Chapter 5. Correlation and Regression

1 Basic Concepts

So far we have been discussing a number of descriptive techniques for describing one variable only, its parameters, expectation, variance, median, symmetry, skewness, etc. However, 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

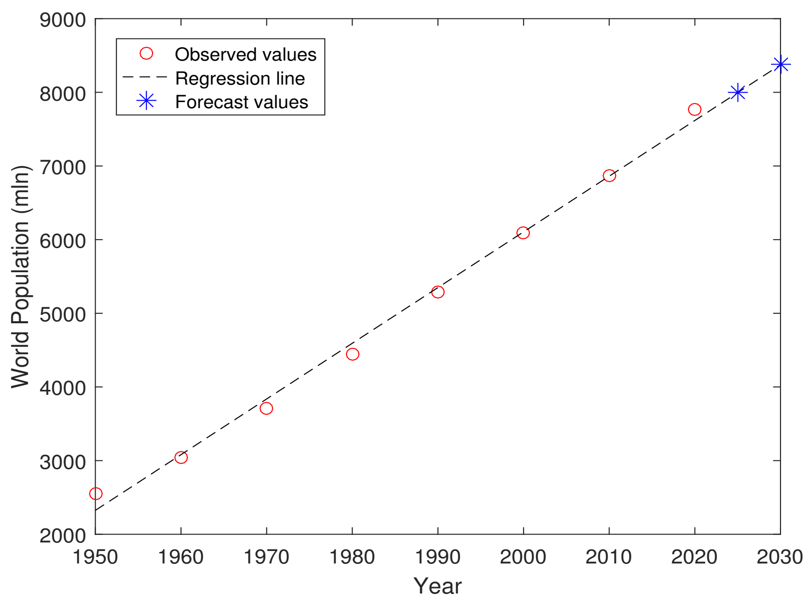


Fig. 1: World population and regression forecast

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 last summer ...). 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.

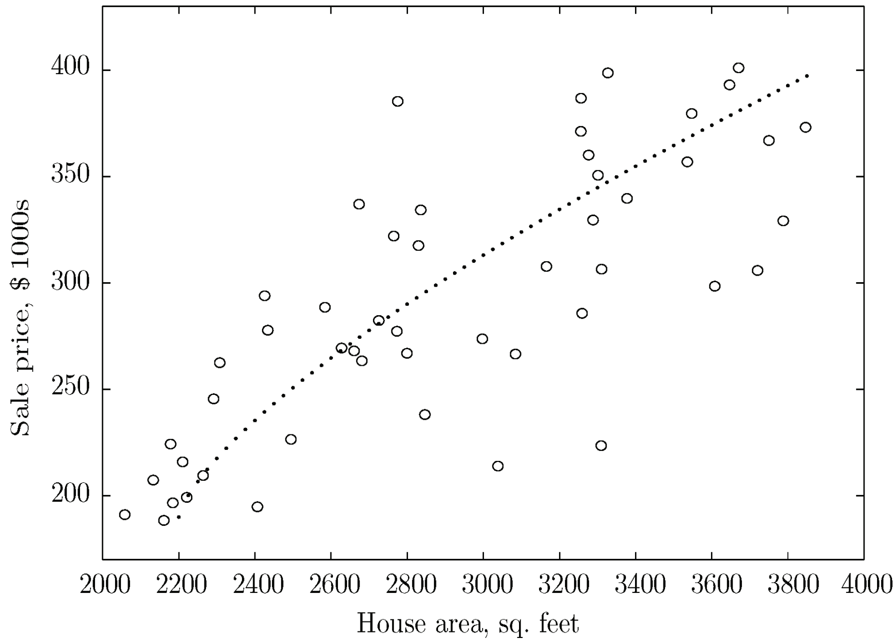


Fig. 2: Ex. 2

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 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*** $\boldsymbol{(k_1, k_2)}$ *of (X, Y) is the value*

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

for primary and for grouped data, respectively.

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

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

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** (*cov*) of* (X, Y) *is the value*

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

for primary and for grouped data

(2) *The **correlation coefficient** (*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

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

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

$$y_1, \ldots, y_n$$

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

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

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

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

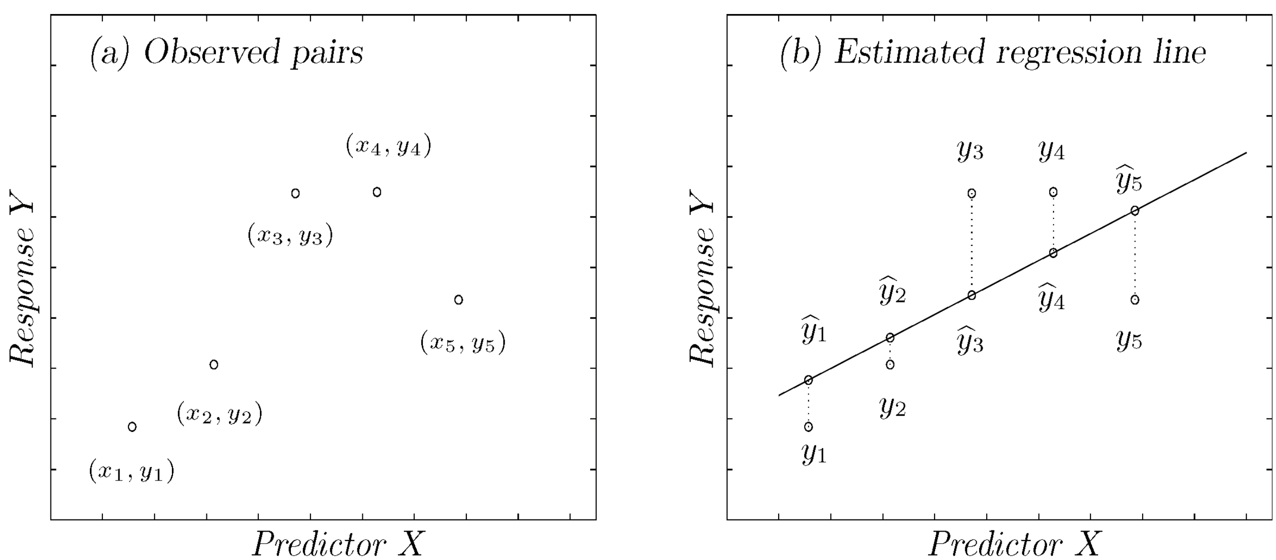


Fig. 3: Least squares estimation of the regression line

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

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

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

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

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.

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

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

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

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

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

$$b_0 = \widehat{\beta}_0 = \overline{\nu}_{01} - \overline{\nu}_{10} b_1 = \overline{y} - b_1 \cdot \overline{x}.$$

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 2023 is 7.9069 billion people.

■

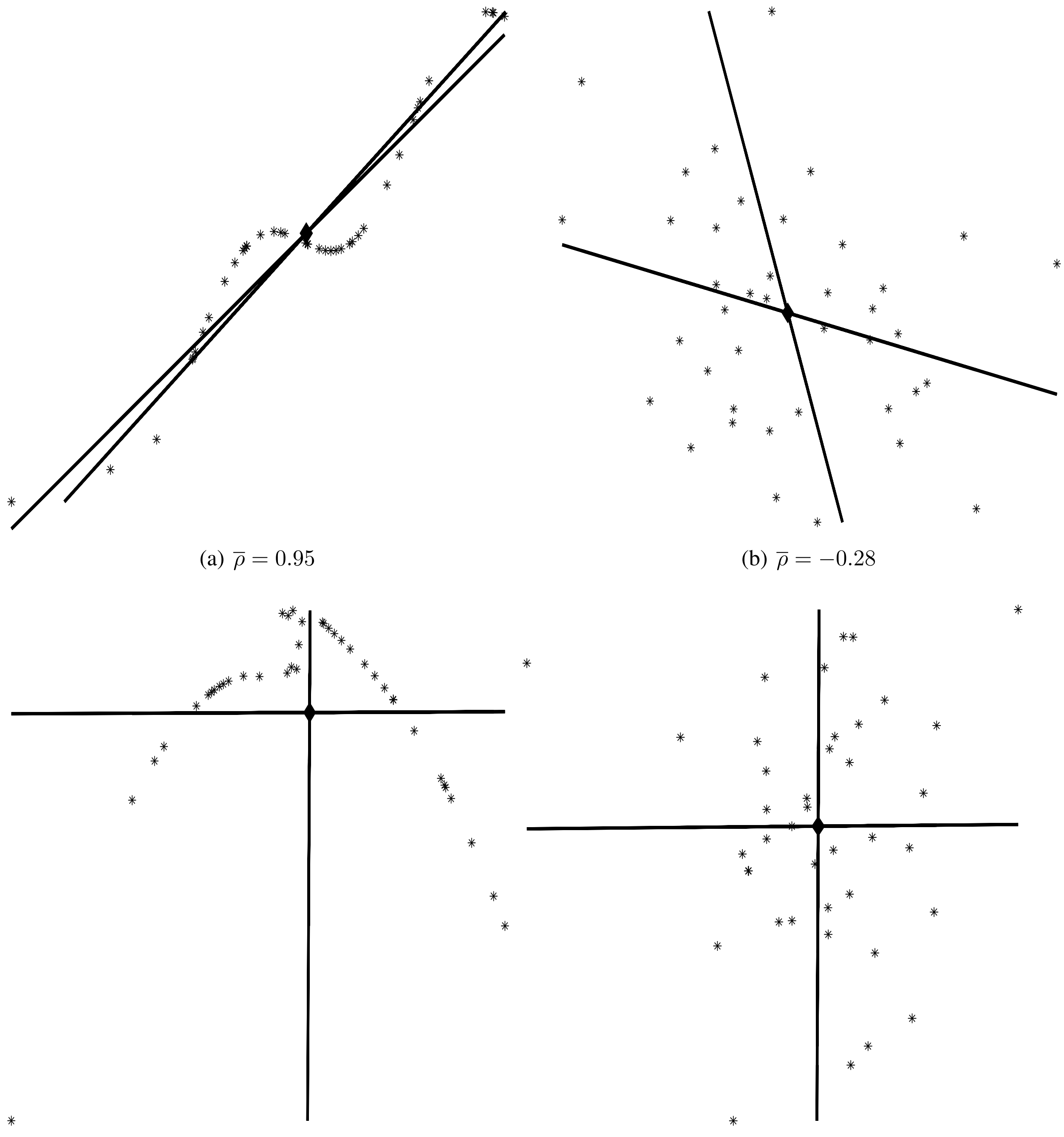
Remark 3.2.
1. The point of intersection of the two lines of regression (3.2) and (3.3) is $(\overline{x},\overline{y})$. This is called the *centroid* of the distribution of the characteristic (X,Y).
2. The slope $\overline{a}_{Y|X}=\overline{\rho}\,\dfrac{s_y}{s_x}$ of the line of regression of Y on 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 $|\,\overline{\rho}\,|=1$, then $\alpha=0$, i.e. the two lines coincide. If $|\,\overline{\rho}\,|=0$ (for instance, if X and Y are independent), then $\alpha=\dfrac{\pi}{2}$, i.e. the two lines are perpendicular.



(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

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.

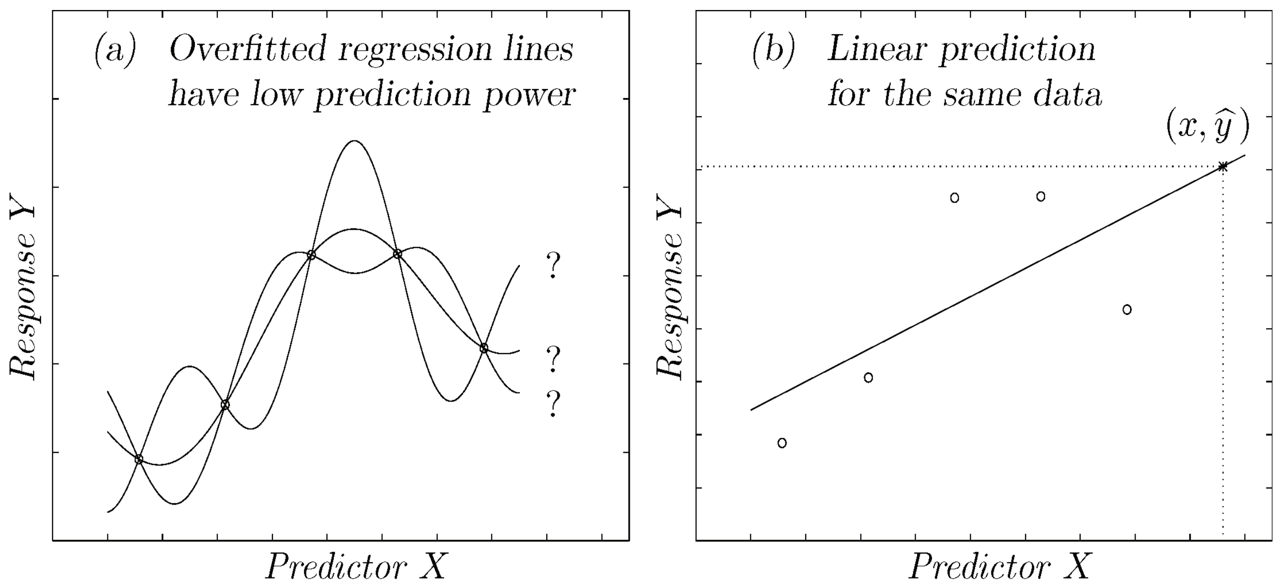


Fig. 5: Regression-based prediction, overfitting

3.3 Polynomial Univariate Regression

We have seen that linear regression with one predictor does not always produce the best approximation and, hence, the best tool for forecasting. What can be done about that? One thing would be considering polynomials of *higher* degree.

Example 3.4 (U.S. Population)**.** One can often reduce variability around the trend and do more accurate analysis by adding nonlinear terms into the regression model. In Example 3.1 we predicted the world population for years 2025–2030 based on the linear model

$$y = 76.72x - 147300.5$$

and we saw that this model has a pretty good fit.

However, a linear model does a poor prediction of the U.S. population between 1790 and 2010 (see Figure 6(a)). The population growth over a longer period of time is clearly nonlinear. On the other hand, a quadratic model in Figure 6(b) gives an amazingly excellent fit! It seems to account for everything except a temporary decrease in the rate of growth during World War II (1939–1945).

For this model, we assume

$$y = \beta_2 x^2 + \beta_1 x + \beta_0, \text{ or}$$

$$\text{population} = \beta_2 \cdot (\text{year})^2 + \beta_1 \cdot (\text{year}) + \beta_0.$$

This equation seems to give a more reliable fit. The coefficients β_0, β_1 and β_2 can be found, as

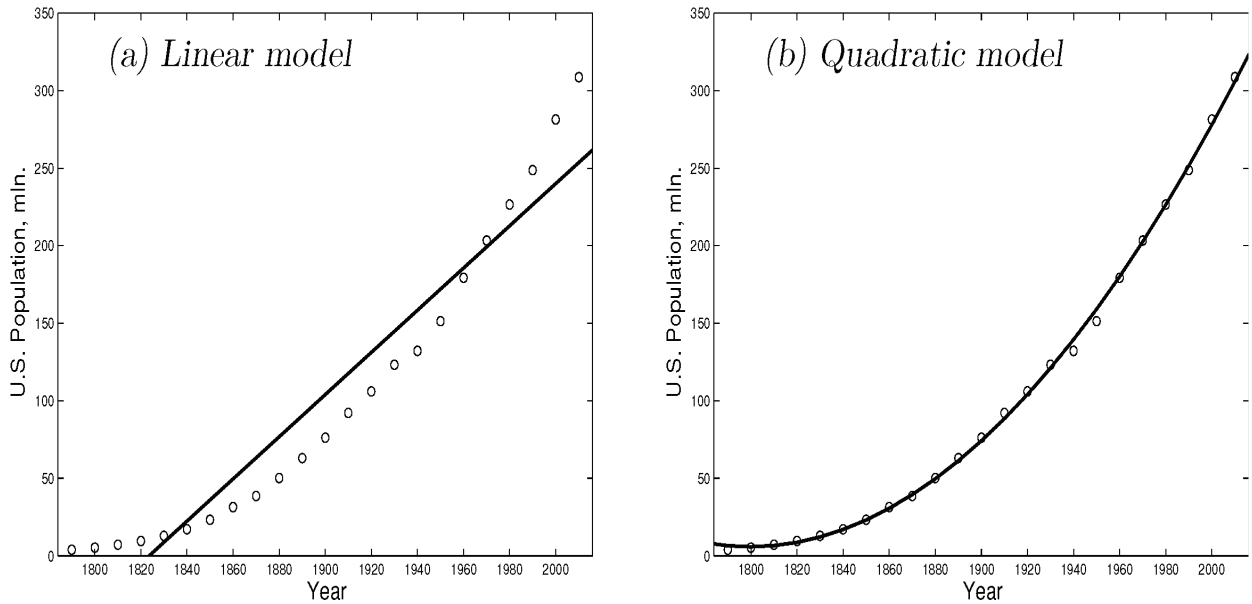


Fig. 6: U.S. population in $1790 - 2010$ (mln. people)

before, using the method of least squares.

Remark 3.5. Other types of curves of regression that are fairly frequently used are
$-$ *exponential* regression $y = ab^x$,
$-$ *logarithmic* regression $y = a \log x + b$,
$-$ *logistic* regression $y = \dfrac{1}{ae^{-x} + b}$,
$-$ *hyperbolic* regression $y = \dfrac{a}{x} + b$.