

Regression analysis

Two variables

(Montgomery and Runger: ch 11

Brani Vidakovic: ch 14)

Reminder

Covariance Defined

Covariance is a number quantifying average dependence between two random variables.

The covariance between the random variables X and Y , denoted as $\text{cov}(X, Y)$ or σ_{XY} is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y \quad (5-14)$$

The units of σ_{XY} are units of X times units of Y .

Unlike the range of variance, $-\infty < \sigma_{XY} < \infty$.

Correlation is “normalized covariance”

- Also called:
Pearson correlation coefficient

$$\rho_{XY} = \sigma_{XY} / \sigma_X \sigma_Y$$

is the covariance
normalized to
be $-1 \leq \rho_{XY} \leq 1$



Karl Pearson (1852– 1936)

English mathematician and biostatistician

Covariance and Scatter Patterns

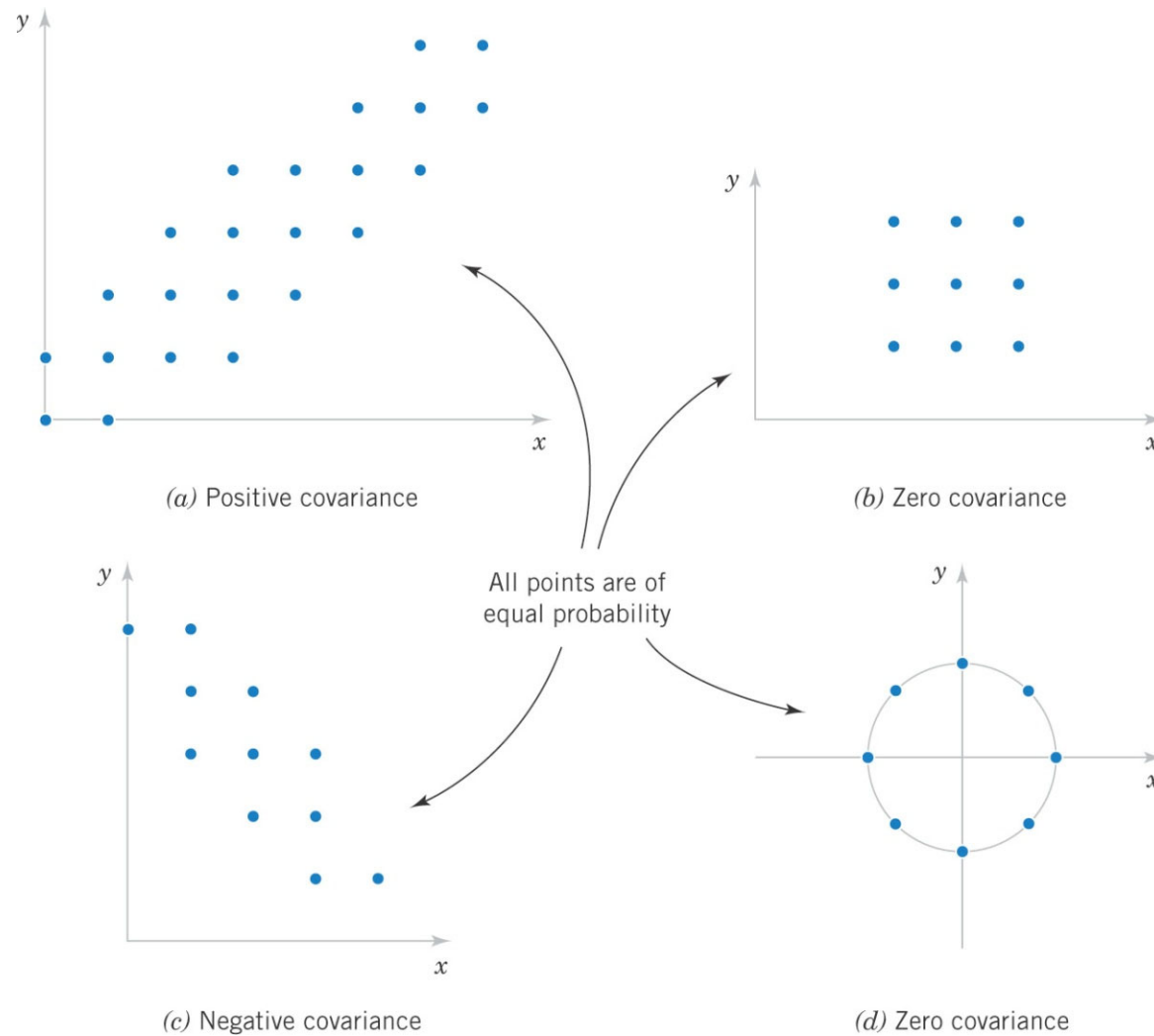
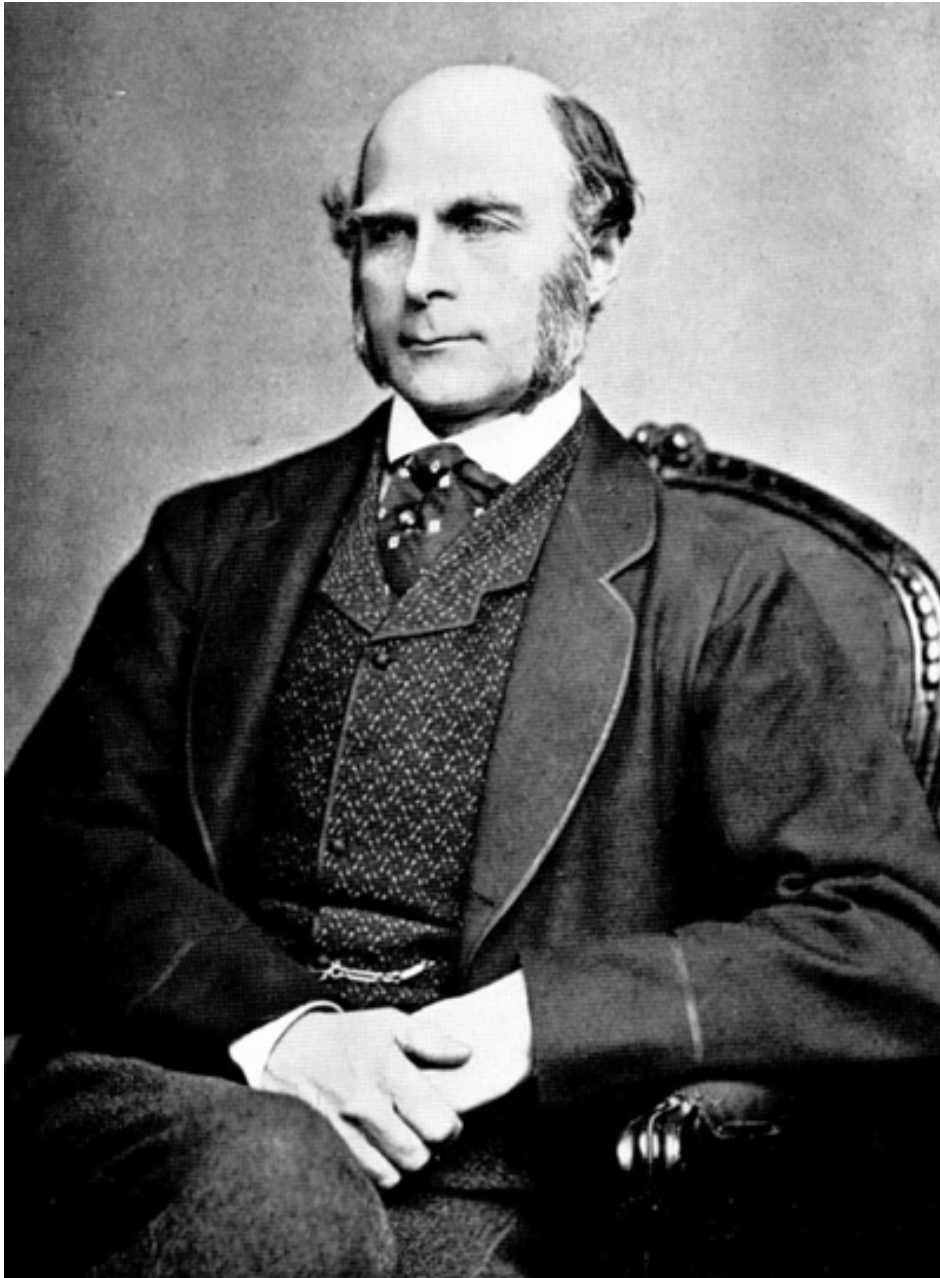


Figure 5-13 Joint probability distributions and the sign of $\text{cov}(X, Y)$. Note that covariance is a measure of linear relationship. Variables with non-zero covariance are **correlated**.

Regression analysis

- Many problems in engineering and science involve sample in which two or more variables were measured. They may not be independent from each other and one (or several) of them can be used to predict another
- Everyday example: in most samples height and weight of people are related to each other
- Biological example: in a cell sorting experiment the copy number of a protein may be measured alongside its volume
- **Regression analysis** uses a sample to build a model to predict protein copy number given a cell volume



Sir Francis Galton, (1822 -1911) was an English **statistician**, anthropologist, proto-geneticist, psychometrician, **eugenicist**, (“Nature vs Nurture”, inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

Invented both **correlation** and **regression analysis** when studied **heights of fathers and sons**

Found that fathers with height above average tend to have sons with height also above average but closer to the average.
Hence **“regression” to the mean**

Two variable samples

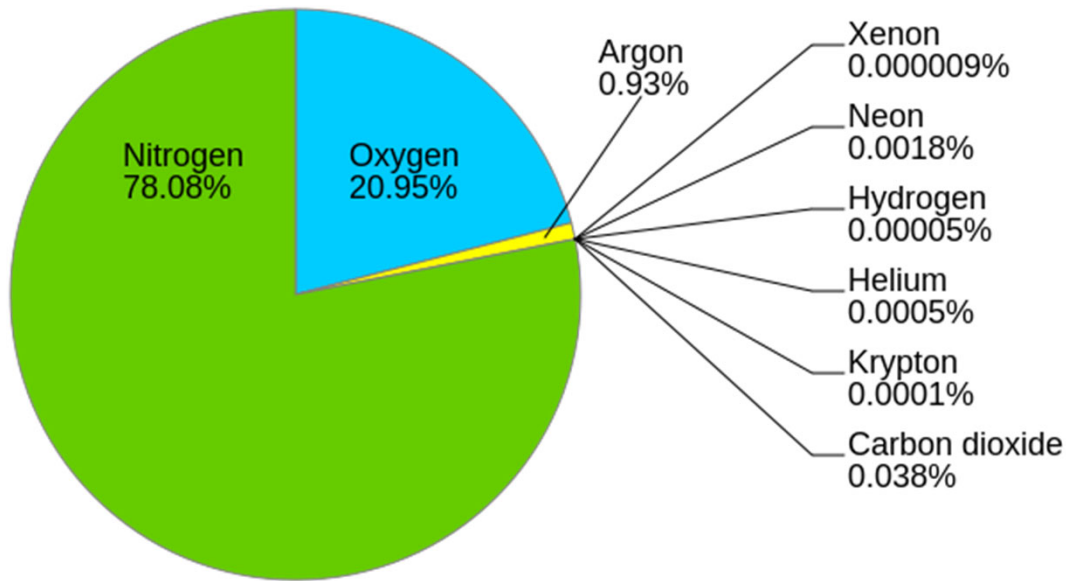


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go **kaboom!!!**
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don't know how dirty was the air to begin with

$$Y = \beta_0 + \beta_1 X + \epsilon$$

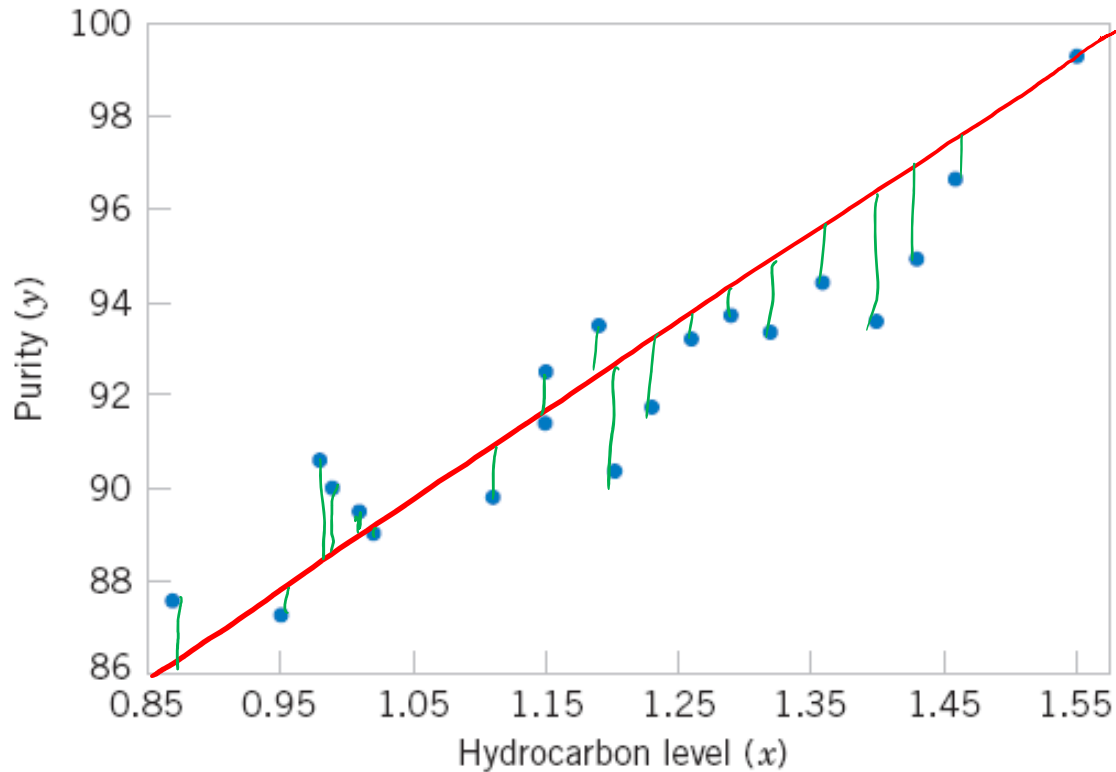


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$Y = 75 + 15 \cdot X + \epsilon$$

Linear regression

The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 X + E = \hat{Y} + E$$

E is the **random error**

slope β_1 and intercept β_0 of the line are called **regression coefficients**

Note: Y , \hat{Y} , X and E are random variables

Let's assume that $E(E | x) = 0 \rightarrow$

$$E(Y | x) = \beta_0 + \beta_1 x + E(E | x) = \beta_0 + \beta_1 x$$

$$Y = \beta_0 + \beta_1 X + \epsilon ; \quad E(\epsilon | x) = 0 \quad \forall x$$

How does one find β_0 & β_1 ?

$$\begin{aligned} \text{Cov}(Y, X) &= \text{Cov}(\beta_0 + \beta_1 X + \epsilon, X) = \\ &= \cancel{\text{Cov}(\beta_0, X)} + \beta_1 \text{Cov}(X, X) + \cancel{\text{Cov}(\epsilon, X)} \end{aligned}$$

$\text{Cov}(\beta_0, X) = 0$ since β_0 is constant

$$\text{Cov}(X, X) = E(X^2) - E(X)^2 = \text{Var}(X)$$

$$\text{Cov}(\epsilon, X) = E(\epsilon \cdot X) - \cancel{E(\epsilon)} \cdot E(X) =$$

$$= E(\epsilon \cdot X) = \sum_{\text{all } x} x \cdot \cancel{E(\epsilon | x)} = 0$$

Thus

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$\beta_0 = E(Y) - \beta_1 E(X)$$

Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

Figure 11-3 Deviations of the data from the estimated regression model.

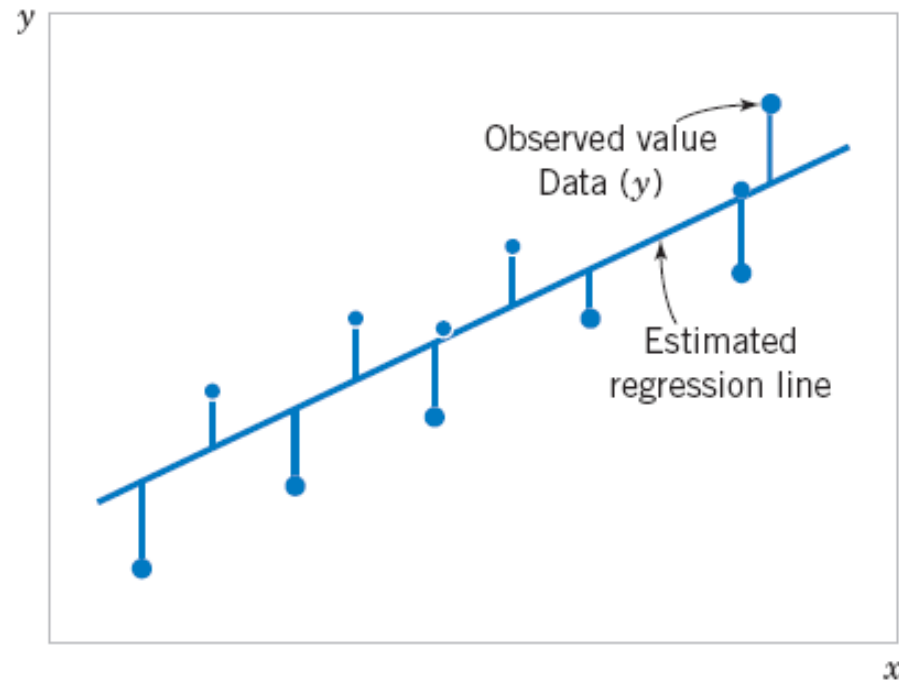


Figure 11-3 Deviations of the data from the estimated regression model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n \quad (11-3)$$

and the sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (11-4)$$

The least squares estimators of β_0 and β_1 , say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} \Big|_{\hat{\beta}_0, \hat{\beta}_1} &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial L}{\partial \beta_1} \Big|_{\hat{\beta}_0, \hat{\beta}_1} &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{aligned} \quad (11-5)$$

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned}$$

$$\hat{\beta}_0 \sum x_i = \hat{\beta}_1 \left(\frac{\sum x_i^2}{n} \right) + \frac{\sum y_i x_i}{n} \quad (11-6)$$

Traditional notation

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

11-2: Simple Linear Regression

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The **analysis of variance** identity is

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11-24)$$

Symbolically,

$$SS_T = SS_R + SS_E \quad (11-25)$$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2) VERY COMMONLY USED

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

Adjusted R-Squared

- Adjusted R-Squared uses unbiased estimates of σ_ε^2 and σ_Y^2 :

$$= 1 - \frac{\sigma_\varepsilon^2}{s_Y^2} = 1 - \frac{\frac{SS_E}{n-2}}{\frac{SS_T}{n-1}}$$

- For small samples it is different from the regular R-Squared: $R^2 = 1 - \frac{SS_E}{SS_T}$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

- For the oxygen purity regression model,

$$\begin{aligned}R^2 &= SS_R/SS_T \\ &= 152.13/173.38 \\ &= 0.877\end{aligned}$$

- Thus, the model accounts for 87.7% of the variability in the data.

11-2: Simple Linear Regression

Estimating σ_ε^2

An **unbiased estimator** of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{n - 2} \quad (11-13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \quad (11-14)$$

11-3: Properties of the Least Squares Estimators

- Slope Properties

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

$$V(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{S_{xx}} = \frac{\hat{\sigma}_\varepsilon^2}{n \hat{\sigma}_x^2}$$

Large $n \rightarrow$ small variance of β_1

- Intercept Properties

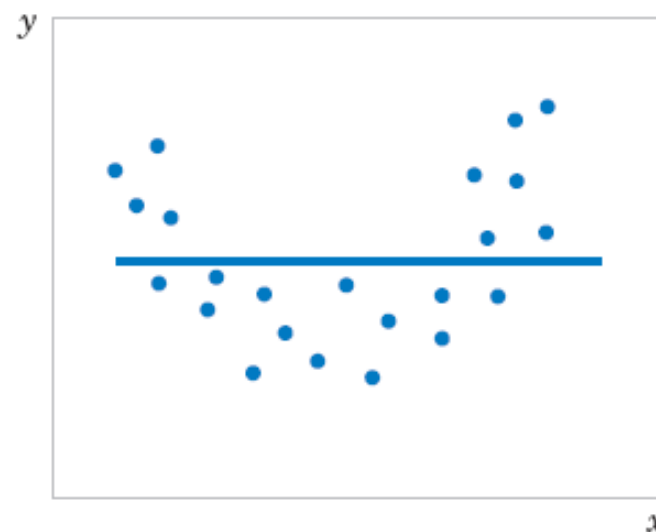
$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] =$$

$$= \hat{\sigma}_\varepsilon^2 \left[1 + \frac{\mu_x^2}{\hat{\sigma}_x^2} \right] \times \frac{1}{n}$$

11-4: Hypothesis Tests in Simple Linear Regression



(a)



(b)

Figure 11-5 The hypothesis $H_0: \beta_1 = 0$ is not rejected.

Figure 11-5 The null hypothesis $H_0: \beta_1 = 0$ is accepted.

11-4: Hypothesis Tests in Simple Linear Regression

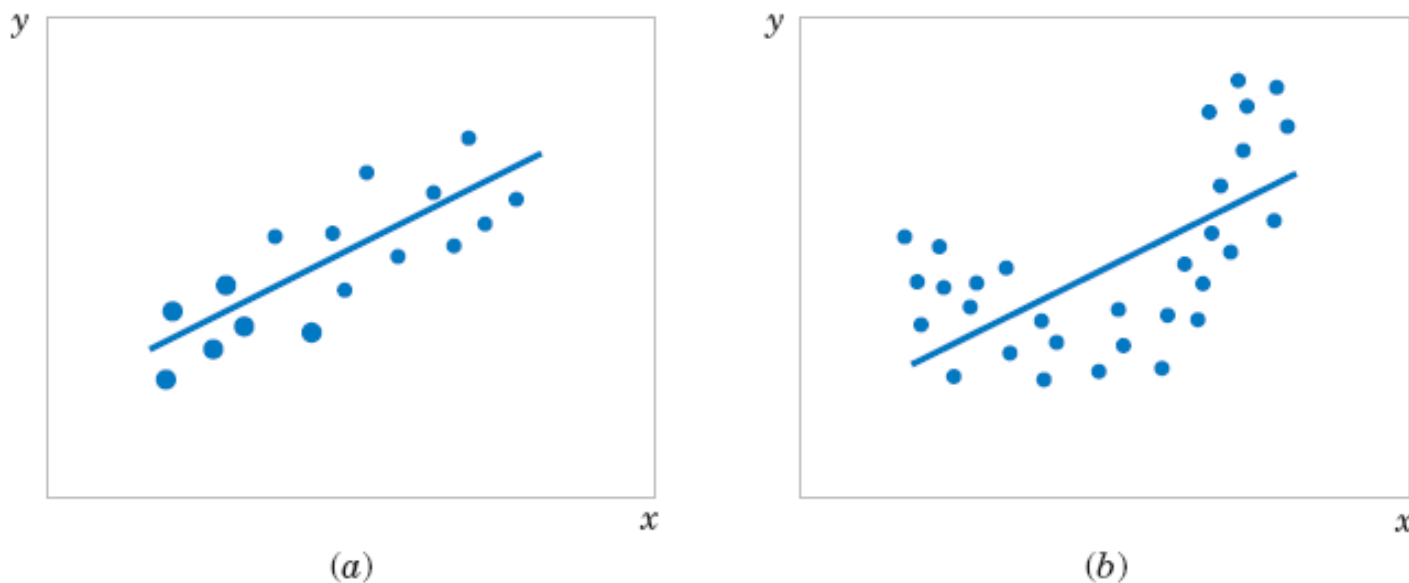


Figure 11-6 The hypothesis $H_0: \beta_1 = 0$ is rejected.

Figure 11-6 The **null hypothesis $H_0: \beta_1 = 0$ is rejected.**

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of Z-tests for large n

An important special case of the hypotheses of Equation 11-18 is

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**. *Failure to reject* H_0 is equivalent to **concluding that there is no linear relationship between X and Y** .

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Choose α

(e.g. $\alpha = 5\%$
for 95%

confidence
in rejecting
 H_0)

$$Z = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}_e}{\hat{\sigma}_x} \cdot \frac{1}{\sqrt{n}}}$$

for $\alpha = 5\%$

Reject H_0 if $|Z| > Z_{\alpha/2} = 1.96$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of t -tests for smaller n .

The number of degrees of freedom in $n-2$

One can always fit a straight line through two points so one needs $n \geq 3$

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$T = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}_e}{\sigma_x} \cdot \frac{1}{\sqrt{n}}}$$

Reject H_0 if $|T| > t_{\alpha/2, n-2}$

Choose α
(e.g. $\alpha = 5\%$
for 95%
confidence
in rejecting
 H_0)

$t_{\alpha/2, n-2}$ is such
 $1 - \frac{\alpha}{2} = \text{cdf}(t_{\alpha/2, n-2}, n-2)$

Credit: XKCD
comics

WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

QUESTIONS FOUND IN GOOGLE AUTOCOMPLETE

WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS

WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE

WHY ARE THERE ANTS IN MY LAPTOP

WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KUNGONS DIFFERENT



WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY



WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS

WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

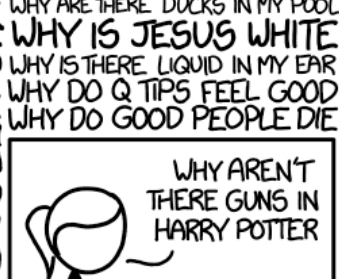
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD



WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY IS LIFE SO BORING



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG