5.5 Tests for Comparing the Parameters of Two Populations (Cont.)

Tests for the difference of means, paired data, $\theta = \mu_1 - \mu_2$ ($\boxed{\textbf{ttest}}$)

We test the hypotheses

$$\begin{aligned} H_0 &: \ \mu_1 - \mu_2 = 0, \\ H_1 &: \ \begin{cases} \mu_1 - \mu_2 < 0 \\ \mu_1 - \mu_2 > 0 \\ \mu_1 - \mu_2 \neq 0, \end{cases} \end{aligned} \quad \text{equivalent to} \quad \begin{aligned} H_0 &: \ \mu_1 = \mu_2, \\ H_1 &: \ \begin{cases} \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 \neq \mu_2. \end{cases} \end{aligned} \tag{5.1}$$

Recall that in many applications, we want to compare the means of two populations, when two random samples (one from each population) are available, which *are not* independent, where each observation in one sample is naturally or by design *paired* with an observation in the other sample.

In such cases, both samples have the same length, n:

$$X_{11}, \ldots, X_{1n} \text{ and } X_{21}, \ldots, X_{2n}$$

and we consider the sample of their differences,

$$D_1, \ldots, D_n,$$

where

$$D_i = X_{1i} - X_{2i}, \ i = \overline{1, n}.$$

For this sample, we have

$$\begin{aligned} \overline{X}_d &= \frac{1}{n} \sum_{i=1}^{n} D_i, \text{ the sample mean and} \\ s_d^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(D_i - \overline{X}_d\right)^2, \text{ the sample variance.} \end{aligned}$$

Then, it is known that when n is large enough ($n > 30$) or the two populations that the samples are drawn from have approximately Normal distributions $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)$, the statistic

$$T = \frac{\overline{X}_d - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}} \tag{5.2}$$

has a Student $T(n-1)$ distribution, so we can use it as a test statistic for testing the hypotheses (5.1). Its observed value is

$$T_0 \;=\; \frac{\overline{X}_d}{\frac{s_d}{\sqrt{n}}}. \tag{5.3}$$

Then, as before, we determine the rejection region corresponding to the three alternatives to be

$$RR: \begin{cases} \{T_0 \le t_\alpha\} \\ \{T_0 \ge t_{1-\alpha}\} \\ \{|T_0| \ge |t_{1-\frac{\alpha}{2}}|\} \end{cases} \tag{5.4}$$

and compute the P-value by

$$P = \begin{cases} P(T \le T_0 \mid H_0) & = & F(T_0) \\ P(T \ge T_0 \mid H_0) & = & 1 - F(T_0) \\ P(|T| \ge |T_0| \mid H_0) & = & 2\left(1 - F(|T_0|)\right), \end{cases} \tag{5.5}$$

where the quantiles and the cdf F refer to the $T(n-1)$ distribution.

Example 5.1. Information about ocean weather can be extracted from radar returns with the aid of special algorithms. A study is conducted to estimate the difference in wind speed as measured on the ground, at 12 specified times, using two methods simultaneously. These data result:

Times	1	2	3	4	5	6	7	8	9	10	11	12
Method I	4.46	3.99	3.73	3.29	4.82	6.71	4.61	3.87	3.17	4.42	3.76	3.3
Method II	4.08	3.94	5.00	5.2	3.92	6.21	5.95	3.07	4.76	3.25	4.89	4.8

Assuming the measurements taken by the two methods are approximately Normally distributed, at the 1% significance level, does the data suggest that, on average, the two sets of measurements differ?

Solution. By looking at the data, we see that at some times the measurement taken by the first method is higher, at others, the one given by the second. So we cannot say if, on average, these differences will cancel each other, to yield about the same mean value.
So, we want to test

$$\begin{aligned} H_0 &: \;\; \mu_1 = \mu_2 \\ H_1 &: \;\; \mu_1 \ne \mu_2, \end{aligned}$$

a two-tailed alternative. The samples yield the following data: sample size $n = 12$, sample mean $\overline{X}_d = -0.4117$ and sample variance $s_d^2 = 1.2973$, so $s_d = 1.139$.
The observed value of the test statistic from (5.3) is

$$T_0 = \frac{\overline{X}_d}{\frac{s_d}{\sqrt{n}}} = -1.2521.$$

For $\alpha = 0.01$, the quantiles for the $T(11)$ distribution are

$$\begin{aligned} t_{\alpha/2} &= t_{0.005} &= -3.1058, \\ t_{1-\alpha/2} &= -t_{\alpha/2} &= 3.1058, \end{aligned}$$

so the rejection region is

$$RR = (-\infty, -3.1058] \cup [3.1058, \infty).$$

Since $T_0 \notin RR$, we cannot reject the null hypothesis, which means we decide that the two population means are approximately equal.
On the other hand, the P-value of this test is

$$P = 2\,(1 - F(|T_0|)) = 0.2365,$$

We have

$$\alpha = 0.01 < 0.2365 = P,$$

the minimum rejection level, so the decision is to *not reject* the null hypothesis. Notice that, again, the P-value is much larger than any conceivable significance level α, so that means that the data strongly suggests that H_0 should not be rejected, i.e., that the two population means *do not* differ.

■

Remark 5.2. The Matlab command **ttest** that performs a T-test for one population mean (in the general case, when σ is not known), can also be used for a paired T-test.

Tests for comparing population proportions, $\theta = p_1 - p_2$

Similarly, in this case, if the samples are large enough ($n_1 + n_2 > 40$), then the variable

$$Z = \frac{\overline{p}_1 - \overline{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \in N(0,1), \tag{5.6}$$

where $\overline{p}_1$ and $\overline{p}_2$ are the two sample proportions. To test

$$\begin{aligned} H_0 :& \quad p_1 - p_2 = 0, \text{ versus} \\ H_1 :& \quad \begin{cases} p_1 - p_2 < 0 \\ p_1 - p_2 > 0 \\ p_1 - p_2 \neq 0, \end{cases} \end{aligned}$$

which is equivalent to

$$\begin{aligned} H_0 :& \quad p_1 = p_2, \text{ versus} \\ H_1 :& \quad \begin{cases} p_1 < p_2 \\ p_1 > p_2 \\ p_1 \neq p_2, \end{cases} \end{aligned} \tag{5.7}$$

we use $TS = Z$ from (5.6) as test statistic. Let us see what the observed value Z_0 would be. Since under the null hypothesis, $p_1 = p_2$, it makes sense to estimate both proportions in (5.6) by the *overall* proportion

$$\hat{p} = \frac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2}, \tag{5.8}$$

called the **pooled proportion** (a proportion that takes into account data from both samples). Then the observed value of the test statistic Z_0 is

$$Z_0 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}. \tag{5.9}$$

The rejection regions for the three alternatives are then given, as before, by

$$RR : \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\} \end{cases} \tag{5.10}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \leq Z_0 \mid H_0) & = \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) & = 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) & = 2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{5.11}$$

since $N(0,1)$ is symmetric, where Φ is Laplace's function, the cdf for the $N(0,1)$distribution.

Example 5.3. Recall Example 5.5 from Lecture 7: A company is receiving a large shipment of items. For quality control purposes, they collect a sample of 200 items and find 24 defective ones. Suppose now that the company is trying a new supplier. A sample of 150 items produced by the second supplier contains 21 defective parts. At the 5% significance level, does the new supplier seem worse than the first one?

Solution. For the first supplier the data was $n_1 = 200$, $\overline{p}_1 = 0.12$, for the new one, we have $n_2 = 150$ and $\overline{p}_2 = 0.14$. Considering that now 14% of items are defective and with the first supplier the percentage was 12%, the company is in a serious bind: it is afraid that the second supplier may be worse than the first one. Now, "worse" would mean that for the entire populations the proportions satisfy $p_1 < p_2$. So, we perform a *left-tailed* test

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_1 &: p_1 < p_2. \end{aligned}$$

For a left-tailed test and significance level $\alpha = 0.05$, the rejection region is

$$RR = (-\infty, z_{0.05}] = (-\infty, -1.654].$$

The pooled proportion from (5.8) is

$$\hat{p} = \frac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2} = \frac{24 + 21}{350} = 0.1286.$$

Then the observed value of the test statistic (from (5.9)) is

$$Z_0 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -0.5531.$$

Since $Z_0 \notin RR$, we do not reject the null hypothesis, i.e. we conclude that overall, the second supplier is *not* worse than the first one.
For significance testing, the P- value of the test is

$$P = P(Z \le Z_0) = P(Z \le -0.5531) = \Phi(-0.5531) = 0.29,$$

again, *very* large, much larger than this (or any reasonable) α, so the decision is to not reject H_0, a decision that seems *strongly* supported by the data.

■

5.6 Summary of hypothesis and significance testing

We can use data to verify statements and *test hypotheses*. Essentially, we measure the evidence provided by the data against the null hypothesis H_0. Then we decide whether it is sufficient for rejecting it or not. Given a significance level $\alpha \in (0, 1)$, we can construct acceptance and rejection regions, compute a suitable test statistic, and make a decision depending on which region it belongs to.

Alternatively, we may compute a P-value of the test. It shows how *significant* the evidence against H_0 is. Low P-values suggest rejection of the null hypothesis. The P-value of a test is the boundary between levels α-to-reject and α-to-accept. It also represents the probability of observing the same or a more extreme sample than the one that was actually observed.

We already mentioned that in practice, *significance* testing is preferred, i.e., computing the P-value and comparing it to the significance level α (and that is how hypothesis testing is implemented in any software). That is much more efficient from the computational perspective, as computation of the quantiles can be rather expensive.

In fact, in practice, a significance level α is *hardly ever specified*. Instead, just the P-value is computed. Since the null hypothesis is *always* in the form of an equality

$$H_0 : \theta \ = \ \theta_0,$$

whichever alternative we are testing (left-, right-, or two-tailed), to reject H_0 (when P is "small") means that the data shows that there are *significant differences* (statistically speaking) from what it states. How "significant"? That depends on how small the P-value is. The following levels are customary for how "significant" the differences are:

$$\begin{aligned} P \ &> \ 0.05 && \Rightarrow \ \textbf{not} \text{ significant}, \\ 0.01 \ < P \ &\leq \ 0.05 && \Rightarrow \ \textbf{significant}, \\ 0.001 \ < P \ &\leq \ 0.01 && \Rightarrow \ \textbf{distinctly} \text{ significant}, \\ P \ &\leq \ 0.001 && \Rightarrow \ \textbf{very} \text{ significant}. \end{aligned}$$

Chapter 5. Correlation and Regression

1 Basic Concepts

In the previous chapter, we were concerned about the distribution of *one random variable*, its parameters, expectation, variance, median, symmetry, skewness, etc., but many variables observed in real life are *related*. A very important part of Statistics is describing *relations* among two (or more) variables, whether or not they are independent, and if they are not, what is the nature of their dependence. One of the most fundamental concepts in statistical research is the concept of *correlation*.

Correlation is a measure of the relationship between one dependent variable and one or more independent variables. If two variables are correlated, this means that one can use information about one variable to predict the values of the other variable.

Definition 1.1.
The independent variables $X^{(1)}, \ldots, X^{(k)}$ are called ***predictors*** *and are used to predict the values and behavior of some other variable Y.*
The dependent variable Y is called ***response*** *and is a variable of interest that we predict based on one or several predictors.*
Regression *is the method or statistical procedure that is used to establish the relationship between response and predictor variables.*

Establishing and testing such a relation enables us:
− to understand interactions, causes, and effects among variables;
− to predict unobserved variables based on the observed ones;
− to determine which variables significantly affect the variable of interest.

Consider several situations when we can predict a dependent variable of interest from independent predictors.

Example 1.2 (World Population)**.** According to the International Data Base of the U.S. Census Bureau, population of the world grows according to Table 1. How can we use these data to predict the world population in years 2025 and 2030?

Year	Pop. (mln. people)	Year	Pop.(mln.people)	Year	Pop.(mln.people)
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6970
1965	3350	1990	5287	2015	7405
1970	3712	1995	5700	2020	7821

Table 1: World Population 1950-2020

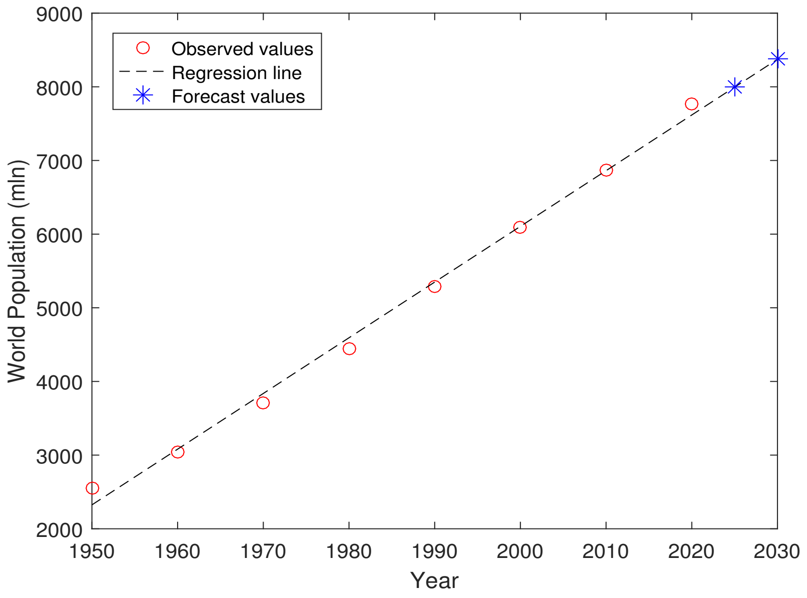


Fig. 1: World population and regression forecast

Figure 1 shows that the population (response) is tightly related to the year (predictor). It increases every year, and its growth is almost linear. If we estimate the regression function relating our response and our predictor (see the dotted line on Figure 1) and extend its graph to the year 2030, the forecast is ready.

A straight line that fits the observed data for years $1950 - 2020$ predicts a population of 8.06 billion in 2025 and 8.444 billion in 2030. It also shows that between 2020 and 2025, the world population reaches the historical mark of 8 billion (which actually happened in 2023 ...). How accurate is the forecast obtained in this example? The observed population during $1950 - 2020$ appears rather close to the estimated regression line in Figure 1. It is reasonable to hope that it will continue to do so through 2030.

Example 1.3 (House Prices)**.** Seventy house sale prices in a certain county (in the USA) are depicted in Figure 2 along with the house area.

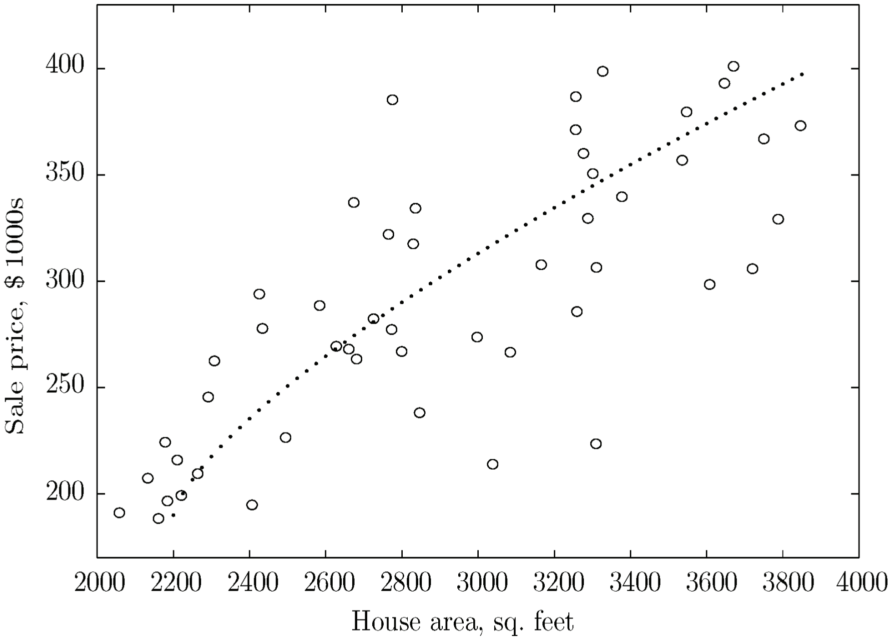


Fig. 2: Ex. 2

First, we see a clear relation between these two variables, and in general, bigger houses are more expensive. However, the trend no longer seems linear.

Second, there is a large amount of variability around this trend. Indeed, area is not the only factor determining the house price. Houses with the same area may still be priced differently. Then, how can we estimate the price of a 3200-square-foot ($\approx$ 300 m^2) house? We can estimate the general trend (the dotted line in Figure 2) and plug 3200 into the resulting formula, but due to obviously high variability, our estimation will not be as accurate as in Example 1.2.

To improve our estimation in the last example, we may take other factors into account: location, the number of bedrooms and bathrooms, the backyard area, the average income of the neighborhood, etc. If all the added variables are relevant for pricing a house, our model will have a closer fit and will provide more accurate predictions.

2 Univariate Regression, Curves of Regression

First we focus on **univariate regression**, predicting response Y based on *one* predictor X.

So, we have two vectors X and Y of the same length. We can get a first idea of the relationship between the two, by plotting them in a **scattergram**, or **scatterplot**, which is a plot of the points with coordinates $(x_i, y_i),\ x_i \in X,\ y_i \in Y,\ i = \overline{1,l}$. Denote by $N = 2l$, the entire data size. If N is large, we can group the data into n^2 classes and denote by (x_i, y_j) the class mark and by f_{ij} the absolute frequency of the class (i,j), $i,j = \overline{1,n}$ (just as in the one-dimensional case). Then we represent the two-dimensional characteristic (X,Y) in a *correlation table*, or *contingency table*, as shown in Table 2.

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	y_n	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1n}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{in}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_n	f_{n1}	$\dots$	f_{nj}	$\dots$	f_{nn}	$f_{n.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.n}$	$f_{..} = N.$

Table 2: Correlation Table

Notice that

$$\sum_{j=1}^{n} f_{ij} = f_{i.},\ \sum_{i=1}^{n} f_{ij} = f_{.j},\ \sum_{i=1}^{n} f_{i.} = \sum_{j=1}^{n} f_{.j} = f_{..} = N = 2l.$$

Now we can define numerical characteristics associated with (X,Y).

Definition 2.1. *Let (X,Y) be a two-dimensional characteristic whose distribution is given by Table 2 and let $k_1, k_2 \in \mathbb{N}$.*

(1) *The **(initial) moment of order $(\boldsymbol{k_1}, \boldsymbol{k_2})$** of (X,Y) is the value*

$$\overline{\nu}_{k_1k_2} = \frac{1}{N}\sum_{i,j=1}^{l} x_i^{k_1} y_j^{k_2},\ \ \overline{\nu}_{k_1k_2} = \frac{1}{N}\sum_{i,j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2}, \tag{2.1}$$

for primary and for grouped data, respectively.

(2) *The **central moment of order $(\boldsymbol{k_1}, \boldsymbol{k_2})$** of (X,Y) is the value*

$$\overline{\mu}_{k_1k_2} = \frac{1}{N}\sum_{i,j=1}^{l} (x_i - \overline{x})^{k_1} (y_j - \overline{y})^{k_2},\ \ \overline{\mu}_{k_1k_2} = \frac{1}{N}\sum_{i,j=1}^{n} f_{ij}(x_i - \overline{x})^{k_1} (y_j - \overline{y})^{k_2}, \tag{2.2}$$

for primary and for grouped data, where $\overline{x} = \overline{\nu}_{10}$ and $\overline{y} = \overline{\nu}_{01}$ are the means of X and Y, respectively.

Remark 2.2. Just as the means of the two characteristics X and Y can be expressed as moments of (X, Y), so can their variances:

$$\begin{aligned} \overline{\sigma}_X^2 &= \overline{\mu}_{20} &= \overline{\nu}_{20} - \overline{\nu}_{10}^2, \\ \overline{\sigma}_Y^2 &= \overline{\mu}_{02} &= \overline{\nu}_{02} - \overline{\nu}_{01}^2. \end{aligned}$$

Definition 2.3. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 2.*

(1) *The **covariance** ($\boxed{\text{cov}}$) of (X, Y) is the value*

$$\operatorname{cov}(X,Y) = \overline{\mu}_{11} = \frac{1}{N}\sum_{i,j=1}^{l}(x_i - \overline{x})(y_j - \overline{y}), \ \operatorname{cov}(X,Y) = \overline{\mu}_{11} = \frac{1}{N}\sum_{i,j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}), \tag{2.3}$$

for primary and for grouped data

(2) *The **correlation coefficient** ($\boxed{\text{corrcoef}}$) of (X, Y) is the value*

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}. \tag{2.4}$$

As we know from random variables, the covariance gives a rough idea of the relationship between X and Y. If X and Y are independent (so there is no relationship, no correlation between them), then the covariance is 0. If large values of X are associated with large values of Y, then the covariance will have a positive value, if, on the contrary, large values of X are associated with small values of Y, then the covariance will have a negative value. Also, an easier computational formula for the covariance is

$$\operatorname{cov}(X,Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}.$$

The correlation coefficient is then

$$\overline{\rho} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y}$$

and it satisfies the inequality

$$-1 \le \overline{\rho} \le 1. \tag{2.5}$$

By its variation between -1 and 1, its value measures the linear relationship between X and Y. If $\overline{\rho}_{XY} = 1$, there is a *perfect positive correlation* between X and Y, if $\overline{\rho}_{XY} = -1$, there is a *perfect negative correlation* between X and Y. In both cases, the linearity is "perfect", i.e there exist $a, b \in \mathbb{R},\ a \neq 0$, such that $Y = aX + b$. If $\overline{\rho}_{XY} = 0$, then there is no linear correlation between X and Y, they are said to be *(linearly) uncorrelated.* However, in this case, they may not be independent, some other type of relationship (not linear) may exist between them.

In what follows, for the simplicity of writing, we assume the data

$$\begin{aligned} X &= (x_1, \ldots, x_n), \\ Y &= (y_1, \ldots, y_n) \end{aligned}$$

is *ungrouped* and use the corresponding formulas (2.1)−(2.3).

Definition 2.4. *The* ***sample variances of*** X ***and*** Y *are given by*

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,\ s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{2.6}$$

and the ***covariance of*** (X, Y) *is*

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}). \tag{2.7}$$

Also, to make the subsequent computations and writing easier, we define the **sums of squares and cross-products**:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i(x_i - \overline{x}), \\ S_{yy} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i(y_i - \overline{y}), \\ S_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i(y_i - \overline{y}) = \sum_{i=1}^{n} y_i(x_i - \overline{x}). \end{aligned} \tag{2.8}$$

The formulas on the right-hand sides of (2.8) follow because $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}(y_i - \overline{y}) = 0$. Then

$$s_x^2 = \frac{S_{xx}}{n-1},\ s_y^2 = \frac{S_{yy}}{n-1},\ s_{xy} = \frac{S_{xy}}{n-1}.$$

Notice that the formula for the correlation coefficient *does not change*, regardless of whether the sums are divided by n or by $n-1$.

$$\overline{\rho} \;=\; \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \;=\; \frac{s_{xy}}{s_x s_y}.$$

To find a relationship between X and Y, we may go the following path: knowing the value of one of the characteristics, try to find a probable, an "expected" value for the other. If the two characteristics are related in any way, then there should be a pattern developing, i.e., the expected value of one of them, *conditioned* by the other one taking a certain value, should be a function of that value that the other variable assumes. In other words, we should consider *conditional means*, defined similarly to regular means, only taking into account the condition.

Definition 2.5.
The ***conditional mean*** *of Y, given $X = x_i$, is the value*

$$\overline{y}_i = \overline{y}(x_i) = E(Y|X = x_i),\; i = \overline{1,n} \tag{2.9}$$

and the curve $y = G(x)$ formed by the points with coordinates $(x_i, \overline{y}_i)$, $i = \overline{1,n}$, is called the ***curve of regression*** *of Y on X.*
The ***conditional mean*** *of X, given $Y = y_j$, is the value*

$$\overline{x}_j = \overline{x}(y_j) = E(X|Y = y_j),\; j = \overline{1,n} \tag{2.10}$$

and the curve $x = H(y)$ formed by the points with coordinates $(y_j, \overline{x}_j)$, $j = \overline{1,n}$, is called the ***curve of regression*** *of X on Y.*

Remark 2.6. The curve of regression of a response Y with respect to the predictor X is then the mean value of Y, $\overline{y}(x)$, given $X = x$. The curve of regression is determined so that it approximates best the scatterplot of (X, Y).

3 Least Squares Estimation

3.1 Least Squares Method

One of the most popular ways of finding curves of regression is the *least squares method*. Assume the curve of regression of Y on X is of the form

$$y = y(x) = G(x; \beta_0, \dots, \beta_k).$$

We are looking for a function $\widehat{G}(x)$ that passes as close as possible to the observed data points. This is achieved by minimizing distances between observed data points

$$y_1, \dots, y_n$$

and **fitted values**, i.e. the corresponding points on the fitted regression line

$$\widehat{y}_1 = \widehat{G}(x_1), \dots, \widehat{y}_n = \widehat{G}(x_n)$$

(see Figure 3). These differences are called **residuals**:

$$e_i \;=\; y_i - \widehat{y}_i, \; i \;=\; 1, \dots, n.$$

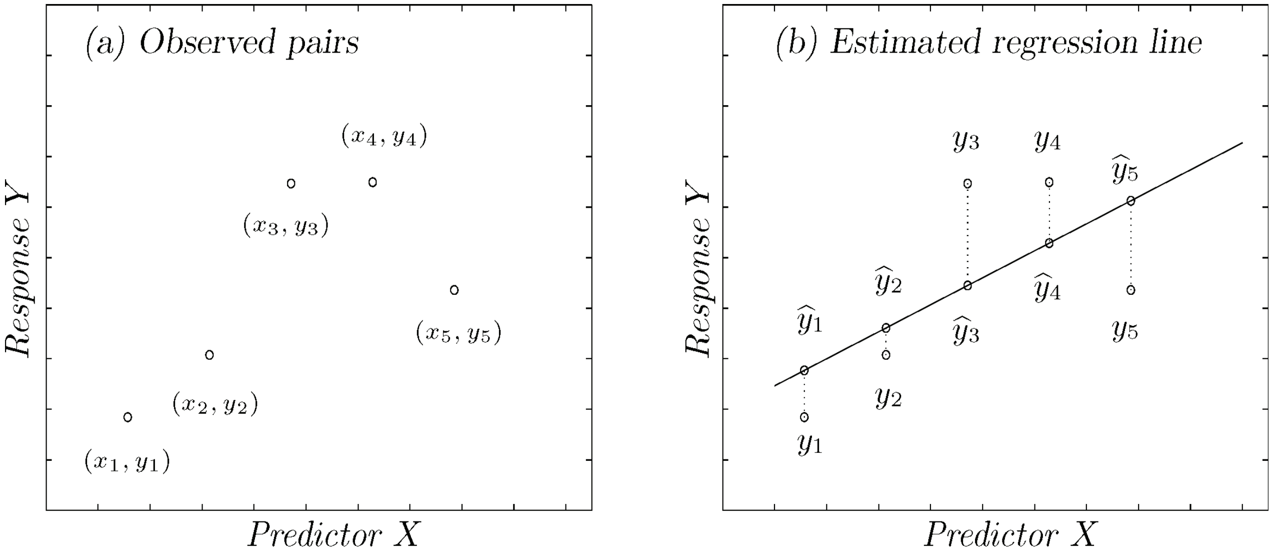


Fig. 3: Least squares estimation of the regression line

Method of least squares finds a regression function $\widehat{G}(x)$ that minimizes the sum of squared residuals

$$\sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2 .$$

Hence, we determine the unknown parameters $\beta_0, \ldots, \beta_s$ so that the *sum of squares error*

$$S = SS_{\text{ERR}} = \sum_{i=1}^{n} \Big(y_i - \widehat{G}(x_i; \beta_0, \ldots, \beta_s)\Big)^2$$

is minimum.
We find the point of minimum $(b_0, \ldots, b_s) = \left(\widehat{\beta}_0, \ldots, \widehat{\beta}_s\right)$ of S by solving the system

$$\frac{\partial S}{\partial \beta_r} = 0,\ r = \overline{0, s},$$

i.e.

$$-2 \sum_{i=1}^{n} \Big(y_i - \widehat{G}(x_i; \beta_0, \ldots, \beta_s)\Big) \frac{\partial \widehat{G}(x_i; \beta_0, \ldots, \beta_s)}{\partial \beta_r} = 0, \tag{3.1}$$

for every $r = \overline{0, s}$. These are called *normal equations*.
Then the equation of the curve of regression of Y on X is

$$y = \widehat{G}\left(x; b_0, \ldots, b_s\right).$$

Function $\widehat{G}$ is usually sought in a convenient (from the computational point of view) form: linear, quadratic, logarithmic, etc. The simplest form is linear.

3.2 Linear Regression

Let us consider the case of *linear regression* and find the equation of the *line of regression* of Y with respect to X.

We are finding a curve

$$y = \widehat{G}(x) = \beta_1 x + \beta_0.$$

The coefficients β_1 and β_0 are called *slope* and *intercept*, respectively. The sum of squared residuals is then

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

for which we find the minimum. We have to solve the 2×2 system

$$\begin{aligned} \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} &= 0 \\ \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} &= 0, \end{aligned}$$

i.e.

$$\begin{aligned} -2 \sum_{i=1}^{n} \Big(y_i - \beta_1 x_i - \beta_0 \Big) x_i &= 0 \\ -2 \sum_{i=1}^{n} \Big(y_i - \beta_1 x_i - \beta_0 \Big) &= 0, \end{aligned}$$

which becomes

$$\begin{cases} \left(\sum_{i=1}^{n} x_i^2 \right) \beta_1 + \left(\sum_{i=1}^{n} x_i \right) \beta_0 = \sum_{i=1}^{n} x_i y_j \\ \\ \left(\sum_{i=1}^{n} x_i \right) \beta_1 + \left(\sum_{i=1}^{n} 1 \right) \beta_0 = \sum_{i=1}^{n} y_j \end{cases}$$

and after dividing both equations by n,

$$\begin{cases} \overline{\nu}_{20} \beta_1 + \overline{\nu}_{10} \beta_0 = \overline{\nu}_{11} \\ \overline{\nu}_{10} \beta_1 + \overline{\nu}_{00} \beta_0 = \overline{\nu}_{01}. \end{cases}$$

Its solution is

$$\begin{aligned} b_1 &= \widehat{\beta}_1 = \frac{\overline{\nu}_{11} - \overline{\nu}_{10} \overline{\nu}_{01}}{\overline{\nu}_{20} - \overline{\nu}_{10}^2} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X^2} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y} \cdot \frac{\overline{\sigma}_Y}{\overline{\sigma}_X} = \overline{\rho} \, \frac{\overline{\sigma}_Y}{\overline{\sigma}_X} = \overline{\rho} \, \frac{s_y}{s_x}, \\ b_0 &= \widehat{\beta}_0 = \overline{\nu}_{01} - \overline{\nu}_{10} b_1 = \overline{y} - b_1 \cdot \overline{x}. \end{aligned}$$

So the equation of the line of regression of Y on X is

$$y - \overline{y} = \overline{\rho} \, \frac{s_y}{s_x} \, (x - \overline{x}) \tag{3.2}$$

and, by analogy, the equation of the line of regression of X on Y is

$$x - \overline{x} = \overline{\rho}\,\frac{s_x}{s_y}\,(y - \overline{y})\,. \tag{3.3}$$

Example 3.1. Let us consider the world population data in Example 1.2 and find the equation of the line of regression.

Solution. For the world population $(1950 - 2020)$ data, we find

$$\begin{aligned}
\overline{x} &= 1985,\ \ \overline{y} \ = \ 4991.5 \\
s_x &= 24.5,\ \ s_y \ = \ 1884.6 \\
\overline{\rho} &= 0.9972 \\
b_1 &= 76.72,\ \ b_0 \ = \ -147300.5
\end{aligned}$$

and the equation of the line of regression

$$y \ = \ 76.72x - 147300.5.$$

With this, we were able to forecast the values of 8.0604 billion for the year 2025 and 8.444 billion for 2030. Also, based on this model, the predicted population for 2024 is 7.9808 billion people (it was actually approximately 8.162 billion).

■

Remark 3.2.
1. The point of intersection of the two lines of regression (3.2) and (3.3) is $(\overline{x}, \overline{y})$. This is called the *centroid* of the distribution of the characteristic (X, Y).
2. The slope $\overline{a}_{Y|X} = \overline{\rho}\,\dfrac{s_y}{s_x}$ of the line of regression of Y on X is called the *coefficient of regression* of Y on X. Similarly, $\overline{a}_{X|Y} = \overline{\rho}\,\dfrac{s_x}{s_y}$ is the coefficient of regression of X on Y and we have the relation

$$\overline{\rho}^2 = \overline{a}_{Y|X}\,\overline{a}_{X|Y}.$$

3. For the angle α between the two lines of regression, we have

$$\tan\alpha \ = \ \frac{1-\overline{\rho}^2}{\overline{\rho}^2}\cdot\frac{s_x s_y}{s_x^2 + s_y^2}.$$

So, if $\mid \overline{\rho} \mid = 1$, then $\alpha = 0$, i.e. the two lines coincide. If $\mid \overline{\rho} \mid = 0$ (for instance, if X and Y are independent), then $\alpha = \frac{\pi}{2}$, i.e. the two lines are perpendicular.

Example 3.3. Let us examine the situations graphed in Figure 4.

- In Figure 4(a) $\overline{\rho} = 0.95$, positive and very close to 1, suggesting a strong positive linear trend. Indeed, most of the points are on or very close to the line of regression of Y on X. The positivity indicates that large values of X are associated with large values of Y. Also, since the correlation coefficient is so close to 1, the two lines of regression almost coincide.

- In Figure 4(b) $\overline{\rho} = -0.28$, negative and fairly small, close to 0. If a relationship exists between X and Y, it does not seem to be linear. In fact, they are very close to being independent, since the points are scattered around the plane, no pattern being visible. The two lines of regression are very distinct and both have negative slopes, suggesting that large values of X are associated with small values of Y.

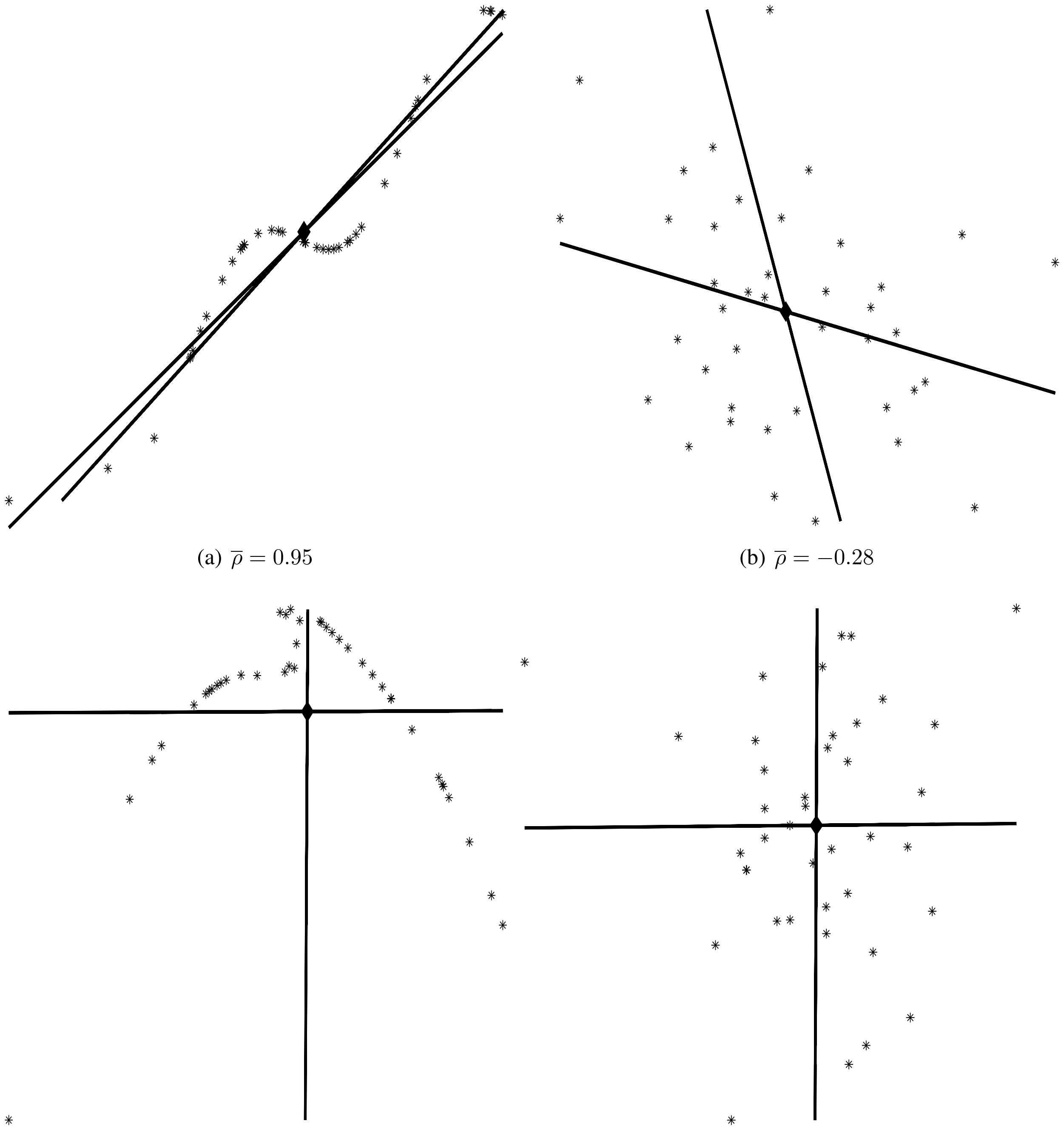
- In Figure 4(c) $\overline{\rho} = 0$, so the two characteristics are uncorrelated, no linear relationship exists between them. However they are not independent, they were chosen so that $Y = -X^2 + \sin\left(\frac{1}{X}\right)$. Notice also, that the two lines of regression are perpendicular.

- Finally, in Figure 4(d) $\overline{\rho} = 0$, again, so no linear relationship exists. In fact the two characteristics are independent, which is suggested by their random scatter inside the plane.

Overfitting a model

Among all possible straight lines, the method of least squares chooses one line that is closest to the observed data. Still, as we see in Figure 4, we can still have have some residuals and some positive sum of squared residuals. The straight line has not accounted for all 100% of variation among the fitted values. Why, one might ask, have we considered only linear models? As long as all x_i's are different, we can always find a regression function $\widehat{G}$ that passes through *all* the observed points without any error. Then, the sum of squared errors S will truly be minimized!

Trying to fit the data perfectly is a rather dangerous habit. Although we can achieve an excellent fit to the observed data, it never guarantees a good prediction. The model will be *overfitted*, too much “attached” to the given data. Using it to predict unobserved responses is very questionable (see Figure 5). Moreover, it will often result in large variances of the fitted values and therefore, unstable regression estimates.



(a) $\overline{\rho} = 0.95$

(b) $\overline{\rho} = -0.28$

(c) $\overline{\rho} = 0$

(d) $\overline{\rho} = 0$

Fig. 4: Scattergram, Lines of Regression and Centroid

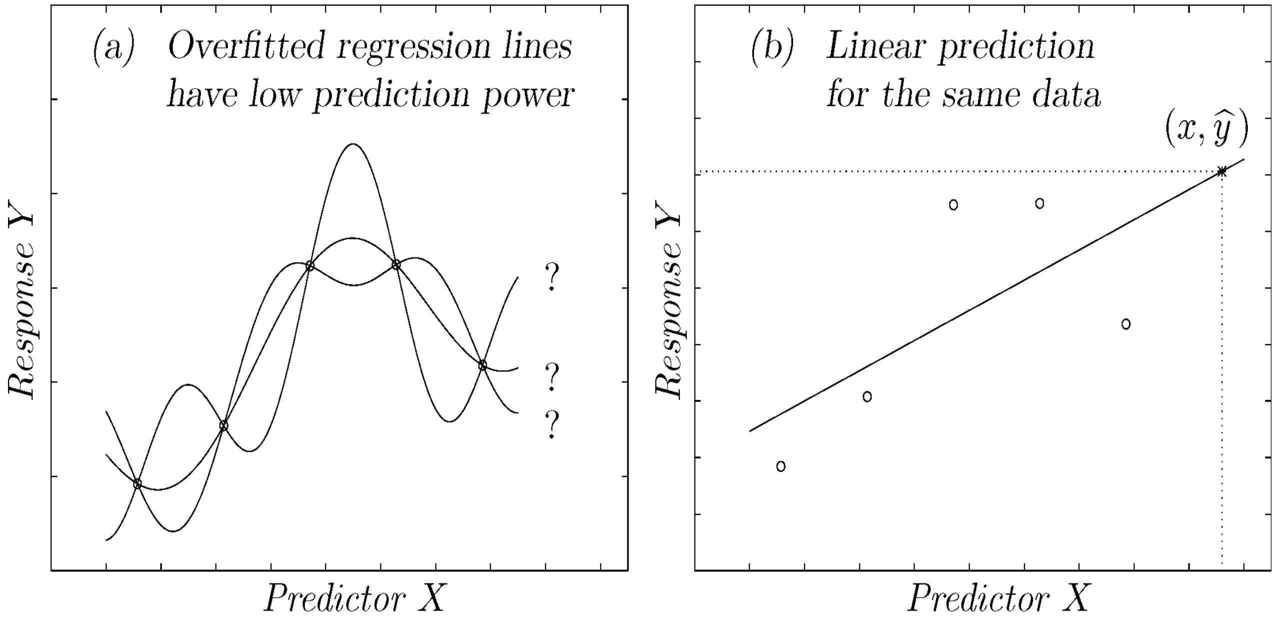


Fig. 5: Regression-based prediction, overfitting