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 3.1 (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 (Lecture 8) 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 1(a)). The population growth over a longer period of time is clearly nonlinear. On the other hand, a quadratic model in Figure 1(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 before, using the method of least squares.

Remark 3.2. 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 $y = \dfrac{a}{x} + b$.

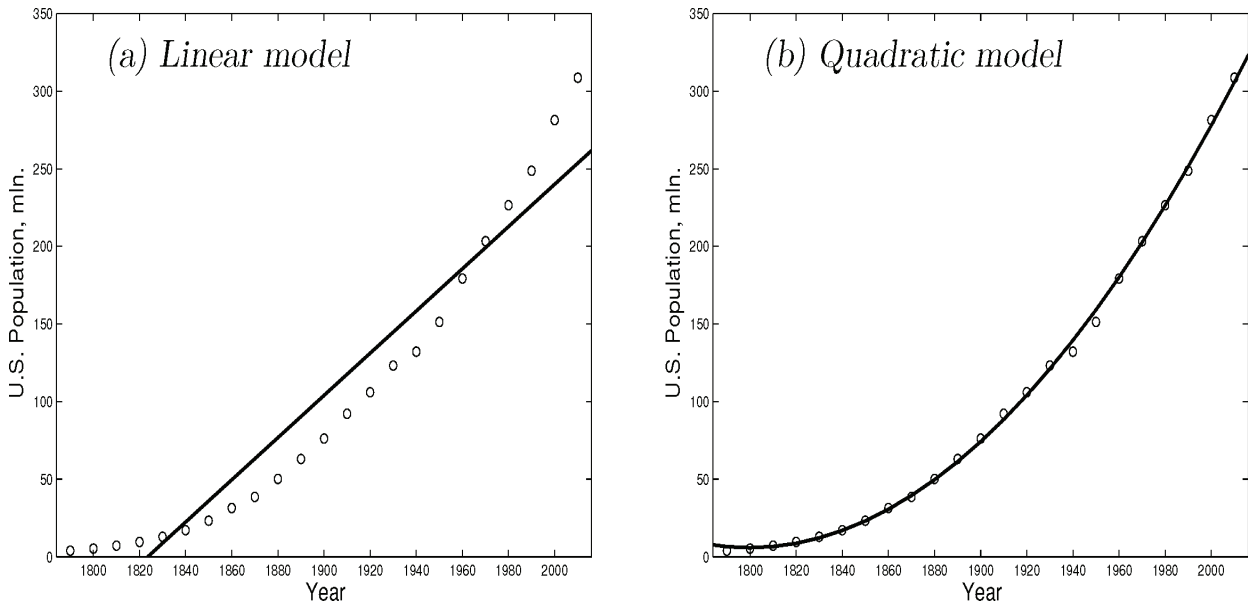

Fig. 1: U.S. population in $1790 - 2010$ (mln. people)

4 ANOVA and R-square

4.1 ANOVA - Preliminaries

Analysis of variance (ANOVA) explores variation among the observed responses. A portion of this variation can be explained by predictors. The rest is attributed to "error".

Let us recall Example 1.3 in Lecture 8. We see on Figure 2 that there exists some variation among the house sale prices on. Why are the houses priced differently? Obviously, the price depends on the house area, and bigger houses tend to be more expensive. So, to some extent, variation among prices is explained by variation among house areas. However, two houses with the same area may still have different prices. These differences cannot be explained by the area.

The total variation among observed responses is measured by the **total sum of squares**

$$SS_{\text{TOT}} \;=\; \sum_{i=1}^{n}(y_i - \overline{y})^2 \;=\; (n-1)s_y^2.$$

This is the variation of y_i about their sample mean *regardless* of our regression model. A portion of this total variation is attributed to predictor X and the regression model connecting the predictor

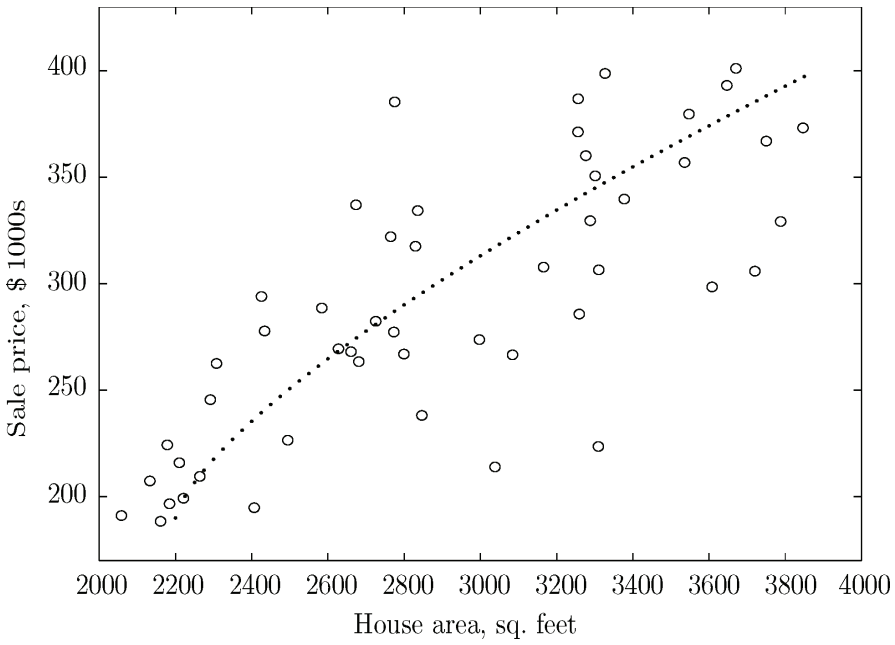

Fig. 2: House prices

with the response. This portion is measured by the **regression sum of squares**

$$SS_{\text{REG}} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2.$$

This is the portion of total variation *explained by the model*. Since the centroid $(\overline{x}, \overline{y})$ belongs to the regression line, we have $\overline{y} = b_1\overline{x} + b_0$, so we can write

$$\begin{aligned} SS_{\text{REG}} &= \sum_{i=1}^{n} (b_0 + b_1 x - \overline{y})^2 \\ &= \sum_{i=1}^{n} (\overline{y} - b_1\overline{x} + b_1 x - \overline{y})^2 \\ &= \sum_{i=1}^{n} b_1^2 (x - \overline{x})^2 \\ &= b_1^2 S_{xx} = b_1^2 (n-1) s_x^2. \end{aligned}$$

The rest of total variation is attributed to "error". It is measured by the sum of squares error SS_{ERR}. This is the portion of total variation *not explained by the model*. It is the sum of squared residuals that the method of least squares minimizes. Regression and error sums of squares partition SS_{TOT}

into two parts,

$$SS_{\text{TOT}} \;=\; SS_{\text{REG}} + SS_{\text{ERR}}.$$

The *goodness of fit*, appropriateness of the predictor and the chosen regression model can be judged by the proportion of SS_{TOT} that the model can explain.

Definition 4.1. ***R-square****, or* ***coefficient of determination*** *is the proportion of the total variation explained by the model,*

$$R^2 \;=\; \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} \;=\; 1 - \frac{SS_{\text{ERR}}}{SS_{\text{TOT}}}. \tag{4.1}$$

It is always between 0 and 1, with high values generally suggesting a good fit.
In univariate regression, R-square also equals the squared sample correlation coefficient

$$R^2 \;=\; \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} \;=\; \frac{b_1^2(n-1)s_x^2}{(n-1)s_y^2} \;=\; \left(b_1\frac{s_x}{s_y}\right)^2 \;=\; \left(\overline{\rho}\frac{s_y}{s_x}\frac{s_x}{s_y}\right)^2 \;=\; \overline{\rho}^2. \tag{4.2}$$

Example 4.2 (World Population, Continued)**.** Let us recall Example 3.1 (Lecture 8). By least square estimation, we found

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

and the equation of the line of regression

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

Now, we further compute

$$\begin{aligned}
SS_{\text{TOT}} \;&=\; (n-1)s_y^2 \;=\; 4.972\cdot 10^7,\\
SS_{\text{REG}} \;&=\; b_1^2(n-1)s_x^2 \;=\; 4.944\cdot 10^7,\\
SS_{\text{ERR}} \;&=\; SS_{\text{TOT}} - SS_{\text{REG}} \;=\; 2.83\cdot 10^5.
\end{aligned}$$

Then

$$R^2 \;=\; \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} \;=\; \overline{\rho}^2 \;=\; 0.9943 \text{ or } 99.43\%,$$

very high! This is a very good fit although some portion of the remaining 0.57% of total variation can still be explained by adding non-linear terms into the model.

4.2 Univariate ANOVA and F-test

For further analysis, we introduce **standard regression assumptions**. We will assume that observed responses y_i are independent Normal random variables with mean

$$E(Y_i) \;=\; \beta_0 + \beta_1 x_i$$

and constant variance σ^2. So, responses $Y_1, \dots, Y_n$ have different means but the same variance. Predictors x_i are considered *non-random*. As a consequence, regression estimates b_0 and b_1 also have Normal distribution.

This variance of the responses, σ^2, is equal to the mean squared deviation of responses from their respective expectations. Let us estimate it.

First, we estimate each expectation

$$E(Y_i) = G(x_i) = \beta_1 x_i + \beta_0 \text{ by } \widehat{G}(x_i) = b_0 + b_1 x_i = \widehat{y}_i.$$

Then, we consider deviations $e_i = y_i - \widehat{y}_i$, square them, and add. We obtain the error sum of squares

$$SS_{\text{ERR}} = \sum_{i=1}^{n} e_i^2.$$

Then, we divide this sum by its number of degrees of freedom, this is how variances are estimated.

Let us compute the degrees of freedom for all three SS in the regression ANOVA.
The total sum of squares

$$SS_{\text{TOT}} = (n-1)s_y^2 \text{ has } \; \text{df}_{\text{TOT}} = n - 1 \text{ degrees of freedom,}$$

because it is computed directly from the sample variance s_y^2.
Out of them, the regression sum of squares

$$SS_{\text{REG}} \text{ has } \; \text{df}_{\text{REG}} = 1 \text{ degree of freedom,}$$

because the regression line, which is just a straight line, has dimension 1.

This leaves $\mathrm{df}_{\mathrm{ERR}} = n - 2$ degrees of freedom for the error sum of squares, so that

$$\mathrm{df}_{\mathrm{TOT}} = \mathrm{df}_{\mathrm{REG}} + \mathrm{df}_{\mathrm{ERR}}.$$

Note that we could find the number of degrees of freedom by subtracting from the sample size n the number of estimated parameters, 2 (β_0 and β_1).

Then the **regression variance** is

$$s^2 \;=\; \frac{SS_{\mathrm{ERR}}}{n-2}$$

and it estimates $\sigma^2 = \mathrm{Var}(Y)$ unbiasedly.

ANOVA F-test

We want to test how significant is a predictor X for the estimate of a response Y, i.e., how much changes in X will produce significant changes in Y, via the linear relationship $G(x) = \beta_1 x + \beta_0$.

A non-zero slope β_1 indicates *significance* of the model, *relevance* of predictor X in the inference about response Y and existence of a linear relation among them. It means that a change in X causes changes in Y. In the absence of such relation, $E(Y) = \beta_0$ remains constant. Therefore, to see if X is significant for the prediction of Y, we test the hypotheses

$$\begin{aligned} H_0 &: \quad \beta_1 = 0 \\ H_1 &: \quad \beta_1 \neq 0. \end{aligned} \tag{4.3}$$

The hypotheses (4.3) can be tested using a T-statistic having a Student $T(n-2)$ distribution.

However, a more universal, and therefore, more popular method of testing significance of a model is is the **ANOVA F-test**. It compares the portion of variation explained by regression with the portion that remains unexplained. *Significant* models explain a relatively *large* portion.

Each portion of the total variation is measured by the corresponding sum of squares, SS_{REG} for the explained portion and SS_{ERR} for the unexplained portion (error). Dividing each sum of squares by the number of degrees of freedom, we obtain the *mean squares*

$$\begin{aligned} MS_{\mathrm{REG}} &= \frac{SS_{\mathrm{REG}}}{\mathrm{df}_{\mathrm{REG}}} = SS_{\mathrm{REG}}\,, \\ MS_{\mathrm{ERR}} &= \frac{SS_{\mathrm{ERR}}}{\mathrm{df}_{\mathrm{ERR}}} = \frac{SS_{\mathrm{ERR}}}{n-2}\,. \end{aligned}$$

We see that the sample regression variance is the mean squared error

$$s^2 \;=\; MS_{\mathrm{ERR}}.$$

Under the null hypothesis

$$H_0 : \beta_1 = 0,$$

both mean squares, MS_{REG} and MS_{ERR} are independent, and their ratio F has an F-distribution with parameters $\mathrm{df}_{\mathrm{REG}} = 1$ and $\mathrm{df}_{\mathrm{ERR}} = n - 2$ degrees of freedom. Then the ratio

$$F \;=\; \frac{MS_{\mathrm{REG}}}{MS_{\mathrm{ERR}}} \;=\; \frac{SS_{\mathrm{REG}}}{s^2}$$

is the test statistic used to test significance of the entire regression model. The ANOVA F-test is always *right-tailed*, because only large values of the F-statistic show a large portion of explained variation and the overall significance of the model.

A standard way to present analysis of variance is the ANOVA Table 1.

Source	Sum of squares SS	Degrees of freedom df	Mean Squares $MS = SS/\mathrm{df}$	F
Model	$SS_{\mathrm{REG}} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$	1	$MS_{\mathrm{REG}} = SS_{\mathrm{REG}}$	$\dfrac{MS_{\mathrm{REG}}}{MS_{\mathrm{ERR}}}$
Error	$SS_{\mathrm{ERR}} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$	$n-2$	$MS_{\mathrm{ERR}} = \dfrac{SS_{\mathrm{ERR}}}{n-2}$	
Total	$SS_{\mathrm{TOT}} = \sum_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$		

Table 1: Univariate ANOVA

Remark 4.3. Let us recall that for a *right*-tailed test, the rejection region is of the form

$$RR = (q_{1-\alpha}, \infty).$$

So, to test significance of the regression model, at the $100\alpha\%$ significance level, we will check if the obtained F-value is *larger* than the quantile $f_{1,n-2,1-\alpha}$.

Example 4.4. A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. Applying the program to data sets of different sizes, she gets the following results:

Data size (gigabytes), x	6	7	7	8	10	10	15
Processed requests, y	40	55	50	41	17	26	16

In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour.
a) Find the equation of the curve of regression and use it to predict the number of requests processed per hour y^*, for $x^* = 16$ gigabytes of data.
b) Construct the ANOVA table and discuss it.

Solution.
a) For our data, we have $n = 7$ and

$$\begin{aligned} \overline{x} &= 9, \quad \overline{y} = 35, \\ s_x &= 3.06, \; s_y = 15.56, \\ \overline{\rho} &= -0.81. \end{aligned}$$

The equation of the line of regression is

$$y = -4.14x + 72.29.$$

Notice the negative slope. It means that *increasing* incoming data sets by 1 gigabyte, we expect to process 4.14 *fewer* requests per hour.

According to this, the predicted value for x^* = 16 gigabytes is

$$y^* \;=\; -4.14 \cdot 16 + 72.29 \;=\; 6 \text{ requests processed within 1 hour.}$$

We already computed

$$b_1 \;=\; -4.14,\; b_0 \;=\; 72.29.$$

b) Further, we compute

$$\begin{aligned} SS_{\text{TOT}} &= S_{yy} \;=\; 1452, \\ SS_{\text{REG}} &= b_1^2 S_{xx} \;=\; 961.14, \\ SS_{\text{ERR}} &= SS_{\text{TOT}} - SS_{\text{REG}} \;=\; 490.86, \\ F &= \frac{(n-2)SS_{\text{REG}}}{SS_{\text{ERR}}} \;=\; 9.79. \end{aligned}$$

We have the ANOVA table

Source	Sum of squares	Degrees of freedom	Mean Squares	F
Model	961.14	1	961.14	9.79
Error	490.86	5	98.17	
Total	1452	6		

The regression variance is estimated by

$$s^2 \;=\; MS_{\text{ERR}} \;=\; 98.17.$$

R-square is

$$R^2 \;=\; \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} \;=\; \frac{961.14}{1452} \;=\; 0.662 \text{ or } 66.2\%.$$

That is, 66.2% of the total variation of the number of processed requests is explained by sizes of data sets only.

The F-statistic of 9.79 is not significant at the 0.025 level (because it is *not* larger than the quantile $f_{1,4,0.975} = 12.22$), but significant at the 0.05 level (it *exceeds* the value of the quantile $f_{1,4,0.950} = 7.71$), so, data size is *moderately significant* in predicting number of processed requests.

■

5 Multivariate Regression

Another thing that may improve a regression model is to (cautiously!) take into consideration more predictors, while still keeping the function linear.

In Example 1.3 in Lecture 8 (about house prices), we discussed predicting price of a house based on its area. We decided that perhaps this prediction is not very accurate due to a high variability among house prices. What is the source of this variability? Why are houses of the same size priced differently? Certainly, area is not the only important parameter of a house. Prices are different due to different design, location, number of rooms and bathrooms, presence of a basement, a garage, a swimming pool, different size of a backyard, etc. When we take all this information into account, we'll have a rather accurate description of a house and hopefully, a rather accurate prediction of its price.

5.1 Multiple Linear Regression - Preliminaries

Now we introduce multiple linear regression that will connect a response Y with several predictors $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$, through the conditional expectation

$$E(Y \mid X^{(1)} = x^{(1)}, \ldots, X^{(k)} = x^{(k)}). \tag{5.1}$$

A **multivariate linear regression** model assumes that the curve of regression of the response Y is of the form

$$\widehat{y} \;=\; \widehat{G}\left(x^{(1)}, \ldots, x^{(k)}; \beta_0, \ldots, \beta_k\right) \;=\; \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)}, \tag{5.2}$$

a linear function of predictors $x^{(1)}, \ldots, x^{(k)}$. Here, the coefficient β_0 is called the **intercept**, while the coefficients $\beta_1, \ldots, \beta_k$ are called **slopes**.

In order to estimate all the parameters of model (5.2), we collect a sample of n *multivariate*

observations

$$\begin{cases} \mathbf{X}_1 &= \left(X_1^{(1)}, X_1^{(2)} \ldots, X_1^{(k)}\right) \\ \mathbf{X}_2 &= \left(X_2^{(1)}, X_2^{(2)} \ldots, X_2^{(k)}\right) \\ \vdots \quad & \vdots \qquad\qquad \vdots \\ \mathbf{X}_n &= \left(X_n^{(1)}, X_n^{(2)} \ldots, X_n^{(k)}\right) \end{cases}.$$

Essentially, we collect a sample of n units (say, houses) and measure all k predictors on each unit (area, number of rooms, etc.). Also, we measure responses, $Y_1, \ldots, Y_n$. We then estimate $\beta_0, \beta_1, \ldots, \beta_k$ by the method of least squares, generalizing it from the univariate case to multivariate regression. So, we minimize the sum of squared errors

$$S \;=\; \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \;=\; \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x^{(1)} - \ldots - \beta_k x^{(k)}\right)^2.$$

To make the writing easier, we put everything in vector-matrix form. We make the following notations for the response vector $\mathbf{Y}$ and the predictor matrix $\mathbf{X}$:

$$\mathbf{Y} \;=\; \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \;\; \mathbf{X} \;=\; \begin{pmatrix} 1 & \mathbf{X}_1 \\ \vdots & \vdots \\ 1 & \mathbf{X}_n \end{pmatrix} \;=\; \begin{pmatrix} 1 & X_1^{(1)} & \ldots & X_1^{(k)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \ldots & X_n^{(k)} \end{pmatrix}$$

Notice that we augmented the predictor matrix with a column of 1's because now the multivariate regression model (5.2) can be written in matrix form as

$$\begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} \;=\; \begin{pmatrix} 1 & X_1^{(1)} & \ldots & X_1^{(k)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \ldots & X_n^{(k)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

If we denote by

$$\boldsymbol{\beta} \;=\; \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

then the multidimensional parameter $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ includes the intercept and all the slopes. In fact, the intercept β_0 can also be treated as one of the slopes that corresponds to the added column of 1's.

Let

$$\widehat{\mathbf{y}} \;=\; \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix}.$$

Then the fitted values will be computed as

$$\widehat{\mathbf{y}} \;=\; \mathbf{X}\boldsymbol{\beta}$$

and the least squares problem reduces to minimizing

$$S(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 \;=\; (\mathbf{Y}-\widehat{\mathbf{y}})^T(\mathbf{Y}-\widehat{\mathbf{y}}) \;=\; (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

The minimum of the function above is attained at the **estimated slopes in multivariate regression**

$$\mathbf{b} \;=\; \widehat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{5.3}$$

Remark 5.1.

1. All the estimated slopes are linear functions of observed responses $(y_1, \ldots, y_n)$.
2. All the estimated slopes are Normally distributed, if the response variable Y is Normal.
3. The vector of slopes in (5.3) satisfies the condition

$$E(\mathbf{b}) \;=\; \boldsymbol{\beta},$$

which makes $\mathbf{b}$ an *unbiased* estimator for $\boldsymbol{\beta}$.

Example 5.2. Consider Example 4.4 again. The computer manager tries to improve the model by adding another predictor. She decides that in addition to the size of data sets, efficiency of the program may depend on the database structure. In particular, it may be important to know how many tables were used to arrange each data set. Putting all this information together, she has the following data:

Data size (gigabytes), x_1	6	7	7	8	10	10	15
Number of tables, x_2	4	20	20	10	10	2	1
Processed requests, y	40	55	50	41	17	26	16

Find the equation of the curve of regression and use it to predict the number of requests processed

per hour y^*, for $x_1^* = 16$ gigabytes of data and $x_2^* = 2$ tables.

Solution. For bivariate linear regression, the predictor matrix and the response vector are

$$\mathbf{X} = \begin{pmatrix} 1 & 6 & 4 \\ 1 & 7 & 20 \\ 1 & 7 & 20 \\ 1 & 8 & 10 \\ 1 & 10 & 10 \\ 1 & 10 & 2 \\ 1 & 15 & 1 \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} 40 \\ 55 \\ 50 \\ 41 \\ 17 \\ 26 \\ 16 \end{pmatrix}.$$

We have

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 7 & 63 & 67 \\ 63 & 623 & 519 \\ 67 & 519 & 1021 \end{pmatrix}, \ (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 3.69 & -0.3 & -0.09 \\ -0.3 & 0.03 & 0.006 \\ -0.09 & 0.006 & 0.004 \end{pmatrix} \text{ and } \mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 245 \\ 1973 \\ 2098 \end{pmatrix}.$$

From (5.3), we get

$$\mathbf{b} \ = \ \widehat{\boldsymbol{\beta}} \ = \ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \ = \ \begin{pmatrix} 52.7 \\ -2.87 \\ 0.85 \end{pmatrix}.$$

Thus, the regression equation is now

$$\widehat{y} \ = \ 52.7 - 2.87x_1 + 0.85x_2,$$

$$\begin{pmatrix} \text{number of} \\ \text{requests} \end{pmatrix} \ = \ 52.7 - 2.87 \begin{pmatrix} \text{size of} \\ \text{data} \end{pmatrix} + 0.85 \begin{pmatrix} \text{number of} \\ \text{tables} \end{pmatrix}.$$

With this new model, the predicted value y^* is

$$y^* \ = \ 52.7 - 2.87 \cdot 16 + 0.85 \cdot 2 \ = \ 8.48 \ \text{ requests processed per hour.}$$

■

Remark 5.3. One could also find a multivariate regression function that is not linear, but polynomial (of higher degree), exponential, logarithmic, etc. When using multivariate regression, for accurate estimation and efficient prediction, it is important to select the right subset of predictors.