Once we have found the intercept b_0 and the slope b_1 of the least squares regression we can obtain the N *fitted values*, these are the approximations to the observed y values made by the regression line, i.e.

$$\hat{y}_i = b_0 + b_1 x_i \qquad (i = 1, 2, \cdots, N).$$

We can also find the N *residuals*, these are the distances from the true y values to the fitted y values, i.e.

$$e_i = y_i - \hat{y}_i \qquad (i = 1, 2, \cdots, N).$$

Example 53. Consider the dataset of $N = 10$ companies producing tables that was introduced in Example 49. In Example 52 we found that $b_0 = -13.02$ and $b_1 = 2.545$, therefore we obtain the fitted values $\hat{y}_i = b_0 + b_1 x_i$ and the residuals $e_i = y_i - \hat{y}_i$ shown in Table 27. □

i	x_i	y_i	$\hat{y}_i$	e_i
1	12	20	17.53	2.47
2	14	21	22.62	-1.62
3	15	27	25.16	1.84
4	18	30	32.80	-2.80
5	19	32	35.35	-3.35
6	24	50	48.07	1.93
7	26	54	53.16	0.84
8	27	57	55.71	1.29
9	28	61	58.25	2.75
10	30	60	63.34	-3.34

Table 27: Fitted values and residuals in the dataset of 10 companies producing tables

From now on we concentrate on least squares regression, i.e. the regression that minimizes the sum of squares error. Therefore we will drop the adjective "least squares" and we will simply refer to it as "regression".

The sum of squares error SSE can be seen as a measure of the goodness of fit of the regression: the smaller the SSE, the better the regression fits the data. It is easy to see that the minimum value that SSE can take is 0. This would happen if the relation between x and y is a perfect straight line, in that case the fitted value $\hat{y}_i$ would be exactly equal to the observed value y_i for all i. On the other hand, the maximum value that SSE can take is obtained if $x_i = 0$ for all i. In that case (making use of a method that is beyond the scope of the course) we would obtain $b_1 = 0$ and $b_0 = \bar{y}_U$, i.e. we would obtain $\hat{y} = \bar{y}_U$, leading to a SSE equal to $\sum_U (y_i - \bar{y}_U)^2$.

So, we have found that the SSE lies between 0 and $\sum_U (y_i - \bar{y}_s)^2$. Values close to the lower boundary indicate that x is adequately explaining y, whereas values close to the upper boundary indicate that x is not so adequate for explaining y. However, as the units of SSE are the squared units of y, it is not so easy to interpret its value. One would like to count with a statistic that is easier to interpret. Before defining such a statistic we will talk about the *analysis of variance* of a variable y in a regression.

Note that the upper boundary $\sum_U (y_i - \bar{y}_U)^2$ is simply the variance of y times a factor. This upper boundary is known as the *sum of squares total*, in other words, the total variation of y (around its mean). It can be shown that

$$\sum_U (y_i - \bar{y}_U)^2 = \sum_U (\hat{y}_i - \bar{y}_U)^2 + \sum_U (y_i - \hat{y}_U)^2.$$

The term on the left hand side is the sum of squares total —SST—, the second term on the right hand side is the sum of squares error —SSE—. The first term on the right hand side is known as the *sum of squares regression* SSR. So we have

$$SST = SSR + SSE,$$

where

- $SST = \sum_U (y_i - \bar{y}_U)^2$ is the total variation of y;
- $SSE = \sum_U (y_i - \hat{y}_U)^2$ is the variation unexplained by the regression;
- $SSR = \sum_U (\hat{y}_i - \bar{y}_U)^2$ is the variation explained by the regression.

So the total variation of y, SST, can be decomposed into two terms: one representing the variation explained by the regression, SSR; another one representing the variation unexplained by the regression, SSE. Having this into account, the following statistic of goodness of fit for the regression is proposed.

Definition 54. The *coefficient of determination* of a regression, denoted by R^2, is

$$R^2 = \frac{SSR}{SST} \qquad \text{or equivalently} \qquad R^2 = 1 - \frac{SSE}{SST}. \qquad \square$$

By construction, R^2 takes values between 0 and 1, where 0 indicates that the regression line does not fit well to the data and 1 indicates a perfect linear association between x and y. R^2 measures the proportion of variability in the dependent variable y that is explained by the independent variable x.

R^2 holds a very interesting relation with the coefficient of correlation between x and y, $r_{xy,U}$.

Result 55. Let R^2 be the coefficient of determination of the least squares regression of a variable y in terms of x and let $r_{xy,U}$ be the coefficient of correlation between x and y. Then, the following relation holds

$$R^2 = r_{xy,U}^2,$$

i.e. the coefficient of determination is equal to the squared correlation. $\square$

Example 56. Let us calculate the coefficient of determination R^2 of the least squares regression of the workers dataset that we have considered in Examples 49 to 52. We found that $SSE = 56$. We have $\bar{y}_U = 41.2$, therefore

$$SST = \sum_U (y_i - \bar{y}_U)^2 = (20 - 41.2)^2 + (21 - 41.2)^2 + \cdots + (60 - 41.2)^2 = 2506$$

and we obtain $R^2 = 1 - 56/2506 = 97.77\%$, which indicates that almost 98% of the variability in the number of tables produced is explained by the number of workers.

Let us verify that Result 55 holds. The correlation between x and y is $r_{xy,U} = 0.9888$. Therefore, $r_{xy,U}^2 = 0.9888^2 = 97.77\%$, which coincides with the coefficient of determination R^2. $\square$

The coefficient of determination R^2 can be used as a threshold for deciding whether an independent variable x is useful for explaining a dependent variable y or not. In order to do this, before fitting the regression, one should define the value of this threshold, for instance one may say that if R^2 is smaller than 5% the variable x will be considered as insufficient for

explaining the variability in y. Later in the course we will define another criterion that is also used for this purpose.

There are mainly two reasons for using linear regression in practice. The first reason is for *descriptive* purposes where we want to describe how x and y are related. In this case the slope b_1 plays an important role, as it indicates the change in y associated to a unit change in x. In our example of the workers, we have already mentioned that a change of one unit in the number of workers is associated to an increase of 2.5 in the production of tables. This information can be used for taking decisions, for instance, the manager may use this information for deciding if it is worth considering hiring one more worker.

The second reason is for *predictive* purposes. In this case we want to predict the value of y that would be observed for an element with value x_0. For instance, in the example of the workers, we would like to know what is the number of tables that is expected to be produced by $x_0 = 20$ workers, we would get

$$\hat{y} = -13.02 + 2.545 \cdot 20 = 37.89,$$

meaning that one would expect around 38 tables being produced by 20 workers.

Dummy variables as independent variables

Up to this point we have tacitly assumed that the independent or explanatory variable is numerical. However, it can also be a dummy variable, i.e. a variable x that takes the value one when the ith element has some characteristic of interest and 0 otherwise. For instance, let U be a set of sold housing units. We want to explain the closing price y in terms of a dummy variable that indicates whether the housing unit has a balcony or not.

In this case, the interpretation of the regression coefficients b_0 and b_1 is slightly different. The intercept b_0 is interpreted as the expected value of y in the absence of the characteristic of interest; and the slope b_1 is interpreted as the expected additional effect due to the presence of the characteristic of interest. In our example, b_0 would be the expected closing price of a housing unit without a balcony and b_1 would be the expected additional price due to the unit having a balcony.

We illustrate this with an example.

Example 57. Let U be a set of $N = 20$ housing units with y_i = closing price of the ith unit and

$$x_i = \begin{cases} 1 & \text{if the } i\text{th unit has a balcony} \\ 0 & \text{otherwise} \end{cases}$$

Table 28 shows the values of x and y for the $N = 20$ housing units, where the closing price y is given in millions of SEK.

Balcony	Closing	Balcony	Closing	Balcony	Closing	Balcony	Closing
0	3.90	1	2.36	1	2.54	1	2.54
1	3.17	1	2.32	0	1.83	1	4.09
1	3.83	1	2.32	1	2.36	1	3.74
0	3.80	0	3.70	0	3.17	1	3.39
0	3.77	1	1.94	1	3.61	1	2.49

Table 28: Housing units dataset

The regression coefficients are

$$b_1 = r_{xy,U}\frac{S_{y,U}}{S_{x,U}} = -0.2885\frac{0.7406}{0.4702} = -0.4545 \quad \text{and} \quad b_0 = \bar{y}_U - b_1\bar{x}_U = 3.044-(-0.4545)\cdot 0.7 = 3.362.$$

This means that, for this dataset, a housing unit without balcony is expected to have a closing price of 3.362 millions of SEK; whereas a unit with balcony is expected to cost 0.4545 millions of SEK less. Admittedly, this result may seem counterintuitive.

Just for completeness, let us find the coefficient of determination R^2 for this regression. We could make use of Definition 54, but instead we will make use of Result 55 which says that R^2 is simply the coefficient of correlation $r_{xy,U}$ squared. We have $R_2 = r^2_{xy,U} = (-0.2885)^2 = 8.33\%$.

6.1 R output

Let us summarize all the quantities that we have defined in the frame of a simple linear regression:

- the intercept b_0 and the slope b_1;
- the fitted values $\hat{y}_i$ and the errors or residuals $e_i = y_i - \hat{y}_i$;
- the sums of squares: sum of squares error —SSE—, sum of squares regression —SSR— and sum of squares total —SST—;
- the coefficient of determination R^2.

Figure 37 shows R's output of the simple linear regression on the set of $N = 10$ companies producing tables that has been used as a running example in this section. There are some quantities that we are not familiar with yet. We will introduce them later in the course. Let us describe what we should know from this output at this point:

- Under *Residuals:* we see some summary statistics of the residuals $e_i = y_i - \hat{y}_i$, namely, the minimum, the three quartiles and the maximum.
- The first and second rows of the table *Coefficients:* show some information about the intercept b_0 and the slope b_1, respectively. By now we are only familiar with the actual coefficients that are found in the column *Estimate*.
- The coefficient of determination R^2 is found under the name *Multiple R–squared*.

6.2 More on regression

In the first part of this section we have learned how to find the least squares regression, its interpretation and one goodness–of–fit measure, namely, the R^2. In this subsection we will discuss some specific issues that are worth having in mind whenever we are fitting a regression.

6.2.1 Transformations

Up to this point we have considered the situation in which we want to fit a straight line to the data. We do this guided either by the knowledge or the belief that the data can be adequately described by a straight line. What if the association between the dependent variable y and the independent variable x is not linear? In this section we will discuss that situation.

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-3.346 -2.505  1.064  1.905  2.746

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.016      3.015  -4.317  0.00256 **
x              2.545      0.136  18.716 6.86e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.645 on 8 degrees of freedom
Multiple R-squared:  0.9777,    Adjusted R-squared:  0.9749
F-statistic: 350.3 on 1 and 8 DF,  p-value: 6.859e-08
```

Figure 37: R output of a simple linear regression on the dataset of tables producers.

We will consider the situation in which we want to express y as a linear function of a function of x, i.e. $y \approx b_0 + b_1 g(x)$ with $g(\cdot)$ a known function. Let us illustrate the idea with an example. Let us consider a set of $N = 392$ automobiles in which we have measured the *autonomy* i.e. the miles per gallon ($= y$) and the weight in pounds ($= x$). We want to explain the autonomy in terms of miles per gallon. Figure 38 shows the scatterplot between both variables. It is clear that there is some association between both variables. Maybe our first intuition is to fit a linear regression of the type $y \approx b_0 + b_1 x$ as we have been doing until now, but on a closer look, one may also consider fitting a curve of the type $y \approx b_0 + b_1(1/x)$, i.e. we want to express y as a linear function of $1/x$.

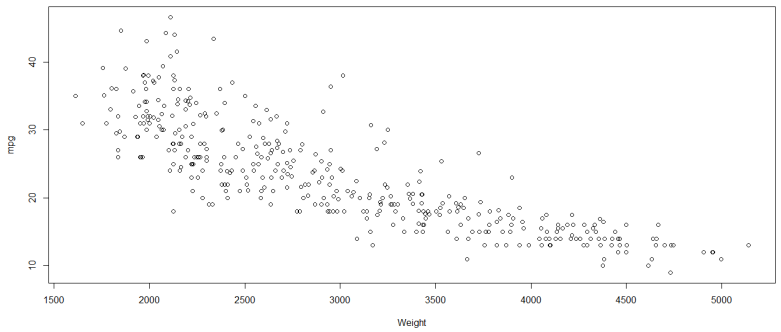

Figure 38: Scatter plot of weight (pounds) vs. autonomy (miles per gallon) of 392 automobiles.

It turns out that all we have to do is to define a new variable, let us call it z as $z = 1/x$ and then we have that the regression of interest can be written as

$$y \approx b_0 + b_1 \frac{1}{x} = b_0 + b_1 z.$$

And now we see that all we have to do is to fit the least squares regression that explains y in terms of z. When we do that we obtain

$$\hat{y} = -0.5083 - 6593z = -0.5083 - 6593\frac{1}{x}.$$

The fitted regression is shown in Figure 39. Apparently this curve is more adequate for describing the set of points that a straight line.

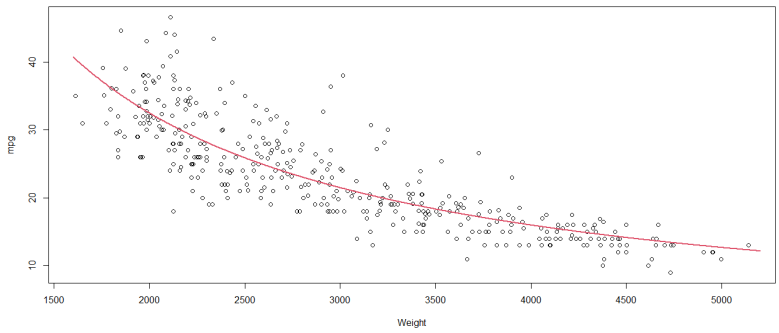

Figure 39: Scatter plot of weight (pounds) vs. autonomy (miles per gallon) of 392 automobiles.

With the method we have just seen it is simple to fit another functional form to the data. The desired functional form, however, must be specified. In our example, for instance, we decided in advance that we wanted to use a transformation of the type $1/x$.

6.2.2 Extrapolation

The least squares regression is the straight line that minimizes the SSE *for the set of available observations*. As we have seen this is a very interesting characteristic: it is simply impossible to find any other straight line that fits *the available data* better than the least squares regression. In other words, *given* the data we have found the best line.

However, more often than not, the interest when we fit a regression is to *extrapolate* the results to elements that we have not observed. For instance, in our example of the companies producing tables, we would like to be able to use the regression to predict the number of tables that will be produced by another company with, let us say, 20 workers, or, to extrapolate the conclusions to a broader population than just what we have observed. Well, unfortunately, strictly speaking, (except by pure luck) our fitted regression will not give the best line for that broader population of interest. It may be that it is close to the best one but it may be that it is far from it.

Let us illustrate this with an example. The third and fourth columns of Table 29 give $y =$ "closing price" and $x =$ "starting price" of the same twenty housing units considered in Example 57. The first two columns give the same measurements on twenty different housing units. Fitting a regression that explains the closing price y in terms of the starting price x using the set of units on the first two columns gives:

$$\hat{y} = 0.3706 + 0.9916 \cdot x$$

with $R^2 = 91.93\%$. The left panel of Figure 40 shows the fitted regression over the scatterplot of these two variables. The regression line fits the data pretty well. Now we would like to use the fitted regression to predict the closing price of another housing unit, which has a starting price of 1.795. We get $\hat{y} = 0.3706 + 0.9916 \cdot 1.795 = 2.15$.

Fitting an analogous regression with the second set of units gives:

$$\hat{y} = -0.1025 + 1.1105 \cdot x$$

with $R^2 = 99.72\%$. The right panel of Figure 40 shows the fitted regression over the scatterplot of these two variables. Using this regression to predict the closing price of the unit with starting price of 1.795 we get $\hat{y} = -0.1025 + 1.1105 \cdot 1.795 = 1.89$.

Both regressions fit the data pretty well, yet their predictions are quite different. Which one is better? In real practice we may never know. In this case I know the true closing

Set one		Set two	
Starting	Closing	Starting	Closing
3.595	3.91	3.59	3.9
3.595	3.86	2.995	3.17
3.595	3.95	3.495	3.83
3.695	4.04	3.495	3.8
3.695	4.00	3.500	3.77
3.695	4.05	2.195	2.36
3.695	4.06	2.195	2.32
3.695	4.06	2.195	2.32
3.695	4.08	3.450	3.7
3.775	4.11	1.895	1.94
3.795	4.10	2.395	2.54
3.795	4.13	1.699	1.83
3.795	4.16	2.195	2.36
3.795	4.14	2.995	3.17
3.795	4.16	3.295	3.61
3.795	4.16	2.395	2.54
3.850	4.18	3.795	4.09
3.875	4.23	3.490	3.74
3.895	4.23	3.095	3.39
4.000	4.29	2.300	2.49

Table 29: Starting and closing prices of two sets of twenty housing units.

price of that housing unit is 1.89, thus although both predictions were really good, the second regression performed extremely well at predicting. One may be tempted to say that as the R^2 of the second one is higher than that for the first one, this is an indication that that regression will work better. That is not necessarily true. Then what can we do in order to guarantee that our extrapolations are reliable? Although this is a topic that is beyond the scope of the course, in a few words, the answer is: sampling adequately. (The reader may be thinking that the second set was chosen purposively just to obtain a "perfect" prediction. In fact, that set was selected through a random experiment.)

A more specific advice is: extrapolating to values of the explanatory variable that are quite different from those that we observed is risky. Note that in the first case the x-values vary within 3.6 and 4 millions, but we are predicting for a housing unit with 1.795. This value is quite far from the observations we have used to fit the regression, then anything can happen. In the second case, the x-values vary between 1.7 and 3.8, a range that covers the value 1.795.