2. The Simple Regression Model

2.1 Definition

Two (observable) variables "y" and "x".

(1) $$y = \beta_0 + \beta_1 x + u.$$

Equation (1) defines the *simple regression model*.

Terminology:

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor
Regressand	Regressor

Error term u_i is a combination of a number of effects, like:

1. Omitted variables: Accounts the effects of variables omitted from the model

2. Nonlinearities: Captures the effects of nonlinearities between y and x. Thus, if the true model is $y_i = \beta_0 + \beta_1 x_i + \gamma x_i^2 + v_i$, and we assume that it is $y_i = \beta_0 + \beta_1 x + u_i$, then the effect of x_i^2 is absorbed to u_i. In fact, $u_i = \gamma x_i^2 + v_i$.

3. Measurement errors: Errors in measuring y and x are absorbed in u_i.

4. Unpredictable effects: u_i includes also inherently unpredictable random effects.

2.2 Estimation of the model, OLS

Given observations (x_i, y_i), $i = 1, \ldots, n$, we estimate the population parameters β_0 and β_1 of (1) making the following

Assumptions (classical assumptions):

1. $y = \beta_0 + \beta_1 x + u$ in the population.
2. $\{(x_i, y_i) : i = 1, \ldots, n\}$ is a random sample of the model above, implying uncorrelated residuals: $\mathbb{C}\text{ov}(u_i, u_j) = 0$ for all $i \neq j$.
3. $\{x_i, \, , i = 1, \ldots, n\}$ are not all identical, implying $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$.
4. $\mathbb{E}[u|x] = 0$ for all x (zero average error), implying $\mathbb{E}[u] = 0$ and $\mathbb{C}\text{ov}(u, x) = 0$.
5. $\mathbb{V}\text{ar}[u|x] = \sigma^2$ for all x, implying $\mathbb{V}\text{ar}[u] = \sigma^2$ (homoscedasticity).

Here $|x$ means "conditional on x", that is, we restrict our sample space to this particular value of x. The practical implication in calculations and derivations is that we can treat x as nonrandon, for the price that our result will hold for that particular value of x only. Otherwise the calucation rules for conditional expectations and variances are identical to their unconditional counterparts.

The goal in the estimation is to find values for β_0 and β_1 that the error terms is as small as possible (in suitable sense).

Under the classical assumptions above, the Ordinary Least Squares (OLS) that minimizes the residual sum of squares of the error terms $u_i = y_i - \beta_0 - \beta_1 x_i$ produces optimal estimates for the parameters (the optimality criteria are discussed later).

Denote the sum of squares as

$$(2) \qquad f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

The first order conditions (foc) for the minimum are found by setting the partial derivatives equal to zero. Denote by $\widehat{\beta}_0$ and $\widehat{\beta}_1$ the values satisfying the foc.

First order conditions:

$$(3)\quad \frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$(4)\quad \frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

These yield so called normal equations

$$(5)\quad \begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i &= \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 &= \sum x_i y_i, \end{aligned}$$

where the summation is from 1 to n.

The explicit solutions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are (OLS estimators of β_0 and β_1)

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{6}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}, \tag{7}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

and

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

are the sample means.

In the solutions (6) and (7) we have used the properties

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \tag{8}$$

and

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2. \tag{9}$$

Fitted regression line:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x. \tag{10}$$

Residuals:

$$\begin{aligned} \widehat{u}_i &= y_i - \widehat{y}_i \\ &= (\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1)x_i + u_i. \end{aligned} \tag{11}$$

Thus the residual component $\widehat{u}_i$ consist of the pure error term u_i and the sample errors due to the estimation of the parameters β_0 and β_1.

Remark 2.1: The slope coefficient $\widehat{\beta}_1$ in terms of sample covariance of x and y and variance of x.

Sample *covariance*:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{12}$$

Sample *variance*:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{13}$$

Thus

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2}. \tag{14}$$

Remark 2.2: The slope coefficient $\widehat{\beta}_1$ in terms of sample correlation and standard deviations of x and y.

Sample *correlation*:

$$(15)\quad r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y},$$

where $s_x = \sqrt{s_x^2}$ and $s_y = \sqrt{s_y^2}$ are the sample *standard deviations* of x and y, respectively.

Thus we can also write the slope coefficient in terms of sample standard deviations and correlation as

$$(16)\qquad \widehat{\beta}_1 = \frac{s_y}{s_x} r_{xy}.$$

Example 2.1: Relationship between wage and education.

wage = average hourly earnings
educ = years of education
Data is collected in 1976, $n = 526$
Excerpt of the data set wage.raw (Wooldridge)

wage (y)	educ (x)
3.10	11
3.24	12
3.00	11
6.00	8
5.30	12
8.75	16
11.25	18
5.00	12
3.60	12
18.18	17
6.25	16
⋮	⋮

Scatterplot of the data with regression line.

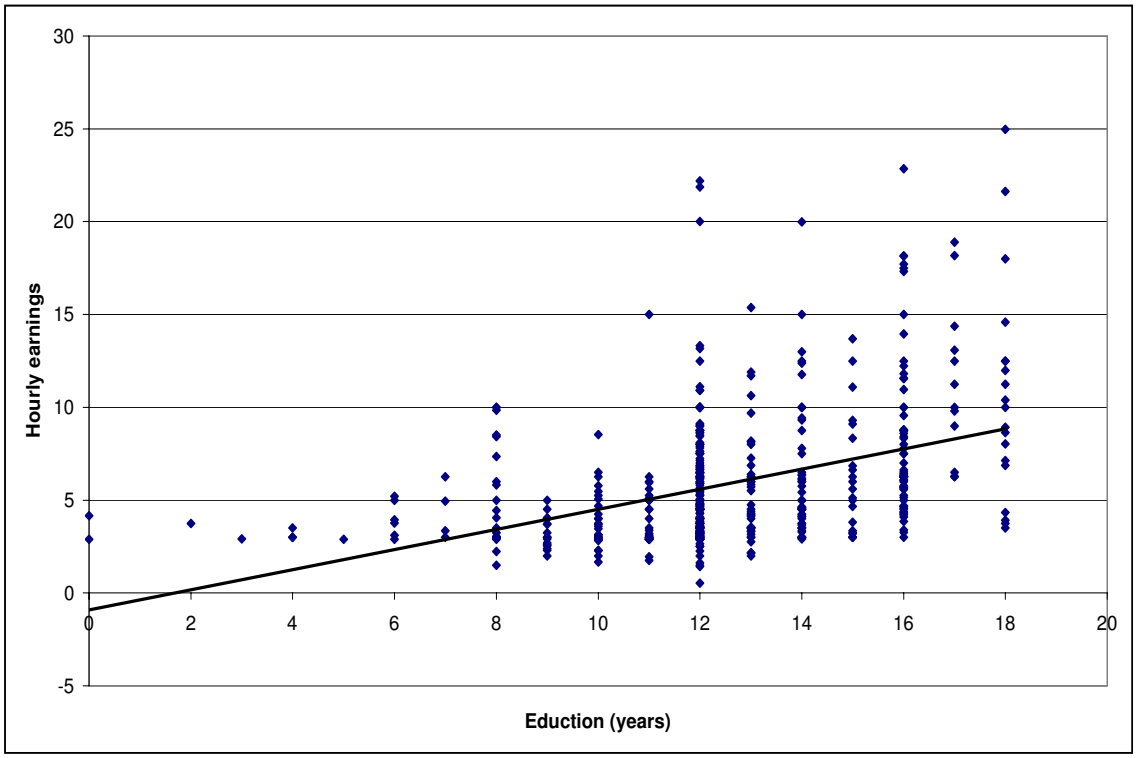

Figure 2.2: Wages and education.

Sample statistics:

	Wage	Educ
Mean	5.90	12.56
Standard deviation	3.69	2.769
n	526	526
Correlation	0.406	

Eviews estimation results:

Dependent Variable: WAGE
Method: Least Squares
Date: 02/07/06 Time: 20:25
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.904852	0.684968	-1.321013	0.1871
EDUC	0.541359	0.053248	10.16675	0.0000

R-squared	0.164758	Mean dependent var	5.896103
Adjusted R-squared	0.163164	S.D. dependent var	3.693086
S.E. of regression	3.378390	Akaike info criterion	5.276470
Sum squared resid	5980.682	Schwarz criterion	5.292688
Log likelihood	-1385.712	F-statistic	103.3627
Durbin-Watson stat	1.823686	Prob(F-statistic)	0.000000

The estimated model is

$$\widehat{y} = -0.905 + 0.541x.$$

Thus the model predicts that an additional year increases hourly wage on average by 0.54 dollars.

Using (16) you can verify the OLS estimate for β_1 can be computed using the correlation and standard deviations. After that, applying (7) you get OLS estimate for the intercept. Thus in all, the estimates can be derived from the basic sample statistics.

2.3 OLS Statistics

Algebraic properties

(17) $$\sum_{i=1}^{n} \hat{u}_i = 0.$$

(18) $$\sum_{i=1}^{n} x_i \hat{u}_i = 0.$$

(19) $$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}.$$

(20) $$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

(21) $$\text{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

(22) $$\text{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

It can be shown that

$$(23) \quad \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2,$$

that is

$$(24) \qquad \text{SST}=\text{SSE}+\text{SSR}.$$

Prove this!

Remark 2.3: It is unfortunate that different books and different statistical packages use different definitions, particularly for SSR and SSE. In many the former means *Regression* sum of squares and the latter *Error* sum of squares. I.e., just the opposite we have here!

Goodness-of-Fit, the R-square

R-square (coefficient of determination)

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}. \tag{25}$$

The positive square root of R^2, denoted as R, is called the multiple correlation.

Remark 2.4: Here in the case of simple regression $R^2 = r_{xy}^2$, i.e. $R = |r_{xy}|$. These do not hold in the general case (multiple regression)!

Prove Remark 2.4 yourself.

Remark 2.5: Generally it holds for the OLS estimation, however, that $R = r_{y\hat{y}}$, i.e. correlation between the observed and fitted (or predicted) values.

Remark 2.6: It is obvious that $0 \leq R^2 \leq 1$ with $R^2 = 0$ representing no linear relation between x and y and $R^2 = 1$ representing a perfect fit.

Adjusted R-square:

(1) $$\bar{R}^2 = 1 - \frac{s_u^2}{s_y^2},$$

where

(2) $$s_u^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

is an estimate of the residual variance $\sigma_u^2 = \mathbb{V}\text{ar}[u]$.

We find easily that

(3) $$\bar{R}^2 = 1 - \frac{n-1}{n-2}(1 - R^2).$$

One finds immediately that $\bar{R}^2 < R^2$.

Example 2.2: In the previous example $R^2 = 0.164758$ and adjusted R-squared, $\bar{R}^2 = 0.163164$. The R^2 tells that about 16.5 percent of the variation in the hourly earnings can be explained by education. However, the rest 83.5 percent is not accounted by the model.

2.4 Units of Measurement and Functional Form

Scaling and translation

Consider the simple regression model

(29) $$y_i = \beta_0 + \beta_1 x_i + u_i$$

with $\sigma_u^2 = \text{Var}[u_i]$.

Let $y_i^* = a_0 + a_1 y_i$ and $x_i^* = b_0 + b_1 x_i$, $a_1 \neq 0$ and $b_1 \neq 0$. Then (29) becomes

(30) $$y_i^* = \beta_0^* + \beta_1^* x_i^* + u_i^*,$$

where

(31) $$\beta_0^* = a_1 \beta_0 + a_0 - \frac{a_1}{b_1} \beta_1 b_0,$$

(32) $$\beta_1^* = \frac{a_1}{b_1} \beta_1,$$

and

(33) $$\sigma_{u^*}^2 = a_1^2 \sigma_u^2.$$

Remark 2.7: Coefficients a_1 and b_1 scale the measurements and a_0 and b_0 shift the measurements.

For example, if y is temperature measured in Celsius, then

$$y^* = 32 + \frac{9}{5}y$$

gives temperature in Fahrenheit.

Example 2.3: Let the estimated model be

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

"Demeaned" observations:

$y_i^* = y_i - \bar{y}$ and $x_i - \bar{x}$. So $a_0 = -\bar{y}$, $b_0 = -\bar{x}$, and $a_1 = b_1 = 1$.

Because $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, we obtain from (31)

$$\hat{\beta}_0^* = \hat{\beta}_0 - (\bar{y} - \hat{\beta}_1 \bar{x}) = 0.$$

So

$$\hat{y}^* = \hat{\beta}_1 x^*.$$

(Note $\hat{\beta}_1$ remains unchanged).

If we further define $a_1 = 1/s_y$ and $b_1 = 1/s_x$, where s_y and s_x are the sample standard deviations of y and x, respectively. Applying the transformation yields standardized observations

$$y_i^* = \frac{y_i - \bar{y}}{s_y}$$

and

$$x_i^* = \frac{x_i - \bar{x}}{s_x}.$$

Then again $\hat{\beta}_0 = 0$. The slope coefficient becomes

$$\hat{\beta}_1^* = \frac{s_x}{s_y} \hat{\beta}_1,$$

which is called *standardized regression coefficient*.

As an exercise show that in this case $\hat{\beta}_1^* = r_{xy}$, the correlation coefficient of x and y.

Nonlinearities

Logarithmic transformation is one of the most applied transformation for economic variables.

Table 2.1 Functional forms including log-transformations

Model	Dependent variable	Independent variable	Interpretation of β_1
level-level	y	x	$\Delta y = \beta_1 \Delta x$
level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Can you find the rationale for the interpretations?

Remark 2.8: Log-transformation can be only applied to variables that assume strictly positive values!

Example 2.4 Consider the wage example (Ex 2.1).

Suppose we believe that instead of the absolute change a better choice is to consider the percentage change of wage (y) as a function of education (x). Then we would consider the model

$$\log(y) = \beta_0 + \beta_1 x + u.$$

Estimation of this model yields

Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.583773	0.097336	5.997510	0.0000
EDUC	0.082744	0.007567	10.93534	0.0000

R-squared	0.185806	Mean dependent var	1.623268
Adjusted R-squared	0.184253	S.D. dependent var	0.531538
S.E. of regression	0.480079	Akaike info criterion	1.374061
Sum squared resid	120.7691	Schwarz criterion	1.390279
Log likelihood	-359.3781	F-statistic	119.5816
Durbin-Watson stat	1.801328	Prob(F-statistic)	0.000000

That is

(34) $$\widehat{\log(y)} = 0.584 + 0.083x$$

$n = 526$, $R^2 = 0.186$. Note that R-squares of this model and the level-level model are not comparable.

The model predicts that an additional year of education increases on average hourly earnings by 8.3%.

Remark 2.9: Typically all models where transformations on y and x are functions of these variables alone can be cast to the form of linear model. That is, if have generally

$$g(y) = \beta_0 + \beta_1 h(x) + u, \tag{35}$$

where g and h are functions, then defining $y^* = g(y)$ and $x^* = h(x)$ we have a linear model

$$y^* = \beta_0 + \beta_1 x^* + u.$$

Note, however, that all models cannot be cast to a linear form. An example is

$$\text{cons} = \frac{1}{\beta_0 + \beta_1 \text{income}} + u.$$

2.5 Expected Values and Variances of the OLS Estimators

Unbiasedness

We say generally that an estimator of $\widehat{\theta}$ of a parameter θ is unbiased if $\mathsf{E}[\widehat{\theta}] = \theta$.

Theorem 2.1: Under the classical assumptions 1–5

$$\mathsf{E}[\widehat{\beta}_0] = \beta_0 \text{ and } \mathsf{E}[\widehat{\beta}_1] = \beta_1. \tag{36}$$

Proof: Given observations $x_1, \ldots, x_n$ the expectations are conditional on the given x_i-values.

We prove first the unbiasedness of $\widehat{\beta}_1$. Now

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ &= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \\ &= \frac{\sum (x_i - \bar{x}) \beta_0}{\sum (x_i - \bar{x})^2} + \beta_1 \frac{\sum (x_i - \bar{x}) x_i}{\sum (x_i - \bar{x})^2} + \frac{\sum (x_i - \bar{x}) u_i}{\sum (x_i - \bar{x})^2} \\ &= \beta_1 + \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) u_i. \end{aligned} \tag{37}$$

That is, conditional on $x = x_i$:

$$\mathbb{E}[\widehat{\beta}_1|x_i] = \beta_1 + \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\mathbb{E}[u_i|x_i] = \beta_1 \tag{38}$$

because $\mathbb{E}[u_i|x_i] = 0$ by assumption 4. Because this holds for all x_i we get also unconditionally

$$\mathbb{E}[\widehat{\beta}_1] = \beta_1 + \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\mathbb{E}[u_i] = \beta_1 \tag{39}$$

Thus $\widehat{\beta}_1$ is unbiased.

Proof of unbiasedness of $\widehat{\beta}_0$ is left to students.

Variances

Theorem 2.2: Under the classical assumptions 1 through 5 and given $x_1, \ldots, x_n$

$$\mathbb{V}\text{ar}[\hat{\beta}_1] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \tag{40}$$

$$\mathbb{V}\text{ar}[\hat{\beta}_0] = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right) \sigma_u^2. \tag{41}$$

and for $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ with given x

$$\mathbb{V}\text{ar}[\hat{y}] = \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right) \sigma_u^2. \tag{42}$$

Proof: Again we prove as an example only (40). Using (37) and the properties of variance with $x_1, \ldots, x_n$ given

$$\begin{aligned}
\mathbb{V}\text{ar}[\hat{\beta}_1] &= \mathbb{V}\text{ar}\left[\beta_1 + \tfrac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) u_i\right] \\
&= \left(\tfrac{1}{\sum (x_i - \bar{x})^2}\right)^2 \sum (x_i - \bar{x})^2 \mathbb{V}\text{ar}[u_i] \\
&= \left(\tfrac{1}{\sum (x_i - \bar{x})^2}\right)^2 \sum (x_i - \bar{x})^2 \sigma_u^2 \qquad (43) \\
&= \frac{\sigma_u^2 \sum (x_i - \bar{x})^2}{\left(\sum (x_i - \bar{x})^2\right)^2} \\
&= \frac{\sigma_u^2}{\sum (x_i - \bar{x})^2}.
\end{aligned}$$

Remark 2.10: (41) can be written equivalently as

$$\text{Var}[\hat{\beta}_0] = \frac{\sigma_u^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}. \tag{44}$$

Estimating the Error Variance

Recalling from (11) the residual $\widehat{u}_i = y_i - \widehat{y}_i$ is of the form

(45) $\widehat{u}_i = u_i - (\widehat{\beta}_0 - \beta_0) - (\widehat{\beta}_1 - \beta_1)x_i.$

This reminds us about the difference between the error term u_i and the residual term $\widehat{u}_i$.

An unbiased estimator of the error variance $\sigma_u^2 = \mathsf{Var}[u_i]$ is

(46) $$\widehat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n} \widehat{u}_i^2.$$

Taking the (positive) square root gives an estimator for the error standard deviation $\sigma_u = \sqrt{\sigma_u^2}$, called usually the *standard error of regression*

(47) $$\widehat{\sigma}_u = \sqrt{\frac{1}{n-2}\sum \widehat{u}_i^2}.$$

Theorem 2.3: Under the assumption (1)–(5)

$$\mathrm{E}[\widehat{\sigma}_u^2] = \sigma_u^2,$$

i.e., $\widehat{\sigma}^2$ is an unbiased estimator of σ_u^2.

Proof: Omitted.

Standard Errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

Replacing in (40) and (41) σ_u^2 by $\hat{\sigma}_u^2$ and taking square roots give the *standard error* of $\hat{\beta}_1$ and $\hat{\beta}_0$

$$\text{se}(\hat{\beta}_1) = \frac{\hat{\sigma}_u}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{48}$$

and

$$\text{se}(\hat{\beta}_0) = \hat{\sigma}_u \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}. \tag{49}$$

These belong to the standard output of regression estimation, see computer print-outs above.