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 9 (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.

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

$$\widehat{y} \;=\; \widehat{G}\left(x^{(1)}, \ldots, x^{(k)}; \boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_k\right) \;=\; \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x^{(1)} + \cdots + \boldsymbol{\beta}_k x^{(k)}, \tag{5.2}$$

a linear function of predictors $x^{(1)}, \ldots, x^{(k)}$. Here, the coefficient $\boldsymbol{\beta}_0$ is called the *intercept*, while the coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ are called *slopes*.

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

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

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 $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ by the method of least squares, generalizing it from the univariate case to multivariate regression. So, we minimize the sum of squared errors

$$S \;=\; \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \;=\; \sum_{i=1}^{n}\left(y_i - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_1 x^{(1)} - \cdots - \boldsymbol{\beta}_k x^{(k)}\right)^2 .$$

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 $\mathbf{X}$:

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

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

$$\begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} \;=\; \begin{pmatrix} 1 & X_1^{(1)} & \dots & X_1^{(k)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \dots & X_n^{(k)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{pmatrix}.$$

If we denote by

$$\widehat{\mathbf{y}} \;=\; \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} \text{ and } \boldsymbol{\beta} \;=\; \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{pmatrix},$$

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

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

Example 5.1. 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) (Univariate linear regression) Find the equation of the regression line. Suppose we need to start processing requests that refer to $x^* = 16$ gigabytes of data. To analyze the program efficiency, use univariate linear regression to predict y^*, the number of requests processed within 1 hour.

b) (Multivariate linear regression) 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.

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

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

b) 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

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

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

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

■

Remark 5.2. 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.

6 ANOVA and R-square

6.1 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".

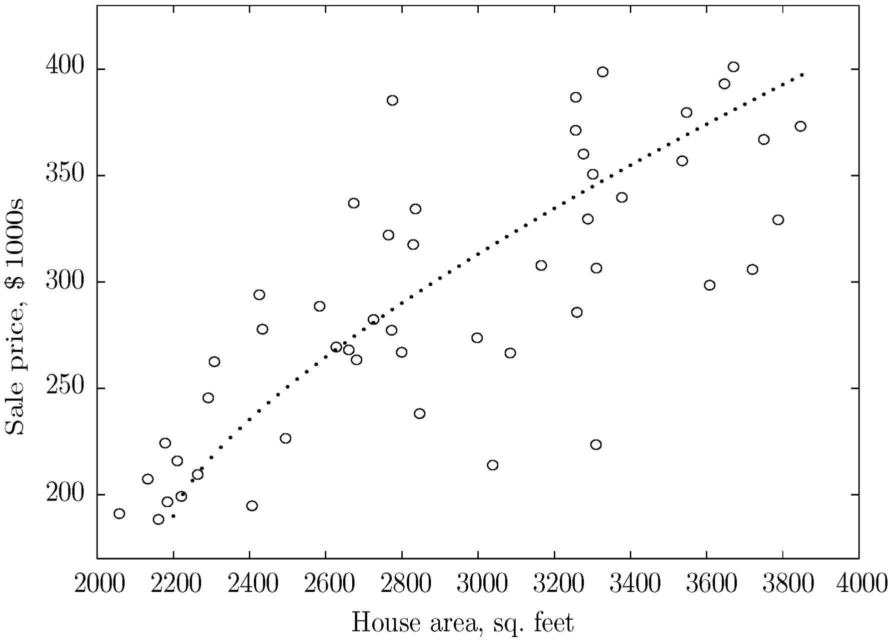

Fig. 1: House prices

Let us recall Example 1.3 in Lecture 9. We see on Figure 1 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 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 6.1. ***R-square****, or* ***coefficient of determination*** *is the proportion of the total variation explained by the model,*

$$R^2 \;=\; \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} \;=\; 1 - \frac{SS_{\text{ERR}}}{SS_{\text{TOT}}}. \tag{6.1}$$

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

$$R^2 \;=\; \frac{SS_{\mathrm{REG}}}{SS_{\mathrm{TOT}}} \;=\; \frac{b_1^2(n-1)s_x^2}{(n-1)s_y^2} \;=\; \left(b_1\frac{s_x}{s_y}\right)^2 \;=\; \left(\overline{\rho}\frac{s_y}{s_x}\frac{s_x}{s_y}\right)^2 \;=\; \overline{\rho}^2. \tag{6.2}$$

Example 6.2 (World Population, Continued)**.** Let us recall Example 1.1 (Lecture 9). 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

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

Then

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

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.

6.2 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 $Y_1, \ldots, Y_n$ 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

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

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

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

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

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

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

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

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

$$\begin{array}{ll} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0. \end{array} \tag{6.3}$$

There are several ways of testing the hypotheses (6.3). One of the most popular method of testing significance of a model 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

Example 6.3. Recall Example 5.1 about the efficiency of a new computer program (number of processed requests per hour, Y, for data sets of different sizes, X). Let us find the ANOVA table and discuss it.

Solution. We already computed

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

Further, we compute

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

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, but significant at the 0.05 level, so, data size is *moderately significant* in predicting number of processed requests.

■

6.3 Multivariate ANOVA and F-test

Next, we consider multiple linear regression that will connect a response Y with several predictors $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$. A **multivariate linear regression** model assumes that the curve of regression of the response Y is of the form

$$\widehat{y} = f\left(x^{(1)}, \ldots, x^{(k)}; \beta_0, \ldots, \beta_k\right) = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_k x^{(k)}, \tag{6.4}$$

We can again partition the *total sum of squares* measuring the total variation of responses into the *regression sum of squares* and the *error sum of squares*. The total sum of squares is still

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

with $\mathrm{df}_{\mathrm{TOT}} = n - 1$ degrees of freedom, where we denote by

$$\overline{\mathbf{y}} = \begin{pmatrix} \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix} = \overline{y} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Again, $SS_{\mathrm{TOT}} = SS_{\mathrm{REG}} + SS_{\mathrm{ERR}}$, where

$$SS_{\mathrm{REG}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = (\widehat{\mathbf{y}} - \overline{\mathbf{y}})^T (\widehat{\mathbf{y}} - \overline{\mathbf{y}})$$

is the **regression sum of squares** and

$$SS_{\mathrm{ERR}} = \sum_{i=1}^{n} (y_i - \hat{y})^2 = (\mathbf{y} - \widehat{\mathbf{y}})^T (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{e}^T \mathbf{e}$$

is the **error sum of squares**, the quantity that we minimized when we applied the method of least squares ($\mathbf{e}$ is the vector of residuals). The multivariate regression model defines a k-dimensional regression plane where the fitted values belong to. Therefore, the regression sum of squares has

$$\mathrm{df}_{\mathrm{REG}} = k$$

degrees of freedom, whereas by subtraction,

$$\mathrm{df}_{\mathrm{ERR}} = \mathrm{df}_{\mathrm{TOT}} - \mathrm{df}_{\mathrm{REG}} = n - k - 1$$

degrees of freedom are left for SS_{ERR}. This is the sample size n minus k estimated slopes and 1 estimated intercept.

For multivariate regression, we can then write the ANOVA Table 2.

The **coefficient of determination**

$$R^2 = \frac{SS_{\mathrm{REG}}}{SS_{\mathrm{TOT}}}$$

again measures the proportion of the total variation explained by regression. When we add new predictors to our model, we explain additional portions of SS_{TOT}. Therefore, R^2 can only go up. Thus, we should expect to increase R^2 and generally, get a better fit by going from univariate to multivariate regression.

	Sum	Degrees	Mean	
Source	of squares	of freedom	Squares	F
	SS	df	$MS = SS/\text{df}$	
Model	$SS_{\text{REG}} = (\widehat{\mathbf{y}} - \overline{\mathbf{y}})^T(\widehat{\mathbf{y}} - \overline{\mathbf{y}})$	k	$MS_{\text{REG}} = \dfrac{SS_{\text{REG}}}{k}$	$\dfrac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	$SS_{\text{ERR}} = (\mathbf{y} - \widehat{\mathbf{y}})^T(\mathbf{y} - \widehat{\mathbf{y}})$	$n-k-1$	$MS_{\text{ERR}} = \dfrac{SS_{\text{ERR}}}{n-k-1}$	
Total	$SS_{\text{TOT}} = (\mathbf{y} - \overline{\mathbf{y}})^T(\mathbf{y} - \overline{\mathbf{y}})$	$n-1$		

Table 2: Multivariate ANOVA

The **regression variance** $\sigma^2 = \text{Var}(Y)$ is then estimated by the mean squared error

$$s^2 \;=\; \frac{SS_{\text{REG}}}{n-k-1}.$$

It is an unbiased estimator of σ^2 that can be used in further inference.

The **ANOVA F-test** in multivariate regression tests significance of the entire model. The model is significant as long as at least one slope is not zero. Thus, we are testing

$$\begin{array}{l} H_0 : \beta_1 = \cdots = \beta_k = 0 \\ H_1 : \text{at least one } \beta_j \neq 0. \end{array}$$

We compute the F-statistic

$$F \;=\; \frac{MS_{\text{REG}}}{MS_{\text{ERR}}} \;=\; \frac{SS_{\text{REG}}/k}{SS_{\text{ERR}}/(n-k-1)}$$

and check it against the F-distribution with k and $(n-k-1)$ degrees of freedom. Again, this is always a right-tailed test. Only large values of F correspond to large SS_{REG} indicating that fitted values $\widehat{y}_i$ are far from the overall mean $\overline{y}$, and therefore, the expected response really changes along the regression plane according to predictors.

With the usual notations,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = \widehat{\boldsymbol{\beta}},$$

recall that the least squares estimate of $\boldsymbol{\beta}$ is given by

$$\mathbf{b} = \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

For a given vector of predictors $\mathbf{X}_* = (X_*^{(1)} = x_*^{(1)}, \ldots, X_*^{(k)} = x_*^{(k)})$, we estimate the expected response by

$$\hat{\mathbf{y}}_* = \mathbf{x}_*\mathbf{b}.$$

Example 6.4. Let us revisit Example 5.1 about the efficiency of a new computer program, where to predict the response Y, the number of processed requests per hour, we consider two predictors, $X^{(1)}$, the data size and $X^{(2)}$, the number of tables. Construct the multivariate ANOVA table.

Solution. The total sum of squares is still

$$SS_{\text{TOT}} = S_{yy} = 1452.$$

It is *the same* for all the models with this response.

Recall the predictor matrix and the response vector

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

for which

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

So, we obtained

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

From here, we can now compute a vector of *fitted values*

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{pmatrix} 38.9 \\ 49.6 \\ 49.6 \\ 38.2 \\ 32.5 \\ 25.7 \\ 10.5 \end{pmatrix}.$$

Then we get (we have already computed $\overline{y} = 35$)

$$SS_{\text{REG}} = (\widehat{\mathbf{y}} - \overline{\mathbf{y}})^T(\widehat{\mathbf{y}} - \overline{\mathbf{y}}) = 1143.3 \text{ and } SS_{\text{ERR}} = (\mathbf{y} - \widehat{\mathbf{y}})^T(\mathbf{y} - \widehat{\mathbf{y}}) = 308.7.$$

We have now 2 degrees of freedom for the model because we now use two predictor variables. The ANOVA table is then completed as

Source	Sum of squares	Degrees of freedom	Mean Squares	F
Model	1143.3	2	571.7	7.41
Error	308.7	4	77.2	
Total	1452	6		

The regression variance σ^2 is now estimated by

$$s^2 = MS_{\text{ERR}} = 77.2.$$

R-square is now

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = 0.787 \text{ or } 78.7\%,$$

which is 12.5% higher than in Example 6.3. These additional 12.5% of the total variation are explained by the new predictor x_2 that is used in the model in addition to x_1. R-square can *only increase* when new variables are added.
The ANOVA F-test statistic is now of 7.41 with 2 and 4 degrees of freedom. It shows that the model is significant at the level of 0.05, but not at the level of 0.025 (as before).

■

6.4 Adjusted R-square

Multivariate regression opens an almost unlimited opportunity for us to improve prediction by adding more and more X-variables into our model. On the other hand, we mentioned the fact that overfitting a model leads to a low prediction power. Moreover, it will often result in large variances $\sigma^2(b_j)$ and therefore, unstable regression estimates. Then, how can we build a model with the right, optimal set of predictors $X^{(j)}$ that will give us a good, accurate fit? One way is to consider the *adjusted R-square criterion*.

It can be shown mathematically that R^2, the coefficient of determination, can only increase when we add predictors to the regression model. No matter how irrelevant it is for the response Y, any new predictor can only increase the proportion of explained variation. Therefore, R^2 is not a fair criterion when we compare models with different numbers of predictors k. Including irrelevant predictors should be penalized whereas R^2 can only reward for this. A fair measure of goodness-of-fit is the *adjusted R-square*.

Definition 6.5. ***Adjusted R-square*** *is the quantity*

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{ERR}}/(n-k-1)}{SS_{\text{TOT}}/(n-1)} = 1 - \frac{SS_{\text{ERR}}/\text{df}_{\text{ERR}}}{SS_{\text{TOT}}/\text{df}_{\text{TOT}}}. \qquad (6.5)$$

The adjusted R-square is a criterion of variable selection that rewards for adding a predictor *only* if it considerably reduces the error sum of squares. Comparing it to

$$R^2 = \frac{SS_{\mathrm{REG}}}{SS_{\mathrm{TOT}}} = \frac{SS_{\mathrm{TOT}} - SS_{\mathrm{ERR}}}{SS_{\mathrm{TOT}}} = 1 - \frac{SS_{\mathrm{ERR}}}{SS_{\mathrm{TOT}}},$$

adjusted R-square includes degrees of freedom into its formula. This adjustment may result in a penalty when a useless X-variable is added to the regression mode. Let us explain that. Indeed, if we add a non-significant predictor, the number of estimated slopes k will increase by 1. However, if this variable is not able to explain any variation of the response, the sums of squares, SS_{REG} and SS_{ERR}, will remain the same. Then, $SS_{\mathrm{ERR}}/(n-k-1)$ will increase and R^2_{adj} will decrease, penalizing us for including such a poor predictor.

So, we choose a model with the *highest* adjusted R-square.

Example 6.6. For our previous example, the adjusted R-square for the model with one predictor x_1 (data size) is

$$R^2_{1,\mathrm{adj}} = 1 - \frac{SS_{\mathrm{ERR}}/(n-k-1)}{SS_{\mathrm{TOT}}/(n-1)} = 1 - \frac{490.86/5}{1452/6} = 0.5943.$$

When an extra predictor was added, x_2 (number of tables), we get an adjusted R-square of

$$R^2_{2,\mathrm{adj}} = 1 - \frac{SS_{\mathrm{ERR}}/(n-k-1)}{SS_{\mathrm{TOT}}/(n-1)} = 1 - \frac{308.7/4}{1452/6} = 0.6811.$$

By adding another predictor, we increased the adjusted R-square with 8.68%. That shows that the number of tables is a significant predictor for the number of processed requests and that it improved the model.

6.5 Categorical predictors and dummy variables

Careful model selection is one of the most important steps in practical statistics. In regression, only a wisely chosen subset of predictors delivers accurate estimates and good prediction. At the same time, any useful information should be incorporated into our model. We conclude this chapter with a note on using *categorical* (non-numerical) predictors in regression modeling.

Often a good portion of the variation of response Y can be explained by *attributes* rather than numbers. Examples are

- computer manufacturer (Dell, IBM, Hewlett Packard, Apple, etc.);
- operating system (Unix, Windows, DOS, etc.);
- color (white, blue, red, green, etc.).

Unlike numerical predictors, attributes have no particular order. For example, it is totally *wrong* to code operating systems with numbers ($1 =$ Unix, $2 =$Windows, $3 =$ DOS), create a new predictor $X^{(k+1)}$, and include it into the regression model. If we do so, it puts Windows right in the middle between Unix and DOS and tells that changing an operating system from Unix to Windows has exactly the same effect on the response Y as changing it from Windows to DOS!

However, performance of a computer really depends on the operating system, manufacturer, type of the processor and other categorical variables. How can we use them in our regression model? We need to create so-called **dummy variables**. A dummy variable is binary, taking values 0 or 1,

$$Z_i^{(j)} = \begin{cases} 1, & \text{if unit } i \text{ in the sample has category } j \\ 0, & \text{otherwise} \end{cases}$$

For a categorical variable with C categories, we create $(C-1)$ dummy predictors, $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(C-1)}$. They carry the entire information about the attribute. Sampled items from category C will be marked by all $(C-1)$ dummies equal to 0.
Notice that if we make the mistake of creating C dummies for an attribute with C categories (one dummy per category), this would cause a linear relation

$$\mathbf{Z}^{(1)} + \cdots + \mathbf{Z}^{(C)} = 1.$$

A column of 1's is already included into the predictor matrix $\mathbf{X}$, and therefore, such a linear relation will cause singularity of $(\mathbf{X}^T\mathbf{X})$ when we compute the least squares estimates $b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Thus, it is necessary and sufficient to have only $(C-1)$ dummy variables.

Fitting the model, all dummy variables are included into the predictor matrix $\mathbf{X}$ as columns.

Example 6.7. Consider the program efficiency study in Example 6.4. The computer manager makes another attempt to improve the prediction power. This time she would like to consider the fact that the first four times the program worked under the operational system A and then switched to the operational system B. Introduce a dummy variable responsible for the operational system and include it into the regression analysis.

Data size (gigabytes), x_1	6	7	7	8	10	10	15
Number of tables, x_2	4	20	20	10	10	2	1
Operational system, x_3	A	A	A	A	B	B	B
Processed requests, y	40	55	50	41	17	26	16

Estimate the new regression equation. Does the new variable improve the goodness of fit?

Solution. Let $z_i = 1$ for the operational system A and $z_i = 0$ for the operational system B. With the addition of this dummy variable, the predictor matrix and the response vector are

$$\mathbf{X} = \begin{pmatrix} 1 & 6 & 4 & 1 \\ 1 & 7 & 20 & 1 \\ 1 & 7 & 20 & 1 \\ 1 & 8 & 10 & 1 \\ 1 & 10 & 10 & 0 \\ 1 & 10 & 2 & 0 \\ 1 & 15 & 1 & 0 \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} 40 \\ 55 \\ 50 \\ 41 \\ 17 \\ 26 \\ 16 \end{pmatrix}.$$

The vector of regression slopes is then

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 24.20 \\ -0.60 \\ 0.57 \\ 18.78 \end{pmatrix}.$$

Then, the estimated regression equation is now

$$\hat{y} = 24.20 - 0.60x^{(1)} + 0.57x^{(2)} + 18.78z.$$

The new adjusted R-square is now

$$R^2_{\text{adj}} = 0.8260,$$

which is higher than the previous adjusted R-square with 14.49%. That shows that including the operating system among predictors improved the goodness of fit.

The ANOVA F-test statistic is now of 10.49 with 3 and 3 degrees of freedom. It shows that the model is significant at the level of 0.05, but not at the level of 0.025.

■

7 Significant Correlation

We briefly mention here just one more procedure, testing if two sets of data (the response and one predictor) are linearly correlated or not. That means, we test the hypotheses

$$\begin{aligned} H_0 &: \quad \rho = 0 \\ H_1 &: \quad \rho \neq 0, \end{aligned}$$

a two-tailed test for the correlation coefficient.

We compute the (absolute value of the) sample correlation coefficient $|\overline{\rho}|$ and compare its value to the ones in the Pearson Table of critical values, with $\mathrm{df} = n - 2$. If the absolute value of the calculated Pearson's correlation coefficient is greater than the critical value from the table, then we reject the null hypothesis that there is no correlation, i.e. we conclude that there is *significant* correlation. How "significant"? We have the same levels as before.

$$\begin{aligned} & & |\overline{\rho}| &< \rho_{0.05} &\Rightarrow \quad & \textbf{not} \text{ significant}, \\ \rho_{0.05} &\le & |\overline{\rho}| &< \rho_{0.01} &\Rightarrow \quad & \textbf{(moderately) significant}, \\ \rho_{0.01} &\le & |\overline{\rho}| &< \rho_{0.001} &\Rightarrow \quad & \textbf{distinctly} \text{ significant}, \\ & & |\overline{\rho}| &\ge \rho_{0.001} &\Rightarrow \quad & \textbf{very} \text{ significant}. \end{aligned}$$

In Example 6.2 about the world population, we found a correlation coefficient of $\overline{\rho} = 0.9972$ between the predictor "year" and the response "world population", for a sample of size $n = 15$. We see from the table that with $\mathrm{df} = 13$, this $\overline{\rho}$ is *very* significant, being larger than $\rho_{0.001} = 0.76$.

In Example 6.3, the correlation coefficient between predictor "data size" and response "number of processed requests" is $\overline{\rho} = -0.81$ for a sample of size $n = 7$. For $\mathrm{df} = 5$, we find from the table that

$$\rho_{0.05} = 0.75 < |\overline{\rho}| = 0.81 < \rho_{0.01} = 0.87,$$

so, this is a *moderately significant* correlation.