Chapter 11

Regression

In Chapters 8, 9, and 10, we were concerned about the distribution of *one random variable*, its parameters, expectation, variance, median, symmetry, skewness, etc. In this chapter, we study *relations* among variables.

Many variables observed in real life are related. The type of their relation can often be expressed in a mathematical form called *regression*. 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.

11.1 Least squares estimation

Regression models relate a *response variable* to one or several predictors. Having observed predictors, we can forecast the response by computing its *conditional expectation*, given all the available predictors.

DEFINITION 11.1

Response or *dependent variable* Y is a variable of interest that we predict based on one or several predictors.

Predictors or *independent variables* $X^{(1)}, \dots, X^{(k)}$ are used to predict the values and behavior of the response variable Y.

Regression of Y on $X^{(1)}, \dots, X^{(k)}$ is the conditional expectation,

$$G(x^{(1)}, \dots, x^{(k)}) = \mathbf{E}\left\{Y \mid X^{(1)} = x^{(1)}, \dots, X^{(k)} = x^{(k)}\right\}.$$

It is a function of $x^{(1)}, \dots, x^{(k)}$ whose form can be estimated from data.

11.1.1 Examples

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

Example 11.1 (World population). According to the International Data Base of the *U.S. Census Bureau*, population of the world grows according to Table 11.1. How can we use these data to predict the world population in years 2015 and 2020?

Figure 11.1 shows that the population (response) is tightly related to the year (predictor),

$$\text{population} \approx G(\text{year}).$$

It increases every year, and its growth is almost linear. If we estimate the *regression function* G relating our response and our predictor (see the dotted line on Figure 11.1) and extend

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287	2015	?
1970	3712	1995	5700	2020	?

TABLE 11.1: Population of the world, 1950–2020.

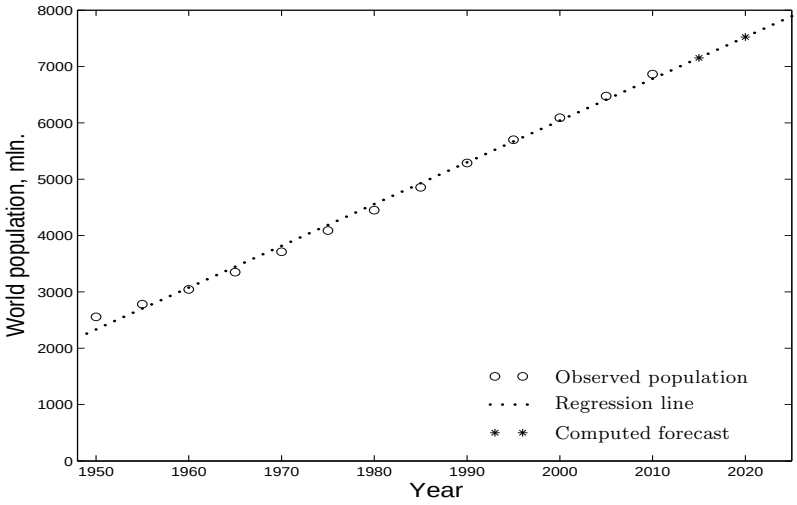

FIGURE 11.1: *World population in 1950–2010 and its regression forecast for 2015 and 2020.*

its graph to the year 2020, the forecast is ready. We can simply compute $G(2015)$ and and $G(2020)$.

A straight line that fits the observed data for years 1950–2010 predicts the population of 7.15 billion in 2015 and 7.52 billion in 2020. It also shows that between 2010 and 2015, around the year 2012, the world population reaches the historical mark of 7 billion. ◊

How accurate is the forecast obtained in this example? The observed population during 1950–2010 appears rather close to the estimated regression line in Figure 11.1. It is reasonable to hope that it will continue to do so through 2020.

The situation is different in the next example.

Example 11.2 (HOUSE PRICES). Seventy house sale prices in a certain county are depicted in Figure 11.2 along with the house area.

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 11.2) and plug 3200 into the resulting formula, but due to obviously high variability, our estimation will not be as accurate as in Example 11.1. ◊

To improve our estimation in the last example, we may take other factors into account: 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. Regression models with multiple predictors are studied in Section 11.3.

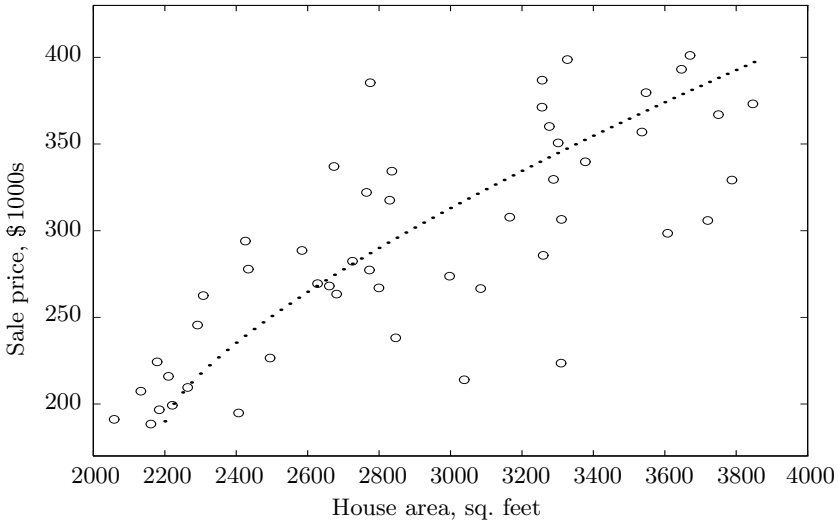

FIGURE 11.2: *House sale prices and their footage.*

11.1.2 Method of least squares

Our immediate goal is to estimate the **regression function** G that connects response variable Y with predictors $X^{(1)}, \ldots, X^{(k)}$. First we focus on *univariate regression* predicting response Y based on one predictor X. The method will be extended to k predictors in Section 11.3.

In univariate regression, we observe *pairs* $(x_1, y_1), \ldots, (x_n, y_n)$, shown in Figure 11.3a.

For accurate forecasting, we are looking for the 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 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 11.3b). Method of least squares minimizes the sum of squared distances.

DEFINITION 11.2

Residuals

$$e_i = y_i - \widehat{y}_i$$

are differences between observed responses y_i and their **fitted values** $\widehat{y}_i = \widehat{G}(x_i)$.

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

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \,. \tag{11.1}$$

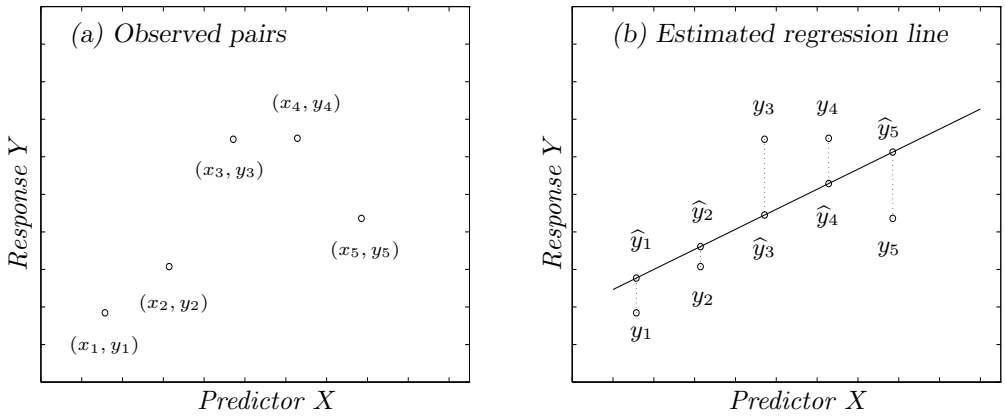

FIGURE 11.3: *Least squares estimation of the regression line.*

Function $\widehat{G}$ is usually sought in a suitable form: linear, quadratic, logarithmic, etc. The simplest form is linear.

11.1.3 Linear regression

Linear regression model assumes that the conditional expectation

$$G(x) = \mathbf{E}\{Y \mid X = x\} = \beta_0 + \beta_1 x$$

is a *linear function* of x. As any linear function, it has an intercept β_0 and a slope β_1.

The **intercept**

$$\beta_0 = G(0)$$

equals the value of the regression function for $x = 0$. Sometimes it has no physical meaning. For example, nobody will try to predict the value of a computer with 0 random access memory (RAM), and nobody will consider the Federal reserve rate in year 0. In other cases, intercept is quite important. For example, according to the *Ohm's Law* ($V = RI$) the voltage across an *ideal* conductor is proportional to the current. A non-zero intercept ($V = V_0 + RI$) would show that the circuit is not ideal, and there is an external loss of voltage.

The **slope**

$$\beta_1 = G(x+1) - G(x)$$

is the predicted change in the response variable when predictor changes by 1. This is a very important parameter that shows how fast we can change the expected response by varying the predictor. For example, customer satisfaction will increase by $\beta_1(\Delta x)$ when the quality of produced computers increases by (Δx).

A zero slope means absence of a linear relationship between X and Y. In this case, Y is expected to stay constant when X changes.

Estimation in linear regression

Let us estimate the slope and intercept by **method of least squares**. Following (11.1), we minimize the sum of squared residuals

$$Q = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - \widehat{G}(x_i)\Big)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

We can do it by taking partial derivatives of Q, equating them to 0, and solving the resulting equations for β_0 and β_1.

The partial derivatives are

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) ; \\ \frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i . \end{aligned}$$

Equating them to 0, we obtain so-called *normal equations*,

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

From the first normal equation,

$$\beta_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n} = \bar{y} - \beta_1 \bar{x}. \tag{11.2}$$

Substituting this into the second normal equation, we get

$$\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} x_i ((y_i - \bar{y}) - \beta_1 (x_i - \bar{x})) = S_{xy} - \beta_1 S_{xx} = 0, \tag{11.3}$$

where

$$S_{xx} = \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{11.4}$$

and

$$S_{xy} = \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{11.5}$$

are *sums of squares and cross-products*. Notice that it is all right to subtract $\bar{x}$ from x_i in the right-hand sides of (11.4) and (11.5) because $\sum (x_i - \bar{x}) = 0$ and $\sum (y_i - \bar{y}) = 0$.

Finally, we obtain the **least squares estimates** of intercept β_0 and slope β_1 from (11.2) and (11.3).

Regression estimates

$$\boxed{\begin{array}{rcl} b_0 &=& \widehat{\beta}_0 \;=\; \bar{y} - b_1\,\bar{x} \\ b_1 &=& \widehat{\beta}_1 \;=\; S_{xy}/S_{xx} \\ \multicolumn{3}{c}{\text{where}} \\ S_{xx} &=& \sum_{i=1}^{n}(x_i-\bar{x})^2 \\ S_{xy} &=& \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \end{array}} \tag{11.6}$$

Example 11.3 (World population). In Example 11.1, x_i is the year, and y_i is the world population during that year. To estimate the regression line in Figure 11.1, we compute

$$\bar{x} = 1980; \quad \bar{y} = 4558.1;$$

$$\begin{aligned} S_{xx} &= (1950-\bar{x})^2 + \ldots + (2010-\bar{x})^2 = 4550; \\ S_{xy} &= (1950-\bar{x})(2558-\bar{y}) + \ldots + (2010-\bar{x})(6864-\bar{y}) = 337250. \end{aligned}$$

Then

$$\begin{aligned} b_1 &= S_{xy}/S_{xx} = 74.1 \\ b_0 &= \bar{y} - b_1\bar{x} = -142201. \end{aligned}$$

The estimated regression line is

$$\widehat{G}(x) = b_0 + b_1\,x = \underline{\text{-142201 + 74.1x}}.$$

We conclude that the world population grows at the average rate of 74.1 million every year.

We can use the obtained equation to predict the future growth of the world population. Regression predictions for years 2015 and 2020 are

$$\begin{aligned} \widehat{G}(2015) &= b_0 + 2015\,b_1 = \underline{\text{7152 million people}} \\ \widehat{G}(2020) &= b_0 + 2020\,b_1 = \underline{\text{7523 million people}} \end{aligned}$$

$\Diamond$

11.1.4 Regression and correlation

Recall from Section 3.3.5 that **covariance**

$$\mathrm{Cov}(X,Y) = \mathbf{E}(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))$$

and **correlation coefficient**

$$\rho = \frac{\mathrm{Cov}(X,Y)}{(\mathrm{Std}X)(\mathrm{Std}Y)}$$

measure the direction and strength of a linear relationship between variables X and Y. From observed data, we estimate $\text{Cov}(X, Y)$ and ρ by the **sample covariance**

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

(it is unbiased for the population covariance) and the **sample correlation coefficient**

$$r = \frac{s_{xy}}{s_x s_y}, \tag{11.7}$$

where

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \text{ and } s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

are sample standard deviations of X and Y.

Comparing (11.3) and (11.7), we see that the estimated slope b_1 and the sample regression coefficient r are proportional to each other. Now we have two new formulas for the regression slope.

Estimated regression slope

$$\boxed{b_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2} = r\left(\frac{s_y}{s_x}\right)}$$

Like the correlation coefficient, regression slope is positive for positively correlated X and Y and negative for negatively correlated X and Y. The difference is that r is dimensionless whereas the slope is measured in units of Y per units of X. Thus, its value by itself does not indicate whether the dependence is weak or strong. It depends on the units, the scale of X and Y. We test significance of the regression slope in Section 11.2.

11.1.5 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 11.3b, we did have some residuals $e_i = (y_i - \widehat{y}_i)$ and some positive sum of squared residuals. The straight line has not accounted for all 100% of variation among y_i.

Why, one might ask, have we considered only linear models? As long as all x_i are different, we can always find a regression function $\widehat{G}(x)$ that passes through all the observed points without any error. Then, the sum $\sum e_i^2 = 0$ 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 11.4a,b).

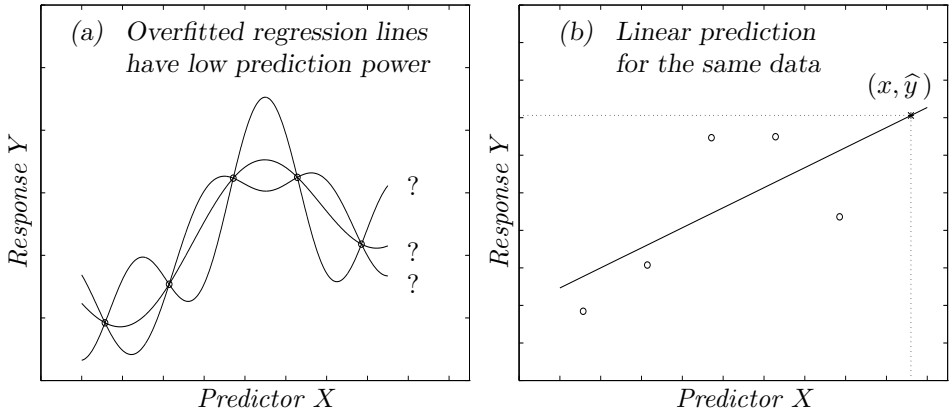

FIGURE 11.4: *Regression-based prediction.*

11.2 Analysis of variance, prediction, and further inference

In this section, we

- evaluate the *goodness of fit* of the chosen regression model to the observed data,
- estimate the response variance,
- test significance of regression parameters,
- construct confidence and prediction intervals.

11.2.1 ANOVA and R-square

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."

For example, there exists some variation among the house sale prices on Figure 11.2. Why are the houses priced differently? Well, 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**

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

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

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

This is the portion of total variation *explained by the model*. It is often computed as

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

The rest of total variation is attributed to "error." It is measured by the **error sum of squares**

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

This is the portion of total variation *not explained by the model*. It equals the sum of squared residuals that the method of least squares minimizes. Thus, applying this method, we minimize the *error sum of squares*.

Regression and error sums of squares partition SS_{TOT} into two parts (Exercise 11.6),

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

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 11.3

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

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

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 (Exercise 11.7),

$$R^2 = r^2.$$

Example 11.4 (World population, continued). Continuing Example 11.3, we find

$$\begin{aligned} SS_{\text{TOT}} &= (n-1)s_y^2 = (12)(2.093 \cdot 10^6) = 2.512 \cdot 10^7, \\ SS_{\text{REG}} &= b_1^2 S_{xx} = (74.1)^2 (4550) = 2.500 \cdot 10^7, \\ SS_{\text{ERR}} &= SS_{\text{TOT}} - SS_{\text{REG}} = 1.2 \cdot 10^5. \end{aligned}$$

A linear model for the growth of the world population has a very high R-square of

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = \underline{0.995} \text{ or } \underline{99.5\%}.$$

This is a very good fit although some portion of the remaining 0.5% of total variation can still be explained by adding non-linear terms into the model. ◊

11.2.2 Tests and confidence intervals

Methods of estimating a regression line and partitioning the total variation do not rely on any distribution; thus, we can apply them to virtually any data.

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

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 x_i$$

and constant variance σ^2. Predictors x_i are considered *non-random*.

As a consequence, regression estimates b_0 and b_1 have Normal distribution. After we estimate the variance σ^2, they can be studied by T-tests and T-intervals.

Degrees of freedom and variance estimation

According to the standard assumptions, responses $Y_1, \dots, Y_n$ have different means but the same variance. This variance equals the mean squared deviation of responses from their respective expectations. Let us estimate it.

First, we estimate each expectation $\mathbf{E}(Y_i) = G(x_i)$ 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*

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

It remains to divide this sum by its number of degrees of freedom (this is how we estimated variances in Section 8.2.4).

Let us compute degrees of freedom for all three SS in the regression ANOVA.

The total sum of squares $SS_{\text{TOT}} = (n-1)s_y^2$ has $\text{df}_{\text{TOT}} = n-1$ degrees of freedom because it is computed directly from the sample variance s_y^2.

Out of them, the regression sum of squares SS_{REG} has $\text{df}_{\text{REG}} = 1$ degree of freedom. Recall (from Section 9.3.4, p. 261) that the number of degrees of is the dimension of the corresponding space. Regression line, which is just a straight line, has dimension 1.

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

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

The error degrees of freedom also follow from formula (9.10),

$$\text{df}_{\text{ERR}} = \text{ sample size } - \begin{array}{c}\text{number of estimated}\\ \text{location parameters}\end{array} = n - 2,$$

with 2 degrees of freedom deducted for 2 estimated parameters, β_0 and β_1.

Equipped with this, we now estimate the variance.

Regression variance

$$\boxed{s^2 = \frac{SS_{\text{ERR}}}{n-2}}$$

It estimates $\sigma^2 = \text{Var}(Y)$ unbiasedly.

Remark: Notice that the usual sample variance

$$s_y^2 = \frac{SS_{\text{TOT}}}{n-1} = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

is biased because $\bar{y}$ no longer estimates the expectation of Y_i.

A standard way to present analysis of variance is the *ANOVA table*.

Univariate ANOVA

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	$SS_{\text{REG}} = \sum(\hat{y}_i - \bar{y})^2$	1	$MS_{\text{REG}} = SS_{\text{REG}}$	$\frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	$SS_{\text{ERR}} = \sum(y_i - \hat{y}_i)^2$	$n-2$	$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{n-2}$	
Total	$SS_{\text{TOT}} = \sum(y_i - \bar{y})^2$	$n-1$		

Mean squares MS_{REG} and MS_{ERR} are obtained from the corresponding sums of squares dividing them by their degrees of freedom. We see that the sample regression variance is the mean squared error,

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

The estimated standard deviation s is usually called *root mean squared error* or *RMSE*.

The *F-ratio*

$$F = \frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$$

is used to test significance of the entire regression model.

Inference about the regression slope

Having estimated the regression variance σ^2, we can proceed with tests and confidence intervals for the regression slope β_1. As usually, we start with the estimator of β_1 and its sampling distribution.

The slope is estimated by

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} = \frac{\sum(x_i - \bar{x})y_i}{S_{xx}}$$

(we can drop $\bar{y}$ because it is multiplied by $\sum(x_i - \bar{x}) = 0$).

According to *standard regression assumptions*, y_i are Normal random variables and x_i are non-random. Being a linear function of y_i, the estimated slope b_1 is also Normal with the expectation

$$\mathbf{E}(b_1) = \frac{\sum(x_i - \bar{x})\,\mathbf{E}(y_i)}{S_{xx}} = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} = \frac{\sum(x_i - \bar{x})^2(\beta_1)}{\sum(x_i - \bar{x})^2} = \beta_1,$$

(which shows that b_1 is an *unbiased estimator* of β_1), and the variance

$$\mathrm{Var}(b_1) = \frac{\sum(x_i - \bar{x})^2\,\mathrm{Var}(y_i)}{S_{xx}^2} = \frac{\sum(x_i - \bar{x})^2\sigma^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

Summarizing the results,

Sampling distribution of a regression slope

b_1 is Normal(μ_b, σ_b),

where

$\mu_b = \mathbf{E}(b_1) = \beta_1$

$\sigma_b = \mathrm{Std}(b_1) = \dfrac{\sigma}{\sqrt{S_{xx}}}$

We estimate the standard error of b_1 by

$$s(b_1) = \frac{s}{\sqrt{S_{xx}}},$$

and therefore, use T-intervals and T-tests.

Following the general principles, a $(1-\alpha)100\%$ **confidence interval** for the slope is

$$\text{Estimator} \pm t_{\alpha/2} \left(\begin{array}{c} \text{estimated} \\ \text{st. deviation} \\ \text{of the estimator} \end{array} \right) = b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}}.$$

Testing hypotheses $H_0:\ \beta_1 = B$ about the regression slope, use the T-statistic

$$t = \frac{b_1 - B}{s(b_1)} = \frac{b_1 - B}{s/\sqrt{S_{xx}}}.$$

P-values, acceptance and rejection regions are computed from Table A5 in the Appendix, T-distribution with $(n-2)$ degrees of freedom. These are degrees of freedom used in the estimation of σ^2.

As always, the form of the alternative hypothesis determines whether it is a two-sided, right-tail, or left-tail test.

A non-zero slope 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, $\mathbf{E}(Y) = \beta_0$ remains constant.

To see if X is significant for the prediction of Y, test the null hypothesis

$$H_0: \ \beta_1 = 0 \ \text{ vs } \ H_a: \ \beta_1 \neq 0.$$

ANOVA F-test

A more universal, and therefore, more 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 SS by the number of degrees of freedom, we obtain **mean squares**,

$$MS_{\text{REG}} = \frac{SS_{\text{REG}}}{\text{df}_{\text{REG}}} = \frac{SS_{\text{REG}}}{1} = SS_{\text{REG}}$$

and

$$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{\text{df}_{\text{ERR}}} = \frac{SS_{\text{ERR}}}{n-2} = s^2.$$

Under the null hypothesis

$$H_0: \ \beta_1 = 0,$$

both mean squares, MS_{REG} and MS_{ERR} are independent, and their ratio

$$F = \frac{MSR}{MSE} = \frac{SSR}{s^2}$$

has **F-distribution** with $\text{df}_{\text{REG}} = 1$ and $\text{df}_{\text{ERR}} = n-2$ degrees of freedom (d.f.).

As we discovered in Section 9.5.4, this F-distribution has two parameters, numerator d.f. and denominator d.f., and it is very popular for testing ratios of variances and significance of models. Its critical values for the most popular significance levels between $\alpha = 0.001$ and $\alpha = 0.25$ are tabulated in Table A7.

ANOVA F-test is always *one-sided* and *right-tail* because only large values of the F-statistic show a large portion of explained variation and the overall significance of the model.

F-test and T-test

We now have two tests for the model significance, a T-test for the regression slope and the ANOVA F-test. For the univariate regression, they are absolutely equivalent. In fact, the F-statistic equals the squared T-statistic for testing $H_0: \beta_1 = 0$ because

$$t^2 = \frac{b_1^2}{s^2/S_{xx}} = \frac{(S_{xy}/S_{xx})^2}{s^2/S_{xx}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} \frac{S_{yy}}{s^2} = \frac{r^2 SS_{\text{TOT}}}{s^2} = \frac{SS_{\text{REG}}}{s^2} = F.$$

Hence, both tests give us the same result.

Example 11.5 (Assumption of independence). Can we apply the introduced methods to Examples 11.1–11.3? For the world population data in Example 11.1, the sample correlation coefficient between residuals e_i and e_{i-1} is 0.78, which is rather high. Hence, we cannot assume independence of y_i, and one of the standard assumptions is violated.

Our least squares regression line is still correct; however, in order to proceed with tests and confidence intervals, we need more advanced *time series* methods accounting not only for the variance but also for covariances among the observed responses.

For the house prices in Example 11.2, there is no evidence of any dependence. These 70 houses are sampled at random, and they are likely to be priced independently of each other. ◊

Remark: Notice that we used residuals $e_i = y_i - \hat{y}_i$ for the correlation study. Indeed, according to our regression model, responses y_i have different expected values, so their sample mean $\bar{y}$ does not estimate the population mean of any of them; therefore, the sample correlation coefficient based on that mean is misleading. On the other hand, if the linear regression model is correct, all *residuals* have the same mean $\mathbf{E}(e_i) = 0$. In the population, the difference between y_i and ε_i is non-random, $y_i - \varepsilon_i = G(x_i)$; therefore, the population correlation coefficients between y_i and y_j and between ε_i and ε_j are the same.

Example 11.6 (Efficiency of computer programs). 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. The response variable here is the number of processed requests (y), and we attempt to predict it from the size of a data set (x).

(a) Estimation of the regression line. We can start by computing

$$n = 7,\ \bar{x} = 9,\ \bar{y} = 35,\ S_{xx} = 56,\ S_{xy} = -232,\ S_{yy} = 1452.$$

Estimate regression slope and intercept by

$$b_1 = \frac{S_{xy}}{S_{xx}} = -4.14 \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x} = 72.3.$$

Then, the estimated regression line has an equation

$$y = 72.3 - 4.14x.$$

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

(b) ANOVA TABLE AND VARIANCE ESTIMATION. Let us compute all components of the ANOVA. We have

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

partitioned into

$$SS_{\text{REG}} = b_1^2 S_{xx} = 961 \quad \text{and} \quad SS_{\text{ERR}} = SS_{\text{TOT}} - SS_{\text{REG}} = 491.$$

Simultaneously, $n-1 = 6$ degrees of freedom of SS_{TOT} are partitioned into $\text{df}_{\text{REG}} = 1$ and $\text{df}_{\text{ERR}} = 5$ degrees of freedom.

Fill the rest of the ANOVA table,

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	961	1	961	9.79
Error	491	5	98.2	
Total	1452	6		

REGRESSION VARIANCE σ^2 is estimated by

$$s^2 = MS_{\text{ERR}} = 98.2.$$

R-SQUARE is

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = \frac{961}{1452} = 0.662.$$

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

(c) INFERENCE ABOUT THE SLOPE. Is the slope statistically significant? Does the number of processed requests really depend on the size of data sets? To test the null hypothesis $H_0: \ \beta_1 = 0$, compute the T-statistic

$$t = \frac{b_1}{\sqrt{s^2/S_{xx}}} = \frac{-4.14}{\sqrt{98.2/56}} = -3.13.$$

Checking the T-distribution table (Table A5) with 5 d.f., we find that the P-value for the *two-sided* test is between 0.02 and 0.04. We conclude that the slope is *moderately significant*. Precisely, it is significant at any level $\alpha \geq 0.04$ and not significant at any $\alpha \leq 0.02$.

(d) ANOVA F-TEST. A similar result is suggested by the F-test. From Table A7, the F-statistic of 9.79 is not significant at the 0.025 level, but significant at the 0.05 level.

$\Diamond$

11.2.3 Prediction

One of the main applications of regression analysis is making forecasts, predictions of the response variable Y based on the known or controlled predictors X.

Let x_* be the value of the predictor X. The corresponding value of the response Y is computed by evaluating the estimated regression line at x_*,

$$\hat{y}_* = \widehat{G}(x_*) = b_0 + b_1 x_*.$$

This is how we predicted the world population in years 2015 and 2020 in Example 11.3. As happens with any forecast, our predicted values are understood as the most intelligent guesses, and not as guaranteed exact sizes of the population during these years.

How reliable are regression predictions, and how close are they to the real true values? As a good answer, we can construct

- a $(1-\alpha)100\%$ **confidence interval** for the expectation

$$\mu_* = \mathbf{E}(Y \mid X = x_*)$$

  and

- a $(1-\alpha)100\%$ **prediction interval** for the actual value of $Y = y_*$ when $X = x_*$.

Confidence interval for the mean of responses

The expectation

$$\mu_* = \mathbf{E}(Y \mid X = x_*) = G(x_*) = \beta_0 + \beta_1 x_*$$

is a population parameter. This is the mean response for the entire subpopulation of units where the independent variable X equals x_*. For example, it corresponds to the average price of all houses with the area $x_* = 2500$ square feet.

First, we estimate μ_* by

$$\begin{aligned}
\hat{y}_* &= b_0 + b_1 x_* \\
&= \bar{y} - b_1 \bar{x} + b_1 x_* \\
&= \bar{y} + b_1 (x_* - \bar{x}) \\
&= \frac{1}{n} \sum y_i + \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}} (x_* - \bar{x}) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{\sum (x_i - \bar{x})}{S_{xx}} (x_* - \bar{x}) \right) y_i .
\end{aligned}$$

We see again that the estimator is a linear function of responses y_i. Then, under standard regression assumptions, $\hat{y}_*$ is Normal, with expectation

$$\mathbf{E}(\hat{y}_*) = \mathbf{E} b_0 + \mathbf{E} b_1 x_* = \beta_0 + \beta_1 x_* = \mu_*$$

(it is unbiased), and variance

$$\begin{aligned}\mathrm{Var}(\hat{y}_*) &= \sum\left(\frac{1}{n}+\frac{\sum(x_i-\bar{x})}{S_{xx}}(x_*-\bar{x})\right)^2 \mathrm{Var}(y_i)\\ &= \sigma^2\left(\sum_{i=1}^{n}\frac{1}{n^2}+2\sum_{i=1}^{n}(x_i-\bar{x})\frac{x_*-\bar{x}}{S_{xx}}+\frac{S_{xx}(x_*-\bar{x})^2}{S_{xx}^2}\right)\\ &= \sigma^2\left(\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}\right) \qquad (11.8)\end{aligned}$$

(because $\sum(x_i-\bar{x})=0$).

Then, we estimate the regression variance σ^2 by s^2 and obtain the following confidence interval.

$(1-\alpha)$100% confidence interval for the mean $\mu_* = \mathbf{E}(Y \mid X = x_*)$ of all responses with $X = x_*$

$$b_0 + b_1 x_* \pm t_{\alpha/2}\, s \sqrt{\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}}$$

Prediction interval for the individual response

Often we are more interested in predicting the actual response rather than the mean of all possible responses. For example, we may be interested in the price of one particular house that we are planning to buy, not in the average price of all similar houses.

Instead of estimating a *population parameter*, we are now predicting the *actual value* of a random variable.

DEFINITION 11.4

An interval $[a, b]$ is a $(1-\alpha)100\%$ **prediction interval** for the individual response Y corresponding to predictor $X = x_*$ if it contains the value of Y with probability $(1-\alpha)$,

$$\boldsymbol{P}\{a \le Y \le b \mid X = x_*\} = 1-\alpha.$$

This time, all three quantities, Y, a, and b, are random variables. Predicting Y by $\hat{y}_*$, estimating the standard deviation

$$\mathrm{Std}(Y-\hat{y}_*) = \sqrt{\mathrm{Var}(Y)+\mathrm{Var}(\hat{y}_*)} = \sigma\sqrt{1+\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}} \qquad (11.9)$$

by

$$\widehat{\mathrm{Std}}(Y-\hat{y}_*) = s\sqrt{1+\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}},$$

and standardizing all three parts of the inequality

$$a \le Y \le b,$$

we realize that the $(1-\alpha)100\%$ prediction interval for Y has to satisfy the equation

$$\boldsymbol{P}\left\{ \frac{a-\hat{y}_*}{\widehat{\text{Std}}(Y-\hat{y}_*)} \le \frac{Y-\hat{y}_*}{\widehat{\text{Std}}(Y-\hat{y}_*)} \le \frac{b-\hat{y}_*}{\widehat{\text{Std}}(Y-\hat{y}_*)} \right\} = 1-\alpha.$$

At the same time, the properly standardized $(Y-\hat{y}_*)$ has T-distribution, and

$$\boldsymbol{P}\left\{ -t_{\alpha/2} \le \frac{Y-\hat{y}_*}{\widehat{\text{Std}}(Y-\hat{y}_*)} \le t_{\alpha/2} \right\} = 1-\alpha.$$

A prediction interval is now computed by solving equations

$$\frac{a-\hat{y}_*}{\widehat{\text{Std}}(Y-\hat{y}_*)} = -t_{\alpha/2} \quad \text{and} \quad \frac{b-\hat{y}_*}{\widehat{\text{Std}}(Y-\hat{y}_*)} = t_{\alpha/2}$$

in terms of a and b.

$(1-\alpha)$100% prediction interval for the individual response Y when $X = x_*$

$$\boxed{b_0 + b_1 x_* \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_*-\bar{x})^2}{S_{xx}}}} \qquad (11.10)$$

Several conclusions are apparent from this.

First, compare the standard deviations in (11.8) and (11.9). Response Y that we are predicting made its contribution into the variance. This is the difference between a confidence interval for the mean of all responses and a prediction interval for the individual response. Predicting the individual value is a more difficult task; therefore, the prediction interval is always wider than the confidence interval for the mean response. More uncertainty is involved, and as a result, the margin of a prediction interval is larger than the margin of a confidence interval.

Second, we get more accurate estimates and more accurate predictions from large samples. When the sample size n (and therefore, typically, S_{xx}), tends to ∞, the margin of the confidence interval converges to 0.

On the other hand, the margin of a prediction interval converges to $(t_{\alpha/2}\sigma)$. As we collect more and more observations, our estimates of b_0 and b_1 become more accurate; however, uncertainty about the individual response Y will never vanish.

Third, we see that regression estimation and prediction are most accurate when x_* is close to $\bar{x}$ so that

$$(x_*-\bar{x})^2 \approx 0.$$

The margin increases as the independent variable x_* drifts away from $\bar{x}$. We conclude that it is easiest to make forecasts under normal and "standard" conditions, and it is hardest to predict anomalies. And this agrees with our common sense.

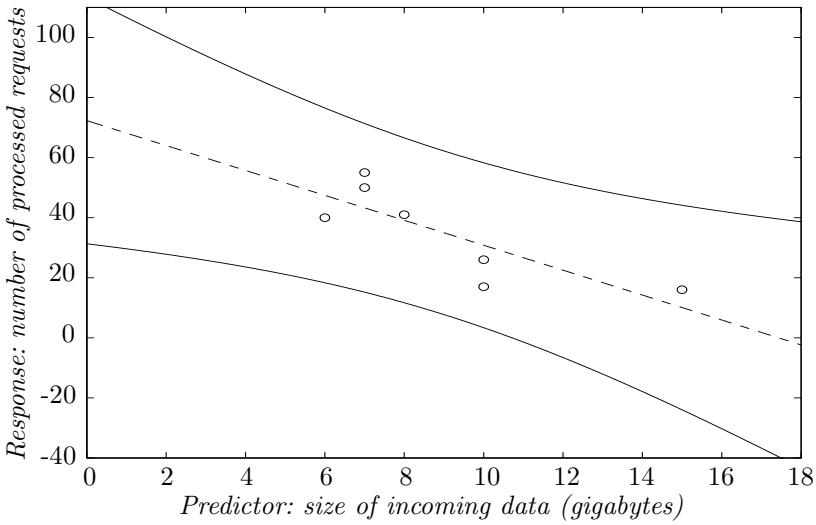

FIGURE 11.5: *Regression prediction of program efficiency.*

Example 11.7 (PREDICTING THE PROGRAM EFFICIENCY). Suppose we need to start processing requests that refer to $x_* = 16$ gigabytes of data. Based on our regression analysis of the program efficiency in Example 11.6, we predict

$$y_* = b_0 + b_1 x_* = 72.3 - 4.14(16) = 6$$

requests processed within 1 hour. A 95% prediction interval for the number of processed requests is

$$\begin{aligned} y_* \pm t_{0.025}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}} &= 6 \pm (2.571)\sqrt{98.2}\sqrt{1 + \frac{1}{7} + \frac{(16-9)^2}{56}} \\ &= 6 \pm 36.2 = [0; 42]. \end{aligned}$$

(using Table A5 with 5 d.f.). We rounded both ends of the prediction interval knowing that there cannot be a negative or fractional number of requests. ♢

Prediction bands

For all possible values of a predictor x_*, we can prepare a graph of $(1-\alpha)$ **prediction bands** given by (11.10). Then, for each value of x_*, one can draw a vertical line and obtain a $100(1-\alpha)\%$ prediction interval between these bands.

Figure 11.5 shows the 95% prediction bands for the number of processed requests in Example 11.7. These are two curves on each side of the fitted regression line. As we have already noticed, prediction is most accurate when x_* is near the sample mean $\bar{x}$. Prediction intervals get wider when we move away from $\bar{x}$.

11.3 Multivariate regression

In the previous two sections, we learned how to predict a response variable Y from a predictor variable X. We hoped in several examples that including more information and using multiple predictors instead of one will enhance our prediction.

Now we introduce **multiple linear regression** that will connect a response Y with several predictors $X^{(1)}$, $X^{(2)}$, ..., $X^{(k)}$.

11.3.1 Introduction and examples

Example 11.8 (Additional information). In Example 11.2, 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. ◊

Example 11.9 (U.S. population and nonlinear terms). One can often reduce variability around the trend and do more accurate analysis by adding nonlinear terms into the regression model. In Example 11.3, we predicted the world population for years 2015–2020 based on the *linear model*

$$\mathbf{E}(\text{population}) = \beta_0 + \beta_1(\text{year}).$$

We showed in Example 11.4 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 11.6a). The population growth over a longer period of time is clearly nonlinear.

On the other hand, a *quadratic model* in Figure 11.6b gives an amazingly excellent fit! It seems to account for everything except a temporary decrease in the rate of growth during the World War II (1939–1945).

For this model, we assume

$$\mathbf{E}(\text{population}) = \beta_0 + \beta_1(\text{year}) + \beta_2(\text{year})^2,$$

or in a more convenient but equivalent form,

$$\mathbf{E}(\text{population}) = \beta_0 + \beta_1(\text{year-1800}) + \beta_2(\text{year-1800})^2.$$

◊

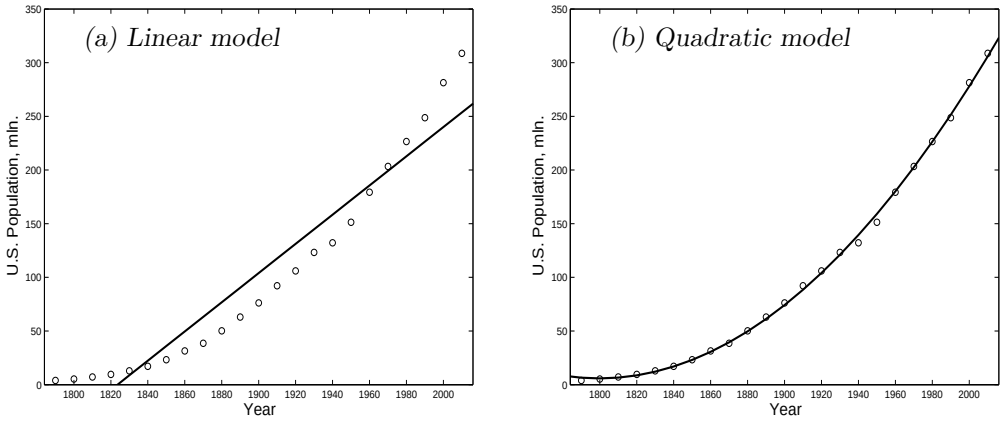

FIGURE 11.6: *U.S. population in 1790–2010 (million people).*

A **multivariate linear regression model** assumes that the conditional expectation of a response

$$\mathbf{E}\left\{Y \mid X^{(1)} = x^{(1)}, \ldots, X^{(k)} = x^{(k)}\right\} = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} \tag{11.11}$$

is a linear function of predictors $x^{(1)}, \ldots, x^{(k)}$.

This regression model has one intercept and a total of k slopes, and therefore, it defines a k-dimensional *regression plane* in a $(k+1)$-dimensional space of $(X^{(1)}, \ldots, X^{(k)}, Y)$.

The **intercept** β_0 is the expected response when all predictors equal zero.

Each **regression slope** β_j is the expected change of the response Y when the corresponding predictor $X^{(j)}$ changes by 1 *while all the other predictors remain constant.*

In order to estimate all the parameters of model (11.11), we collect a sample of n *multivariate observations*

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

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 $\beta_0, \beta_1, \ldots, \beta_k$ by the method of least squares, generalizing it from the univariate case of Section 11.1 to multivariate regression.

11.3.2 Matrix approach and least squares estimation

According to the *method of least squares*, we find such slopes $\beta_1, \ldots, \beta_k$ and such an intercept β_0 that will minimize the sum of squared "errors"

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

Minimizing Q, we can again take partial derivatives of Q with respect to all the unknown parameters and solve the resulting system of equations. It can be conveniently written in a *matrix form* (which requires basic knowledge of linear algebra; if needed, refer to Appendix, Section 12.4).

Matrix approach to multivariate linear regression

We start with the data. Observed are an $n \times 1$ response vector $\boldsymbol{Y}$ and an $n \times (k+1)$ predictor matrix $\boldsymbol{X}$,

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} 1 & \boldsymbol{X}_1 \\ \vdots & \vdots \\ 1 & \boldsymbol{X}_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^{(1)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{(1)} & \cdots & X_n^{(k)} \end{pmatrix}.$$

It is convenient to augment the predictor matrix with a column of 1's because now the multivariate regression model (11.11) can be written as

$$\mathbf{E} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^{(1)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{(1)} & \cdots & X_n^{(k)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

or simply

$$\mathbf{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}.$$

Now the multidimensional parameter

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \in \mathbb{R}^{k+1}$$

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.

Our goal is to estimate $\boldsymbol{\beta}$ with a vector of **sample regression slopes**

$$\boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

Fitted values will then be computed as

$$\widehat{\boldsymbol{y}} = \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} = \boldsymbol{X}\boldsymbol{b}.$$

Thus, the least squares problem reduces to minimizing

$$\begin{aligned} Q(\boldsymbol{b}) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T (\boldsymbol{y} - \widehat{\boldsymbol{y}}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}). \end{aligned} \tag{11.12}$$

with T denoting a transposed vector.

Least squares estimates

In the matrix form, the minimum of the sum of squares

$$Q(\boldsymbol{b}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{y}^T\boldsymbol{y}$$

is attained by

Estimated slopes in multivariate regression

$$\boxed{\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$$

As we can see from this formula, all the estimated slopes are

– *linear* functions of observed responses $(y_1, \dots, y_n)$,

– *unbiased* for the regression slopes because

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

– *Normal* if the response variable Y is Normal.

This is a multivariate analogue of $b = S_{xy}/S_{xx}$ that we derived for the univariate case.

11.3.3 Analysis of variance, tests, and prediction

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

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

with $\text{df}_{\text{TOT}} = (n-1)$ degrees of freedom, where

$$\bar{\boldsymbol{y}} = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \bar{y}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

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

$$SS_{\text{REG}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^T(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$$

is the **regression sum of squares**, and

$$SS_{\text{ERR}} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{e}^T\boldsymbol{e}$$

is the **error sum of squares**, the quantity that we minimized when we applied the method of least squares.

The multivariate regression model (11.11) 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 again the sample size n minus k estimated slopes and 1 estimated intercept.

We can then write the ANOVA table,

Multivariate ANOVA

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	$SS_{\mathrm{REG}} = (\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^T(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$	k	$MS_{\mathrm{REG}} = \dfrac{SS_{\mathrm{REG}}}{k}$	$\dfrac{MS_{\mathrm{REG}}}{MS_{\mathrm{ERR}}}$
Error	$SS_{\mathrm{ERR}} = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T(\boldsymbol{y} - \widehat{\boldsymbol{y}})$	$n - k - 1$	$MS_{\mathrm{ERR}} = \dfrac{SS_{\mathrm{ERR}}}{n - k - 1}$	
Total	$SS_{\mathrm{TOT}} = (\boldsymbol{y} - \bar{\boldsymbol{y}})^T(\boldsymbol{y} - \bar{\boldsymbol{y}})$	$n - 1$		

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 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.

Testing significance of the entire model

Further inference requires **standard multivariate regression assumptions** of Y_i being independent Normal random variables with means

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 X_i^{(1)} + \ldots + \beta_k X_i^{(k)}$$

and constant variance σ^2 while all predictors $X_i^{(j)}$ are non-random.

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

$$H_0: \ \beta_1 = \ldots = \beta_k = 0 \quad \text{vs} \quad H_A: \ \text{not } H_0; \ \text{at least one } \beta_j \neq 0.$$

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 in Table A7.

This is always a one-sided right-tail test. Only large values of F correspond to large SS_{REG} indicating that fitted values $\hat{y}_i$ are far from the overall mean $\bar{y}$, and therefore, the expected response really changes along the regression plane according to predictors.

Variance estimator

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

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

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

Testing individual slopes

For the inference about **individual regression slopes** β_j, we compute all the variances $\text{Var}(\beta_j)$. Matrix

$$\text{VAR}(\boldsymbol{b}) = \begin{pmatrix} \text{Var}(b_1) & \text{Cov}(b_1, b_2) & \cdots & \text{Cov}(b_1, b_k) \\ \text{Cov}(b_2, b_1) & \text{Var}(b_2) & \cdots & \text{Cov}(b_2, b_k) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(b_k, b_1) & \text{Cov}(b_k, b_2) & \cdots & \text{Var}(b_k) \end{pmatrix}$$

is called a **variance-covariance matrix** of a vector $\boldsymbol{b}$. It equals

$$\begin{aligned} \text{VAR}(\boldsymbol{b}) &= \text{VAR}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \,\text{VAR}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned}$$

Diagonal elements of this $k \times k$ matrix are variances of individual regression slopes,

$$\sigma^2(b_1) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{11}, \ldots, \ \sigma^2(b_k) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{kk}.$$

We estimate them by sample variances,

$$s^2(b_1) = s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{11}, \ldots,\ s^2(b_k) = s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{kk}.$$

Now we are ready for the inference about individual slopes. Hypothesis

$$H_0:\ \beta_j = B$$

can be tested with a T-statistic

$$t = \frac{b_j - B}{s(b_j)}.$$

Compare this T-statistic against the T-distribution with $\mathrm{df}_{\mathrm{ERR}} = n - k - 1$ degrees of freedom, Table A5. This test may be two-sided or one-sided, depending on the alternative.

A test of

$$H_0:\ \beta_j = 0 \qquad \text{vs} \quad H_A:\ \beta_j \neq 0$$

shows whether predictor $X^{(j)}$ is relevant for the prediction of Y. If the alternative is true, the expected response

$$\mathbf{E}(Y) = \beta_0 + \beta_1 X^{(1)} + \ldots + \beta_j X^{(j)} + \ldots + \beta_k X^{(k)}$$

changes depending on $X^{(j)}$ even if all the other predictors remain constant.

Prediction

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

$$\hat{y}_* = \widehat{\mathbf{E}}\,\{Y \mid \boldsymbol{X}_* = \boldsymbol{x}_*\} = \boldsymbol{x}_*\boldsymbol{b}$$

and predict the individual response by the same statistic.

To produce confidence and prediction intervals, we compute the variance,

$$\mathrm{Var}(\hat{y}_*) = \mathrm{Var}(\boldsymbol{x}_*\boldsymbol{b}) = \boldsymbol{x}_*^T\,\mathrm{Var}(\boldsymbol{b})\boldsymbol{x}_* = \sigma^2\boldsymbol{x}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_*,$$

where $\boldsymbol{X}$ is the matrix of predictors used to estimate the regression slope $\boldsymbol{\beta}$.

Estimating σ^2 by s^2, we obtain a $(1-\alpha)100\%$ **confidence interval** for $\mu_* = \mathbf{E}(Y)$.

$(1-\alpha)$100% confidence interval for the mean $\mu_* = \mathbf{E}(Y \mid \boldsymbol{X}_* = \boldsymbol{x}_*)$ of all responses with $X_* = x_*$

$$\boldsymbol{x}_*\boldsymbol{b} \pm t_{\alpha/2}\, s \sqrt{\boldsymbol{x}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_*}$$

Accounting for the additional variation of the individual response y_*, we get a $(1-\alpha)100\%$ **prediction interval** for y_*.

$(1-\alpha)$**100% prediction interval for the individual response** Y **when** $\boldsymbol{X}_* = \boldsymbol{x}_*$	$\boldsymbol{x}_*\boldsymbol{b} \pm t_{\alpha/2}\, s \sqrt{1 + \boldsymbol{x}_*^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_*}$

In both expressions, $t_{\alpha/2}$ refers to the T-distribution with $(n-k-1)$ degrees of freedom.

Example 11.10 (Database structure). The computer manager in Examples 11.6 and 11.7 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, we have

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

(a) Least squares estimation. The *predictor matrix* and the *response vector* are

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

We then compute

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 7 & 63 & 67 \\ 63 & 623 & 519 \\ 67 & 519 & 1021 \end{pmatrix} \quad \text{and} \quad \boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} 245 \\ 1973 \\ 2908 \end{pmatrix},$$

to obtain the estimated *vector of slopes*

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

Thus, the regression equation is

$$y = 52.7 - 2.87x_1 + 0.85x_2,$$

or

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

(b) ANOVA AND F-TEST. The total sum of squares is still $SS_{\text{TOT}} = S_{yy} = 1452$. It is the same for all the models with this response.

Having figured a vector of *fitted values*

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

we can immediately compute

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

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		

Notice 2 degrees of freedom for the model because we now use two predictor variables.

R-SQUARE is now $R^2 = SS_{\text{REG}}/SS_{\text{TOT}} = 0.787$, which is 12.5% higher than in Example 11.6. 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.

ANOVA F-TEST statistic of 7.41 with 2 and 4 d.f. shows that the model is significant at the level of 0.05 but not at the level of 0.025.

REGRESSION VARIANCE σ^2 is estimated by $s^2 = 77.2$.

(c) INFERENCE ABOUT THE NEW SLOPE. Is the new predictor variable x_2 significant? It is, as long as the corresponding slope β_2 is proved to be non-zero. Let us test $H_0: \ \beta_2 = 0$.

The vector of slopes $\boldsymbol{b}$ has an estimated variance-covariance matrix

$$\widehat{\text{VAR}}(\boldsymbol{b}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 284.7 & -22.9 & -7.02 \\ -22.9 & 2.06 & 0.46 \\ -7.02 & 0.46 & 0.30 \end{pmatrix}.$$

From this, $s(b_2) = \sqrt{0.30} = 0.55$. The T-statistic is then

$$t = \frac{b_2}{s(b_2)} = \frac{0.85}{0.55} = 1.54,$$

and for a two-sided test this is not significant at any level up to 0.10. This suggests that adding the data structure into the model does not bring a significant improvement.

$\Diamond$

11.4 Model building

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 saw in Section 11.1.5 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?

Two methods of variable selection are introduced here. One is based on the *adjusted R-square* criterion, the other is derived from the *extra sum of squares principle*.

11.4.1 Adjusted R-square

It is 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 11.5

Adjusted R-square

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

is a criterion of variable selection. It rewards for adding a predictor only if it considerably reduces the error sum of squares.

Comparing with

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

adjusted R-square includes degrees of freedom into this formula. This adjustment may result in a penalty when a useless X-variable is added to the regression mode.

Indeed, imagine adding a non-significant predictor. The number of estimated slopes k increments 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, $SS_{\text{ERR}}/(n-k-1)$ will increase and R^2_{adj} will decrease, penalizing us for including such a poor predictor.

Adjusted R-square criterion: *choose a model with the highest adjusted R-square.*

11.4.2 Extra sum of squares, partial F-tests, and variable selection

Suppose we have K predictors available for predicting a response. Technically, to select a subset that maximizes adjusted R-square, we need to fit all 2^K models and choose the one with the highest R^2_{adj}. This is possible for rather moderate K, and such schemes are built in some statistical software.

Fitting all models is not feasible when the total number of predictors is large. Instead, we consider a *sequential* scheme that will follow a reasonable path through possible regression models and consider only a few of them. At every step, it will compare some set of predictors

$$\boldsymbol{X}(Full) = \left(X^{(1)}, \ldots, X^{(k)}, X^{(k+1)}, \ldots, X^{(m)}\right)$$

and the corresponding *full model*

$$\mathbf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \beta_{k+1} x^{(k+1)} + \ldots + \beta_m x^{(m)}$$

with a subset

$$\boldsymbol{X}(Reduced) = \left(X^{(1)}, \ldots, X^{(k)}\right)$$

and the corresponding *reduced model*

$$\mathbf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)}.$$

If the full model is significantly better, expanding the set of predictors is justified. If it is just as good as the reduced model, we should keep the smaller number of predictors in order to attain lower variances of the estimated regression slopes, more accurate predictions, and a lower adjusted R-square.

DEFINITION 11.6

A model with a larger set of predictors is called a **full model**.

Including only a subset of predictors, we obtain a **reduced model**.

The difference in the variation explained by the two models is the **extra sum of squares**,

$$\begin{aligned} SS_{\text{EX}} &= SS_{\text{REG}}(Full) - SS_{\text{REG}}(Reduced) \\ &= SS_{\text{ERR}}(Reduced) - SS_{\text{ERR}}(Full). \end{aligned}$$

Extra sum of squares measures the *additional* amount of variation explained by additional predictors $X^{(k+1)}, \ldots, X^{(m)}$. By subtraction, it has

$$\text{df}_{\text{EX}} = \text{df}_{\text{REG}}(Full) - \text{df}_{\text{REG}}(Reduced) = m - k$$

degrees of freedom.

Significance of the additional explained variation (measured by SS_{EX}) is tested by a **partial F-test statistic**

$$F = \frac{SS_{\text{EX}}/\text{df}_{\text{EX}}}{MS_{\text{ERR}}(Full)} = \frac{SS_{\text{ERR}}(Reduced) - SS_{\text{ERR}}(Full)}{SS_{\text{ERR}}(Full)} \left(\frac{n-m-1}{m-k}\right).$$

As a set, $X^{(k+1)}, \ldots, X^{(m)}$ affect the response Y if at least one of the slopes $\beta_{k+1}, \ldots, \beta_m$ is not zero in the full model. The partial F-test is a test of

$$H_0: \ \beta_{k+1} = \ldots = \beta_m = 0 \qquad \text{vs} \qquad H_A: \ \text{not } H_0.$$

If the null hypothesis is true, the partial F-statistic has the *F-distribution* with

$$\text{df}_{\text{EX}} = m - k \quad \text{and} \quad \text{df}_{\text{ERR}}(Full) = n - m - 1$$

degrees of freedom, Table A7.

The *partial F-test* is used for sequential selection of predictors in multivariate regression. Let us look at two algorithms that are based on the partial F-test: *stepwise selection* and *backward elimination*.

Stepwise (forward) selection

The **stepwise selection** *algorithm* starts with the simplest model that excludes all the predictors,

$$G(\boldsymbol{x}) = \beta_0.$$

Then, predictors enter the model sequentially, one by one. Every new predictor should make the most significant contribution, among all the predictors that have not been included yet.

According to this rule, the first predictor $X^{(s)}$ to enter the model is the one that has the most significant univariate ANOVA F-statistic

$$F_1 = \frac{MS_{\text{REG}}(X^{(s)})}{MS_{\text{ERR}}(X^{(s)})}.$$

All F-tests considered at this step refer to the same F-distribution with 1 and $(n-2)$ d.f. Therefore, the largest F-statistic implies the lowest P-value and the most significant slope β_s

The model is now

$$G(\boldsymbol{x}) = \beta_0 + \beta_s x^{(s)}.$$

The next predictor $X^{(t)}$ to be selected is the one that makes the most significant contribution, in addition to $X^{(s)}$. Among all the remaining predictors, it should maximize the partial F-statistic

$$F_2 = \frac{SS_{\text{ERR}}(Reduced) - SS_{\text{ERR}}(Full)}{MS_{\text{ERR}}(Full)}$$

designed to test significance of the slope β_t when the first predictor $X^{(s)}$ is already included. At this step, we compare the "full model" $G(\boldsymbol{x}) = \beta_0 + \beta_s x^{(s)} + \beta_t x^{(t)}$ against the "reduced model" $G(\boldsymbol{x}) = \beta_0 + \beta_s x^{(s)}$. Such a partial F-statistic is also called **F-to-enter**.

All F-statistics at this step are compared against the same F-distribution with 1 and $(n-3)$ d.f., and again, the largest F-statistic points to the most significant slope β_t.

If the second predictor is included, the model becomes

$$G(\boldsymbol{x}) = \beta_0 + \beta_s x^{(s)} + \beta_t x^{(t)}.$$

The algorithm continues until the F-to-enter statistic is not significant for all the remaining

predictors, according to a pre-selected significance level α. The final model will have all predictors significant at this level.

Backward elimination

The **backward elimination algorithm** works in the direction opposite to stepwise selection.

It starts with the full model that contains all possible predictors,

$$G(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_m x^{(m)}.$$

Predictors are *removed* from the model sequentially, one by one, starting with the *least significant* predictor, until all the remaining predictors are statistically significant.

Significance is again determined by a partial F-test. In this scheme, it is called **F-to-remove**.

The first predictor to be removed is the one that *minimizes* the F-to-remove statistic

$$F_{-1} = \frac{SS_{\text{ERR}}(Reduced) - SS_{\text{ERR}}(Full)}{MS_{\text{ERR}}(Full)}.$$

Again, the test with the lowest value of F_{-1} has the highest P-value indicating the least significance.

Suppose the slope β_u is found the least significant. Predictor $X^{(u)}$ is removed, and the model becomes

$$G(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_{u-1} x^{(u-1)} + \beta_{u+1} x^{(u+1)} + \ldots + \beta_m x^{(m)}.$$

Then we choose the next predictor to be removed by comparing all F_{-2} statistics, then go to F_{-3}, etc. The algorithm stops at the stage when all F-to-remove tests reject the corresponding null hypotheses. It means that in the final resulting model, all the remaining slopes are significant.

Both sequential model selection schemes, stepwise and backward elimination, involve fitting at most K models. This requires much less computing power than the adjusted R^2 method, where all 2^K models are considered.

Modern statistical computing packages (SAS, Splus, SPSS, JMP, and others) are equipped with all three considered model selection procedures.

Example 11.11 (Program efficiency: choice of a model). How should we predict the program efficiency in Examples 11.6, 11.7, and 11.10 after all? Should we use the size of data sets x_1 alone, or the data structure x_2 alone, or both variables?

(a) Adjusted R-square criterion. For the *full model*,

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{ERR}}/\text{df}_{\text{ERR}}}{SS_{\text{TOT}}/\text{df}_{\text{TOT}}} = 1 - \frac{308.7/4}{1452/6} = 0.681.$$

Reduced model with only one predictor x_1 (Example 11.6) has

$$R^2_{\text{adj}} = 1 - \frac{491/5}{1452/6} = 0.594,$$

and another *reduced model* with only x_2 has $R^2_{\text{adj}} = 0.490$ (Exercise 11.9).

How do we interpret these R^2_{adj}? The price paid for including both predictors x_1 and x_2 is the division by 4 d.f. instead of 5 when we computed R^2_{adj} for the full model. Nevertheless, the full model explains such a large portion of the total variation that fully compensates for this penalty and makes the full model preferred to reduced ones. According to the *adjusted R-square criterion*, *the full model is best*.

(b) PARTIAL F-TEST. How significant was addition of a new variable x_2 into our model? Comparing the *full model* in Example 11.10 with the *reduced model* in Example 11.6, we find the *extra sum of squares*

$$SS_{\text{EX}} = SS_{\text{REG}}(Full) - SS_{\text{REG}}(Reduced) = 1143 - 961 = 182.$$

This is the additional amount of the total variation of response explained by x_2 when x_1 is already in the model. It has 1 d.f. because we added only 1 variable. The *partial F-test statistic* is

$$F = \frac{SS_{\text{EX}}/\text{df}_{\text{EX}}}{MS_{\text{ERR}}(Full)} = \frac{182/1}{309} = 0.59.$$

From Table A7 with 1 and 4 d.f., we see that this F-statistic is *not significant* at the 0.25 level. It means that a relatively small additional variation of 182 that the second predictor can explain does not justify its inclusion into the model.

(c) SEQUENTIAL MODEL SELECTION. What models should be selected by stepwise and backward elimination routines?

Stepwise model selection starts by including the first predictor x_1. It is significant at the 5% level, as we know from Example 11.6, hence we keep it in the model. Next, we include x_2. As we have just seen, it fails to result in a significant gain, $F_2 = 0.59$, and thus, we do not keep it in the model. The resulting model predicts the program efficiency y based on the size of data sets x_1 only.

Backward elimination scheme starts with the full model and looks for ways to reduce it. Among the two reduced models, the model with x_1 has a higher regression sum of squares SS_{REG}, hence the other variable x_2 is the first one to be removed. The remaining variable x_1 is significant at the 5% level; therefore, we again arrive to the reduced model predicting y based on x_1.

Two different model selection criteria, adjusted R-square and partial F-tests, lead us to two different models. Each of them is best in a different sense. Not a surprise. ◊

11.4.3 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* (that is, 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, etc.);

– operating system (Unix, Windows, DOS, etc.);

– major (Statistics, Computer Science, Electrical Engineering, etc.);

– gender (female, male);

– color (white, blue, 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

$$G(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \beta_{k+1} x^{(k+1)}.$$

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, $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(C-1)}$. They carry the entire information about the attribute. Sampled items from category C will be marked by all $(C-1)$ dummies equal to 0.

Example 11.12 (Dummy variables for the operating system). In addition to numerical variables, we would like to include the operating system into the regression model. Suppose that each sampled computer has one of three operating systems: Unix, Windows, or DOS. In order to use this information for the regression modeling and more accurate forecasting, we create *two* dummy variables,

$$Z_i^{(1)} = \begin{cases} 1 & \text{if computer } i \text{ has Unix} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_i^{(2)} = \begin{cases} 1 & \text{if computer } i \text{ has Windows} \\ 0 & \text{otherwise} \end{cases}$$

Together with numerical predictors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k)}$, the regression model will be

$$G(\boldsymbol{x}, \boldsymbol{z}) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \gamma_1 z^{(1)} + \gamma_2 z^{(2)}.$$

$\Diamond$

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

Avoid singularity by creating only $(C-1)$ dummies

Notice that if we make a mistake and create C dummies for an attribute with C categories, one dummy per category, this would cause a linear relation

$$\boldsymbol{Z}^{(1)} + \ldots + \boldsymbol{Z}^{(C)} = 1.$$

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

Interpretation of slopes for dummy variables

Each slope γ_j for a dummy predictor $Z^{(j)}$ is the expected change in the response caused by incrementing $Z^{(j)}$ by 1 while keeping all other predictors constant. Such an increment occurs when we compare the last category C with category j.

Thus, the slope γ_j is the difference in the expected response comparing category C with category j. The difference of two slopes $(\gamma_j - \gamma_C)$ compares category j with category C.

To test significance of a categorical variable, we test all the corresponding slopes γ_j simultaneously. This is done by a partial F-test.

Matlab notes

MATLAB (MATrix LABoratory) is great for matrix computations, so all the regression analysis can be done by writing the matrix formulas, `b=inv(X'*X)*(X*Y)` for the regression slope, `Yhat=X*b` for the fitted values, `e=Y-Yhat` for residuals, etc., given a vector of responses Y and a matrix of predictors X. For more instructions on that, see the last paragraph of Section 12.4.

Also, MATLAB's Statistics Toolbox has special tools for regression. The general command `regress(Y,X)` returns a sample regression slope with a response Y and predictor X. Notice that if you'd like to fit regression with an *intercept*, then the vector of ones has to be included into matrix X. For example,

```
X0 = ones(size(Y));
regress(Y,[X0,X1,X2]);
```

will create a vector of ones of the same size as the vector of responses and use it to fit a regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. To get $(1-\alpha)100\%$ confidence intervals for all the regression slopes, write `[b bint] = regress(Y,[X0,X1,X2],alpha)`.

Many components of the regression analysis are available by the command `regstats(Y,X)`. A list of options appears

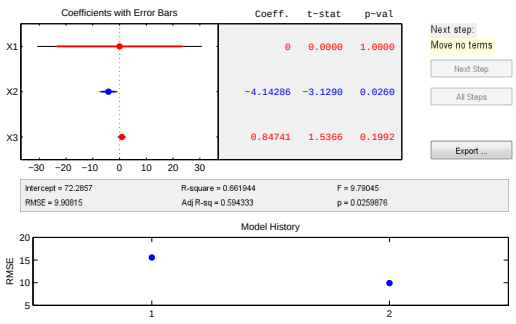

FIGURE 11.7: *Stepwise variable selection* to model program efficiency, Example 11.11.

where you can mark which statistics you'd like to obtain – mean squared error s^2, ANOVA F-statistic, R^2, R^2_{adj}, etc.

You can also use `stepwise(X,Y)` to select variables for the multivariate regression. A window opens such as on Figure 11.7, and the stepwise variable selection algorithm runs, reporting the vital regression statistics at each step.

Summary and conclusions

This chapter provides methods of estimating mathematical relations between one or several predictor variables and a response variable. Results are used to explain behavior of the response and to predict its value for any new set of predictors.

Method of least squares is used to estimate regression parameters. Coefficient of determination R^2 shows the portion of the total variation that the included predictors can explain. The unexplained portion is considered as "error."

Analysis of variance (ANOVA) partitions the total variation into explained and unexplained parts and estimates regression variance by the mean squared error. This allows further statistical inference, testing slopes, constructing confidence intervals for mean responses and prediction intervals for individual responses. ANOVA F-test is used to test significance of the entire model.

For accurate estimation and efficient prediction, it is important to select the right subset of predictors. Sequential model selection algorithms are based on partial F-tests comparing full and reduced models at each step.

Categorical predictors are included into regression modeling by creating dummy variables.

Exercises

Suppose that the standard regression assumptions, univariate or multivariate, hold in Exercises 11.2, 11.3, 11.4, 11.5, 11.9, 11.14, and 11.15.

11.1. The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86.

Based on this data, fit a linear regression model and predict the time it will take to transmit a 400 Kbyte file.

11.2. The following statistics were obtained from a sample of size $n = 75$:
– the predictor variable X has mean 32.2, variance 6.4;