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 7 (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

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

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

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

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

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 $\beta_0, \beta_1, \ldots, \beta_k$ by the method of least squares, generalizing it from the univariate case to multivari-

ate regression. So, we minimize the sum of squared errors

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

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

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

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

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

If we denote by

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

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

Let

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

Then the fitted values will be computed as

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

and the least squares problem reduces to minimizing

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

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

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

Remark 5.1.

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

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

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

Example 5.2. 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

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

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

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

Thus, the regression equation is now

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

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

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

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

■

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

6 ANOVA and Adjusted R-square

6.1 Multivariate ANOVA and F-test

So, for the multivariate linear regression model, we assume 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 $\text{df}_{\text{TOT}} = n-1$ 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, $SS_{\text{TOT}} \;=\; SS_{\text{REG}} + SS_{\text{ERR}}$, where

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

is the **regression sum of squares** and

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

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

$$\text{df}_{\text{REG}} \;=\; k$$

degrees of freedom, whereas by subtraction,

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

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 1.

Source	Sum of squares SS	Degrees of freedom df	Mean Squares $MS = SS/\text{df}$	F
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 1: Multivariate ANOVA

The **coefficient of determination**

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

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.

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

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

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 6.1. Let us revisit Example 5.2 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 the case with one predictor only (Example 4.3 from last time). 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).

■

6.2 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 6.2. ***Adjusted R-square*** *is the quantity*

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

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

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

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, $SS_{\rm REG}$ and $SS_{\rm ERR}$, will remain the same. Then, $SS_{\rm ERR}/(n-k-1)$ will increase and $R^2_{\rm adj}$ will decrease, penalizing us for including such a poor predictor.

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

6.3 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 $R^2_{\rm adj}$. This is possible for rather moderate values of K and such schemes are built in some statistical software. But 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

$$\mathbf{X}(full) \;=\; \left(X^{(1)},\ldots,X^{(k)},X^{(k+1)},\ldots,X^{(m)}\right)$$

and the corresponding *full* model

$$E(\mathbf{Y}\mid\mathbf{X}=\mathbf{x})=\beta_0+\beta_1x^{(1)}+\cdots+\beta_kx^{(k)}+\beta_{k+1}x^{(k+1)}+\cdots+\beta_mx^{(m)}$$

with a subset

$$\mathbf{X}(reduced) \;=\; \left(X^{(1)},\ldots,X^{(k)}\right)$$

and the corresponding *reduced* model

$$E(\mathbf{Y}\mid\mathbf{X}=\mathbf{x})=\beta_0+\beta_1x^{(1)}+\cdots+\beta_kx^{(k)}.$$

If the full model is *significantly* better, then 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 6.3.
*– 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**

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

The predictors $X^{(k+1)}, \ldots, X^{(m)}$ 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

$$\begin{aligned} H_0 &: \ \beta_{k+1} = \cdots = \beta_m = 0 \\ H_1 &: \ \text{at least one } \beta_j \neq 0. \end{aligned}$$

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

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

degrees of freedom.
The *partial F-test* is used for sequential selection of predictors in multivariate regression. There are two algorithms that are based on the partial F-test: *stepwise selection* and *backward elimination*.
The **stepwise (forward) selection** algorithm starts with the simplest model that excludes all the predictors,

$$G(x) = \beta_0.$$

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 $(n-2)$ degrees of freedom. Therefore, the largest F-statistic implies the lowest P-value and the most significant slope β_s.
The model is now

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

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"

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

against the "reduced model"

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

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.
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.

Remark 6.4.
1. The **backward elimination** algorithm works in the direction *opposite* to stepwise selection. It starts with the full model that contains all possible predictors and then removes predictors 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, now called **F-to-remove**. The first predictor to be removed is the one that *minimizes* the F-to-remove statistic

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

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

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

Example 6.5. Consider again the previous example, about program efficiency. Let us discuss the choice of a model. Should we use the size of data sets x_1 alone, or the data structure x_2 alone, or both variables?

Solution.

a) **Adjusted R-square criterion**

The adjusted R-square for the *reduced model* with one predictor x_1 (data size) is

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

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

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

By adding another predictor, we increased the adjusted R-square with 8.68%.

Another *reduced model*, with only x_2, has

$$R^2_{\text{adj}} \;=\; 0.490.$$

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 with the reduced model with one predictor x_1, 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

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

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.

Stepwise model selection starts by including the first predictor x_1. It is significant at the 5% level, as we know from Example 4.3 from last time, 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 only.

The 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.

■

6.4 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, $\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(C-1)}$. 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 $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. 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 6.6. Consider the program efficiency study in Example 6.1. 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.

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

In Example 4.2 (Lecture 7) 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 4.3 (Lecture 7), 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.