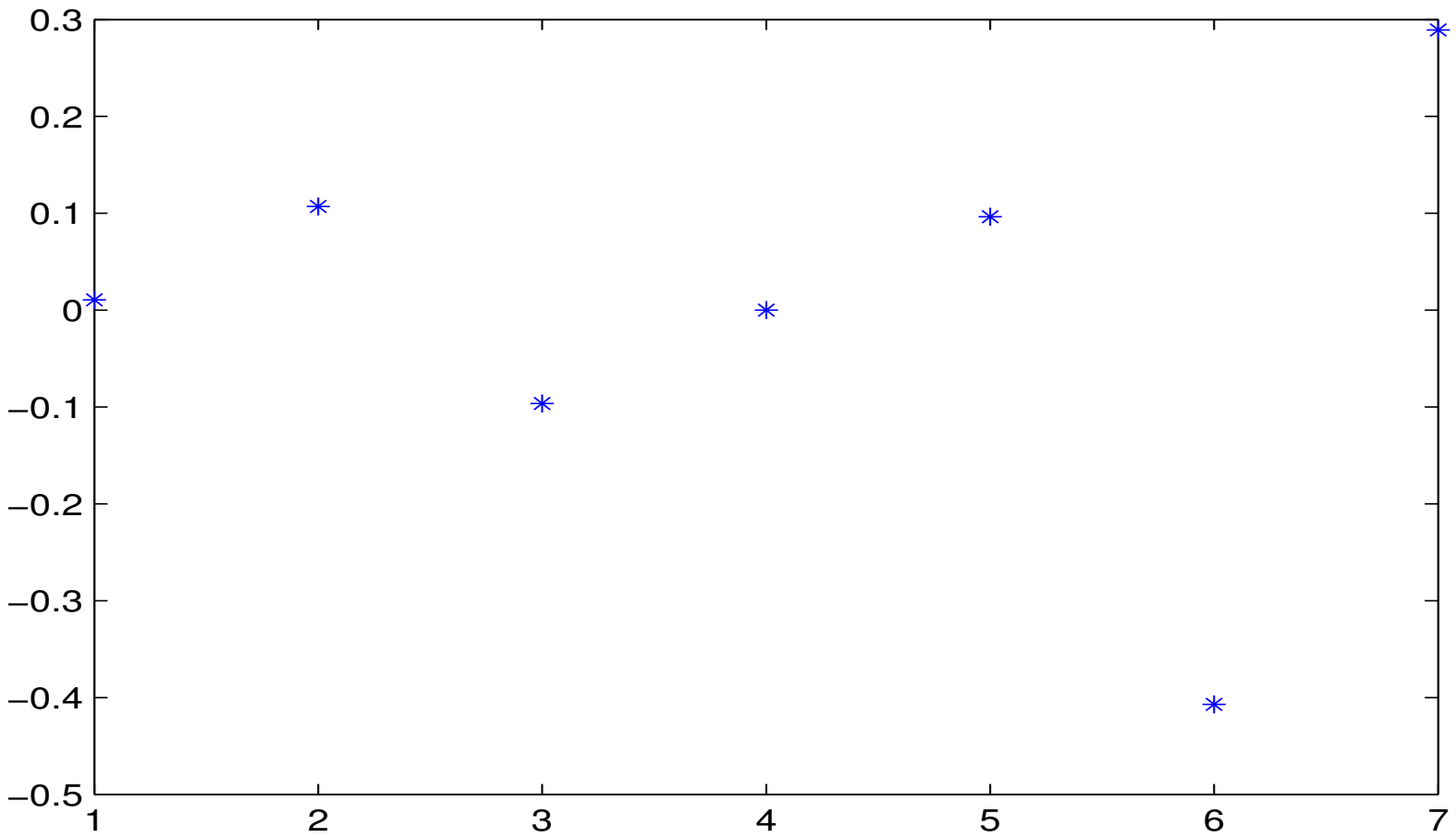
Simple linear regression: A toy example

$$\hat{y} = 0.3143 + 0.5964x, \quad Q_0 = 0.2796, \quad s^2 := \frac{\sum_{i=1}^{n} Q_0}{7-2} = 0.0559$$

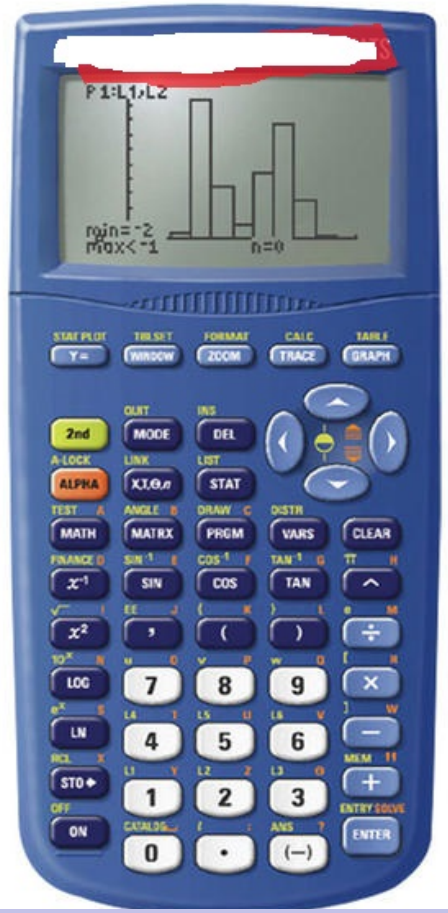


The error scatter plot in this toy example

The error scatter plot displays the points $(x_1, e_1), \ldots, (x_7, e_7)$. We see no visible pattern in the plot here.



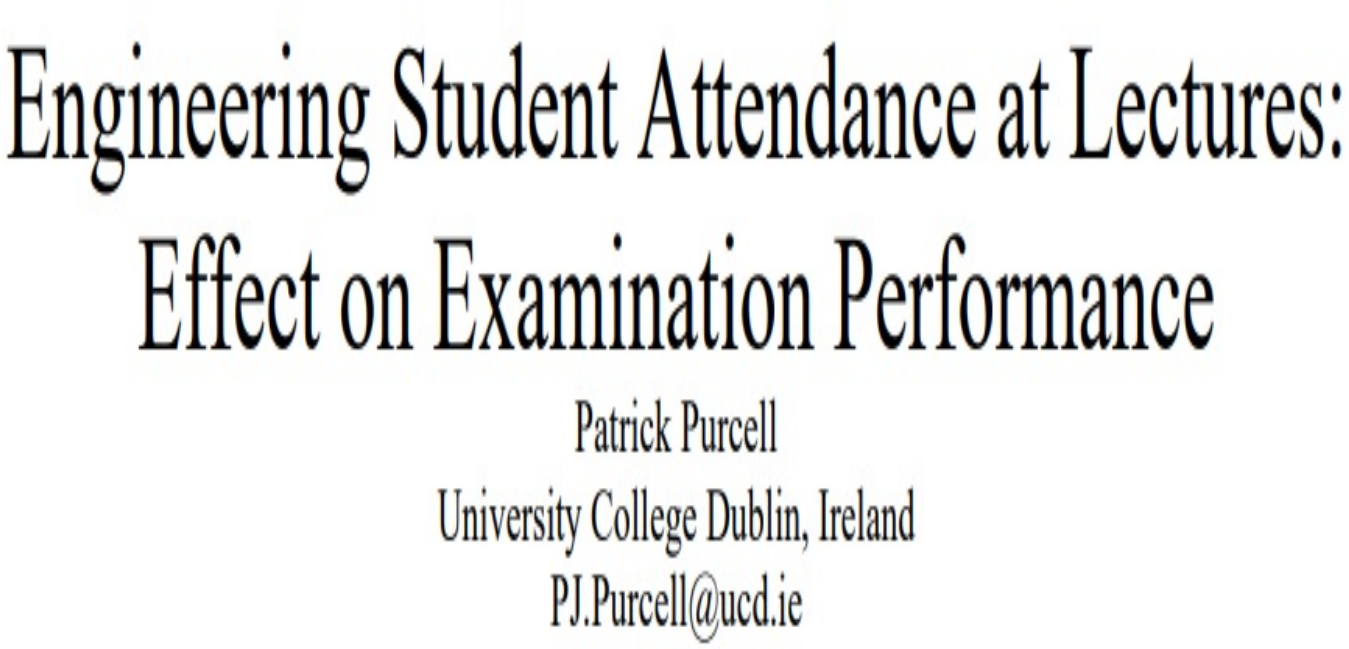
LSE is a Standard Function in Handheld Scientific Calculators (Not a Product Placement, as the Manufacturer Name is Erased)



Simple Linear Regression model: An Observational Study

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*. International Conference on Engineering Education – ICEE 2007 Coimbra, Portugal September 3 – 7, 2007

Linear Regression: Prediction of Success at an Exam by Attendance at Lectures



Engineering Student Attendance at Lectures:
Effect on Examination Performance

Patrick Purcell
University College Dublin, Ireland
PJ.Purcell@ucd.ie

- This study had two principal objectives: to establish the levels of attendance at lectures by civil engineering students at University College Dublin and to ascertain whether lecture attendance influenced the examination performance of these students.

- This study had two principal objectives: to establish the levels of attendance at lectures by civil engineering students at University College Dublin and to ascertain whether lecture attendance influenced the examination performance of these students.
- Lecture attendance for two classes of engineering students was monitored and analysed. The average lecture attendance rate for these students was found to be 68 %, which is in line with attendance rates in US studies, but higher than comparable Irish studies in other disciplines.

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*

- A linear regression analysis of the data showed a strong correlation between lecture attendance and examination performance.

evident between class attendance and examination performance. Examination of these figures also shows that the pass mark of 40% can be attained at relatively low attendance levels (< 20% attendance).

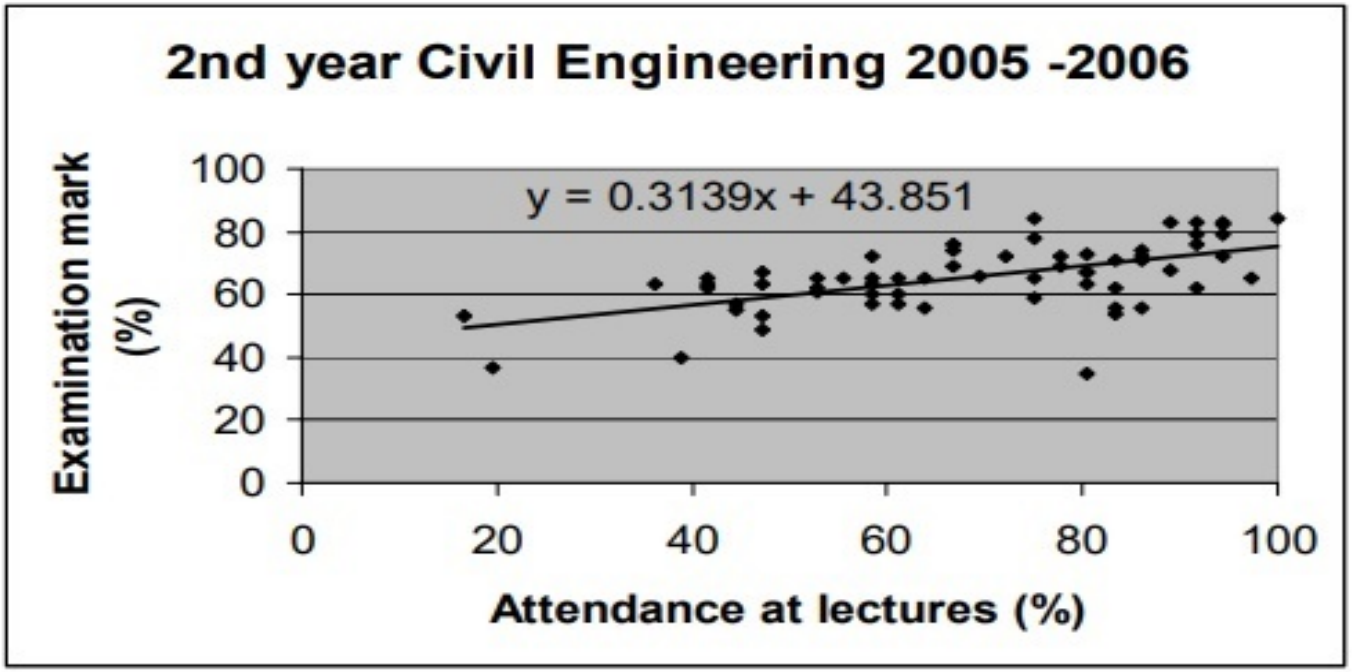


FIGURE I
PERFORMANCE OF SECOND-YEAR CIVIL ENGINEERING STUDENTS

PATRIC PURCELL: *Engineering Student Attendance at Lectures: Effect on Examination Performance*

towards continuous assessment.

Examination of Figures 1 and 2 clearly shows that students who have chosen to attend lectures regularly perform significantly better in their examinations than students that have chosen not to attend lectures. The best-fit equations ($y = 0.31x + 43.90$ and $y = 0.32x + 33.91$) indicate that each 10% increase in lecture attendance results in an approximate 3% improvement in examination performance. These correlation equations compare favourably to other studies, for example, Lockwood et al. [9].

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*

- Each 10% increase in student attendance at lectures improved examination performance by about 3%, which is again in line with that found by other studies.

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*

- Each 10% increase in student attendance at lectures improved examination performance by about 3%, which is again in line with that found by other studies.
- Let us write the predictors as

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \hat{y}(x+h) = \hat{\beta}_0 + \hat{\beta}_1 (x+h)$$

Improvement equals $\hat{y}(x+h) - \hat{y}(x) = \hat{\beta}_1 h$. Hence, in the study by Purcell, $0.3 \cdot 10 = 3$.

Question: A Policy Recommendation?

There are other studies confirming the findings of Purcell, more of this later.

- Question: Are we now in our right to request an audience with Rektor Anders Söderholm at Rektor's office in Brinellvägen 8 to deliver the (unsolicited) advice:

  Bäste Anders! Statistical research has shown that exam results will be substantially improved, were KTH to introduce mandatory attendance at lectures, or, KTH is to require, say, at least 90% mandatory attendance!

[3] Use and Abuse of Regression & Causality

Question: A Policy Recommendation?

There are other studies confirming the findings of Purcell, more of this later.

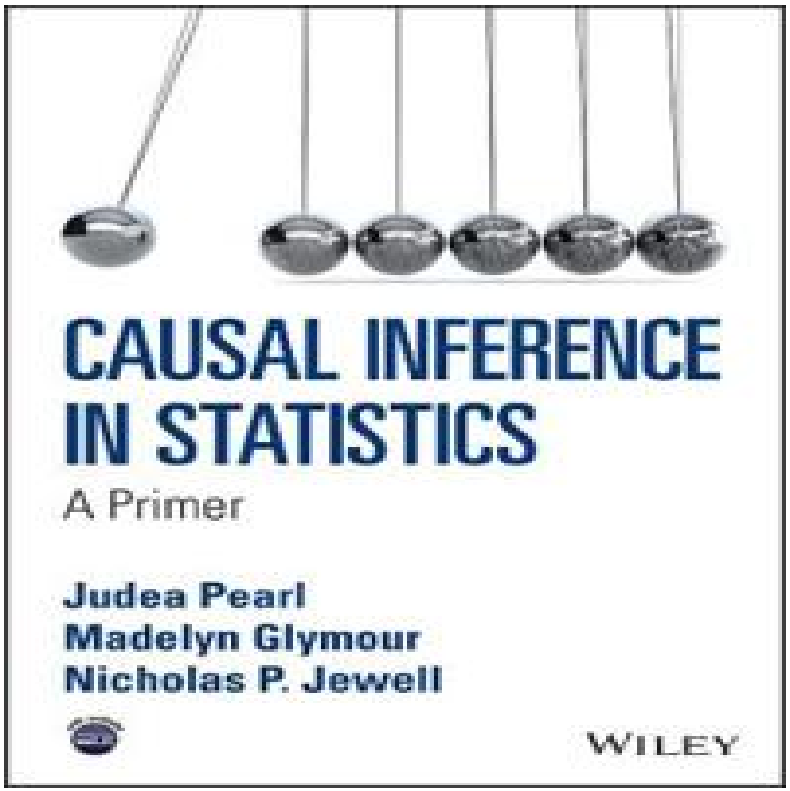
- Question: Are we now in our right to request an audience with Rektor Anders Söderholm at Rektor's office in Brinellvägen 8 to deliver the (unsolicited) advice:

  Bäste Anders! Statistical research has shown that exam results will be substantially improved, were KTH to introduce mandatory attendance at lectures, or, KTH is to require, say, at least 90% mandatory attendance!

- This question (which an instance of a more general and a very important issue[3]) will be adressed later by means of a mathematical study of multiple regression and intervention.

[3] Use and Abuse of Regression & Causality

Policy Interventions and Regression will be analyzed by means of this booklet published in 2021



The Learning Outcomes

- - Curve fitting, training set, theoretical line of regression $y = \beta_0 + \beta_1 x$

- Part 1: Analysis of simple linear regression
  - Fundamental Analysis of Variance Identity
  - What does the word 'regression' signify here?
- Part 2: The true data generating process: Normal linear regression model
  - Additional expressions $\hat{\beta}_0$ and $\hat{\beta}_1$
  - Normal distribution, expectation and variance for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Part 3.: Linear Regression is about Conditional Expectation
  - Mathematics about conditional expectations and mean square estimates.
- Part 4.: Period-Luminosity Law (?)

The Learning Outcomes

- - Curve fitting, training set, theoretical line of regression $y = \beta_0 + \beta_1 x$
  - LSE $\leftrightarrow \hat{\beta}_0$ and $\hat{\beta}_1$

- Part 1: Analysis of simple linear regression
  - Fundamental Analysis of Variance Identity
  - What does the word 'regression' signify here?
- Part 2: The true data generating process: Normal linear regression model
  - Additional expressions $\hat{\beta}_0$ and $\hat{\beta}_1$
  - Normal distribution, expectation and variance for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Part 3.: Linear Regression is about Conditional Expectation
  - Mathematics about conditional expectations and mean square estimates.
- Part 4.: Period-Luminosity Law (?)

The Learning Outcomes

- - Curve fitting, training set, theoretical line of regression $y = \beta_0 + \beta_1 x$
  - LSE $\leftrightarrow \hat{\beta}_0$ and $\hat{\beta}_1$
  - We are here now
- Part 1: Analysis of simple linear regression
  - Fundamental Analysis of Variance Identity
  - What does the word 'regression' signify here?
- Part 2: The true data generating process: Normal linear regression model
  - Additional expressions $\hat{\beta}_0$ and $\hat{\beta}_1$
  - Normal distribution, expectation and variance for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Part 3.: Linear Regression is about Conditional Expectation
  - Mathematics about conditional expectations and mean square estimates.
- Part 4.: Period-Luminosity Law (?)

Auxiliaries for Technical Details

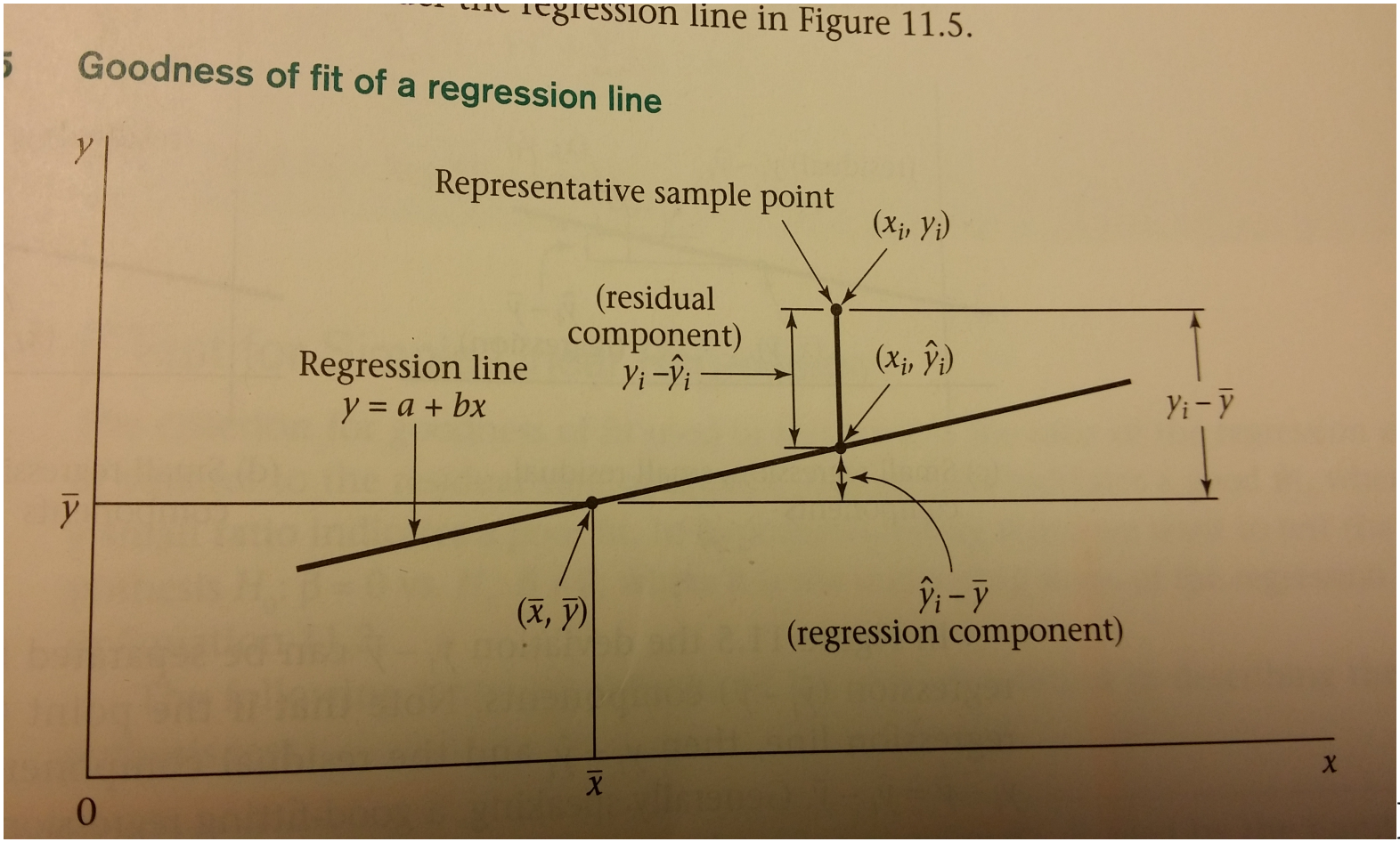
- Appendix A
- Appendix B
- Appendix C
- Appendix D
- Appendix E
- Appendix F Link to Simple Linear regression in Stanford CS229: Machine Learning Lecture 1 (Autumn 2018)

We have found in the preceding $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Thus $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$. Hence if $x = \bar{x}$, we have

$$\hat{y} = \bar{y} + \hat{\beta}_1 (\bar{x} - \bar{x}) = \bar{y}.$$

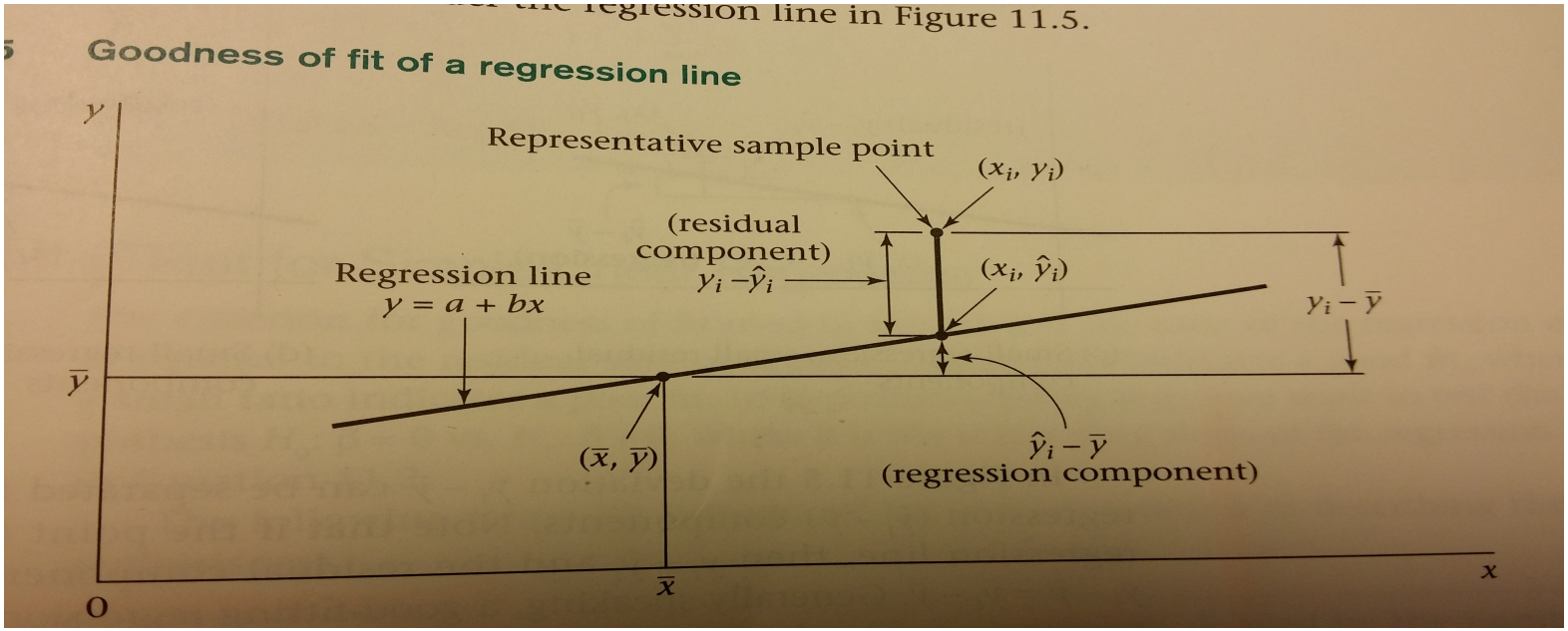
This means that the estimated regression line always passes through the point $(\bar{x}, \bar{y})$.

Fundamental Analysis of Variance Identity



For any (x_i, y_i) the **explained deviation** of that pair about the regression line is $\hat{y}_i - \bar{y}$.

Fundamental Analysis of Variance Identity



For any (x_i, y_i) the **regression component** of that pair about the regression line is $\hat{y}_i - \bar{y}$.

A good-fitting regression line will have regression components large in absolute value relative to the residuals.

Fundamental Analysis of Variance Identity

For any (x_i, y_i) the residual or the **unexplained deviation** of that pair about the regression line is $e_i = y_i - \hat{y}_i$.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \leftrightarrow \textbf{Total Variation}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \leftrightarrow \quad \textbf{Explained Variation}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftrightarrow \textbf{Unexplained Variation}$$

Fundamental Analysis of Variance Identity & Coefficient of determination

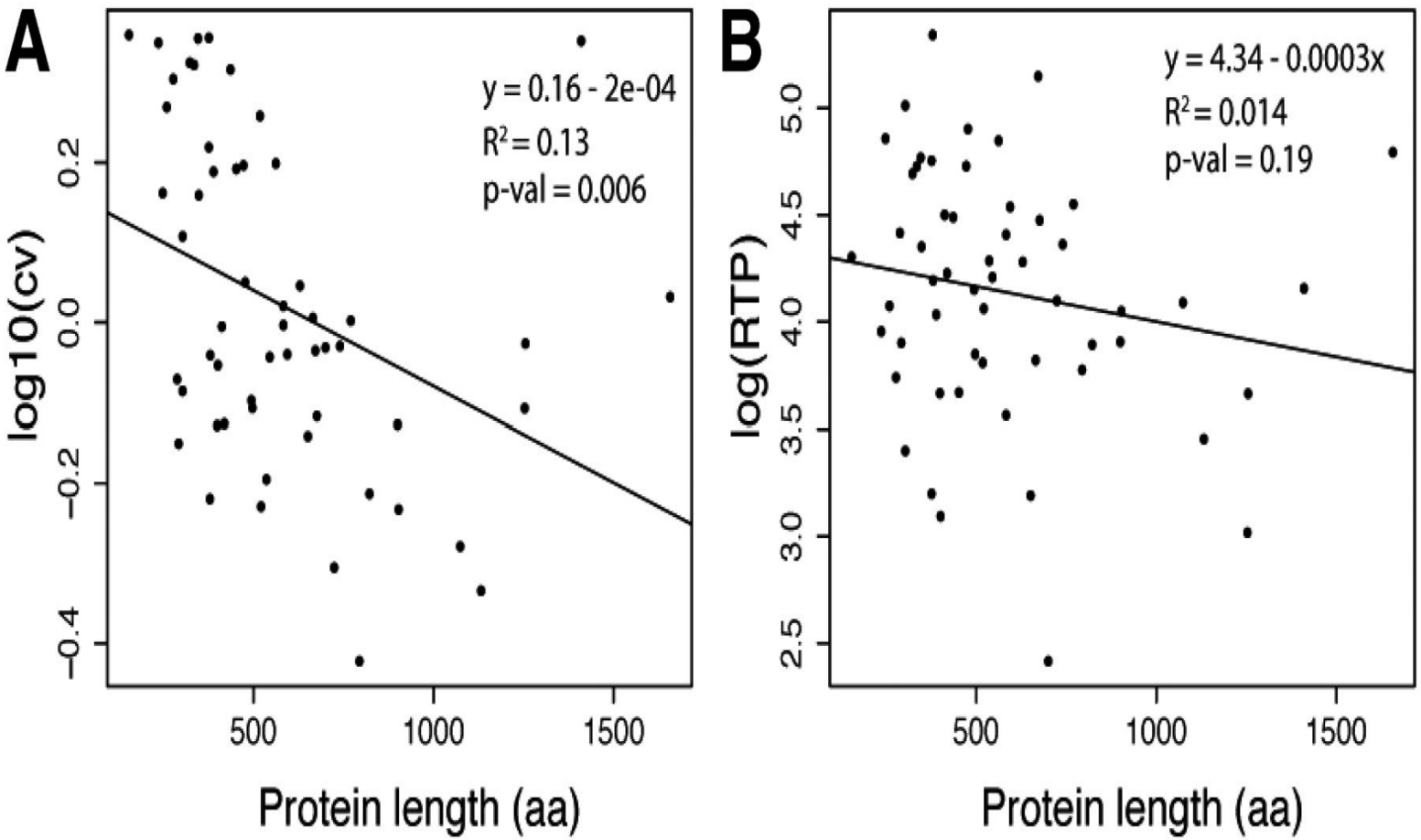
The **coefficient of determination** R^2 is the amount of variation in y that is explained by the regression line.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2}$$

LSE & and R^2: A Case in Biotechnology

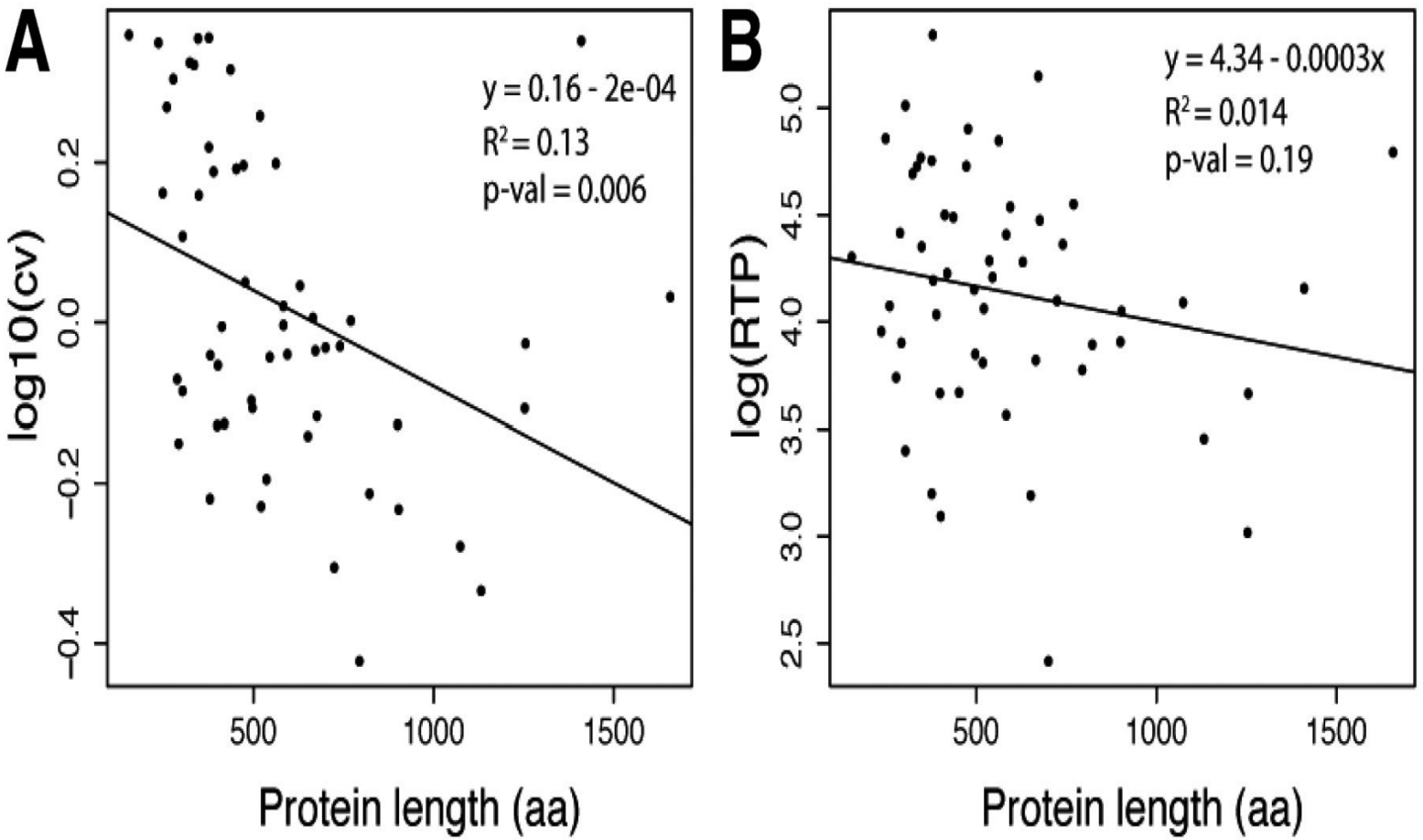
Edfors, Fredrik and Danielsson, Frida and Hallström, Björn M and Käll, Lukas and Lundberg, Emma and Pontén, Fredrik and Forsström, Björn and Uhlén, Mathias: *Gene-specific correlation of RNA and protein levels in human cells and tissues*, **Molecular Systems Biology**, 12, 883 , 2016.

LSE & and R^2: A Case in Biotechnology



In the left hand field (**A**) we see the variation measured as coefficient of variation (CV) across samples plotted against the protein length.

LSE & and R^2: A Case in Biotechnology



In the right hand field (**B**) the protein lengths for the 55 target proteins are plotted against the RNA-to-protein (RTP) ratio.

LSE: A Case in Biotechnology

Data information from the paper cited: **Test** based on correlation coefficient and follows a **t-distribution with length(x)-2 degrees of freedom**, if the samples follow **independent normal distributions**. An asymptotic **confidence interval** (the boldfaced stuff to be established below) is given based on **Fisher′s z-transform**.

Fundamental Analysis of Variance Identity

It will later be established by a piece of matrix calculus that

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{3}$$

This can be found (very tediously) by the rules in Appendix C.

On p. 26 of MVP (3) is called the **Fundamental Analysis of Variance Identity**

Renaming

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \leftrightarrow \; SS_{\mathrm{T}}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \leftrightarrow \; SS_{\mathrm{R}}$$ **regression or model sum of squares**

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftrightarrow \; SS_{\mathrm{Res}}$$ **Residual Sum of Squares**

Fundamental Analysis of Variance Identity

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

On p. 26 of MVP this is also expressed as

$$SS_{\mathrm{T}}=SS_{\mathrm{R}}+SS_{\mathrm{Res}} \tag{4}$$

Hence in (2)

$$R^2=\frac{SS_{\mathrm{R}}}{SS_{\mathrm{T}}}=1-\frac{SS_{\mathrm{Res}}}{SS_{\mathrm{T}}}. \tag{5}$$

$$Q_0 := Q_0(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{e}_i^2 \leq Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)))^2$$

i.e., in another notation

$$SS_{\text{Res}} = Q_0$$

MVP introduces (for the first time on p.21) the 'universal' symbol M to write e.g.

$$\frac{SS_{\text{Res}}}{n-k-1} = MSS_{\text{Res}}, \frac{SS_{\text{R}}}{k} = MSS_{\text{R}}$$

This leads to expressions like (see MVP p. 85)

$$\frac{SS_{\text{R}}/k}{SS_{\text{Res}}/(n-k-1)} = \frac{MSS_{\text{R}}}{MSS_{\text{Res}}}$$

Awkward ?

One-Way Analysis of Variance: ANOVA Table for LSE in Simple Linear Regression

Source	df	Sum of Squares	MSS
Regression	1	SS_{R}	SS/df
Residual	$n-2$	SS_{Res}	$\hat{\sigma}^2$=SS/df
Total	n	SS_{T}	

Source = source of variation, df= degrees of freedom, SS= sum of squares, MSS= mean sum of squares. $SS_{\mathrm{RES}} = Q_0$. The rationale for $\hat{\sigma}^2 = \frac{Q_0}{n-2}$ will given later in a lecture on multiple regression). $\hat{\sigma}^2$ is an estimate of the variance of the deviations from a population theoretic regression line (c.f. below). The meaning of this will made clear in Lecture 5.

One-Way Analysis of Variance: ANOVA Table for LSE in Simple Linear Regression

Source	df	Sum of Squares	MSS
Regression	1	SS_{R}	SS/df
Residual	$n-2$	SS_{Res}	$\hat{\sigma}^2$=SS/df
Total	n	SS_{T}	

ONE KEY TOPIC in SF2930 is the development of the properties and diagnostic use of the ratio (in the Table $k = 1$)

$$\frac{SS_{\mathrm{R}}/k}{SS_{\mathrm{Res}}/(n-k-1)}. \tag{6}$$

What does the word regression refer to ?

Sir Francis Galton, 1822-1911 **defined regression was as the process of returning to the mean**.

1. Introduction

Regression, as we know it today, was born from Galton's investigations into the laws of heredity. The phenomenon that Galton discovered is best described in his own words:

> ...offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small ... (Galton 1886)

That Galton came to this conclusion almost single-handedly and not by drawing on the contributions from his predeces-

Gorroochurn, P.: On Galton's change from "reversion" to "regression", vol. 70, 3, 227–231, The American Statistician, 2016

How is 'returning to the mean" expressed mathematically by what we have found so far ? Rewrite $\hat{\beta}_1$

A)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \Leftrightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{7}$$

where we have set

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \text{ and } S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{8}$$

Regression, what is in the name ?

B) Introduce $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Then by (7) above

$$\hat{\beta}_1 = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} \frac{S_{xy}}{\sqrt{S_{yy}} \cdot \sqrt{S_{xx}}} = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} r_{xy},$$

i.e.,

$$\hat{\beta}_1 = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} r_{xy}, \tag{9}$$

where r_{xy} is the (sample) coefficient of correlation (well known from the first course sannstat), i.e.,

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{yy}} \cdot \sqrt{S_{xx}}} = \frac{\frac{1}{n} S_{xy}}{\sqrt{\frac{1}{n} S_{yy}} \cdot \sqrt{\frac{1}{n} S_{xx}}}$$

Regression, what is in the name ?

c) We write the predictor as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1(x - \bar{x})$ and use (9) to get

$$\hat{y} - \bar{y} = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} r_{xy}(x - \bar{x}) \Leftrightarrow \frac{\hat{y} - \bar{y}}{\sqrt{S_{yy}}} = r_{xy} \frac{x - \bar{x}}{\sqrt{S_{xx}}}.$$

c) By sannstat we know that $|r_{xy}| \leq 1$. Hence, if $|r_{xy}| < 1$,

$$\frac{|\,\hat{y} - \bar{y}\,|}{\sqrt{S_{yy}}} < \frac{|\,x - \bar{x}\,|}{\sqrt{S_{xx}}}.$$

Regression, what is in the name ?

$$\frac{|\hat{y} - \bar{y}|}{\sqrt{S_{yy}}} < \frac{|x - \bar{x}|}{\sqrt{S_{xx}}}.$$

In words, the predicted value of y is for given x is always closer, scaled by its standard deviation, to its mean $\bar{y}$ than x is to its mean $\bar{x}$, when scaled by its standard deviation, i.e., regression to mean, as soon as there is not a perfect correlation.

EXQ

Show that the squared sample coefficient of correlation equals the coefficient of determination R^2 defined in (2), i.e.,

$$r_{xy}^2 = R^2. \tag{10}$$

Aid: *Use (9) above, and the results (22) and (19) in Appendix E, see also (4).*

Quoting Gorroochurn, P. loc.cit

Indeed, the above explanation is one of the most intuitive ways of understanding the regression effect: In general, suppose a first measurement X is made on a given subject, followed by a second measurement Y. Assume X is exceptionally high. As long as X and Y are imperfectly correlated, X can be thought to be made up of two components:

1. The first component, which is usually extreme and is expected to remain extreme.
2. The second component, which is not extreme and is expected to remain near the center of the distribution.

The first measurement X is extreme because both components are high. However, for the second measurement Y, the first component is expected to remain high, but the second component is expected to be near the center. Hence, the average value of Y will be less extreme than X and closer to the center of the distribution (e.g., Wallis and Roberts 1956, p. 61; Stigler 1997).

Quoting Gorroochurn, P. loc.cit

Notice that the above explanation does not require any biological, economic, or other force to be present for regression to occur. The regression effect (or regression to the mean) is a purely statistical artifact arising from imperfect correlation between X and Y (of course, the concept of correlation was not known to Galton when he first discovered the regression effect). As such, it is also symmetric in the sense that the same reasoning as above can made to argue that, for a given extreme Y, the value of X is expected to be less extreme and closer to the center.

Statistical Artefact

MVP writes on p. 55:

> *While regression and correlation are closely related, regression is more powerful tool in many situations. Correlation is only a measure of association and is of little use in prediction. However, regression methods are useful in developing quantitative relationships between variables, which can be used in prediction.*

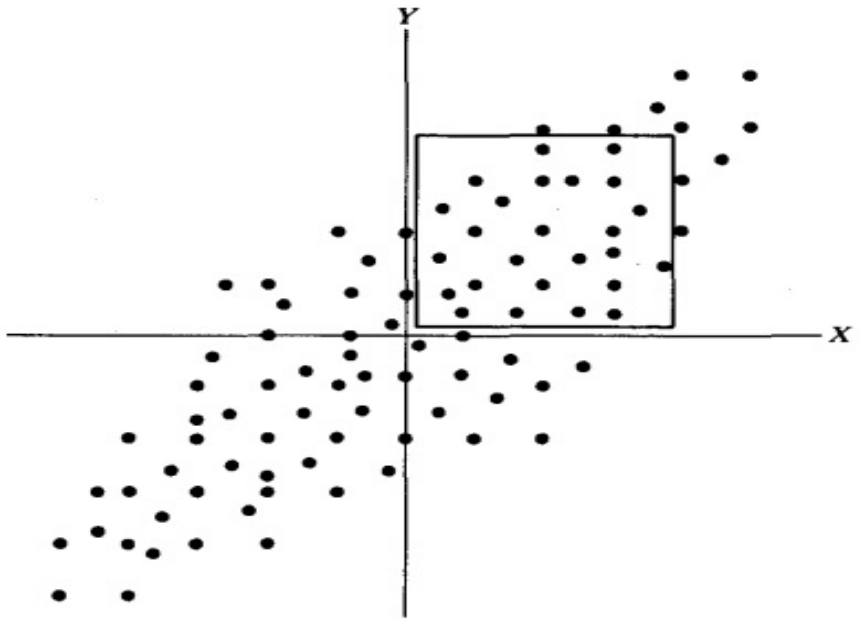
Does this make sense ?

Statistical Artefact

In the study by Purcell, with $X =$ attendance at lectures (percentage of total time), $Y =$ examination performance $\in [0, 100]$. Purcell did a regression of observations of Y on observations of X and got the slope $0.3 = \hat{\beta}_1 = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} r_{xy}$, see (9).
If we had the raw data, could run the **inverse regression** of observations of X on observations of Y. That would give us the slope $\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} r_{yx} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} r_{xy}$. If $S_{xx} = S_{yy}$, this inverse regression has the same slope $= 0.3$.
However, we do not think that examination performance 'explains' or is the 'cause' of attendance at lectures. On the other hand, we can think of explaining examination performance by attendance. This is not in the mathematics, but a notion we entertain from elsewhere.

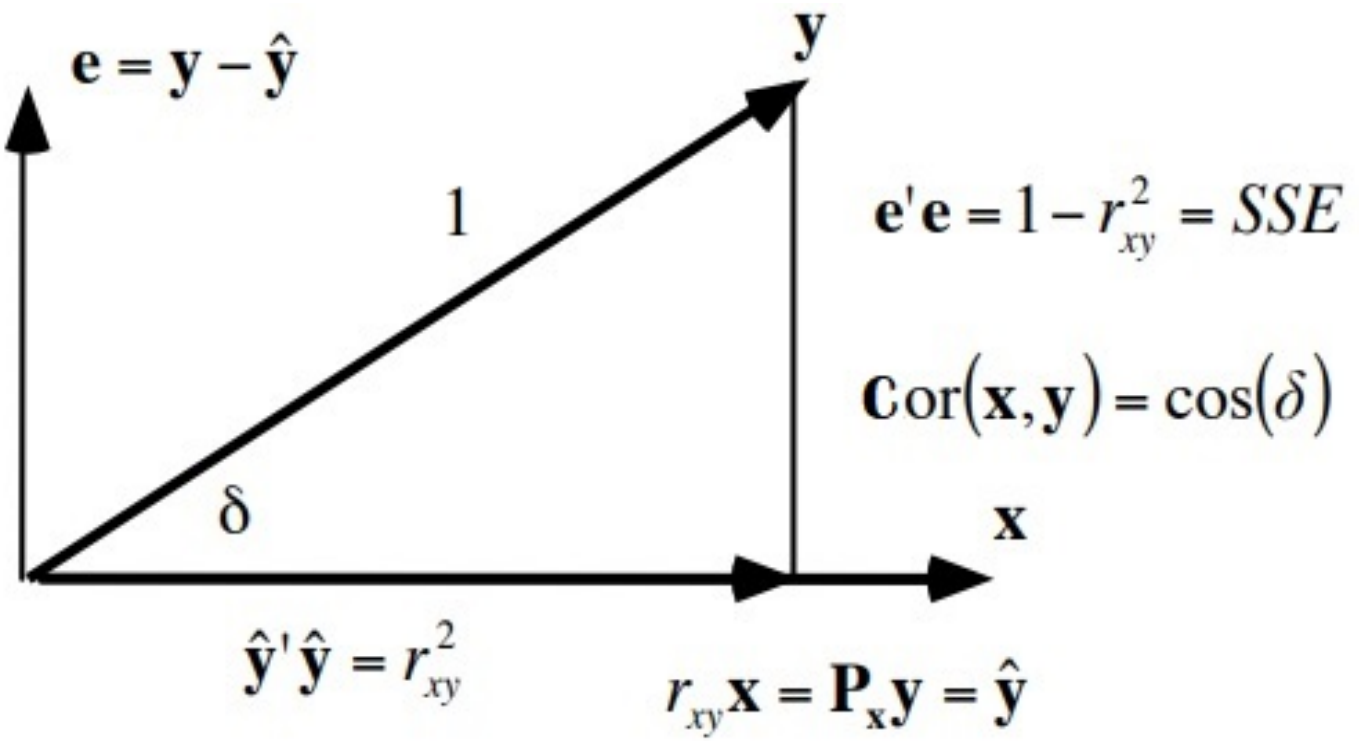
D. Heise: Causal Analysis, 1975, p. 150

4.18 REDUCED CORRELATION DUE TO RESTRICTION OF RANGE



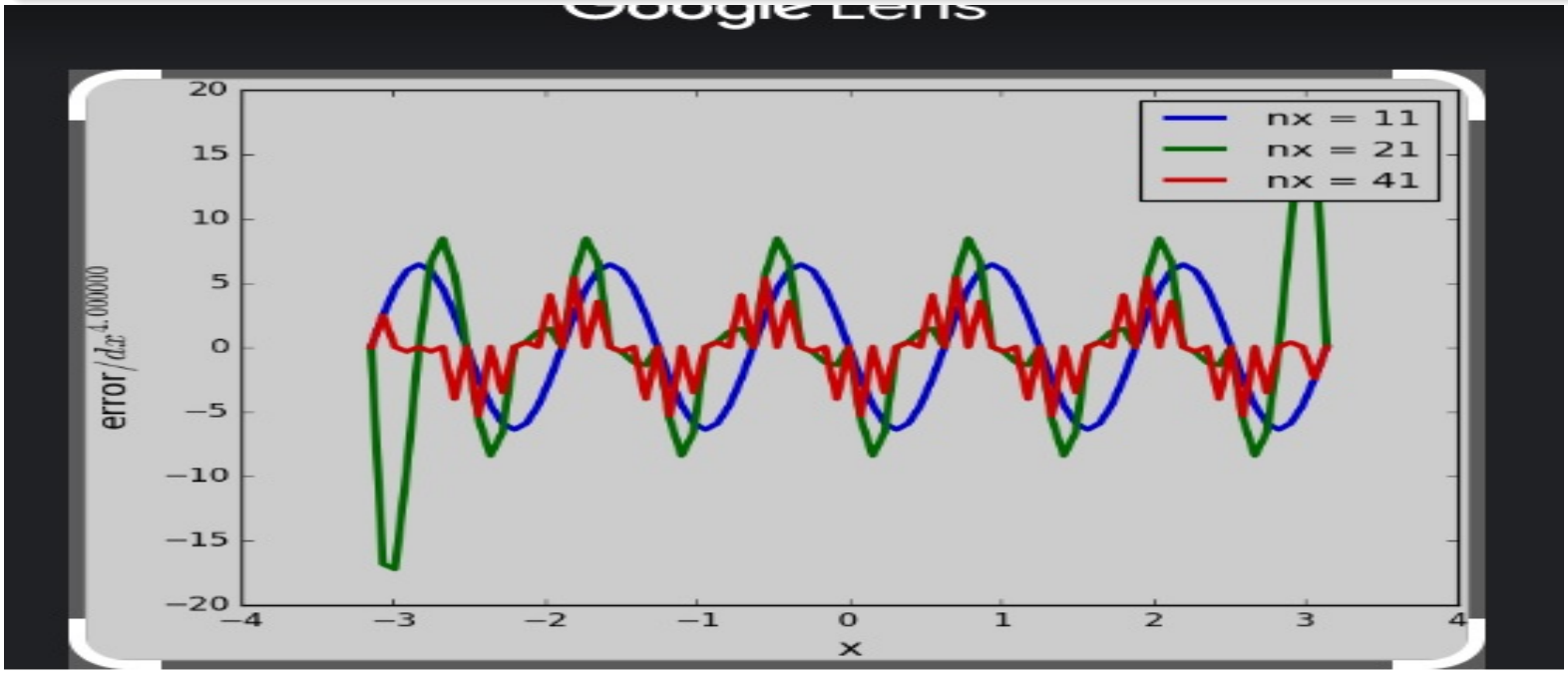
X and *Y* show a moderate degree of correlation when their full ranges are considered and almost no correlation when the range of both variables is restricted to values between 0 and + 2.0.

LSE Geometrically with $S_{xx} = S_{yy} = 1$



Overfitting

$\mathcal{D}_{tr} = \left\{ (x_j, y_j)_{j=i}^{n} \right\}$. *The Lagrange theorem (1795) says that there is a polynomial* $L(x)$ *of degree* $\leq n-1$ *such that* $L(x_j) = y_j$ *for all* j. *That is,* $L(x)$*gives a perfect fit on the training set. But this is* **overfitting**: *a perfect description of* $\mathcal{D}_{tr}$ *but unlikely to predict well the response* y *at a new point* x.



Part 2

Purcell's study is an example of an observational study: there was no preplanned design of the levels of attendance at lectures. We provide certain mathematical ideas for observational simple linear regressions.

The true data generating process

The *true data generating process* provides us with observations of a pair (X, Y) such that there is a theoretical population line of regression,

$$Y = \beta_0^* + \beta_1^* X + \varepsilon \tag{11}$$

where β_0^* and β_1^* are true parameter values, and ε is an unobserved random variable, a random disturbance.

- The true line of regression exists in simulations, where tβ_0^* and β_1^* are fixed by the simulating agent, who draws samples of ε from pseudorandom number generator.
- What does the true line of regression mean w.r.t real data?[4]
- If $\varepsilon \sim N(0, \sigma^2)$ and $X = x$, then $Y \sim N(\beta_0^* + \beta_1^* x, \sigma^2)$. Note that σ is supposed not to depend on x

[4] We assume things unseen (?)

Simple linear regression & Least Squares & Normal Distribution

Corresponding to $X_1 = x_1, \ldots, X_n = x_n$, $y_1, \ldots, y_n$ are observations of the random variables $Y_1, \ldots, Y_n$ by (11). respectively, where $Y_i \sim N(\mu_i, \sigma)$, where

$$\mu_i^* = \beta_0^* + \beta_1^* x_i, \quad i = 1, \ldots, n.$$

By the true model

$$y_i = \mu_i^* + \varepsilon_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

In addition the true data generating process gives $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. (independent, identically distributed)

- We can do LSE without assuming that ε is a random variable. LSE is then a mere description of the training set.
- But under the assumption $\varepsilon \sim N(0, \sigma^2)$ we can find the (normal) distributions of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$
- Then we can find find explicit and exact confidence intervals and test hypotheses about β_0^* andβ_1^*. (Later in the course as a special case of multiple regression).

Given observed $X = x$

$$Y = \beta_0^* + \beta_1^* x + \varepsilon$$

$\varepsilon \sim N(0, \sigma^2) \Rightarrow Y \sim N(\beta_0^* + \beta_1^* x, \sigma^2)$.
Y is a random variable, y is a real number (sample or outcome of Y). Note the syntax of the cumulative distribution function

$$F_Y(y) = P(Y \leq y) = P\left(\varepsilon \leq y - (\beta_0^* + \beta_1^* x)\right) = P\left(\frac{\varepsilon}{\sigma} \leq \frac{y - (\beta_0^* + \beta_1^* x)}{\sigma}\right)$$

$$= \Phi\left(\frac{y - (\beta_0^* + \beta_1^* x)}{\sigma}\right)$$

since $\frac{\varepsilon}{\sigma} \sim N(0, 1)$.

The true normal statistical regression model & distributions of LSE

- $$Y = \beta_0^* + \beta_1^* x + \varepsilon$$
- $$\varepsilon \sim N(0, \sigma^2)$$

  The variance σ^2 does not depend on x.
- β_0^* and β_1^* are unknown constants.

Rewriting the LSE 1.

S_{xy} is now written (by the rules in Appendix B, especially (6)) as

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) \tag{12}$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

c.f. (7) & (8) in Appendix C. We have used that $\sum_{i=1}^{n}(x_i - \bar{x})$ (and $\sum_{i=1}^{n}(y_i - \bar{y})$ are both equal to $= 0$, c.f., (6) in Appendix C.
Similarly, see (9) in Appendix C:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Rewriting the LSE 2.

Now we see that LSE are linear combinations of y_i (use repeatedly (4) in Appendix C)

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

$$= \frac{1}{S_{xx}} \left(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i \right) = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}} \right) y_i = \sum_{i=1}^{n} c_i y_i$$

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

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} c_i \right) y_i = \sum_{i=1}^{n} d_i y_i.$$

Rewriting the LSE 3.

That is,

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \text{and} \quad \hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i$$

where

$$c_i = (x_i - \bar{x})/S_{xx} \quad \text{and} \quad d_i = \frac{1}{n} - c_i \bar{x}. \tag{13}$$

In terms of the statistical model we have thus the random variables $\hat{\beta}_1$ and $\hat{\beta}_0$ given by

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \quad \text{and} \quad \hat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i$$

The variance of ε, σ^2, is estimated by

$$\widehat{\sigma}^2 := \frac{\sum_{i=1}^{n} Q_0}{n-2} = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}$$

This will be explained an studied in multiple regression, where

$$\widehat{\sigma}^2 := \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-k-1}$$

for appropriate residuals.

Normal Distribution

The coefficients c_i and d_i are functions of $x_1, \ldots, x_n$ which are are given numbers.

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \quad \text{and} \quad \hat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i$$

Here for $i = 1, 2, \ldots, n$, Y_i are independent and $Y_i \sim N(\beta_0^* x_i + \beta_1^* x_i, \sigma^2)$ and independent. Hence both $\hat{\beta}_1$ and $\hat{\beta}_0$ are normally distributed r.v.s ! We need to find the means and variances.

$E\left(\hat{\beta}_0\right) = \beta_0^*$

We have from (70)

$$E\left(\hat{\beta}_0\right) = \sum_{i=1}^{n} d_i E\left(Y_i\right).$$

and, since $Y_i \in N(\beta_0^* + \beta_1^* x_i, \sigma)$, we have

$$\sum_{i=1}^{n} d_i E\left(Y_i\right) = \sum_{i=1}^{n} d_i \left(\beta_0^* + \beta_1^* x_i\right) = \beta_0^* \sum_{i=1}^{n} d_i + \beta_1^* \sum_{i=1}^{n} d_i x_i$$

It is shown in Appendix A that $\sum_{i=1}^{n} d_i = 1$ and $\sum_{i=1}^{n} d_i x_i = 0$. Hence

$$E\left(\hat{\beta}_0\right) = \beta_0^* \sum_{i=1}^{n} d_i + \beta_1^* \sum_{i=1}^{n} d_i x_i = \beta_0^*.$$

In other words, $\hat{\beta}_0$ is an **unbiased** estimator.

$E(\hat{\beta}_1) = \beta_1^*$

Vi har $\hat{\beta}_1 = \sum_{i=1}^n c_i Y_i$ and thus

$$E\left(\hat{\beta}_1\right) = \sum_{i=1}^n c_i E\left(Y_i\right).$$

Since $Y_i \in N(\beta_0^* + \beta_1^* x_i, \sigma)$,

$$\sum_{i=1}^n c_i E\left(Y_i\right) = \sum_{i=1}^n c_i \left(\beta_0^* + \beta_1^* x_i\right) = \beta_0^* \sum_{i=1}^n c_i + \beta_1^* \sum_{i=1}^n c_i x_i$$

The auxiliary results in (*I*) and (*II*) in Appendix A entail

$$E\left(\hat{\beta}_1\right) = \beta_1^*$$

and $\hat{\beta}_1$ is an unbiased estimator, too.

Variance $\mathrm{Var}\left(\beta_0^*\right)$

Since Y_i are independent it holds that

$$\mathrm{Var}\left(\beta_0^*\right) = \sum_{i=1}^{n} d_i^2 \mathrm{Var}\left(Y_i\right).$$

But $Y_i \in N(\beta_0^* + \beta_1^* x_i, \sigma)$, hence

$$\sum_{i=1}^{n} d_i^2 \mathrm{Var}\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} d_i^2$$

We find in Appendix A that $\sum_{i=1}^{n} d_i^2 = \frac{1}{n} + \frac{\bar{x}}{S_{xx}}$. Hence

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right).$$

Variance $\mathbf{Var}\left(\hat{\beta}_1\right)$

By the above $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$ and as Y_i are independent

$$\mathbf{Var}\left(\hat{\beta}_1\right) = \sum_{i=1}^{n} c_i^2 \mathbf{Var}\left(Y_i\right).$$

Again, by $Y_i \in N(\beta_0^* + \beta_1^* x_i, \sigma)$, we get

$$\sum_{i=1}^{n} c_i^2 \mathbf{Var}\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{S_{xx}}$$

according to the auxiliary (IV) in Appendix A.

Probability Distributions of LSE

PROPOSITION

$$\hat{\beta}_0 \sim N\left(\beta_0^*, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$
$$\hat{\beta}_1 \sim N\left(\beta_1^*, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right). \tag{14}$$

Probability Distributions of LSE

Another (not so tedious) excercise in the algebra of the relevant finite sums shows that

$$Q_0 = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{e}_i^2 =$$

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

Square, sum and simplify to get

$$Q(\hat{\beta}_0, \hat{\beta}_1) = S_{yy} \left(1 - r_{xy}^2 \right) .$$

ExQ *C.f. Problem 2.25 in MVP*

- Show that

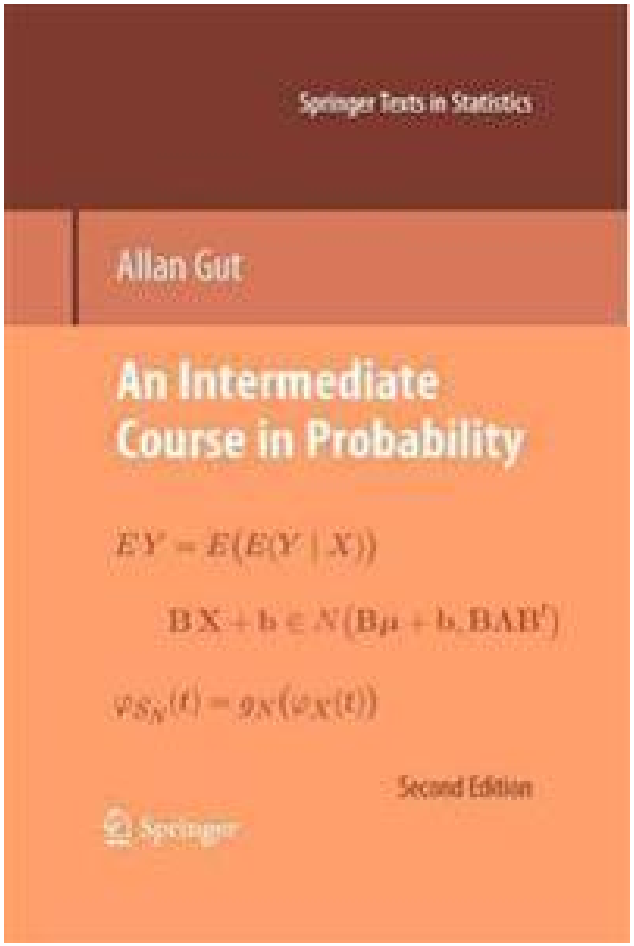
$$\mathbf{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{-\bar{x}\sigma^2}{S_{xx}} \tag{15}$$

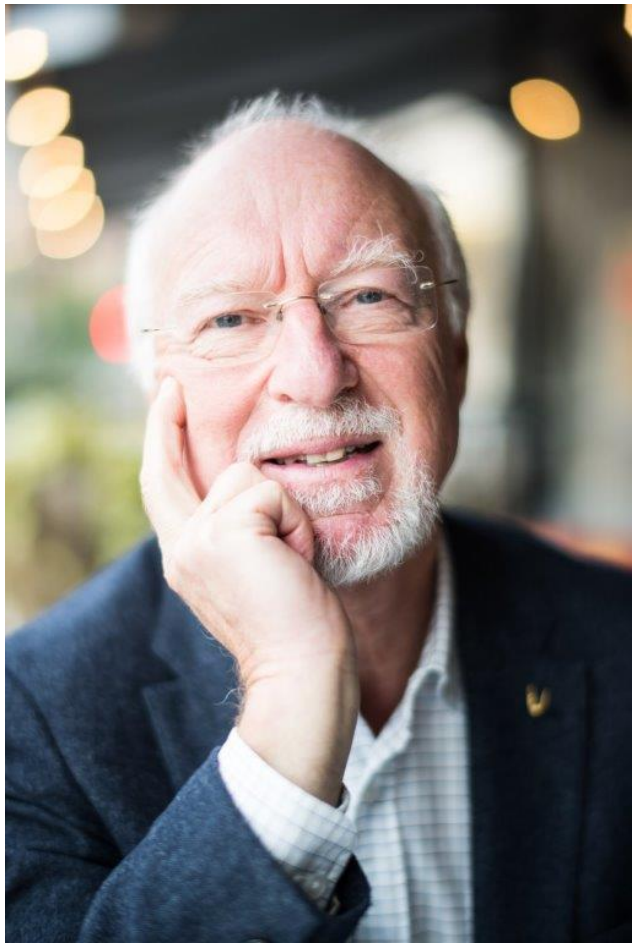
Aid: the definition of **Cov** is

$$\mathbf{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = E\left[\left(\hat{\beta}_0 - \beta_0^*\right)\left(\hat{\beta}_1 - \beta_1^*\right)\right]$$

- What does (15) tell us ?

Part 3.: Conditonal Expectation & True Data Generating Process





Let $(X, Y) \sim f_{X,Y}(x, y)$, a joint probability density function (PDF) such that $E[|Y|] < +\infty, E\left[Y^2\right] < +\infty$.

The conditional expectation of Y given $X = x$ is

$$E[Y|X = x] := \int_{-\infty}^{+\infty} y f_{Y|X=x}(y) dy.$$

$E[Y|X]$ is a random variable, which is a function of the random variable X and takes the value $E[Y|X = x]$, when $X = x$.

We have the rule of double expectation Theorem 2.1. p.34 in Gut, Allan: *An Intermediate Course in Probability. Second Edition*, Springer, 2009

$$E[Y] = E\left[E\left[Y|X\right]\right]. \tag{16}$$

A sketch of proof of the double expectation

$$E\left[E\left[Y|X\right]\right] = \int_{-\infty}^{+\infty} E\left[Y|X=x\right] f_X(x)dx$$

$$= \int_{-\infty}^{+\infty} y \int_{-\infty}^{+\infty} f_{Y|X=x}(y) f_X(x)dxdy$$

$$= \int_{-\infty}^{+\infty} y \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{+\infty} y f_Y(y)dy = E[Y]$$

Theorem 2.3. p.36 in Gut, Allan: *An Intermediate Course in Probability. Second Edition*, Springer, 2009.

Proposition

$g(x)$ is a real valued function. Assume in addition that $E\left[(g(X))^2\right] < +\infty$. Then

$$E[(Y - g(X))^2] = E\left[\mathrm{Var}\left[Y|X\right]\right] + E\left[\left(E\left[Y|X\right] - g(X)\right)^2\right]. \tag{17}$$

The choice $g_{\mathrm{opt}}(X) = E\left[Y|X\right]$ minimizes the mean squared error (MSE), since

$$E[(Y - g(X))^2] = E\left[\mathrm{Var}\left[Y|X\right]\right] + E\left[\left(E\left[Y|X\right] - g(X)\right)^2\right]$$

$$\geq E\left[\mathrm{Var}\left[Y|X\right]\right] = E[\left(Y - g_{\mathrm{opt}}(X)\right)^2]$$

Linear Regression is about Conditional Expectation

The optimal predictor of Y given X in MSE is $g_{\text{opt}}(X) = E[Y|X]$. This can be difficult to find. We can approximate with a straight line

$$g_{\text{opt}}(X) \approx \hat{Y}(X) = \beta_0 + \beta_1 X$$

However, take $g(X) = \beta_0 + \beta_1 X$ in (17), which gives

$$E[(Y - (\beta_0 + \beta_1 X))^2] = E[\mathbf{Var}[Y|X]] + E\left[(E[Y|X] - (\beta_0 + \beta_1 X))^2\right].$$

Since the first term on the right hand does not depend on β_0 and β_1, it follows that if β_0^* and β_1^* minimize $E[(Y - (\beta_0 + \beta_1 X))^2]$, these values will also minimize $E\left[(E[Y|X] - (\beta_0 + \beta_1 X))^2\right]$.

Linear Regression is about Conditional Expectation

The optimal predictor of Y given X in MSE is

$$g_{\text{opt}}(X) = E[Y|X].$$

The optimal linear predictor of $E[Y|X]$ in MSE is

$$\hat{Y}(X) = \beta_0^* + \beta_1^* X$$

$\hat{Y}(X)$ is also the best linear predictor of Y in MSE.

EXQ: Find β_0^* and β_1^* such that

$$MSE(\beta_0, \beta_1) = E[(Y - \beta_0 + \beta_1 X)^2$$

is minimized.

Set $\mu_X = E[X]$,$\mu_Y = E[Y]$, $\sigma_x^2 = \mathbf{Var}[X]$, $\sigma_y^2 = \mathbf{Var}[Y]$, $\sigma_{xy} = \mathbf{Cov}[X, Y]$, $\rho = \sigma_{xy}/\sigma_y\sigma_x$. Then, by partial differentiations, the minimizers of $MSE(\beta_0, \beta_1)$ are

$$\begin{aligned} \beta_0^* &= \mu_Y - \beta_1^* \mu_X \\ \beta_1^* &= \frac{\sigma_{xy}}{\sigma_x^2} = \rho \frac{\sigma_y}{\sigma_x} \end{aligned} \tag{18}$$

LSE look like sample equivalents of the true population parameters β_0^* and β_1^* (?)

- $\overline{x} \approx \mu_x$. $\overline{y} \approx \mu_y$.
- $\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) \approx \sigma_{xy}$
- $\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2 \approx \sigma_x^2$
- Therefore:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2} \approx \beta_1^* = \frac{\sigma_{xy}}{\sigma_x^2}$$

Or, if this was to be true, in which sense? And under which conditions?

LSE look like sample equivalents of the true population parameters β_0^* and β_1^* (?)

$$Q(\hat{\beta}_0, \hat{\beta}_1) = S_{yy}\left(1 - r_{xy}^2\right).$$

This estimates obviously the residual variance around the theoretical line, i.e.,

$$MSE(\beta_0^*, \beta_1^*) = \sigma_y^2\left(1 - \rho^2\right)$$

where $\rho = \sigma_{xy}/\sigma_y\sigma_x$.

Recall $\mathbf{Var}[X] = E[X^2] - (E[X])^2$. Then

$$E\left[\left(\hat{\beta}_1 - \beta_1^*\right)^2\right] = \mathbf{Var}\left[\left(\hat{\beta}_1 - \beta_1^*\right)\right] + \left(E\left[\left(\hat{\beta}_1 - \beta_1^*\right)\right]\right)^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

by (14). In the same way

$$E\left[\left(\hat{\beta}_0 - \beta_0^*\right)^2\right] = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

We see that there is convergence in mean square, i.e., $\hat{\beta}_1 \xrightarrow{2} \beta_1^*$, and $\hat{\beta}_0 \xrightarrow{2} \beta_0^*$, if $S_{xx} \to +\infty$, and $\bar{x} \to c < +\infty$, as $n \to +\infty$.

EXQ Consistency of the Statistical Predictor

Show that for any fixed x_o

$$E\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_o - \left(\beta_0^* + \beta_1^* x_o\right)\right)^2\right] =$$

$$= \frac{1}{S_{xx}}\left(x_o^2\sigma^2 + 2x_o\bar{x}\sigma^2 + \frac{S_{xx}}{n} + \bar{x}^2\right).$$

Aid: Recall (15), too. Question: What about conditions for convergence in mean square, a.k.a. consistency in mean square?

Part 4.:

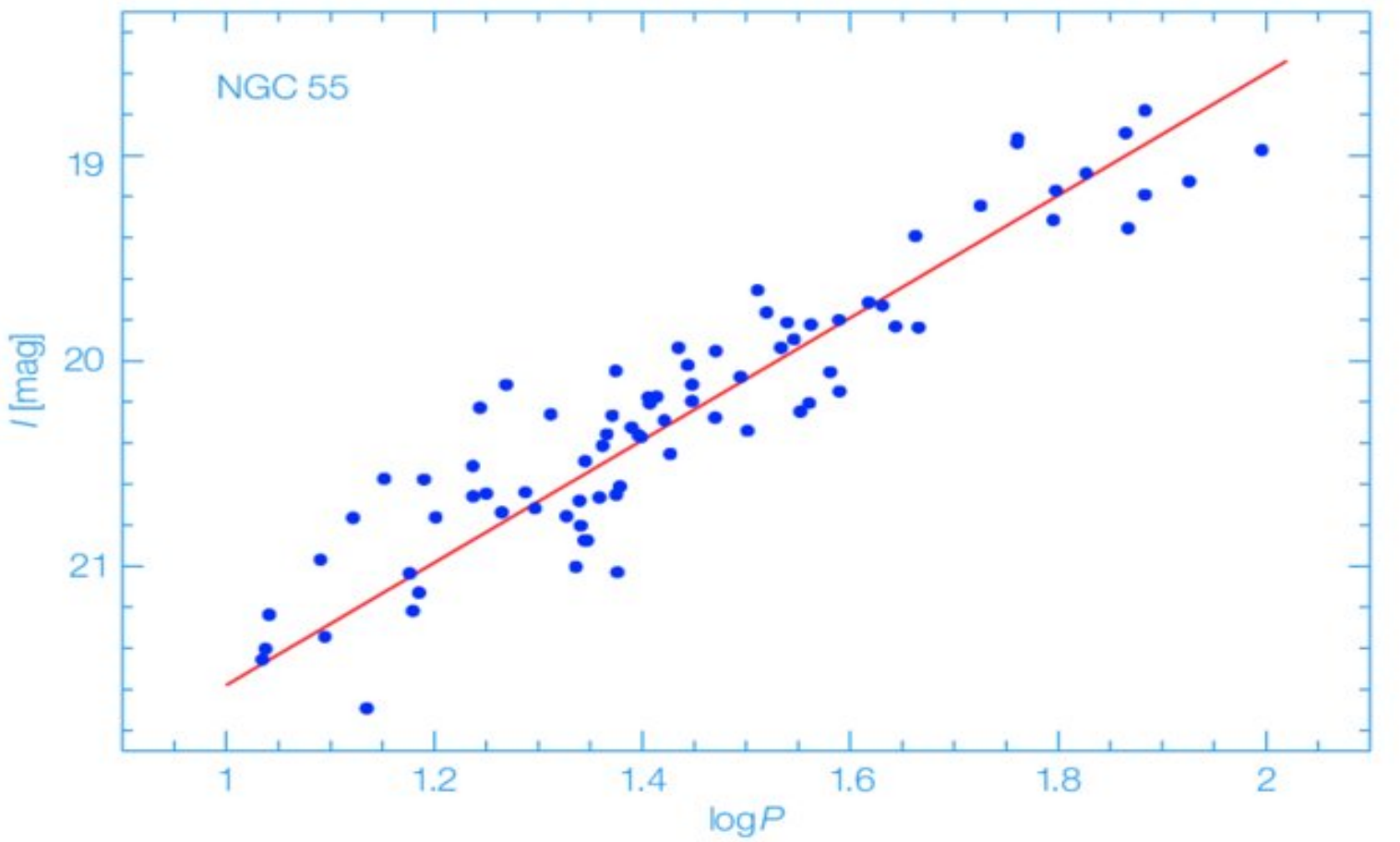
Period-Luminosity Law a.k.a. Leavitt's Law

https:/ www.atnf.csiro.au/outreach/education/senior/astrophysics/variable_cepheids.html

Australia Telescope National Facility Outreach

Some types of pulsating variable stars such as Cepheids exhibit a definite relationship between their period and their intrinsic luminosity. Such period-luminosity relationships are invaluable to astronomers as they are a vital method in calculating distances within and beyond our galaxy.

Period-Luminosity Plot



Measuring Improved Distances to Nearby Galaxies: The Araucaria Project

Predictor

The radially pulsating Cepheid supergiant stars exhibit a well-known relation between their mean intrinsic luminosity, and their pulsation periods – the period-luminosity (PL) relation, which is normally written in the form

$$M = a \log P + b,$$

where M is the mean absolute magnitude (in a given photometric band), and P the period(in days). With the PL relation calibrated, the mean luminosities of Cepheids, and thus their distances, can be inferred from their periods (the detailed procedure available on the Australia Telescope National Facility Outreach page).

Period-Luminosity Plot

The following relationship between a Population I Cepheid's period P and its mean absolute magnitude M_V was in 2007 established from Hubble Space Telescope trigonometric parallaxes for 10 nearby Cepheids:

$$M_{\mathrm{v}} = (-2.43 \pm 0.12)\,(\log_{10} P - 1) - (4.05 \pm 0.02)$$

P is measured in days. (Wikipedia)

https:/ www.atnf.csiro.au/outreach/education/senior/astrophysics/variable_cepheids.html

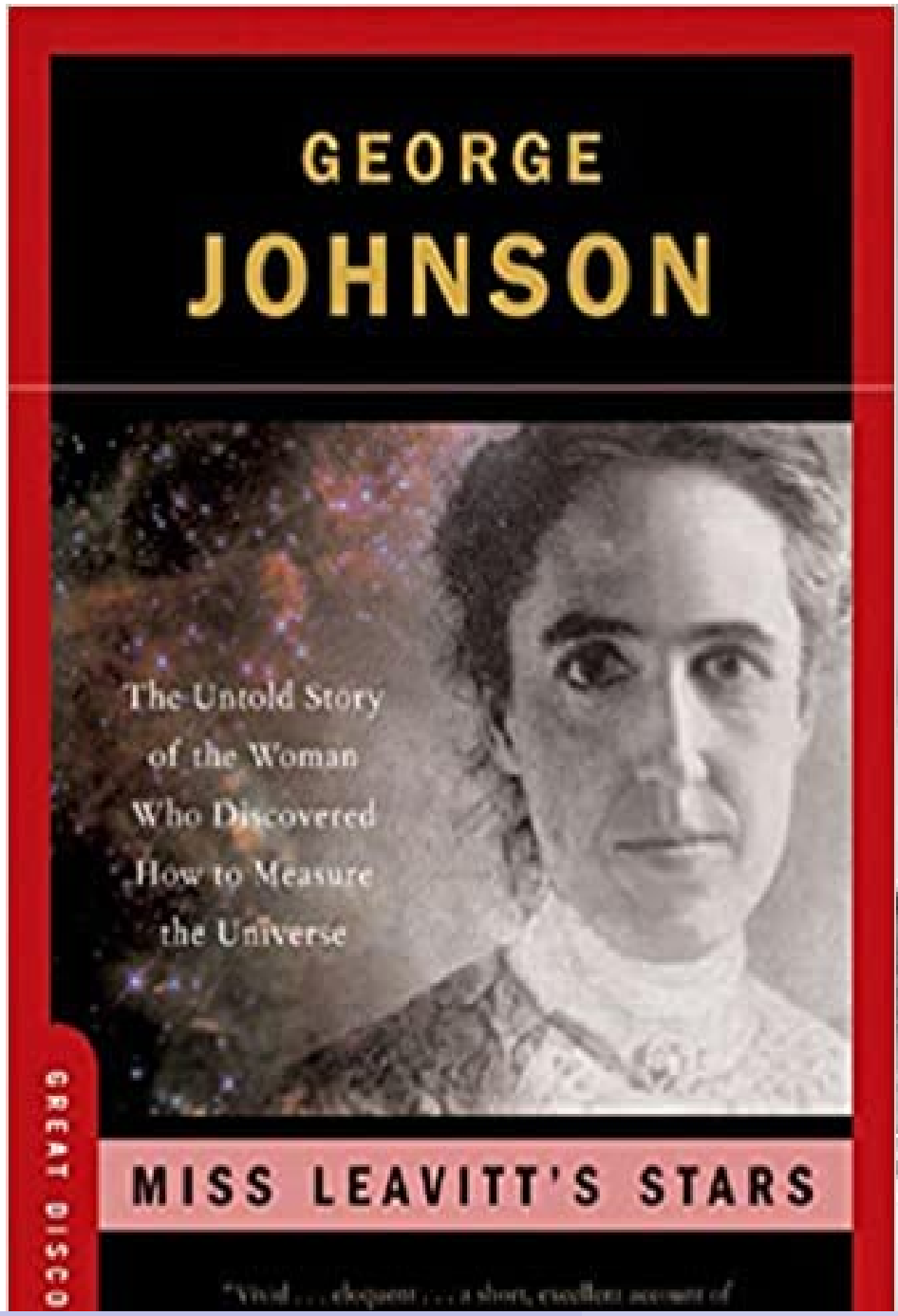
In 1924 Edwin Hubble detected Cepheids in the Andromeda nebula, M31 and the Triangulum nebula M33. Using these he determined that their distances were 900,000 and 850,000 light years respectively. He thus established conclusively that these "spiral nebulae" were in fact other galaxies and not part of our Milky Way. This was a momentous discovery and dramatically expanded the scale of he known Universe.

You have heard about Hubble but have you heard about Leavitt?

- The accomplishments of Edwin Hubble, the American astronomer who established that the universe is expanding, also were made possible by Leavitt's groundbreaking research. Hubble often said that Leavitt deserved the Nobel Prize for her work.

You have heard about Hubble but have you heard about Leavitt?

- The accomplishments of Edwin Hubble, the American astronomer who established that the universe is expanding, also were made possible by Leavitt's groundbreaking research. Hubble often said that Leavitt deserved the Nobel Prize for her work.
- Henrietta Swan Leavitt 1868 – 1921. A graduate of Radcliffe College, she worked at the Harvard College Observatory as a "computer", tasked with examining photographic plates in order to measure and catalog the brightness of stars. This work led her to discover the relation between the luminosity and the period of Cepheid variables. In the early 1900's Henrietta Leavitt fit the the predictor $M = a \log P + b$ using paper and pencil by applying Legendre's formula for LSE from 1805.


GEORGE
JOHNSON
The Untold Story
of the Woman
Who Discovered
How to Measure
the Universe
MISS LEAVITT'S STARS

Appendix A : Algebra for the variance and expectation of $\hat{\beta}_0$ and $\hat{\beta}_1$

(1) $\sum_{i=1}^n d_i = 1$

We have $c_i = (x_i - \bar{x})/S_{xx}$. Then

$$\sum_{i=1}^n d_i = \sum_{i=1}^n (\frac{1}{n} - c_i \bar{x}) = \sum_{i=1}^n \frac{1}{n} - \sum_{i=1}^n c_i \bar{x}$$

$$= 1 - \bar{x} \sum_{i=1}^n c_i = 1 - \bar{x} \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})$$

$$= 1 - \bar{x} \frac{1}{S_{xx}} \cdot 0 = 1.$$

since $\sum_{i=1}^n (x_i - \bar{x}) = 0$ by (6) in Appendix C. Note that we got also

$$\sum_{i=1}^n c_i = 0.$$

(2) $\sum_{i=1}^n d_i x_i$, $\sum_{i=1}^n c_i x_i = 1$

$$\sum_{i=1}^n d_i x_i = \sum_{i=1}^n (\frac{1}{n} - c_i \bar{x}) x_i$$

$$= \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n c_i x_i.$$

But

$$\sum_{i=1}^n c_i x_i = (1/S_{xx}) \sum_{i=1}^n (x_i - \bar{x}) x_i = (1/S_{xx}) \sum_{i=1}^n (x_i^2 - \bar{x} x_i)$$

$$= (1/S_{xx}) \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) =$$

$$= (1/S_{xx}) \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = (1/S_{xx}) \cdot S_{xx} = 1.$$

(2) $\sum_{i=1}^n d_i x_i = 0$

Since $\sum_{i=1}^n c_i x_i = 1$, we get

$$\sum_{i=1}^n d_i x_i = \sum_{i=1}^n \left(\frac{1}{n} - c_i \bar{x} \right) x_i$$

$$= \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} \underbrace{\sum_{i=1}^n c_i x_i}_{=1}.$$

$$= \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} = 0.$$

Summary of auxiliaries

(I)

$$\sum_{i=1}^{n} c_i x_i = 1$$

(II)

$$\sum_{i=1}^{n} c_i = 0,$$

(III)

$$\sum_{i=1}^{n} d_i x_i = 0$$

(IV)

$$\sum_{i=1}^{n} c_i^2 = \frac{1}{S_{xx}}$$

((IV) follows immediately by the definitions of c_i and S_{xx}.)

$\sum_{i=1}^n d_i^2 = \frac{1}{n} + \frac{\bar{x}}{S_{xx}}$

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n \left(\frac{1}{n} - c_i \bar{x} \right)^2$$

$$= \sum_{i=1}^n \frac{1}{n^2} - \frac{2\bar{x}}{n} \sum_{i=1}^n c_i + \bar{x}^2 \sum_{i=1}^n c_i^2$$

By the auxiliary (*II*), $\sum_{i=1}^n c_i = 0$. By (*IV*) we get

$$\sum_{i=1}^n c_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / S_{xx}^2 = \frac{1}{S_{xx}}$$

Note that $\sum_{i=1}^n \frac{1}{n^2} = 1/n$. In summary:

$$\sum_{i=1}^n d_i^2 = \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}.$$

Appendix C : Rules of computation with finite sums

Definition

(1) $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$

Proposition

(2) $\sum_{i=1}^{n} a \cdot x_i = a \sum_{i=1}^{n} x_i.$

Proposition

(3) $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

Proposition

(4) $\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$

Proposition

(5) $\sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$

$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$. Then it follows that:

Proposition

(6) $\underline{\sum_{i=1}^{n}(x_i - \overline{x}) = 0.}$

Proposition

(7) $\underline{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})\, y_i = \sum_{i=1}^{n} x_i\,(y_i - \overline{y})\,.}$

Proposition

(8) $\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}.$

Proposition

(9) $\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$

Proofs:

(1) $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$

(2) $\sum_{i=1}^{n} a \cdot x_i = a \sum_{i=1}^{n} x_i.$

Proof: Definition (1) entails $\sum_{i=1}^{n} a \cdot x_i = ax_1 + ax_2 + \ldots + ax_n$
$= a(x_1 + x_2 + \ldots + x_n) = a \sum_{i=1}^{n} x_i.$
Example: $x_i = 1, i = 1 \ldots, n$
$\sum_{i=1}^{n} a = a + a + \ldots + a = a(1 + 1 + \ldots + 1) = a \cdot n.$

(3) $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

Proof: Definition (1) entails
$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n)$
$= x_1 + x_2 + \ldots + x_n + y_1 + y_2 + \ldots + y_n = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

(4) $\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$

Proof: This follows by (3) and (2).

(5) $\sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$

Bevis: Use $(x_i + y_i)^2 = x_i^2 + 2x_i y_i + y_i^2$ and (4) as well as (2) with $a = 2$.

(6) $\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$

Proof: $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x}$ according to (4). But here we have with $a = \overline{x}$ in (2) that $\sum_{i=1}^{n} \overline{x} = \overline{x} \sum_{i=1}^{n} 1 = \overline{x} \cdot n$ according to the Example in (2). But $\overline{x} \cdot n = \sum_{i=1}^{n} x_i$ and this entails the assertion in (6).

(7) $\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})\, y_i = \sum_{i=1}^{n} x_i\,(y_i - \overline{y})\,.$

Proof: $(x_i - \overline{x}) \cdot (y_i - \overline{y}) = (x_i - \overline{x}) \cdot y_i - (x_i - \overline{x}) \cdot \overline{y}$. Then we get by (4) that
$\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x}) \cdot y_i - \sum_{i=1}^{n}(x_i - \overline{x}) \cdot \overline{y}$.
With $a = \overline{y}$ in (2) we obtain $\sum_{i=1}^{n}(x_i - \overline{x}) \cdot \overline{y} = \overline{y} \cdot \sum_{i=1}^{n}(x_i - \overline{x})$ and then (6) give that $\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})\, y_i$.
The other equality follows analogously.

(8) $\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})=\sum_{i=1}^{n}x_iy_i-n\overline{x}\overline{y}$.

Proof: Expand $\sum_{i=1}^{n}x_i(y_i-\overline{y})$ in the right hand side of (7) and use (2) and the definition on $\overline{x}$.

(9) $\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2$.

Proof: From (5) we get that
$\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\overline{x}+\sum_{i=1}^{n}\overline{x}^2$. Then (2) with $a=\overline{x}$ and the Example in (2) entail that
$\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\overline{x}+\sum_{i=1}^{n}\overline{x}^2=\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2$.
The definition of $\overline{x}$ gives $\sum_{i=1}^{n}x_i=n\overline{x}$, so that
$\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2=\sum_{i=1}^{n}x_i^2-2\overline{x}\cdot n\overline{x}+n\overline{x}^2$
$=\sum_{i=1}^{n}x_i^2-n\overline{x}^2$.

(10) $\sum_{i=1}^{n}(x_i-a)^2=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-a)^2.$

Proof: We note the identity

$$\sum_{i=1}^{n}(x_i-a)^2=\sum_{i=1}^{n}[(x_i-\overline{x})-(a-\overline{x})]^2$$

and by (5)

$$=\sum_{i=1}^{n}(x_i-\overline{x})^2-2\sum_{i=1}^{n}(x_i-\overline{x})\cdot(a-\overline{x})+\sum_{i=1}^{n}(a-\overline{x})^2\,.$$

With (2)

$$=\sum_{i=1}^{n}(x_i-\overline{x})^2-2\,(a-\overline{x})\sum_{i=1}^{n}(x_i-\overline{x})+\sum_{i=1}^{n}(a-\overline{x})^2$$

- and the Example in (2) again

$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 - 2(a - \overline{x}) \sum_{i=1}^{n} (x_i - \overline{x}) + n(a - \overline{x})^2$$

and (6) gives $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$ so that

$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + n(a - \overline{x})^2 ,$$

which is the right hand side of (10), as was claimed.

- We have thus by (10)

$$\sum_{i=1}^{n}(y_i-\beta_0)^2=\sum_{i=1}^{n}(y_i-\overline{y})^2+n(\overline{y}-\beta_0)^2\geq\sum_{i=1}^{n}(y_i-\overline{y})^2,$$

since $n(\overline{y}-\beta_0)^2\geq 0$. Hence $\hat{\beta}_0=\overline{y}$ is the LSE of β_0, when the regression model does not have a covariate, or, $\beta_1^*=0$ in the true model.

Appendix D: LSE verified

We check that

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

where $\overline{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\overline{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Derivative of $\sum_{i=1}^{n} (x_i - \mu_i(\theta))^2$ w.r.t. θ

Lemma

$g(\theta) = \sum_{i=1}^{n} (x_i - \mu(\theta))^2$.

(11) *Then* $\frac{d}{d\theta} g(\theta) = -2 \sum_{i=1}^{n} (x_i - \mu(\theta)) \mu_i'(\theta)$, *where* $\mu_i'(\theta) = \frac{d}{d\theta} \mu_i(\theta)$.

Proof: We have for each i that

$$\frac{d}{d\theta} (x_i - \mu_i(\theta))^2 = 2 (x_i - \mu_i(\theta)) \left(-\mu_i'(\theta) \right).$$

The derivative of a sum with a finite number of terms is the sum of the derivatives of the constituent terms, i.e., $\frac{d}{d\theta} g(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} (x_i - \mu_i(\theta))^2$. The number -2 does not depend on the index of summation, and can hence be moved outside the sum, rule (2). The claimed expression for $\frac{d}{d\theta} g(\theta)$ follows.

LSE

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

We seek $(\hat{\beta}_0, \hat{\beta}_1)$ by solving the system of equations

$$\begin{cases} \frac{\partial}{\partial \beta_0} Q(\hat{\beta}_0, \hat{\beta}_1) & = 0 \\ \frac{\partial}{\partial \beta_1} Q(\hat{\beta}_0, \hat{\beta}_1) & = 0. \end{cases}$$

Set $\theta = (\beta_0, \beta_1)$. By (11), with $\mu_i(\theta) = \beta_0$, $\frac{\partial}{\partial \beta_0} \mu_i(\theta) = 1$
$\frac{\partial}{\partial \beta_0} Q(\hat{\beta}_0, \hat{\beta}_1) = 0 \Leftrightarrow \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$, which gives by (2),(3) and (4), that $\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \Leftrightarrow \underline{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$, as expected.

LSE

By (11), with $\mu_i(\theta) = \beta_1 x_i$, $\frac{\partial}{\partial \beta_1}\mu_i(\theta) = x_i$ and this cannot moved outside summation. Hence $\frac{\partial}{\partial \beta_1} Q(\hat{\beta}_0, \hat{\beta}_1) = 0 \Leftrightarrow$
$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$. We substitute $\hat{\beta}_0$ as establilshed above, to obtain $\sum_{i=1}^n \left(y_i - \overline{y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i\right) x_i = 0 \Leftrightarrow$
$\sum_{i=1}^n (y_i - \overline{y})\, x_i - \hat{\beta}_1 \sum_{i=1}^n (x_i - \overline{x})\, x_i = 0$. This gives

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

By (12) above we have $\sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. Furthermore, $\sum_{i=1}^n (x_i - \overline{x})\, x_i = \sum_{i=1}^n x_i^2 - \overline{x} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i^2 - n\overline{x}^2$ =by the rule (9) $= \sum_{i=1}^n (x_i - \overline{x})^2$. Hence $\hat{\beta}_1$ equals the asserted expression.

Appendix E: Another Expression for SS_{Res}

$$SS_{\text{Res}} = SS_{\text{T}} - \hat{\beta}_1 S_{xy} \tag{19}$$

Check: $SS_{\text{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} ((y_i - \bar{y}) - (\hat{y}_i - \bar{y}))^2$.
Squaring and rule (5) yield

$$SS_{\text{Res}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{20}$$

Here, by definition of $\hat{y}_i$,

$$\hat{y}_i - \bar{y} = \bar{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i - \bar{y} = \hat{\beta}_1 (x_i - \overline{x}).$$

We insert these expressions in (20) to get

$$SS_{\text{Res}} = SS_{\text{T}} - 2\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \overline{x}) + \hat{\beta}_1{}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{21}$$

We simplify the last two terms in the right hand side.

Appendix E: Another Expression for SS_R

Use (7) and (8) to get

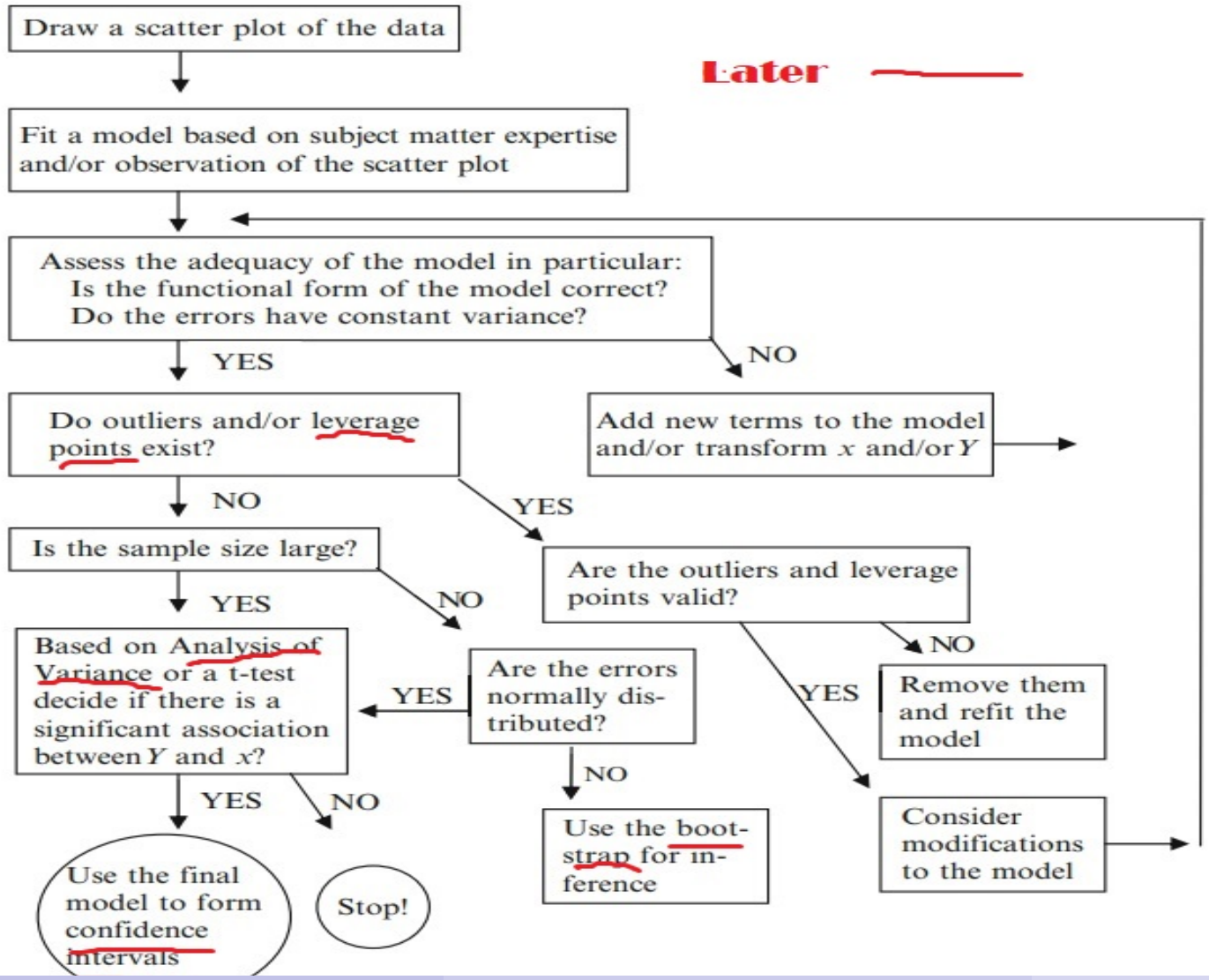
$$\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \hat{\beta}_1 S_{xy}$$

$$\hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx} = \hat{\beta}_1 \frac{S_{xy}}{S_{xx}} S_{xx} = \hat{\beta}_1 S_{xy} \tag{22}$$

When these results are substituted back into (21), (19) follows. □
In view of (4), i.e., $SS_T = SS_R + SS_{Res}$ we have

$$\boxed{SS_R = \hat{\beta}_1 S_{xy}. \tag{23}}$$

Appendix F: Flowchart for Simple Linear Regression



Appendix G: Stanford Online

Stanford CS229: Machine Learning Lecture 1 (Autumn 2018) Andrew Ng (Adjunct Professor of Computer Science) lecturing on simple linear regression starting 40:41 in

`https://www.youtube.comwatch?v=jGwO_UgTS7I&t=4050s`