MICHIGAN STATE UNIVERSITY

AFRE 802
Statistical Methods for Agricultural, Food, & Resource Economists

Linear models & estimation by least squares – Part 2 of 3
(WMS Ch. 11.4 & Wooldridge pp. 38-60, 101-102)

November 30, 2017

Nicole Mason
Michigan State University
Fall 2017

GAME PLAN

- Return Ch. 10 HW
- Collect take-home graded exercise and distribute extra practice problem (not graded; we will go over the answers on Tuesday)
- Review

- Linear models & estimation by least squares – Part 2 of 3
  - The simple linear regression model (cont'd)
    - Using Stata to estimate a simple linear regression model
    - Algebraic properties of OLS
    - Statistical properties of OLS
    - Gauss-Markov Theorem

1

Review: the simple linear regression (SLR) model

- Suppose y and x are two variables that represent some population. *What does a SLR model look like?*

$$y = \beta_0 + \beta_1 x + u$$

- *What are the population parameters we want to estimate?*
  - β_0 and β_1
- *Why is u included in the model? What does it represent?*
  - u: error term (unobserved; all factors other than x that affect y)
- *What are some names for y and x?*

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

Review: to get unbiased estimates of β_0 and β_1, we need to restrict the relationship b/w x and u

$$y = \beta_0 + \beta_1 x + u$$

1. $E(u) = 0$ (not restrictive if have an intercept, β_0)
2. *** $E(u|x) = E(u)$. *What does this mean?*

#1 & #2 → $\boxed{E(u|x) = E(u) = 0 \quad \text{(zero conditional mean)}}$

What does this imply about E(y|x)?

$$\boxed{E(y|x) = \beta_0 + \beta_1 x}$$

Interpretation of β_1?

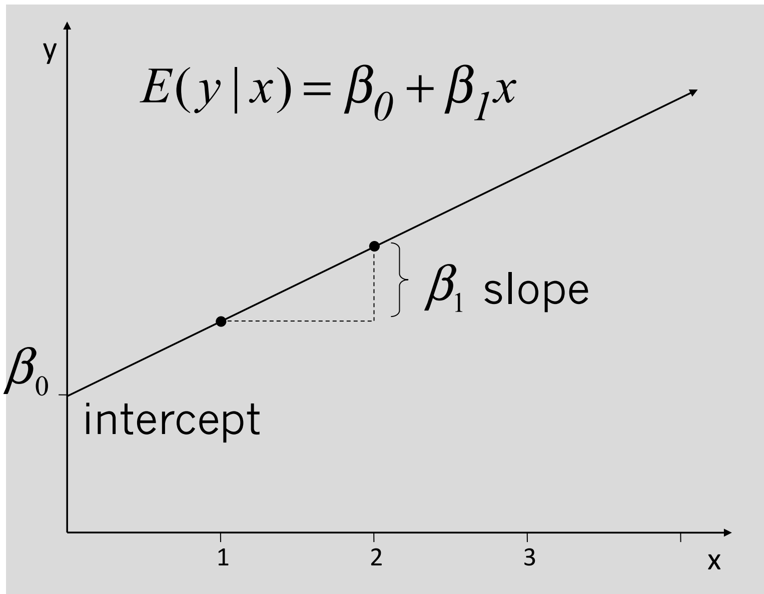
$$\frac{\partial E(y|x)}{\partial x} = \beta_1$$

The expected or average change in y given a one unit increase in x, ceteris paribus (slope)

Interpretation of β_0?

β_0 is the expected value of y when $x = 0$ (intercept)

3



Aside: NPR "Hidden Brain" example of a natural experiment, and when it might be reasonable to assume E(u|x)=E(u)

- Listen for the following:
  - *What is the dependent variable?*
  - *What is the main explanatory variable of interest?*
  - *Why might it be reasonable to assume E(u|x)=E(u) here?*
  - *What is a natural experiment?*
- Dependent variable: cognitive function of elderly
- Main explanatory variable: wealth
- E(u|x)=E(u) might be reasonable – Congress computational mistake – people in one cohort got higher benefits that next cohort (level of benefits shouldn't be correlated with unobservables)

5

MICHIGAN STATE UNIVERSITY

Aside: Natural experiments

A natural experiment occurs when some exogenous event—often a change in government policy—changes the environment in which individuals, families, firms, or cities operate. A natural experiment always has a control group, which is not affected by the policy change, and a treatment group, which is thought to be affected by the policy change. Unlike with a true experiment, where treatment and control groups are randomly and explicitly chosen, the control and treatment groups in natural experiments arise from the particular policy change. (Wooldridge, 2003: 417)

6

Review: Ordinary least squares (OLS) approach

What is the OLS approach to estimating β_0 and β_1 ?

- Minimize the sum of squared residuals (difference b/w observed & estimated values of y_i)

What are the estimated values of y_i called? Formula?

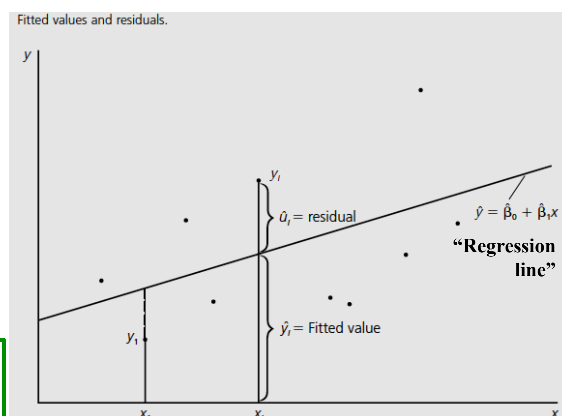
Fitted (or predicted) values of y_i:
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

What are the residuals?

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

OLS:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} \hat{u}_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$



Source: Wooldridge (2003)

MICHIGAN STATE UNIVERSITY

Review: the OLS estimators for β_0 and β_1

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}}$$

8

Obtaining OLS estimates – example (Stata)

Wooldridge (2003) Example 2.4: Wage and education
Use Stata to run the simple linear regression of wage (*y*) on educ (*x*).

Command: regress wage educ (or: reg wage educ)

reg wage educ

Source	SS	df	MS
Model	1179.73204	1	1179.73204
Residual	5980.68225	524	11.4135158
Total	7160.41429	525	13.6388844

Number of obs	=	526
F(1, 524)	=	103.36
Prob > F	=	0.0000
R-squared	=	0.1648
Adj R-squared	=	0.1632
Root MSE	=	3.3784

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

$\hat{\beta}_1$ $\hat{\beta}_0$

MICHIGAN STATE UNIVERSITY

Basic Stata commands

- **regress** *y x* Linear regression of *y* on *x*
  - EX) regress wage educ OR reg wage educ
- **predict** *newvar1*, **xb** Compute fitted values
  - EX) predict wagehat, xb (I just made up the name wagehat)
- **predict** *newvar2*, **resid** Compute residuals
  - EX) predict uhat, resid (I just made up the name uhat)

10

MICHIGAN STATE UNIVERSITY

Algebraic properties of OLS

Recall: $\hat{u}_i = y_i - \hat{y}_i$
$= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

1. $\sum_{i=1}^{N} \hat{u}_i = 0$ Follows from F.O.C. w.r.t. $\hat{\beta}_0$: $\sum_{i=1}^{N} y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$

2. $\sum_{i=1}^{N} x_i \hat{u}_i = 0$ Follows from F.O.C. w.r.t. $\hat{\beta}_1$: $\sum_{i=1}^{N} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

3. The point $(\bar{x}, \bar{y})$ is always on the OLS regression line

You proved this on HW question 11.1.

4. $y_i = \hat{y}_i + \hat{u}_i$ Because $\hat{u}_i = y_i - \hat{y}_i$.

11

Total, explained, & residual sum of squares, R^2

Total sum of squares: $$\boxed{SST \equiv \sum_{i=1}^{N} (y_i - \overline{y})^2}$$

Explained sum of squares: $$\boxed{SSE \equiv \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2}$$

Residual sum of squares: $$\boxed{SSR \equiv \sum_{i=1}^{N} \hat{u}_i^2}$$

$$\boxed{SST = SSE + SSR}$$

Proof is on p. 39 of Wooldridge (2003)

Coefficient of determination or R^2: *Interpretation?*

$$\boxed{R^2 = SSE / SST = 1 - (SSR / SST)}$$

The proportion of the sample variation in y that is explained by x

SST (total SS), SSE (explained SS), SSR (residual SS), and R^2 in Stata

```
use "/Users/nicolemason/Documents/AEC802/data/WAGE1_Stata13.dta"

reg wage educ
```

Source	SS	df	MS
Model	1179.73204	1	1179.73204
Residual	5980.68225	524	11.4135158
Total	7160.41429	525	13.6388844

SSE → Model SS; SSR → Residual SS; SST → Total SS

Number of obs	=	526
F(1, 524)	=	103.36
Prob > F	=	0.0000
R-squared	=	0.1648
Adj R-squared	=	0.1632
Root MSE	=	3.3784

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

$$\boxed{SST = SSE + SSR} \qquad \boxed{R^2 = SSE / SST = 1 - (SSR / SST)}$$

13

Why is it called R^2?

- Letter R sometimes used to refer to correlation coefficient (we used ρ)

R^2 is the squared sample correlation coefficient between y_i and $\hat{y}_i$

```
. reg wage educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

```
. predict wagehat, xb

.
. corr wage wagehat
(obs=526)

             |     wage  wagehat
-------------+------------------
        wage |   1.0000
     wagehat |   0.4059   1.0000

.
. display 0.4059^2
.16475481
```

My R^2 is too low!

Does a low R^2 mean the regression results are useless? Why or why not?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{\beta}_1$ may still be <u>good (unbiased) estimate of *ceteris paribus* effect</u> of x on y even if R^2 is low

15

MICHIGAN STATE UNIVERSITY

Statistical properties of OLS

- **In order to draw inferences** about population parameters β_0 *and* β_1 from our OLS estimates, **need to know the sampling distributions** thereof:
  - Expected value
  - Variance
  - Etc.

- Once we know the sampling distribution and have an estimate of the variance, then we can compute **Z and T statistics, construct confidence intervals, and do hypothesis testing**

16

MICHIGAN STATE UNIVERSITY

Unbiasedness of OLS (simple linear regression)

If the following 4 assumptions hold, then OLS is unbiased. (OLS is also consistent under these assumptions, and under slightly weaker assumptions → AFRE 835.)

SLR.1. Linear in parameters: $y = \beta_0 + \beta_1 x + u$

SLR.2. Random sampling

****SLR.3. Zero conditional mean (exogeneity):**

$$\boxed{E(u\,|\,x) = E(u) = 0}$$

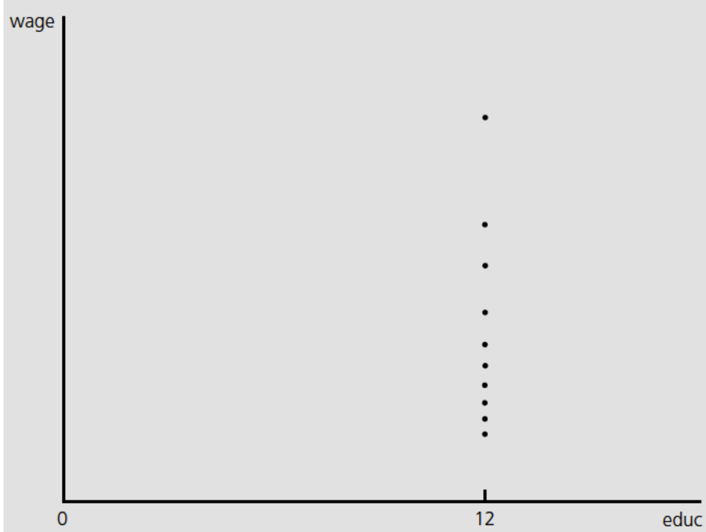
SLR.4. Sample variation in x

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Why necessary?

Can't estimate slope parameter if no variation in x

A scatterplot of wage against education when $educ_i$ = 12 for all i.



Source: Wooldridge (2003)

OLS estimators for β_0 and β_1 are unbiased under SLR.1-SLR.4

$$\boxed{E(\hat{\beta}_1) = \beta_1 \quad \text{and} \quad E(\hat{\beta}_0) = \beta_0}$$

- Leave proof for AFRE 835
  - WMS pp. 577-578 and Wooldridge pp. 46-50
- The key assumption is E(u|x) = E(u) = 0 (zero conditional mean / exogeneity) – SLR.3
  - Under SLR.1-SLR.4, OLS estimate of β_1 is the **causal effect (*ceteris paribus* effect)** of x on y
  - If E(u|x) ≠ E(u), then x is **endogenous** to y → OLS estimates biased

MICHIGAN STATE UNIVERSITY

Variance of the OLS estimators

Let $V(u) = \sigma^2$

SLR.5. Homoskedasticity (constant variance):

$$\boxed{V(u\,|\,x) = V(u) = \sigma^2}$$

What does this imply about V(y|x)?
(Hint: Plug in $y = \beta_0 + \beta_1 x + u$ and use the rules for conditional variances.)

$$\boxed{V(y\,|\,x) = V(u\,|\,x) = \sigma^2}$$

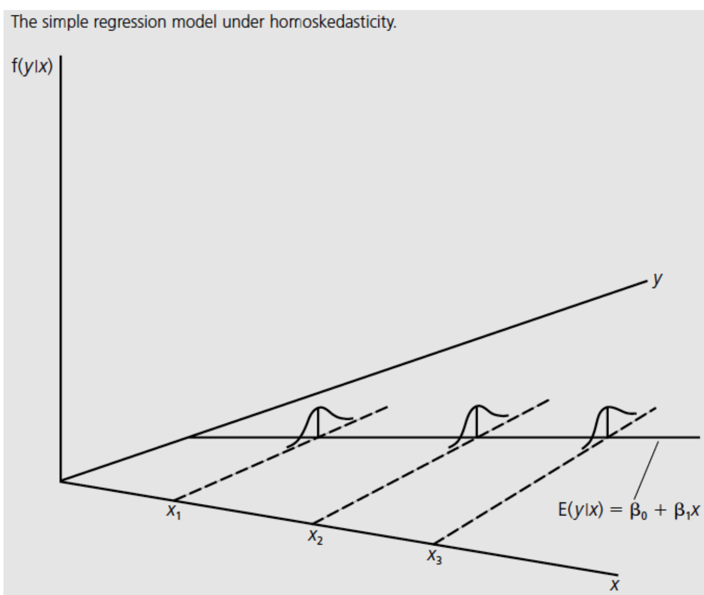
Reminder: What is E(y|x) given our earlier assumptions?

$$\boxed{E(y\,|\,x) = \beta_0 + \beta_1 x}$$

20

The simple regression model under homoskedasticity

MICHIGAN STATE UNIVERSITY

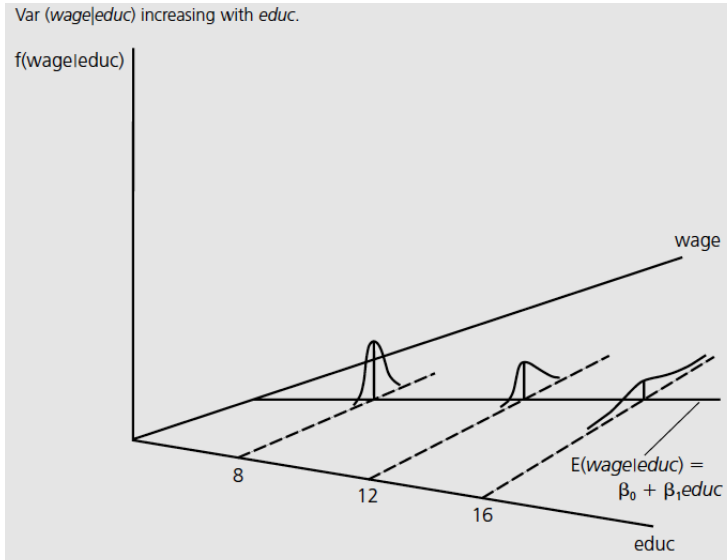


Source: Wooldridge (2003)

21

The simple regression model under heteroskedasticity

Heteroskedasticity: $V(u \mid x) \neq V(u) \Rightarrow V(y \mid x)$ is a function of x



Source: Wooldridge (2003)

22

Under SLR.1-SLR.5, the variances of the OLS estimators for β_0 and β_1 *are:*

$$\boxed{V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

$$\boxed{V(\hat{\beta}_0) = \frac{\sigma^2 N^{-1} \sum_{i=1}^{N} x_i^{\,2}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

- Leave proof for AFRE 835
  - WMS pp. 578-579 and Wooldridge pp. 55
- With heteroskedasticity, these formulas are incorrect (and the correct formulas a more complicated)
- Note: **SLR.5 NOT needed for unbiasedness**

23

$$\boxed{V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

- *What happens to this variance as:*
  - *V(u) increases?*
  - *The sample variation in x increases?*
- *What happens to the standard errors of our OLS estimator when their variances increases?*
  - *Then what happens to our T and Z stats?*
  - *Then what happens to our probability of rejecting H0 in favor of H1?*

24

Gauss-Markov Theorem
(simple linear regression, cross-sectional data case)

Under SLR.1-SLR.5, OLS is BLUE

Best (most efficient, i.e., smallest variance)

Linear (linear function of y_i)

Unbiased

Estimator

25

MICHIGAN STATE UNIVERSITY

Estimating $V(u) = \sigma^2$

- Almost have all the pieces we need to do inference. Need to estimate $V(u) = \sigma^2$.
- Starting point:

$$\sigma^2 = V(u) = E(u^2) - [E(u)]^2 = E(u^2) \quad \textit{Why?}$$

- *What would the method of moments estimator be here?* $\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2$ Consistent but biased estimator of σ^2
- **Unbiased & consistent estimator for σ^2:**

$$\boxed{\hat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{SSR}{N-2}}$$

B/c N-2 d.o.f. = # of observations minus # of estimated parameters

Leave proof for AFRE 835. See WMS pp. 580-581, Wooldridge pp. 57

Putting it all together: simple linear regression

$$y = \beta_0 + \beta_1 x + u$$

OLS estimators for β_0 and β_1:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \qquad \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}}$$

Expected values (under SLR.1-SLR.4):

$$\boxed{E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0}$$

Sample variances (under SLR.1-SLR.5):

$$\boxed{\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \qquad \boxed{\hat{V}(\hat{\beta}_0) = \frac{\hat{\sigma}^2 N^{-1}\sum_{i=1}^{N}x_i^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \qquad \boxed{\text{where } \hat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{SSR}{N-2}}$$

$\hat{\sigma}$ is the **standard error** of the regression

In Stata

```
use "/Users/nicolemason/Documents/AEC802/data/WAGE1_Stata13.dta"

reg wage educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

$\hat{\sigma}^2$ (Residual MS 11.4135158)

$\hat{\sigma}$ (Root MSE = 3.3784)

$\sqrt{\hat{V}(\hat{\beta}_i)},\ i = 0,1$ (Std. Err. .053248, .6849678)

28

Homework:

- WMS Ch. 11 (cont'd):
  - Find the total, explained and residual sum of squares (SST, SSE, SSR), R^2 (and interpretation), estimate of σ^2, and estimates of the variances of the OLS estimators for β_0 and β_1 for WMS 11.3 (Excel), 11.4 (Stata), and 11.5 (Stata)
- Complete all Ch. 11 HW before last day of class so that we can go over it then (you won't turn in Ch. 11)

Next class:

Linear regression part 3 of 3

- Classical Linear Model
- Inference

Reading for next class:

- WMS Ch. 11: sections 11.5
- Wooldridge *Introductory Econometrics* (2003): pp. 113-136

Estimating the variances of the OLS estimators for β_0 and β_1 - example 11.1 (cont'd)

We found: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 1 + 0.7x_i$

Table 11.1 Data for Example 11.1

x	y
−2	0
−1	0
0	1
1	1
2	3

$$\boxed{\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}}$$

$$\boxed{\hat{V}(\hat{\beta}_0) = \frac{\hat{\sigma}^2 N^{-1} \sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}}$$

$$\boxed{\text{where } \hat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{SSR}{N-2}}$$

30