

THE TWO-VARIABLE LINEAR REGRESSION MODEL

Herman J. Bierens

Pennsylvania State University

November 20, 2008

1. *Introduction*

Suppose you are an economics or business major in a college close to the beach in the southern part of the US, for example southern California¹, where the weather is almost always nice the whole year around. In order to support yourself through college, you have started your own (weekend) business: an ice cream parlor on the beach. You have experienced that on hot weekends you usually sell more ice cream than on cold weekends. Also, you have recorded the average temperature and the sales of ice cream during eight weekends. Let Y_j be the sales of ice cream on weekend j , measured in \$100, and let X_j be the average temperature on weekend j , measured in units of 10 degrees Fahrenheit:

Table 1: *Ice cream data*

| Sales (unit = \$100) | Temperature (unit = 10 degrees) |
|----------------------|---------------------------------|
| $Y_1 = 8$ | $X_1 = 5$ |
| $Y_2 = 10$ | $X_2 = 7$ |
| $Y_3 = 8$ | $X_3 = 6$ |
| $Y_4 = 13$ | $X_4 = 8$ |
| $Y_5 = 15$ | $X_5 = 10$ |
| $Y_6 = 14$ | $X_6 = 9$ |
| $Y_7 = 11$ | $X_7 = 7$ |
| $Y_8 = 9$ | $X_8 = 8$ |

You want to use this information to forecast next weekend's sales of ice cream, given a good forecast of next weekend's temperature. Such a forecast of the sales will enable you to

¹ These lecture notes are based on lecture notes that I wrote while teaching at the University of California, San Diego, in the winter of 1987.

reduce your cost by adjusting your purchase of ice cream to the expected demand, because the ice cream you don't sell has to be thrown away.

Let your forecasting scheme be

$$\hat{Y} = \hat{\alpha} + \hat{\beta}.X,$$

i.e., given the temperature of X times 10 degrees and given the values of $\hat{\alpha}$ and $\hat{\beta}$, \hat{Y} times \$100 will be your forecast of the sales of ice cream. This forecasting scheme together with the points (X_j, Y_j) , $j = 1, 2, \dots, 8$, is plotted in Figure 1:

Y=Sales (unit: \$100) (t=1->8)

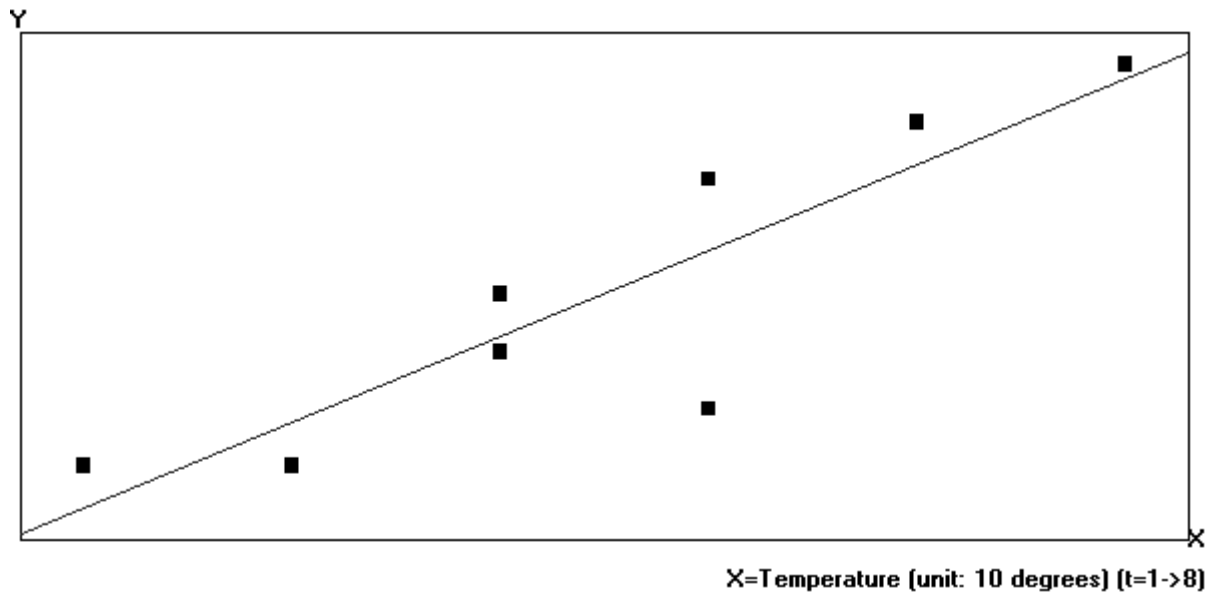


Figure 1 Scatter plot of (X_j, Y_j) , $j = 1, 2, \dots, 8$, together with the line $\hat{Y} = \hat{\alpha} + \hat{\beta}.X$.

The best values for $\hat{\alpha}$ and $\hat{\beta}$ are those for which the forecast error (= actual sales minus forecasted sales) is minimal. However, you do not know yet the actual sales in the next weekend, but you do know the actual sales in the eight weekends for which you have recorded your sales and the corresponding temperature. So what you could do is to forecast the sales of ice cream on each of these eight weekends and to determine $\hat{\alpha}$ and $\hat{\beta}$ such that the forecast errors are minimal. Because forecast errors can be positive and negative, as can be seen from Figure 1, the sum of the forecast errors is not a good measure of the performance of your forecasting

scheme, because large positive errors can be offset by large negative errors. Therefore, use the sum of squared errors as your measure of the accuracy of your forecasts:

$$Q(\hat{\alpha}, \hat{\beta}) = \sum_{j=1}^n (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta}X_j)^2,$$

where n is the sample size ($n = 8$ in our example), and minimize $Q(\hat{\alpha}, \hat{\beta})$ to $\hat{\alpha}$ and $\hat{\beta}$. The first-order conditions for a minimum are:

$$\begin{aligned} \frac{\partial Q(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = 0 &\Leftrightarrow \sum_{j=1}^n 2(Y_j - \hat{\alpha} - \hat{\beta}X_j)(-1) = 0 \\ \Leftrightarrow \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta}X_j) = 0 &\Leftrightarrow \sum_{j=1}^n Y_j - \sum_{j=1}^n \hat{\alpha} - \sum_{j=1}^n (\hat{\beta}X_j) = 0 \\ \Leftrightarrow \sum_{j=1}^n Y_j = n\hat{\alpha} + \hat{\beta}\sum_{j=1}^n X_j = 0 &\Leftrightarrow \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}, \end{aligned} \quad (1)$$

and

$$\begin{aligned} \frac{\partial Q(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = 0 &\Leftrightarrow \sum_{j=1}^n 2(Y_j - \hat{\alpha} - \hat{\beta}X_j)(-X_j) = 0 \\ \Leftrightarrow \sum_{j=1}^n (Y_jX_j - \hat{\alpha}X_j - \hat{\beta}X_j^2) = 0 &\Leftrightarrow \sum_{j=1}^n X_jY_j - \hat{\alpha}\sum_{j=1}^n X_j - \hat{\beta}\sum_{j=1}^n X_j^2 = 0 \\ \Leftrightarrow \sum_{j=1}^n X_jY_j = \hat{\alpha}\sum_{j=1}^n X_j + \hat{\beta}\sum_{j=1}^n X_j^2 &\Leftrightarrow \frac{1}{n}\sum_{j=1}^n X_jY_j = \hat{\alpha}\bar{X} + \hat{\beta}\frac{1}{n}\sum_{j=1}^n X_j^2 \end{aligned} \quad (2)$$

where $\bar{X} = (1/n)\sum_{j=1}^n X_j$ and $\bar{Y} = (1/n)\sum_{j=1}^n Y_j$ are the sample means of the X_j 's and Y_j 's, respectively.

The last equations in (1) and (2) are called the *normal equations*:

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}, \quad (3)$$

$$\frac{1}{n}\sum_{j=1}^n X_jY_j = \hat{\alpha}\bar{X} + \hat{\beta}\frac{1}{n}\sum_{j=1}^n X_j^2. \quad (4)$$

To solve these normal equations, substitute $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ in (4). Then we get

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^n X_jY_j &= (\bar{Y} - \hat{\beta}\bar{X})\frac{1}{n}\sum_{j=1}^n X_j + \hat{\beta}\frac{1}{n}\sum_{j=1}^n X_j^2 = \bar{Y}\bar{X} - \hat{\beta}\bar{X}^2 + \hat{\beta}\frac{1}{n}\sum_{j=1}^n X_j^2 \\ &= \bar{X}\bar{Y} + \hat{\beta}\left(\frac{1}{n}\sum_{j=1}^n X_j^2 - \bar{X}^2\right) \end{aligned}$$

hence

$$\frac{1}{n} \sum_{j=1}^n X_j Y_j - \bar{X} \cdot \bar{Y} = \hat{\beta} \left(\frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2 \right). \quad (5)$$

Equation (5) can also be written as

$$\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) = \hat{\beta} \left(\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 \right), \quad (6)$$

because

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) &= \frac{1}{n} \sum_{j=1}^n (X_j Y_j - \bar{X} \cdot Y_j - X_j \cdot \bar{Y} + \bar{X} \cdot \bar{Y}) \\ &= \frac{1}{n} \sum_{j=1}^n X_j Y_j - \bar{X} \cdot \frac{1}{n} \sum_{j=1}^n Y_j - \bar{Y} \cdot \frac{1}{n} \sum_{j=1}^n X_j + \bar{X} \cdot \bar{Y} = \frac{1}{n} \sum_{j=1}^n X_j Y_j - \bar{X} \cdot \bar{Y} \end{aligned} \quad (7)$$

and similarly

$$\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2. \quad (8)$$

Thus it follows from (5) that

$$\hat{\beta} = \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} \quad (9)$$

and from (3) and (9) that

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \cdot \bar{X} = \bar{Y} - \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} \cdot \bar{X}. \quad (10)$$

In the ice cream parlor case we have

$$n = 8, \bar{X} = 7.5, \bar{Y} = 11, \sum_{j=1}^n X_j^2 = 468, \sum_{j=1}^n X_j Y_j = 687,$$

$$\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) = \sum_{j=1}^n X_j Y_j - n \cdot \bar{X} \cdot \bar{Y} = 27,$$

$$\sum_{j=1}^n (X_j - \bar{X})^2 = \sum_{j=1}^n X_j^2 - n \cdot \bar{X}^2 = 18,$$

so that

$$\hat{\beta} = 1.5, \quad \hat{\alpha} = -0.25.$$

Thus, our best forecasting scheme is $\hat{Y} = -0.25 + 1.5X$. This is the straight line in Figure 1.

Now suppose that the forecast of next weekend's temperature is 75 degrees. Then $X = 7.5$, hence $\hat{Y} = -0.25 + 1.5(7.5) = 11$. Therefore, the best forecast of next weekend's sales is:
 $\hat{Y} \times \$100 = \$1,100$.

2. *The two-variable linear regression model.*

In order to answer the question how good this forecast is, we have to make assumptions about the true relationship between the *dependent variable* Y_j and the *independent variable* X_j , (also called the *explanatory variable*). The true relationship we are going to assume is the two-variable linear regression model:

$$Y_j = \alpha + \beta \cdot X_j + U_j, \quad j = 1, 2, \dots, n. \quad (11)$$

The U_j 's are random error variables, called *error terms*, for which we assume:

Assumption I: *The U_j 's are independent and identically distributed (i.i.d) random variables.*

Assumption II: *The mathematical expectation of U_j equals zero: $E(U_j) = 0$ for $j = 1, 2, \dots, n$.*

Assumption III: *The variance $\sigma^2 = \text{var}(U_j) = E[U_j - E(U_j)]^2 = E(U_j^2)$ of the U_j 's is constant and finite.*

Regarding the explanatory variables X_j we shall assume for the time being that

Assumption IV: *The independent variables X_j are non-random.*

This assumption is not strictly necessary, and is actually quite unrealistic in economics, but will be made for the sake of convenience, as it will ease the argument. Finally, we will assume that the errors are normally distributed:

Assumption V: *The errors U_j 's are $N(0, \sigma^2)$ distributed.*

In particular, we shall need this assumption in order to say something about the reliability of the forecast. These assumption will be relaxed later on.

3. *The properties of $\hat{\alpha}$ and $\hat{\beta}$.*

Although we have motivated model (11) by the need to forecast out-of-sample values of the dependent variables Y_j , a linear regression model is more often used for testing economic hypotheses. For example, let Y_j be the hourly wage of wage earner j in a random sample of size n of wage earners, and let X_j be a gender indicator, say $X_j = 1$ if person j is a female, and $X_j = 0$ if person j is a male. If you suspect gender discrimination in the workplace, you can test this suspicion by testing the null hypothesis that $\beta = 0$ (no gender discrimination) against one of three possible alternative hypotheses:

- (a) $\beta \neq 0$: women are paid different hourly wages than men, either higher or lower;
- (b) $\beta > 0$: women are paid higher hourly wages than men;
- (c) $\beta < 0$: women are paid lower hourly wages than men.

The last hypothesis is usually what is meant by “gender discrimination.” A test for the null hypothesis $\beta = 0$ against one of these alternative hypotheses can be based on the estimate $\hat{\beta}$ of β , provided that we know how $\hat{\beta}$ is related to β .

It will be shown below that $\hat{\alpha}$ and $\hat{\beta}$ are indeed reasonable approximations of α and β , respectively, possessing particular desirable properties.

In general an *estimator* of an unknown parameter is a function of the data that serves as an approximation of the parameter involved. It follows from (9) and (10) that $\hat{\alpha}$ and $\hat{\beta}$ are functions of the data, $(Y_1, X_1), \dots, (Y_n, X_n)$. Because $\hat{\alpha}$ and $\hat{\beta}$ will be used as approximations of α and β , respectively, and were obtained by minimizing the squared errors, we will call $\hat{\alpha}$ and $\hat{\beta}$ the

Ordinary² Least Squares (OLS) estimators of α and β , respectively.

3.1 Unbiasedness

The first property of $\hat{\alpha}$ and $\hat{\beta}$ is that they are *unbiased* estimators of α and β :

Proposition 1. *Under Assumptions II and IV the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are unbiased, which means that $E[\hat{\alpha}] = \alpha$ and $E[\hat{\beta}] = \beta$.*

Proof: Before we turn to the actual proof, we need some results regarding summations:

$$\sum_{j=1}^n (X_j - \bar{X}) = \sum_{j=1}^n X_j - \sum_{j=1}^n \bar{X} = n \cdot \bar{X} - n \cdot \bar{X} = 0, \quad (12)$$

and therefore

$$\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}) = \sum_{j=1}^n (X_j - \bar{X})Y_j - \bar{Y} \cdot \sum_{j=1}^n (X_j - \bar{X}) = \sum_{j=1}^n (X_j - \bar{X})Y_j. \quad (13)$$

Now consider the formula (9) for $\hat{\beta}$. It follows from (9) and (13) that

$$\hat{\beta} = \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{\sum_{j=1}^n (X_j - \bar{X})Y_j}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (14)$$

Next, substitute model (11) in (14). Then

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{j=1}^n (X_j - \bar{X})(\alpha + \beta X_j + U_j)}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{\alpha \sum_{j=1}^n (X_j - \bar{X}) + \beta \sum_{j=1}^n (X_j - \bar{X})X_j + \sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \beta \cdot \frac{\sum_{j=1}^n (X_j - \bar{X})X_j}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} = \beta + \frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2}, \end{aligned} \quad (15)$$

where the last step follows from the fact that similar to (13),

² The estimators $\hat{\alpha}$ and $\hat{\beta}$ are called "Ordinary" least squares estimators to distinguish them from "Nonlinear" least squares estimators.

$$\sum_{j=1}^n (X_j - \bar{X})^2 = \sum_{j=1}^n (X_j - \bar{X})X_j. \quad (16)$$

Now take the mathematical expectation at both sides of (15). Then,

$$E[\hat{\beta}] = \beta + E\left(\frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2}\right) = \beta + \frac{\sum_{j=1}^n (X_j - \bar{X})E(U_j)}{\sum_{j=1}^n (X_j - \bar{X})^2} = \beta, \quad (17)$$

because taking the mathematical expectation of a constant (β) does not effect that constant, and taking the mathematical expectation of a linear function of random variables is equal to taking the linear function of the mathematical expectation of these random variables. The last conclusion in (17) follows from assumption II, and the second step in (17) can be taken because we have assumed that the X_j 's are non-random (assumption IV).

Now consider $\hat{\alpha}$. We have already established that $\hat{\alpha} = \bar{Y} - \hat{\beta} \cdot \bar{X}$. Substituting the right-hand side of (15) for $\hat{\beta}$ in this equation yields

$$\hat{\alpha} = \bar{Y} - \left(\beta + \frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \right) \cdot \bar{X} = \bar{Y} - \beta \cdot \bar{X} - \frac{\sum_{j=1}^n \bar{X}(X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (18)$$

Next, substitute

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j = \frac{1}{n} \sum_{j=1}^n (\alpha + \beta X_j + U_j) = \alpha + \beta \cdot \bar{X} + \frac{1}{n} \sum_{j=1}^n U_j$$

in (18). Then,

$$\hat{\alpha} = \alpha + \frac{1}{n} \sum_{j=1}^n U_j - \frac{\sum_{j=1}^n \bar{X}(X_j - \bar{X})U_j}{\sum_{i=1}^n (X_i - \bar{X})^2} = \alpha + \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \cdot U_j. \quad (19)$$

Similar as for $\hat{\beta}$ we therefore have:

$$E[\hat{\alpha}] = \alpha + \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) E[U_j] = \alpha. \quad (20)$$

This completes the proof of Proposition 1.

3.2 The variances of $\hat{\alpha}$ and $\hat{\beta}$.

Our next issue concerns the variances of $\hat{\alpha}$ and $\hat{\beta}$. For deriving these variances the following two lemmas are convenient.

Lemma 1. Let U_1, U_2, \dots, U_n be independent random variables with zero mathematical expectation (thus $E(U_j) = 0$) and variance σ^2 . (Thus $E(U_j - E(U_j))^2 = E(U_j)^2 = \sigma^2$). Let v_1, v_2, \dots, v_n and w_1, w_2, \dots, w_n be given constants. Then $E[(\sum_{j=1}^n v_j U_j)(\sum_{j=1}^n w_j U_j)] = \sigma^2 \sum_{j=1}^n v_j w_j$.

Proof. We have

$$\begin{aligned} E\left[\left(\sum_{j=1}^n v_j U_j\right)\left(\sum_{j=1}^n w_j U_j\right)\right] &= E\left[\left(\sum_{i=1}^n v_i U_i\right)\left(\sum_{j=1}^n w_j U_j\right)\right] = E\left[\sum_{i=1}^n \sum_{j=1}^n v_i w_j U_i U_j\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n v_i w_j E(U_i U_j) = \sum_{j=1}^n v_j w_j \sigma^2, \end{aligned} \quad (21)$$

where the last equality in (21) follows from

$$\begin{aligned} E(U_i U_j) &= E(U_i)E(U_j) = 0 \text{ if } i \neq j, \\ &= E(U_j^2) = \sigma^2 \text{ if } i = j. \end{aligned} \quad (22)$$

Note that if we choose $v_j = w_j$ for $j = 1, 2, \dots, n$ in Lemma 1 then it reads:

Lemma 2. Let U_1, U_2, \dots, U_n be independent random variables with zero mathematical expectation and variance σ^2 . Let w_1, w_2, \dots, w_n be given constants. Then

$$E[(\sum_{j=1}^n w_j U_j)^2] = \sigma^2 \sum_{j=1}^n w_j^2.$$

We shall now prove:

Proposition 2. Under the assumptions I - IV,

$$\begin{aligned} \text{var}(\hat{\alpha}) &= \frac{\sigma^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j - \bar{X})^2} = \sigma_{\hat{\alpha}}^2, \text{ say, } \text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{j=1}^n (X_j - \bar{X})^2} = \sigma_{\hat{\beta}}^2, \text{ say, and} \\ \text{cov}(\hat{\alpha}, \hat{\beta}) &= \frac{-\sigma^2 \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}. \end{aligned} \quad (23)$$

Proof. From formula (15) and Lemma 2 it follows:

$$\begin{aligned}
\text{var}(\hat{\beta}) &= E[(\hat{\beta}-\beta)^2] = E\left[\left(\sum_{j=1}^n \left(\frac{X_j-\bar{X}}{\sum_{i=1}^n (X_i-\bar{X})^2}\right) U_j\right)^2\right] = \sigma^2 \sum_{j=1}^n \left(\frac{X_j-\bar{X}}{\sum_{i=1}^n (X_i-\bar{X})^2}\right)^2 \\
&= \sigma^2 \frac{\sum_{j=1}^n (X_j-\bar{X})^2}{\left(\sum_{i=1}^n (X_i-\bar{X})^2\right)^2} = \sigma^2 \frac{\sum_{j=1}^n (X_j-\bar{X})^2}{\left(\sum_{j=1}^n (X_j-\bar{X})^2\right)^2} = \frac{\sigma^2}{\sum_{j=1}^n (X_j-\bar{X})^2}.
\end{aligned} \tag{24}$$

Similarly, from formula (19) and Lemma 2 it follows that

$$\begin{aligned}
\text{var}(\hat{\alpha}) &= E[(\hat{\alpha}-\alpha)^2] = E\left[\left(\sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j-\bar{X})}{\sum_{i=1}^n (X_i-\bar{X})^2}\right) U_j\right)^2\right] = \sigma^2 \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j-\bar{X})}{\sum_{i=1}^n (X_i-\bar{X})^2}\right)^2 \\
&= \sigma^2 \sum_{j=1}^n \left(\frac{1}{n^2} - \frac{2\bar{X}(X_j-\bar{X})}{n \sum_{i=1}^n (X_i-\bar{X})^2} + \frac{\bar{X}^2 (X_j-\bar{X})^2}{\left(\sum_{i=1}^n (X_i-\bar{X})^2\right)^2}\right) \\
&= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{X}(1/n)\sum_{j=1}^n (X_j-\bar{X})}{\sum_{i=1}^n (X_i-\bar{X})^2} + \frac{\bar{X}^2 \sum_{j=1}^n (X_j-\bar{X})^2}{\left(\sum_{i=1}^n (X_i-\bar{X})^2\right)^2}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^n (X_j-\bar{X})^2}\right) = \sigma^2 \left(\frac{(1/n)\sum_{j=1}^n (X_j-\bar{X})^2 + \bar{X}^2}{\sum_{j=1}^n (X_j-\bar{X})^2}\right) = \frac{\sigma^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j-\bar{X})^2},
\end{aligned} \tag{25}$$

where the last equality follows from the fact that $(1/n)\sum_{j=1}^n (X_j-\bar{X})^2 = (1/n)\sum_{j=1}^n X_j^2 - \bar{X}^2$.

Finally, it follows from Lemma 1 and the formulas (15) and (19) that

$$\begin{aligned}
\text{cov}(\hat{\alpha}, \hat{\beta}) &= E[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)] = E\left[\left(\sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j-\bar{X})}{\sum_{i=1}^n (X_i-\bar{X})^2}\right) U_j\right) \left(\sum_{j=1}^n \left(\frac{X_j-\bar{X}}{\sum_{i=1}^n (X_i-\bar{X})^2}\right) U_j\right)\right] \\
&= \sigma^2 \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j-\bar{X})}{\sum_{i=1}^n (X_i-\bar{X})^2}\right) \left(\frac{(X_j-\bar{X})}{\sum_{i=1}^n (X_i-\bar{X})^2}\right)
\end{aligned}$$

$$= \sigma^2 \left(\frac{(1/n) \sum_{j=1}^n (X_j - \bar{X}) - \bar{X} \sum_{j=1}^n (X_j - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} \right) = \frac{-\sigma^2 \cdot \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}.$$

3.3 Normality of $\hat{\alpha}$ and $\hat{\beta}$.

If we also assume normality of the error terms U_j then $\hat{\alpha}$ and $\hat{\beta}$ are also normally distributed. This result follows from the following lemma.

Lemma 3. *Let Z_1, Z_2, \dots, Z_n be independent $N(\mu, \sigma^2)$ distributed random variables and let w_1, \dots, w_n be constants. Then $\sum_{j=1}^n w_j Z_j$ is distributed $N[(\sum_{j=1}^n w_j) \mu, (\sum_{j=1}^n w_j^2) \sigma^2]$.*

The proof of this lemma requires advanced probability theory and is therefore omitted.

It follows now straightforwardly from Proposition 2, Lemma 3, and the formulas (15) and (19) that:

Proposition 3. *Under the assumptions I - V,*

$$\hat{\alpha} - \alpha \sim N \left[0, \frac{\sigma^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j - \bar{X})^2} \right], \quad \hat{\beta} - \beta \sim N \left[0, \frac{\sigma^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right], \quad (26)$$

where “ \sim ” is the symbol for “is distributed as.”

Moreover, applying Lemma 3 again for $n = 1$ it follows from (26) (Exercise: Why?) that

Proposition 4. *Under the assumptions I - V,*

$$\frac{(\hat{\alpha} - \alpha) \sqrt{n \sum_{j=1}^n (X_j - \bar{X})^2}}{\sigma \cdot \sqrt{\sum_{j=1}^n X_j^2}} \sim N[0, 1], \quad \frac{(\hat{\beta} - \beta) \sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\sigma} \sim N[0, 1]. \quad (27)$$

These results play a key-role in testing hypotheses about α and β . The only problem that prevents

us from using these results for testing is that σ is unknown. This problem will be addressed in the next section.

4. *How to estimate the error variance σ^2 ?*

If α and β were known then we could estimate σ^2 by

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (Y_j - \alpha - \beta \cdot X_j)^2 = \frac{1}{n} \sum_{j=1}^n U_j^2. \quad (28)$$

However, α and β are not known, but we do have OLS estimators of α and β . This suggests to replace α and β in (28) by their OLS estimators:

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta} \cdot X_j)^2 = \frac{1}{n} \sum_{j=1}^n \hat{U}_j^2, \quad (29)$$

where

$$\hat{U}_j = Y_j - \hat{\alpha} - \hat{\beta} \cdot X_j \quad (30)$$

is called the regression *residual*.

In order to derive the properties of the estimator (29), observe first from (3) and (30) that

$$\frac{1}{n} \sum_{j=1}^n \hat{U}_j = \bar{Y} - \hat{\alpha} - \hat{\beta} \cdot \bar{X} = 0 \quad (31)$$

so that we can write

$$\hat{U}_j = \hat{U}_j - \frac{1}{n} \sum_{i=1}^n \hat{U}_i = (Y_j - \bar{Y}) - \hat{\beta} \cdot (X_j - \bar{X}). \quad (32)$$

Next, observe from (11) that $Y_j - \bar{Y} = U_j - \bar{U} + \beta \cdot (X_j - \bar{X})$, where $\bar{U} = (1/n) \sum_{j=1}^n U_j$.

Substituting the former equation in (32) yields

$$\hat{U}_j = (U_j - \bar{U}) - (\hat{\beta} - \beta)(X_j - \bar{X}), \quad (33)$$

hence

$$\begin{aligned}
\sum_{j=1}^n \hat{U}_j^2 &= \sum_{j=1}^n \left((U_j - \bar{U}) - (\hat{\beta} - \beta)(X_j - \bar{X}) \right)^2 \\
&= \sum_{j=1}^n (U_j - \bar{U})^2 - 2(\hat{\beta} - \beta) \sum_{j=1}^n (X_j - \bar{X})(U_j - \bar{U}) + (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2 \\
&= \sum_{j=1}^n (U_j - \bar{U})^2 - 2(\hat{\beta} - \beta) \sum_{j=1}^n (X_j - \bar{X})U_j + (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2,
\end{aligned} \tag{34}$$

where the last equality follows from the fact that $\sum_{j=1}^n (X_j - \bar{X})\bar{U} = 0$. It follows from (15), (34) and the equality $\sum_{j=1}^n (U_j - \bar{U})^2 = \sum_{j=1}^n U_j^2 - n\bar{U}^2$ that

$$\begin{aligned}
\sum_{j=1}^n \hat{U}_j^2 &= \sum_{j=1}^n (U_j - \bar{U})^2 - (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2 = \sum_{j=1}^n U_j^2 - n\bar{U}^2 - (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2. \\
&= \sum_{j=1}^n U_j^2 - \frac{1}{n} \left(\sum_{i=1}^n U_i \right)^2 - (\hat{\beta} - \beta)^2 \sum_{j=1}^n (X_j - \bar{X})^2.
\end{aligned} \tag{35}$$

Taking expectations and using Lemma 2 and Proposition 2 it follows now from (35) that

$$\begin{aligned}
E[\sum_{j=1}^n \hat{U}_j^2] &= \sum_{j=1}^n E[U_j^2] - \frac{1}{n} E\left[\left(\sum_{i=1}^n U_i \right)^2 \right] - (E(\hat{\beta} - \beta)^2) \sum_{j=1}^n (X_j - \bar{X})^2 \\
&= n\sigma^2 - \sigma^2 - \sigma^2 = (n-2)\sigma^2.
\end{aligned} \tag{36}$$

This result suggests to use

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^n \hat{U}_j^2 \tag{37}$$

as an estimator of σ^2 instead of (29), because by (36), $\hat{\sigma}^2$ is an unbiased estimator of σ^2 :

$$E[\hat{\sigma}^2] = \sigma^2. \tag{38}$$

The sum $\sum_{j=1}^n \hat{U}_j^2$ is called the Sum of Squares Residuals, shortly *SSR*, and $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is called the Standard Error of the Residuals, shortly *SER*. Thus,

$$SSR = \sum_{j=1}^n \hat{U}_j^2, \quad SER = \sqrt{\frac{\sum_{j=1}^n \hat{U}_j^2}{n-2}} = \sqrt{\frac{SSR}{n-2}} (= \hat{\sigma}). \tag{39}$$

Finally, note that the sum of squared residuals can be computed as follows:

$$\begin{aligned}
SSR &= \sum_{j=1}^n \hat{U}_j^2 = \sum_{j=1}^n (Y_j - \hat{\alpha} - \hat{\beta} \cdot X_j)^2 = \sum_{j=1}^n (Y_j - (\bar{Y} - \hat{\beta} \cdot \bar{X}) - \hat{\beta} \cdot X_j)^2 \\
&= \sum_{j=1}^n \left((Y_j - \bar{Y}) - \hat{\beta} \cdot (X_j - \bar{X}) \right)^2 \\
&= \sum_{j=1}^n (Y_j - \bar{Y})^2 - 2\hat{\beta} \sum_{j=1}^n (Y_j - \bar{Y})(X_j - \bar{X}) + \hat{\beta}^2 \sum_{j=1}^n (X_j - \bar{X})^2 \\
&= \sum_{j=1}^n (Y_j - \bar{Y})^2 - \hat{\beta}^2 \sum_{j=1}^n (X_j - \bar{X})^2.
\end{aligned} \tag{40}$$

5. *Standard errors, t-values and p-values of the OLS estimators*

The variances of $\hat{\alpha}$ and $\hat{\beta}$ can now be estimated by replacing σ^2 in (23) by $\hat{\sigma}^2$:

$$\begin{aligned}
\text{Estimated var}(\hat{\alpha}) &= \frac{\hat{\sigma}^2 \sum_{j=1}^n X_j^2}{n \sum_{j=1}^n (X_j - \bar{X})^2} = \hat{\sigma}_{\hat{\alpha}}^2, \text{ say,} \\
\text{Estimated var}(\hat{\beta}) &= \frac{\hat{\sigma}^2}{\sum_{j=1}^n (X_j - \bar{X})^2} = \hat{\sigma}_{\hat{\beta}}^2, \text{ say.}
\end{aligned} \tag{41}$$

Then $\hat{\sigma}_{\hat{\alpha}} = \sqrt{\hat{\sigma}_{\hat{\alpha}}^2}$ is called the standard error of $\hat{\alpha}$, also denoted by $SE(\hat{\alpha})$, and $\hat{\sigma}_{\hat{\beta}} = \sqrt{\hat{\sigma}_{\hat{\beta}}^2}$ is called the standard error of $\hat{\beta}$, also denoted by $SE(\hat{\beta})$.

If we replace σ in Proposition 4 by the SER, $\hat{\sigma}$, the standard normality results involved change:

Proposition 5. *Under the assumptions I - V,*

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} = \frac{(\hat{\alpha} - \alpha) \sqrt{n \sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma} \cdot \sqrt{\sum_{j=1}^n X_j^2}} \sim t_{n-2}, \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} = \frac{(\hat{\beta} - \beta) \sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma}} \sim t_{n-2}. \tag{42}$$

The proof of Proposition 5 is based on the fact that under these assumptions, SSR/σ^2 is distributed χ_{n-2}^2 and is independent of $\hat{\alpha}$ and $\hat{\beta}$, but the proofs of these propositions is beyond the scope of

this course.

Because for large degrees of freedom the t distribution is approximately equal to the standard normal distribution, and due to the central limit theorem Proposition 4 holds if n is large and the errors are not normally distributed, we also have:

Proposition 6. *If the sample size n is large then under the assumptions I - IV we have approximately,*

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} = \frac{(\hat{\alpha} - \alpha)\sqrt{n\sum_{j=1}^n(X_j - \bar{X})^2}}{\hat{\sigma} \cdot \sqrt{\sum_{j=1}^n X_j^2}} \sim N(0,1), \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} = \frac{(\hat{\beta} - \beta)\sqrt{\sum_{j=1}^n(X_j - \bar{X})^2}}{\hat{\sigma}} \sim N(0,1). \quad (43)$$

The results in Proposition 5 now enable us to test hypothesis about α and β . In particular the null hypothesis that $\beta = 0$ is of importance, because this hypothesis implies that X has no effect on Y . The test statistic for testing this hypothesis is the t -value (or t -statistic) of $\hat{\beta}$:

$$\hat{t}_{\hat{\beta}} (= t\text{-value of } \hat{\beta}) \stackrel{\text{def.}}{=} \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}\sqrt{\sum_{j=1}^n(X_j - \bar{X})^2}}{\hat{\sigma}} \sim t_{n-2} \text{ if } \beta = 0. \quad (44)$$

If $\beta > 0$ and the sample size $n \rightarrow \infty$ then the t -value of $\hat{\beta}$ converges in probability to $+\infty$, and if $\beta < 0$ and $n \rightarrow \infty$ then the t -value of $\hat{\beta}$ converges in probability to $-\infty$. Moreover, if the sample size n is large we may use the standard normal distribution instead of the t distribution to find critical values of the test.

Similarly,

$$\hat{t}_{\hat{\alpha}} (= t\text{-value of } \hat{\alpha}) \stackrel{\text{def.}}{=} \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}} \sim t_{n-2} \text{ if } \alpha = 0. \quad (45)$$

However, the hypothesis $\alpha = 0$ is often of no interest.

In the ice cream example,

$$\sum_{j=1}^n(X_j - \bar{X})^2 = 18 \Rightarrow \sqrt{\sum_{j=1}^n(X_j - \bar{X})^2} = \sqrt{18} \approx 4.24264,$$

$$\sum_{j=1}^n(Y_j - \bar{Y})^2 = \sum_{j=1}^n Y_j^2 - n \cdot \bar{Y}^2 = 1020 - 8 \times 11^2 = 52$$

and by (40),

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{j=1}^n \hat{U}_j^2 = \frac{1}{n-2} \sum_{j=1}^n (Y_j - \bar{Y})^2 - \hat{\beta}^2 \frac{1}{n-2} \sum_{j=1}^n (X_j - \bar{X})^2 \\ &= \frac{52 - (1.5)^2 \cdot 18}{8-2} = \frac{11.5}{6} \approx 1.916667 \Rightarrow \hat{\sigma} \approx 1.384437\end{aligned}$$

Hence,

$$\hat{t}_{\hat{\beta}} = \frac{\hat{\beta} \sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}{\hat{\sigma}} = \frac{1.5 \times 4.24264}{1.384437} \approx 4.597 \quad (46)$$

Assuming that the conditions of Proposition 5 hold, the null hypothesis $H_0: \beta = 0$ can be tested against the alternative hypothesis $H_1: \beta \neq 0$ using the two-sided t-test at say the 5% significance level, as follows. Under the null hypothesis, (46) is a random drawing from the t distribution with $n-2 = 6$ degrees of freedom. Look up in the table of the t distribution the value t_* such that for $T \sim t_6$, $P[|T| > t_*] = 0.05$. This value is $t_* = 2.447$. Then accept the null hypothesis if $-t_* = -2.447 \leq \hat{t}_{\hat{\beta}} \leq 2.447 = t_*$, and reject the null hypothesis in favor of the alternative hypothesis if $|\hat{t}_{\hat{\beta}}| > t_* = 2.447$. Thus, in the ice cream example we reject the null hypothesis $H_0: \beta = 0$ because $|\hat{t}_{\hat{\beta}}| = 4.597 > 2.447 = t_*$.

This test is illustrated in Figure 2 below. The curved line in Figure 2 is the density of the t distribution with 6 degrees of freedom. The grey areas are each 0.025, so that the total grey area is 0.05.

The null hypothesis $H_0: \beta = 0$ can be tested against the alternative hypothesis $H_1: \beta > 0$ at the 5% significance level by the right-sided t-test. Now look up in the table of the t distribution the value t_* such that for $T \sim t_6$, $P[T > t_*] = 0.05$. This value corresponds to the critical value of the two-sided t-test at the 10% significance level: $t_* = 1.943$. Then accept the null hypothesis if $\hat{t}_{\hat{\beta}} \leq t_* = 1.943$, and reject the null hypothesis in favor of the alternative hypothesis if $\hat{t}_{\hat{\beta}} > t_* = 1.943$. Thus, in the ice cream case we reject the null hypothesis $H_0: \beta = 0$ in favor of the alternative hypothesis $H_1: \beta > 0$.

This right-sided t-test is illustrated in Figure 3 below. Again, the curved line in Figure 3 is the density of the t distribution with 6 degrees of freedom, and the grey area is 0.05.

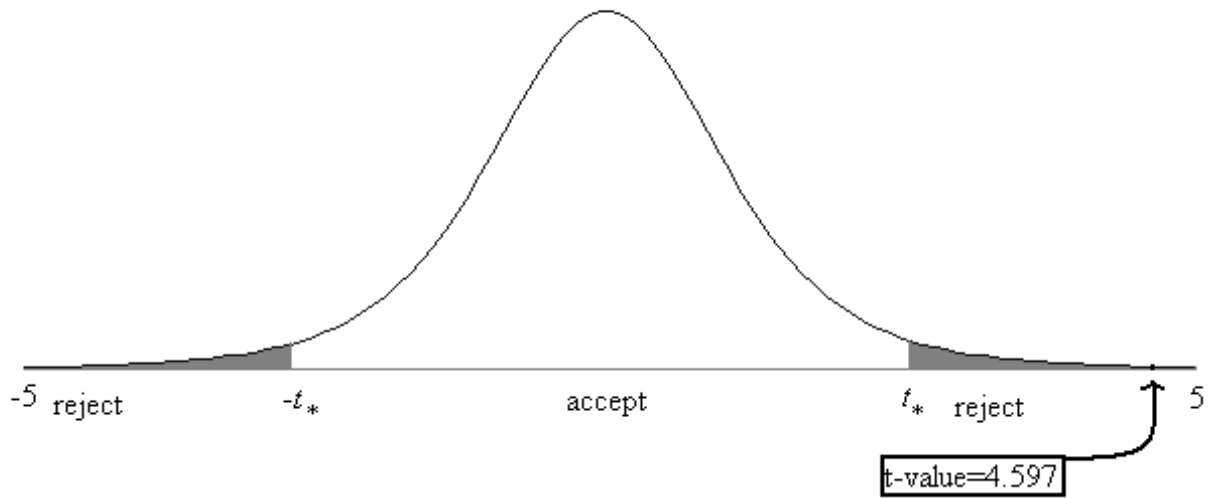


Figure 2 Two-sided t-test of $H_0: \beta = 0$ against the alternative hypothesis $H_1: \beta \neq 0$.

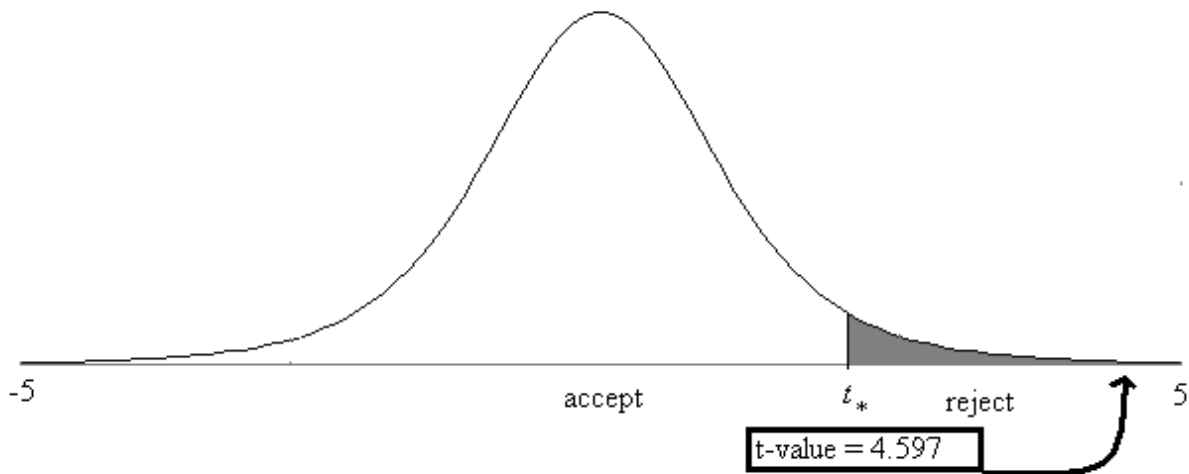


Figure 3 Right-sided t-test of $H_0: \beta = 0$ against the alternative hypothesis $H_1: \beta > 0$.

If the sample size n is large, so that $\hat{t}_{\hat{\beta}} \sim N(0,1)$ if $\beta = 0$, then an alternative way of testing the null hypothesis $\beta = 0$ against the alternative hypothesis $\beta \neq 0$ is to use the (two-sided) p-value:

$$\hat{p}_{\hat{\beta}} (= p\text{-value of } \hat{\beta}) \stackrel{\text{def.}}{=} P[|U| > |\hat{t}_{\hat{\beta}}|], \text{ where } U \sim N(0,1). \quad (47)$$

For example, if $\hat{p}_{\hat{\beta}} < 0.05$ we reject the null hypothesis $\beta = 0$ in favor of the alternative hypothesis $\beta \neq 0$ at the 5% significance level, and if $\hat{p}_{\hat{\beta}} \geq 0.05$ we accept the null hypothesis $\beta = 0$. The p-value for $\hat{\alpha}$ is defined and used similarly.

Although a t-value is a test statistics of the null hypothesis that the corresponding coefficient in the regression model is zero, it is quite easy to rebuild the t-value for testing other null hypotheses, as follows. Suppose you want to test the null hypothesis that $\beta = \beta_0$, where β_0 is a given number, for example $\beta_0 = 1$. Then

$$\frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} - \frac{\beta_0}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} - \frac{\beta_0 \hat{\beta}}{\hat{\beta} \hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \left(1 - \frac{\beta_0}{\hat{\beta}} \right) = \frac{\hat{\beta} - \beta_0}{\hat{\beta}} \cdot \hat{t}_{\hat{\beta}}, \quad (48)$$

so that by Proposition 5,

$$\hat{t}_{\hat{\beta}, \beta = \beta_0} = \frac{\hat{\beta} - \beta_0}{\hat{\beta}} \cdot \hat{t}_{\hat{\beta}} \sim t_{n-2}. \quad (49)$$

For example, suppose we want to test the null hypothesis $H_0: \beta = 1$ in the ice cream case. Then

$$\hat{t}_{\hat{\beta}, \beta = 1} = \frac{\hat{\beta} - \beta_0}{\hat{\beta}} \cdot \hat{t}_{\hat{\beta}} = \frac{1.5 - 1}{1.5} \times 4.597 \approx 1.532, \quad (50)$$

which under the null hypothesis $H_0: \beta = 1$ is a random drawing from the t distribution with 6 degrees of freedom. Note that the value of this test statistic is in the acceptance regions in Figures 2 and 3.

This trick is useful if the econometric software you are using only reports the t-values but not the standard errors.³ If the standard errors are reported, you can compute $\hat{t}_{\hat{\beta}, \beta = \beta_0}$ directly as $\hat{t}_{\hat{\beta}, \beta = \beta_0} = (\hat{\beta} - \beta_0) / \hat{\sigma}_{\hat{\beta}}$. Of course, if only the standard errors are reported and not the t-values you can compute the t-value of $\hat{\beta}$ as $\hat{t}_{\hat{\beta}} = \hat{\beta} / \hat{\sigma}_{\hat{\beta}}$.

³ This was the case in *EasyReg International*. However, in the February 15, 2004, upgrade the standard errors of the OLS estimators are now also reported.

6. *The R^2*

The R^2 of a regression model compares the sum of squared residuals (SSR) of the model with the SSR of a “regression model” without regressors:

$$Y_j = \alpha + U_j, \quad j = 1, 2, \dots, n. \quad (51)$$

It is easy to verify that the OLS estimator $\tilde{\alpha}$ of α is just the sample mean of the Y_j 's:

$$\tilde{\alpha} = \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j.$$

Therefore, the SSR of “regression model” (51) is $\sum_{j=1}^n (Y_j - \bar{Y})^2$, which is called the Total Sum of Squares (TSS):

$$TSS = \sum_{j=1}^n (Y_j - \bar{Y})^2. \quad (53)$$

The R^2 is now defined as:

$$R^2 \stackrel{\text{def.}}{=} 1 - \frac{SSR}{TSS}. \quad (54)$$

The R^2 is always between zero and one, because $SSR \leq TSS$. (*Exercise: Why?*) If $SSR = TSS$, so that $R^2 = 0$, then model (51) explains the dependent variable Y_j 's equally well as model (11). In other words, the explanatory variables X_j in (11) do not matter: $\beta = 0$. The other extreme is where $R^2 = 1$, which corresponds to $SSR = 0$. Then the dependent variable Y_j in model (11) is completely explained by X_j , without error: $Y_j \equiv \alpha + \beta X_j$. Thus, the R^2 measures how well the explanatory variables X_j are able to explain the corresponding dependent variables Y_j . For example, in the ice cream case, $SSR = 11.5$ and $TSS = 52$, hence $R^2 = 0.778846$. Loosely speaking, this means that about 78% of the variation of ice cream sales can be explained by the variation in temperature.

7. *Presenting regression results*

When you need to report regression results you should include, next to the OLS estimates of course, either the corresponding t-values or the standard errors, the sample size n , the standard error of the residuals (SER), and the R^2 , because this information will enable the reader to judge your results. For example our ice cream estimation results should be displayed as either

$$\text{Sales} = -0.25 + 1.5\text{Temp.}, \quad n = 8, \quad \text{SER} = 1.384437, \quad R^2 = 0.778846$$

$$(-0.100) \quad (4.597)$$

(t-values between brackets)

or

$$\text{Sales} = -0.25 + 1.5\text{Temp.}, \quad n = 8, \quad \text{SER} = 1.384437, \quad R^2 = 0.778846$$

$$(2.49583) \quad (0.32632)$$

(standard errors between brackets)

It is helpful to the reader if you would indicate whether you have displayed the t-values between brackets or the standard errors, but you only need to mention this once.

8. *Out-of-sample forecasting*

The linear regression model was introduced as a forecasting scheme. The question we now address is: How reliable is an out-of-sample forecast?

Consider the linear regression model (11), and suppose we observe X_{n+1} . Then the forecast of Y_{n+1} is $\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} \cdot X_{n+1}$, where the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are computed on the basis of the observations for $j = 1, 2, \dots, n$. The actual but unknown value of Y_{n+1} is $Y_{n+1} = \alpha + \beta \cdot X_{n+1} + U_{n+1}$, so that the forecast error is:

$$Y_{n+1} - \hat{Y}_{n+1} = U_{n+1} - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) \cdot X_{n+1} = U_{n+1} - \sum_{j=1}^n \left(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \cdot U_j$$

$$- \sum_{j=1}^n \left(\frac{X_{n+1}(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j = U_{n+1} - \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \cdot U_j. \quad (55)$$

where the second equality in (55) follows from (15) and (19). It follows now from Lemma 3 that under assumptions I - V, $Y_{n+1} - \hat{Y}_{n+1} \sim N[0, \sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2]$, where

$$\begin{aligned} \sigma_{Y_{n+1} - \hat{Y}_{n+1}}^2 &= \sigma^2 + \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 \cdot \sigma^2 \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{2}{n} \cdot \frac{(X_{n+1} - \bar{X}) \sum_{j=1}^n (X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{(X_{n+1} - \bar{X})^2 \sum_{j=1}^n (X_j - \bar{X})^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \right) \\ &= \sigma^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right). \end{aligned} \tag{56}$$

Denoting,

$$\hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}}^2 = \hat{\sigma}^2 \left(\frac{n+1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right), \tag{57}$$

it follows now similar to Proposition 5 that

Proposition 7. *Under the assumptions I - V, $(Y_{n+1} - \hat{Y}_{n+1}) / \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}} \sim t_{n-2}$.*

This result can be used to construct a 95% confidence interval, say, of Y_{n+1} . Look up in the table of the t distribution the critical value t_* of the two-sided t-test with $n-2$ degrees of freedom. Then it follows from Proposition 7 that

$$\begin{aligned} 0.95 &= P[-t_* \leq (Y_{n+1} - \hat{Y}_{n+1}) / \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}} \leq t_*] \\ &= P[-t_* \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}} \leq Y_{n+1} - \hat{Y}_{n+1} \leq t_* \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}}] \\ &= P[\hat{Y}_{n+1} - t_* \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}} \leq Y_{n+1} \leq \hat{Y}_{n+1} + t_* \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}}] \end{aligned} \tag{58}$$

Thus, the 95% confidence interval of Y_{n+1} is $[\hat{Y}_{n+1} - t_* \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}}, \hat{Y}_{n+1} + t_* \hat{\sigma}_{Y_{n+1} - \hat{Y}_{n+1}}]$.

Observe from (57) that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}$ increases with $(X_{n+1}-\bar{X})^2$, and so does the width of the confidence interval. Thus, the farther X_{n+1} is away from \bar{X} , the more unreliable the forecast \hat{Y}_{n+1} of Y_{n+1} becomes. Also observe from (57) that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}} \geq \hat{\sigma}$, and that $\hat{\sigma}_{Y_{n+1}-\hat{Y}_{n+1}}$ gets close to $\hat{\sigma}$ if n is large because $\lim_{n \rightarrow \infty} \sum_{j=1}^n (X_j - \bar{X})^2 = \infty$.

9. *Relaxing the non-random regressor assumption*

As said before, the assumption that the regressors X_j are non-random is too strong an assumption in economics. Therefore, we now assume that the X_j 's are random variables. This requires the following modifications of the Assumptions I-V:

Assumption I*: *The pairs (X_j, Y_j) , $j = 1, 2, 3, \dots, n$, are independent and identically distributed.*

Assumption II*: *The conditional expectations $E[U_j | X_j]$ are equal to zero: $E[U_j | X_j] \equiv 0$.*

Assumption III*: *The conditional expectations $E[U_j^2 | X_j]$ do not depend on the X_j 's and are finite, constant and equal: $E[U_j^2 | X_j] \equiv \sigma^2 < \infty$. (This is called the **homoscedasticity** assumption.)*

Assumption IV*: *Conditional on X_j , U_j is $N(0, \sigma^2)$ distributed.*

The Assumptions I* and II* imply that for $j = 1, \dots, n$,

$$E[U_j | X_1, X_2, \dots, X_n] \equiv 0, \quad (59)$$

and similarly the Assumptions I* and III* imply that for $j = 1, \dots, n$,

$$E[U_j^2 | X_1, X_2, \dots, X_n] \equiv \sigma^2. \quad (60)$$

Because (loosely speaking) conditioning on X_1, X_2, \dots, X_n is effectively the same as treating them as given constants, most of the previous proposition carry over:

Proposition 8. Under Assumptions I^* - IV^* , Propositions 1, 4, 5, 6 and 7 carry over, and the results in Propositions 2 and 3 now hold conditional on X_1, X_2, \dots, X_n .

However, without the conditional normality assumption IV^* we need an additional condition in Proposition 6 in order to use the central limit theorem, namely:

Proposition 9. If the sample size n is large then under the assumptions I^* - III^* and the additional condition $E[X_j^2] < \infty$ the approximate normality results in Proposition 6 carry over.

10. Heteroscedasticity⁴

We say that the errors U_j of regression model (11) are heteroskedastic if assumption III^* does not hold:

$$E[U_j^2 | X_j] = \psi(X_j) \text{ for some function } \psi(\cdot). \quad (61)$$

Heteroscedasticity often occurs in practice. It is actually the rule rather than the exception. The main consequence of heteroscedasticity is that the conditional variance formulas in Propositions 2 and 3 do no longer hold, although the unbiasedness result in Proposition 1 is not affected by heteroscedasticity. Therefore, the Propositions 4-7 are no longer valid. To see this, let us compute the conditional variance of $\hat{\beta}$ [see (24)] under heteroscedasticity:

$$\begin{aligned} \text{var}(\hat{\beta} | X_1, \dots, X_n) &= E[(\hat{\beta} - \beta)^2 | X_1, \dots, X_n] \\ &= E \left[\left(\sum_{j=1}^n \left(\frac{X_j - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) U_j \right)^2 \middle| X_1, \dots, X_n \right] = \sum_{j=1}^n \left(\frac{X_j - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 E[U_j^2 | X_j] \\ &= \frac{\sum_{j=1}^n (X_j - \bar{X})^2 \psi(X_j)}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right)^2}. \end{aligned} \quad (62)$$

⁴ Also spelled as "Heteroskedasticity."

A cure for the heteroscedasticity problem is to replace the standard error of $\hat{\beta}$ by

$$\tilde{\sigma}_{\hat{\beta}} = \sqrt{\left(\frac{n}{n-2}\right) \frac{\sum_{j=1}^n (X_j - \bar{X})^2 \hat{U}_j^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2}}. \quad (63)$$

This is known as the Heteroscedasticity Consistent (H.C.) standard error. The H.C. t-value then becomes $\tilde{t}_{\hat{\beta}} = \hat{\beta} / \tilde{\sigma}_{\hat{\beta}}$. Under the null hypothesis $\beta = 0$ this t-value is no longer t distributed, but the standard normal approximation remains valid if the sample size n is large.

A popular test for heteroscedasticity is the Breusch-Pagan⁵ test. Given that

$$E[U_j^2 | X_j] = g(\gamma_0 + \gamma_1 X_j) \text{ for some unknown function } g(.). \quad (64)$$

the Breusch-Pagan test tests the null hypothesis

$$H_0: \gamma_1 = 0 \Leftrightarrow E[U_j^2 | X_j] = g(\gamma_0) = \sigma^2, \text{ say} \quad (65)$$

against the alternative hypothesis

$$H_0: \gamma_1 \neq 0 \Leftrightarrow E[U_j^2 | X_j] = g(\gamma_0 + \gamma_1 X_j) = \psi(X_j), \text{ say}. \quad (66)$$

Under the null hypothesis (65) of homoskedasticity the test statistic of the Breusch-Pagan test has a χ_1^2 distribution⁶, and the test is conducted right-sided.

11. How close are OLS estimators?

The ice cream data in Table 1 is not based on any actual observations on sales and temperature; I have picked the numbers for X_j and Y_j quite arbitrarily. Therefore, there is no way to find out how close the OLS estimates $\hat{\alpha} = -0.25$, $\hat{\beta} = 1.5$ are to the unknown parameters α and β . Actually, we do not know either whether the linear regression model (11) and its

⁵ Breusch, T. and A. Pagan (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica* 47, 1287-1294.

⁶ In the multiple regression case the degrees of freedom is equal to the number of parameters minus 1 for the intercept.

assumptions are applicable to this artificial data.

In order to show how well OLS estimators approximate the corresponding parameters I have generated random samples⁷ $(Y_1, X_1), \dots, (Y_n, X_n)$ for three sample sizes: $n = 10$, $n = 100$ and $n = 1000$, as follows. The explanatory variables X_j have been drawn independently from the χ_1^2 distribution, the regression errors U_j have been drawn independently from the $N(0,1)$ distribution, and the Y_j 's have been generated by

$$Y_j = 1 + X_j + U_j, j = 1, 2, \dots, n. \quad (67)$$

Thus, in this case the parameters α and β in model (11) are $\alpha = 1$ and $\beta = 1$, and the standard error of U_j is $\sigma = 1$. Moreover, note that the Assumptions $I^* - IV^*$ hold for model (67).

The true R^2 can be defined by

$$R_0^2 = 1 - \frac{E[SSR]}{E[TSS]} = 1 - \frac{(n-2)\sigma^2}{\sum_{j=1}^n E[(Y_j - \bar{Y})^2]}.$$

In the case (67), $\sigma^2 = 1$, $\mu_Y = E(Y_j) = 1 + E(X_j) = 2$,

$$\sum_{j=1}^n E[(Y_j - \bar{Y})^2] = E\left[\sum_{j=1}^n \left((Y_j - \mu_Y) - (\bar{Y} - \mu_Y)\right)^2\right] = E\left[\sum_{j=1}^n (Y_j - \mu_Y)^2 - n(\bar{Y} - \mu_Y)^2\right] = (n-1)\text{var}(Y_j)$$

and

$$\text{var}(Y_j) = E[(X_j - 1 + U_j)^2] = E[(X_j - 1)^2] + E[U_j^2] = E[(X_j - 1)^2] + 1 = 3,$$

because X_j is χ_1^2 distributed and therefore has the same distribution as U_j^2 , and it can be shown⁸ that for standard normal random variables U_j , $E[(U_j^2 - 1)^2] = 2$. Thus, the true R^2 in this case is

$$R_0^2 = 1 - \frac{n-2}{3(n-1)} = \frac{2n-1}{3n-3} \approx \begin{cases} 0.7037 & \text{for } n = 10 \\ 0.6700 & \text{for } n = 100 \\ 0.6670 & \text{for } n = 1000 \end{cases}$$

The estimation results involved are given in Table 2:

⁷ Via the *EasyReg International* menus File → Choose an input file → Create artificial data. Rather than generating one random sample of size $n = 1000$ and then using subsamples of sizes $n = 10$ and $n = 100$, these samples have been generated separately for $n = 10$, $n = 100$ and $n = 1000$.

⁸ But that is beyond the level of this course.

Table 2: *Artificial regression estimation results*

| | $\hat{\beta}$ | $\hat{\alpha}$ | $SER (= \hat{\sigma})$ | R^2 | n |
|-------------------|---------------|----------------|------------------------|--------|------|
| <i>estimate:</i> | 1.11748 | 0.55912 | 0.919045 | 0.8842 | 10 |
| <i>(t-value):</i> | (7.817) | (1.675) | | | |
| <i>estimate:</i> | 1.03309 | 0.96028 | 0.992502 | 0.8284 | 100 |
| <i>(t-value):</i> | (21.753) | (8.237) | | | |
| <i>estimate:</i> | 1.02360 | 0.98518 | 0.983608 | 0.6899 | 1000 |
| <i>(t-value):</i> | (47.124) | (26.037) | | | |

Even for a sample size of $n = 10$ the OLS estimator $\hat{\beta}$ is already pretty close to its true value 1, and the same applies to $\hat{\sigma}$, but $\hat{\alpha}$ is too far away from the true value $\alpha = 1$. However, for $n = 100$ the OLS estimators $\hat{\beta}$ and $\hat{\alpha}$ deviate only about $\pm 4\%$ from their true values $\alpha = \beta = 1$, and $\hat{\sigma}$ deviates about -1% from its true value 1. In the case $n = 1000$ these deviations reduce to about $\pm 2\%$. The R^2 's are too high, and only for $n = 1000$ is the R^2 reasonably close to its true value. However, the R^2 is only a descriptive statistic; it does not play a role in hypotheses testing, so that the unreliability of the R^2 in small samples is harmless.

Notice the quite dramatic increase of the t-values. Recall that these t-values are the test statistics of the null hypotheses that the corresponding parameters are zero. Because the true parameters are equal to 1, what you see in Table 2 is the increase of the power of the t-test with the sample size.

12. *Empirical homework*

12.1 *The data*

The data for the empirical homework assignment is available in two formats.

- (a) EasyReg (former) default format, as file

<http://econ.la.psu.edu/~hbierens/ECON490/REALDATA1.TXT>

Type this link in the address box of Microsoft Internet Explorer, and save the file as a text file (TXT) on your hard disk via the "Save as .." option. Note that Internet Explorer saves by default text files as web pages (*.HTM), so you have to overrule that.

- (b) Excel CSV format, as file

<http://econ.la.psu.edu/~hbierens/ECON490/REALDATA1.CSV>

Type this link in the address box of Internet Explorer. Then this file will (likely) be automatically imported in Excel. Use the "Save as .." option of Excel to save the file on your hard disk. If Internet Explorer only allows you to download the CSV file, follow the previous procedure. Other web browsers may open this file as a text file. If so, follow the previous procedure.

The source of this data is the 1980 Dutch Budget Survey. This data set is actually a random sample of size 2000 from this survey, containing the following variables:

Observation number (ranging from 1 to 2000)

Gender (F=1,M=0)

Hourly wage (in 1980 Dutch guilders)

LN[Hourly wage] (the natural log of Hourly wage)

The observations are ordered according to gender. The first 1612 observations are males: Gender = 0, and the remaining 388 observations are females: Gender = 1. There are no missing values.

12.2 *Assignments*

1. Import one of these files in EasyReg (or in any other econometric software package of your choice) as cross-section data. Then estimate the following two linear regression models, using all 2000 observations:

Model 1:

$$Y_j = \alpha + \beta X_j + U_j,$$

where U_j is the error term, $X_j = \text{Gender (F=1, M=0)}$ of individual j , and $Y_j = \text{LN[Hourly wage]}$ of individual j .

Model 2:

$$Z_j = \gamma + \delta X_j + V_j,$$

where V_j is the error term, $X_j = \text{Gender (F=1, M=0)}$ of individual j , and $Z_j = \text{Hourly wage}$ of individual j .

2. Present the estimation results in the form as you would do in an article, book or report, with t-values between brackets. Which t-values would you use, the ones that apply to the homoscedastic case, or the heteroscedastic consistent (H.C.) t-values ?
3. Determine the 95% confidence intervals of the parameters β and δ .
4. The gender wage gap is the difference in average hourly wages for men and women, either expressed as a percentage difference
 - (1) Relative gender wage gap
= 100% (Average hourly wage of men - Average hourly wage of women)/(Average hourly wage of women),
or in Dutch guilders:
 - (2) Absolute gender wage gap
= Average hourly wage of men - Average hourly wage of women.
 - 4.1 The gender wage gap of type (2) can only be determined on the basis of the estimation result for model 2. The gender wage gap of type (1) can be determined on the basis of the estimation result for model 1 as well as model 2, but model 1 is preferred. Explain why.
 - 4.2 Determine these gender wage gaps.
 - 4.3 Determine the 95% confidence intervals of the these gender wage gaps.
5. Suppose you estimate models 1 and 2 for the sub-sample of males only. What would happen, and why?

6. Now estimate each of the following models

$$Y_j = \alpha + U_j,$$

where U_j is the error term and $Y_j = \text{LN}[\text{Hourly wage}]$ of individual j , and

$$Z_j = \gamma + V_j,$$

where V_j is the error term, and $Z_j = \text{Hourly wage}$ of individual j , separately for the sub-samples of males ($j = 1, \dots, 1612$) and females ($j = 1613, \dots, 2000$).

6.1 Determine the gender wage gaps of types (1) and (2).

6.2 Which method is better, this one or the one in problem 4? Explain your answer.

7. A conditional expectation $E[Y|X]$ is always a function of X : $E[Y|X] = g(X)$, say. Defining $U = Y - g(X)$, we can therefore always write $Y = g(X) + U$, where $E[U|X] = 0$. In general this function $g(X)$ is nonlinear, but in the case of models 1 and 2 this function is linear:

$$E[Y_j | X_j] = \alpha + \beta X_j,$$

$$E[Z_j | X_j] = \gamma + \delta X_j,$$

for particular values of α , β , γ and δ , due to the nature of the explanatory variable X_j . Therefore, models 1 and 2 are both correctly specified linear regression models. Explain why.

Remark: Although both models 1 and 2 are correct linear regression models, it is not possible that both are homoscedastic. To see this, assume that the error term U_j in model 1 is independent of X_j and therefore homoscedastic. Then model 2 is related to model 1 in the following way.

$$\begin{aligned} Z_j &= \exp(Y_j) = \exp(\alpha + \beta X_j) \exp(U_j) \\ &= \exp(\alpha + \beta X_j) E[\exp(U_j)] + \exp(\alpha + \beta X_j) (\exp(U_j) - E[\exp(U_j)]) \\ &= \exp(\alpha) E[\exp(U_j)] + (\exp(\alpha + \beta) - \exp(\alpha)) E[\exp(U_j)] X_j \\ &\quad + \exp(\alpha + \beta X_j) (\exp(U_j) - E[\exp(U_j)]) = \gamma + \delta X_j + V_j, \end{aligned}$$

where $\gamma = \exp(\alpha) E[\exp(U_j)]$, $\delta = (\exp(\alpha + \beta) - \exp(\alpha)) E[\exp(U_j)]$ and

$$V_j = \exp(\alpha + \beta X_j) (\exp(U_j) - E[\exp(U_j)]).$$

Note that $E[V_j | X_j] = 0$, but $E[V_j^2 | X_j] = (\exp(\alpha + \beta X_j))^2 E[(\exp(U_j) - E[\exp(U_j)])^2]$, which depends on X_j .

13. *Answer keys*

1. EasyReg estimation results:

Model 1:

Dependent variable:

$$Y = \text{LN}[\text{Hourly wage}]$$

X variables:

$$X(1) = \text{Gender (F=1,M=0)}$$

$$X(2) = 1$$

Model:

$$Y = b(1)X(1) + b(2)X(2) + U,$$

where U is the error term, satisfying

$$E[U|X(1),X(2)] = 0.$$

OLS estimation results

| Parameters | Estimate | t-value (S.E.) [p-value] | H.C. t-value (H.C. S.E.) [H.C. p-value] |
|------------|----------|-----------------------------------|---|
| b(1) | -0.33580 | -16.614 (0.02021) [0.00000] | -19.132 (0.01755) [0.00000] |
| b(2) | 2.82955 | 317.848 (0.00890) [0.00000] | 306.328 (0.00924) [0.00000] |

Notes:

1: S.E. = Standard error

2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.

3: The two-sided p-values are based on the normal approximation.

| | |
|--|------------|
| Effective sample size (n): | 2000 |
| Variance of the residuals: | 0.127749 |
| Standard error of the residuals (SER): | 0.35742 |
| Residual sum of squares (RSS): | 255.243257 |
| (Also called SSR = Sum of Squared Residuals) | |
| Total sum of squares (TSS): | 290.506519 |

R-square: 0.121385

Breusch-Pagan test = 25.087041

Null hypothesis: The errors are homoskedastic

Null distribution: Chi-square(1)

p-value = 0.00000

Significance levels: 10% 5%

Critical values: 2.71 3.84

Conclusions: reject reject

Model 2:

Dependent variable:

Y = Hourly wage

X variables:

X(1) = Gender (F=1,M=0)

X(2) = 1

Model:

$Y = b(1)X(1) + b(2)X(2) + U,$

where U is the error term, satisfying

$E[U|X(1),X(2)] = 0.$

OLS estimation results

| Parameters | Estimate | t-value (S.E.) [p-value] | H.C. t-value (H.C. S.E.) [H.C. p-value] |
|------------|----------|-----------------------------------|---|
| b(1) | -5.56503 | -13.172 (0.42250) [0.00000] | -18.577 (0.29957) [0.00000] |
| b(2) | 18.24756 | 98.056 (0.18609) [0.00000] | 91.188 (0.20011) [0.00000] |

Notes:

1: S.E. = Standard error

2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity

consistent variance matrix.

3: The two-sided p-values are based on the normal approximation.

| | |
|--|---------------|
| Effective sample size (n): | 2000 |
| Variance of the residuals: | 55.824209 |
| Standard error of the residuals (SER): | 7.47156 |
| Residual sum of squares (RSS): | 111536.769695 |
| (Also called SSR = Sum of Squared Residuals) | |
| Total sum of squares (TSS): | 121221.818113 |
| R-square: | 0.079895 |

Breusch-Pagan test = 103.017134

Null hypothesis: The errors are homoskedastic

Null distribution: Chi-square(1)

p-value = 0.00000

Significance levels: 10% 5%

Critical values: 2.71 3.84

Conclusions: reject reject

2. Because in both cases the Breusch-Pagan test strongly rejects the homoscedasticity hypothesis, you have to use the H.C. t-values.

Model 1:

$$\text{LN[Hourly wage]} = 2.82955 - 0.33580 \cdot \text{Gender}, \quad n = 2000, \quad \text{SER} = 0.35742, \\ (306.328) \quad (-19.132) \quad R^2 = 0.121385 \\ (t\text{-values between brackets})$$

Model 2:

$$\text{Hourly wage} = 18.24756 - 5.56503 \cdot \text{Gender}, \quad n = 2000, \quad \text{SER} = 7.47156, \\ (91.188) \quad (-18.577) \quad R^2 = 0.079895 \\ (t\text{-values between brackets})$$

3. We have $\hat{\beta} = -0.33580$, with standard error $\hat{\sigma}_{\hat{\beta}} = \hat{\beta}/\hat{t}_{\hat{\beta}} = 0.01755$. Because n is large,

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim N(0,1),$$

hence

$$\begin{aligned}
0.95 &= P\left[-1.96 \leq \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\beta}} \leq 1.96\right] = P\left[\hat{\beta} - 1.96 \cdot \hat{\sigma}_{\beta} \leq \beta \leq \hat{\beta} + 1.96 \cdot \hat{\sigma}_{\beta}\right] \\
&= P[-0.370198 \leq \beta \leq -0.301402].
\end{aligned} \tag{68}$$

Similarly, we have $\hat{\delta} = -5.56503$, with standard error $\hat{\sigma}_{\delta} = \hat{\delta}/t_{\delta} = 0.29957$, and

$$\frac{\hat{\delta} - \delta}{\hat{\sigma}_{\delta}} \sim N(0,1),$$

hence

$$\begin{aligned}
0.95 &= P\left[-1.96 \leq \frac{\hat{\delta} - \delta}{\hat{\sigma}_{\delta}} \leq 1.96\right] = P\left[\hat{\delta} - 1.96 \cdot \hat{\sigma}_{\delta} \leq \delta \leq \hat{\delta} + 1.96 \cdot \hat{\sigma}_{\delta}\right] \\
&= P[-6.1521872 \leq \delta \leq -4.9778728].
\end{aligned} \tag{69}$$

4.1 The relative gender wage gap on the basis of model 2 is

$$\frac{-\hat{\delta}}{\hat{\gamma} + \hat{\delta}} \times 100\% = \frac{5.56503}{18.24756 - 5.56503} \times 100\% = \frac{5.56503}{12.68253} \times 100\% \approx 43.88\%$$

However, to determine the 95% confidence interval you have to determine the distribution of

$$\left(\frac{-\hat{\delta}}{\hat{\gamma} + \hat{\delta}} - \frac{-\delta}{\gamma + \delta} \right) \times 100\%,$$

which is too difficult.

4.2 Let \hat{W}_f be the hourly wage of females and let \hat{W}_m be the hourly wage of males, as predicted by the models. Then it follows from model 1 that

$$\ln(\hat{W}_m / \hat{W}_f) = \ln(\hat{W}_m) - \ln(\hat{W}_f) = -\hat{\beta} = 0.33580,$$

hence

$$100(\hat{W}_m - \hat{W}_f) / \hat{W}_f = 100(\hat{W}_m / \hat{W}_f - 1) = 100(\exp(0.33580) - 1) = 39.91.$$

Moreover, it follows from model 2 that

$$\hat{W}_m - \hat{W}_f = -\hat{\delta} = 5.56503.$$

4.3 Let W_f and W_m be the “true” values of the hourly wages of females and males,

respectively . Then it follows from model 1 that

$$\ln(W_m/W_f) = \ln(W_m) - \ln(W_f) = -\beta,$$

hence it follows from (68) that

$$\begin{aligned} 0.95 &= P[0.301402 \leq \ln(W_m/W_f) \leq 0.370198] \\ &= P[\exp(0.301402) \leq W_m/W_f \leq \exp(0.370198)] \\ &= P[\exp(0.301402)-1 \leq (W_m - W_f)/W_f \leq \exp(0.370198)-1] \\ &= P[0.3517526369 \leq (W_m - W_f)/W_f \leq 0.4480212945] \\ &= P[35.17526369 \leq 100(W_m - W_f)/W_f \leq 44.80212945] \end{aligned}$$

Moreover, it follows from model 2 that

$$W_m - W_f = -\delta,$$

hence it follows from (69) that

$$0.95 = P[4.9778728 \leq W_m - W_f \leq 6.1521872].$$

5. For the sub-sample of males, $X_j = 0$, so that β and δ can not be estimated. To see this, observe that in the case of model 1,

$$\min_{\hat{\alpha}, \hat{\beta}} \sum_{j=1}^{1612} (Y_j - \hat{\alpha} - \hat{\beta}X_j)^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum_{j=1}^{1612} (Y_j - \hat{\alpha})^2 = \sum_{j=1}^{1612} (Y_j - \bar{Y}_m)^2$$

for $\hat{\alpha} = \bar{Y}_m = (1/1612)\sum_{j=1}^{1612} Y_j$, and any value of $\hat{\beta}$. A similar result holds for model 2.

6. *OLS estimation results*

$Y = \text{LN}[\text{Hourly wage}], \text{ males only}$

$$\begin{aligned} \text{LN}[\text{Hourly wage}] &= 2.82955, n = 1612, \text{SER} = 0.370977 \\ &\quad (0.00924) \end{aligned} \tag{70}$$

(standard error between brackets)

$Y = \text{LN}[\text{Hourly wage}], \text{females only}$

$$\begin{aligned} \text{LN}[\text{Hourly wage}] &= 2.49375, n = 388, \text{SER} = 0.294352 \\ &\quad (0.01494) \\ &\quad (\text{standard error between brackets}) \end{aligned} \tag{71}$$

$Z = \text{Hourly wage}, \text{males only}$

$$\begin{aligned} \text{Hourly wage} &= 18.24756, n = 1612, \text{SER} = 8.036797 \\ &\quad (0.20017) \\ &\quad (\text{standard error between brackets}) \end{aligned} \tag{72}$$

$Z = \text{Hourly wage}, \text{females only}$

$$\begin{aligned} \text{Hourly wage} &= 12.68253, n = 388, \text{SER} = 4.397002 \\ &\quad (0.22322) \\ &\quad (\text{standard error between brackets}) \end{aligned} \tag{73}$$

6.1 Let again \hat{W}_f be the hourly wage of females and let \hat{W}_m be the hourly wage of males, as predicted by the models. Then it follows from (70) and (71) that

$$\ln(\hat{W}_m/\hat{W}_f) = 2.82955 - 2.49375 = 0.3358$$

hence

$$100(\hat{W}_m - \hat{W}_f)/\hat{W}_f = 100(\exp(0.3358) - 1) = 39.91$$

Moreover, it follows from (72) and (73) that

$$\hat{W}_m - \hat{W}_f = 18.24756 - 12.68253 = 5.56503$$

6.2 Note that these results are exactly the same as in 4.2. The reason is that in the case of model 1,

$$\begin{aligned} \min_{\hat{\alpha}, \hat{\beta}} \sum_{j=1}^{2000} (Y_j - \hat{\alpha} - \hat{\beta}X_j)^2 &= \min_{\hat{\alpha}, \hat{\beta}} \sum_{j=1}^{1612} (Y_j - \hat{\alpha})^2 + \min_{\hat{\alpha}, \hat{\beta}} \sum_{j=1613}^{2000} (Y_j - \hat{\alpha} - \hat{\beta})^2 \\ &= \min_{\hat{\alpha}} \sum_{j=1}^{1612} (Y_j - \hat{\alpha})^2 + \min_{\hat{\alpha} + \hat{\beta}} \sum_{j=1613}^{2000} (Y_j - (\hat{\alpha} + \hat{\beta}))^2 \end{aligned}$$

so that $\hat{\alpha} = 2.82955$ and $\hat{\alpha} + \hat{\beta} = 2.49375$. Similarly, in the case of model 2 we have

$\hat{\gamma} = 18.24756$ and $\hat{\gamma} + \hat{\delta} = 12.68253$. Thus, as far as the computations of the gender wage gaps are concerned there is no difference between the two methods.

7. We have $E[Y/X=0] = g(0)$, $E[Y/X=1] = g(1) = g(0) + (g(1)-g(0))$. Because X takes only two values, 0 and 1, these results can be combined into a single statement:

$$E[Y|X] = g(0) + (g(1)-g(0))X.$$

Thus, $\alpha = g(0)$ and $\beta = g(1)-g(0)$. Similarly, if $E[Z|X] = f(X)$ then

$$E[Z|X] = f(0) + (f(1)-f(0))X,$$

hence $\gamma = f(0)$ and $\delta = f(1)-f(0)$.