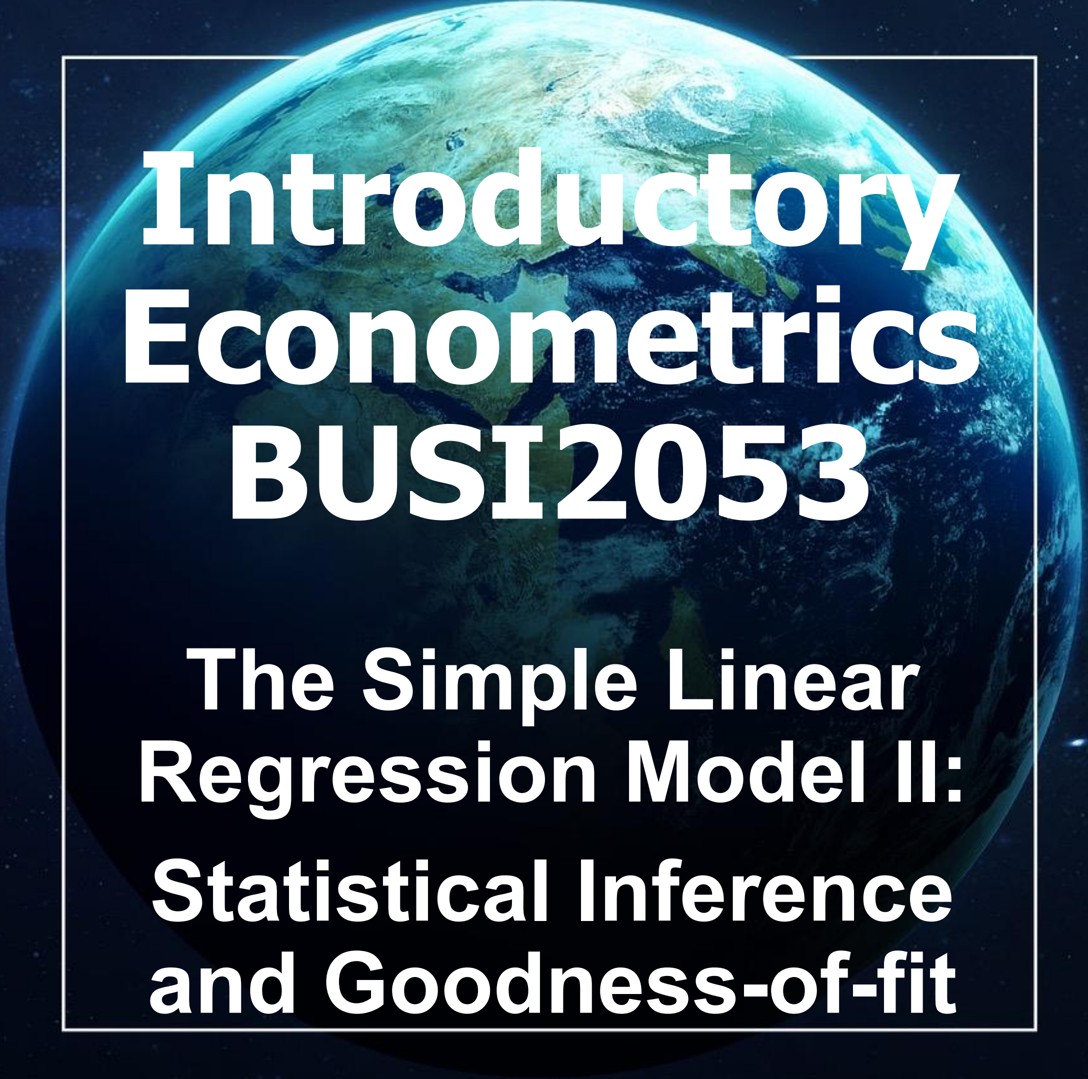
University of Nottingham

UK | CHINA | MALAYSIA



Introductory Econometrics BUSI2053

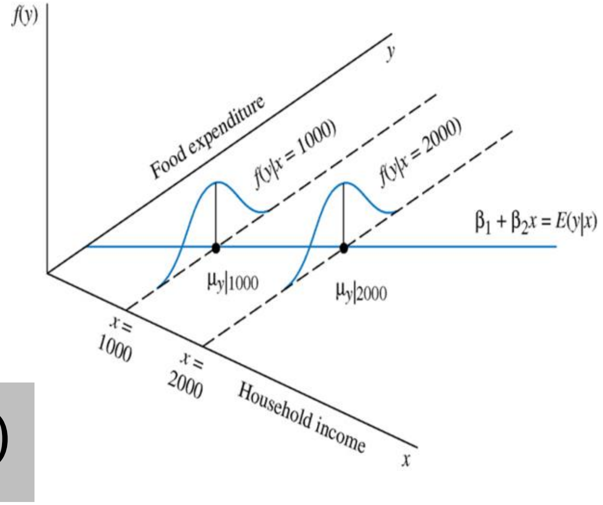
The Simple Linear Regression Model II: Statistical Inference and Goodness-of-fit

Lecture Outline

- Assumptions of OLS and G-M condition
- Properties of the OLS estimator
- Variance of OLS Estimator
- Statistical Inference
- Measuring Goodness-of-fit

- **Suggested Reading:**
- Chapter 2, 3 & 4.1, 4.2: Hill, R.C., Griffiths W.E. and Lim, G.C. Principles of Econometrics, fourth edition, Wiley, 2012 (pp. 39-110)
- Chapter 7-8, 15.2: Westhoff (2013) An introduction to econometrics: a self-contained approach, MIT Press, 2013
- Chapter 1-3: Gujarati, D.N. and Porter D.C. Basic econometrics, 5th ed., McGraw-Hill, 2009 (pp. 34-117)
- Chapter 1 & 2: Dougherty, Christopher. Introduction to econometrics. 4th ed. Oxford University Press, 2011 (pp.83-147)

Assumptions of the Simple Linear Regression



SR1:The value of y, for each value of x, is given by the linear regression: $y = \beta_1 + \beta_2 x + e$

SR2: The expected value of the random error e is: $E(e_i) = 0$

SR3: The variance of the random error e is: $\mathrm{var}(e_i) = \sigma^2 = \mathrm{var}(y_i)$

SR4: The covariance between any pair of random errors, e_i and e_j is:

$$\mathrm{cov}(e_i, e_j) = \mathrm{cov}(y_i, y_j) = 0$$

SR5: The variable x is not random, and must take at least two different values

SR6: Optional $e \sim N(0, \sigma^2)$

Properties of least square estimator

Gauss-Markov condition

Under the assumptions SR1-SR5 of the linear regression model, the OLS estimators b_1 and b_2 have the smallest variance of all linear and unbiased estimators of β_1 and β_2. They are the **B̲est L̲inear U̲nbiased E̲stimators (BLUE)** of β_1 and β_2.

1. The estimators b_1 and b_2 are “best” compared to similar estimators because they have the minimum variance.

2. The **estimators** b_1 and b_2 (not specific b_1 and b_2) are linear (linear combination of the dependent variable) and unbiased ($E(b_1) = \beta$).

$$b_2 = \frac{\square(x_i - \overline{x})(y_i - \overline{y})}{\square(x_i - \overline{x})^2}$$

3. If any of these assumptions are *not* true, then b_1 and b_2 are *not* **BLUE**

See optional slides on Moodle for the proof of the properties

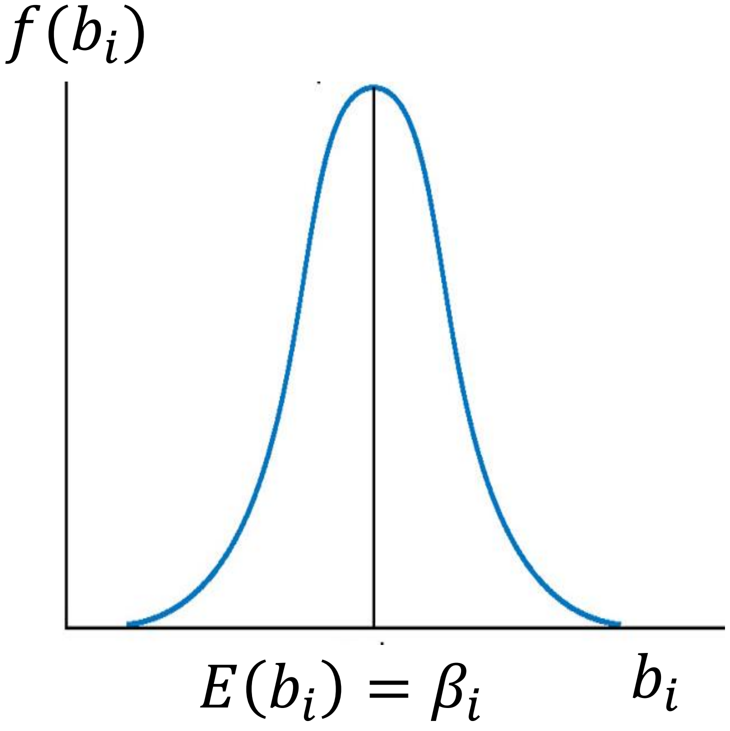
Sampling variation of a least squares estimator

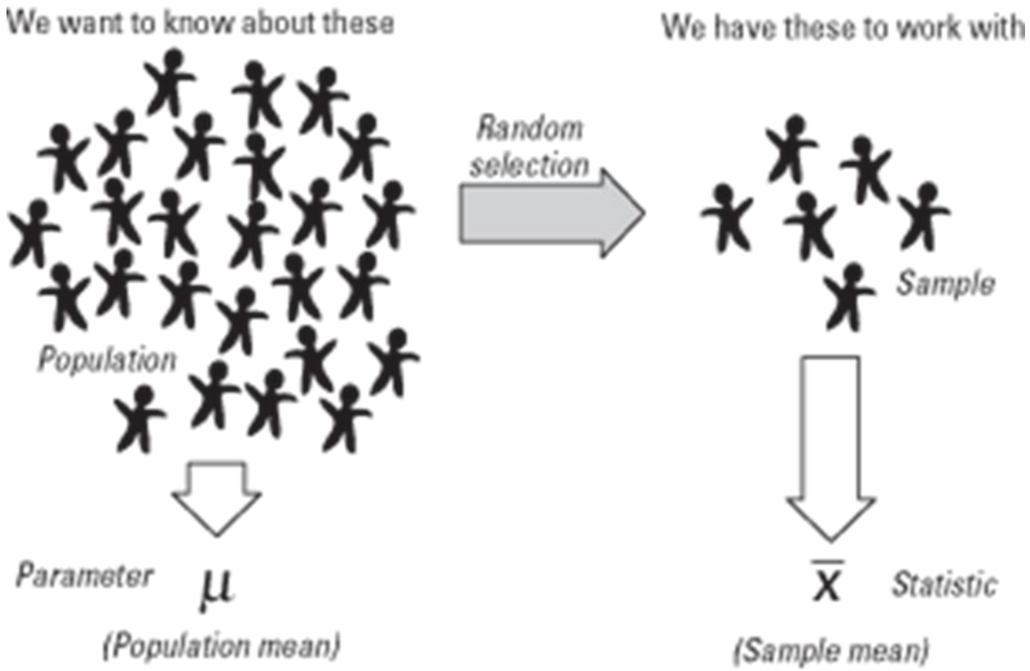
- We are interested to estimate the relationship between the dependent variable y and independent variable x from the population regression model: $y = \beta_1 + \beta_2 x + e$

- We can estimate the population parameters β_1 and β_2 using least squares method and obtain sample (fitted) regression equation: $\hat{y}_i = b_1 + b_2 x_i$

- the OLS estimators b_1 and b_2 are estimators, hence these are random variables that varies according to its associated probability distribution.

- We need to look at the variation of estimators (variance) which represents the accuracy of our estimations.




We want to know about these
We have these to work with
Random selection
Population
Sample
Parameter
μ
(Population mean)
$\bar{X}$
Statistic
(Sample mean)

Variance of least squares estimator

Variances of b_1 and b_2

- If the assumptions SR1-SR5 are correct (SR6 is not required), then the variances of b_1 and b_2 are:

$$\operatorname{var}(b_1) = \sigma_e^2 \left[\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right] \qquad \operatorname{var}(b_2) = \frac{\sigma_e^2}{\sum (x_i - \bar{x})^2}$$

σ_e^2=variance of random error term $= \operatorname{var}(e)$

- The larger the sum of squares, $\sum (x_i - \bar{x})^2$, the smaller the variances
- The larger the sample size n, the smaller the variances and covariance of the least squares estimators.

See optional slides on Moodle for the derivation of var(b_2)

Estimating the variances of b_1 and b_2

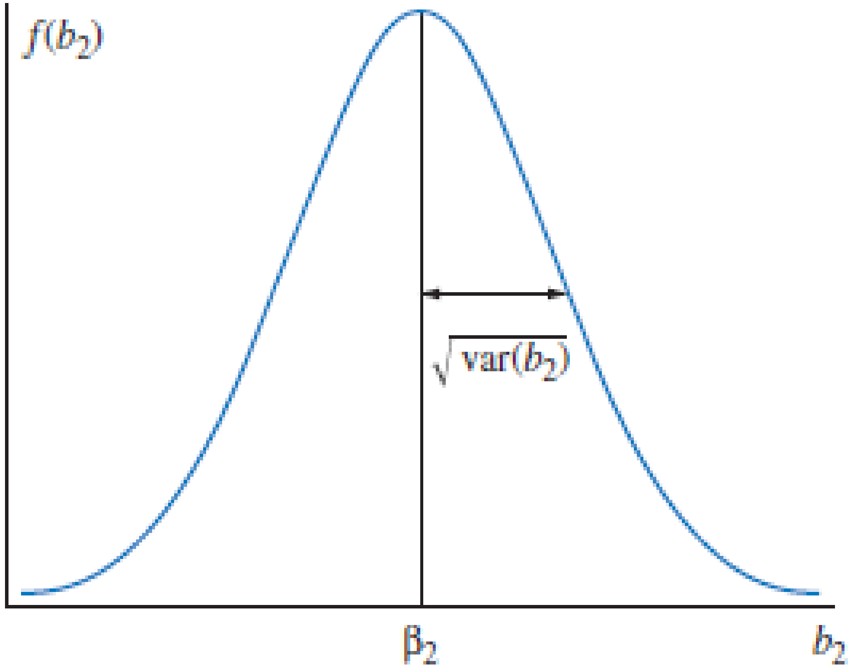
- Replace the unknown error variance $\sigma_e{}^2$ by $\hat{\sigma}_e^2$ to obtain:

$$\widehat{\text{var}}(b_1) = \hat{\sigma}_e^2 \left[\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}\right] \qquad \widehat{\text{var}}(b_2) = \frac{\hat{\sigma}_e^2}{\sum(x_i - \bar{x})^2}$$

- The "standard errors" of b_1 and b_2

$$\text{s.e}(b_1) = \sqrt{\widehat{\text{var}}(b_1)};$$

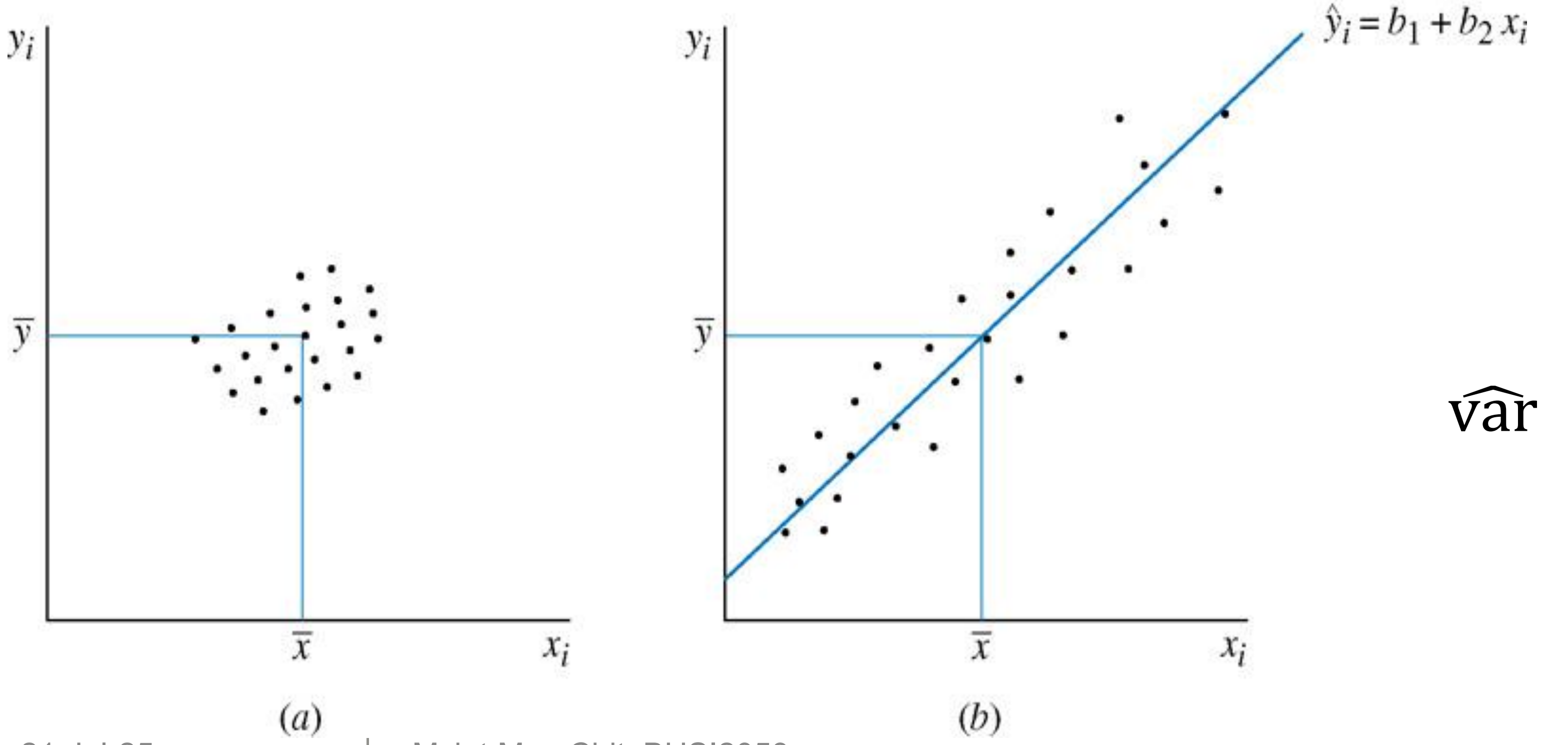
$$\text{s.e}(b_2) = \sqrt{\widehat{\text{var}}(b_2)}$$



Variance of least squares estimator

Variances of b_1 and b_2

- The influence of variation in the explanatory variable x on precision of estimation (a) Low x variation, low precision (b) High x variation, high precision



$$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}_e^2}{\sum(x_i - \bar{x})^2}$$

Statistical Inference (Hypothesis testing)

- Statistical inference: Making a conclusion (testing hypothesis) about a population using the information contained in a sample of data
  - Given an economic and statistical model, hypotheses are formed about economic behavior.
  - Hypothesis tests use the information about a parameter that is contained in a sample of data, its least squares point estimate, and its standard error, to draw a conclusion about the population parameter

Statistical Inference (Hypothesis testing)

Testing Hypotheses

When testing the null hypothesis H_0: $\beta_i = c$ against the alternative hypothesis H_1:$\beta_i \neq c$ (2-tailed); H_1:$\beta_i > c$ or $< c$ (1-tailed),

Test statistic: $t = \dfrac{b_i - c}{s.e(b_i)}$

Reject the null hypothesis and accept the alternative hypothesis if the absolute value of the test statistic $|\, t| \geq t_{(\alpha;n\text{-}k)}$ (k = number of parameters)

Alternatively, reject the null hypothesis if the observed *P*-value of the test is less than the level of significance (α).

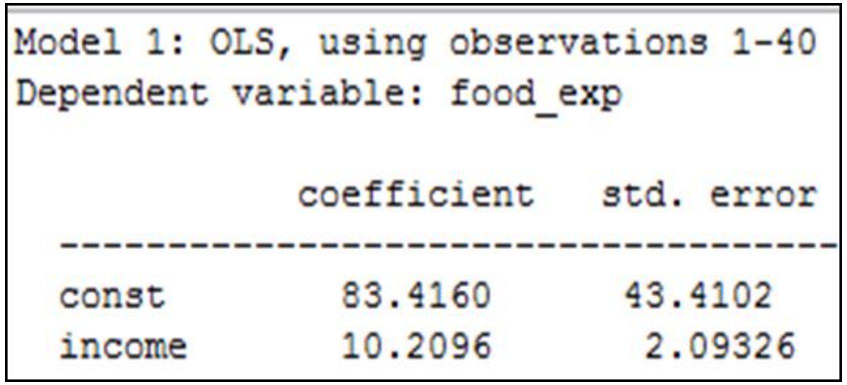
Statistical Inference (Hypothesis testing)

Testing Hypotheses (Food Expenditure example)

1. The null H_0:$\beta_2 = 0$; The alternative H_1:$\beta_2 > 0$ One-tailed test
2. Using the food expenditure data, $b_2 = 10.21$ with standard error se(b_2) = 2.09; n=40

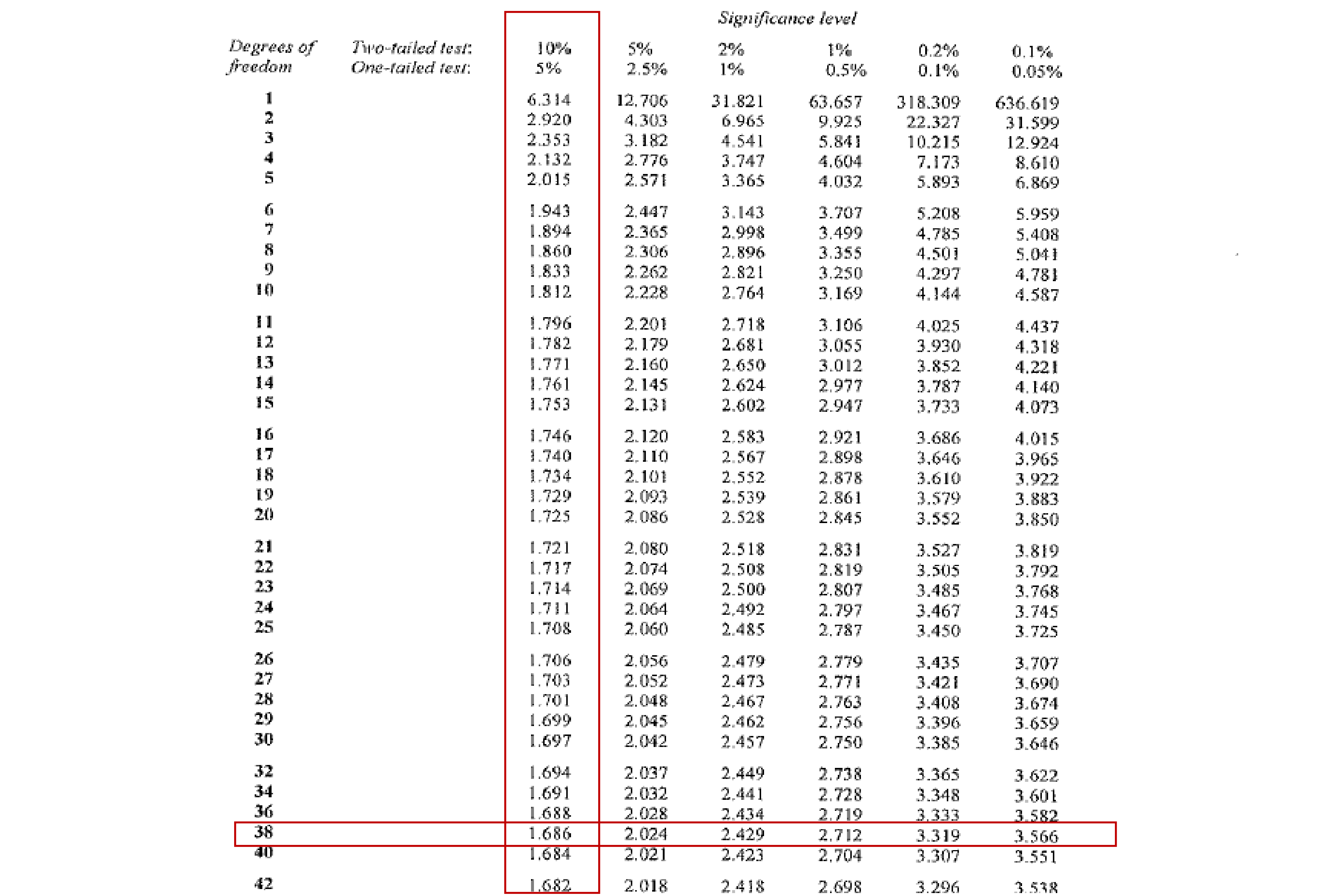
$$t = \frac{b_i - c}{s.e(b_i)} = \frac{10.21 - 0}{2.09} = 4.88$$

3. Select $\alpha = 0.05$; the critical value with $n - k$ =40-2= 38 degrees of freedom, $t_{(0.05,38)} = 1.686$.
4. Decision: Since the calculated t-value 4.88>1.686, reject the H_0 in favour of the alternative that $\beta_2 > 0$
5. Conclusion: there is a statistically significant positive effect of household income on food expenditure



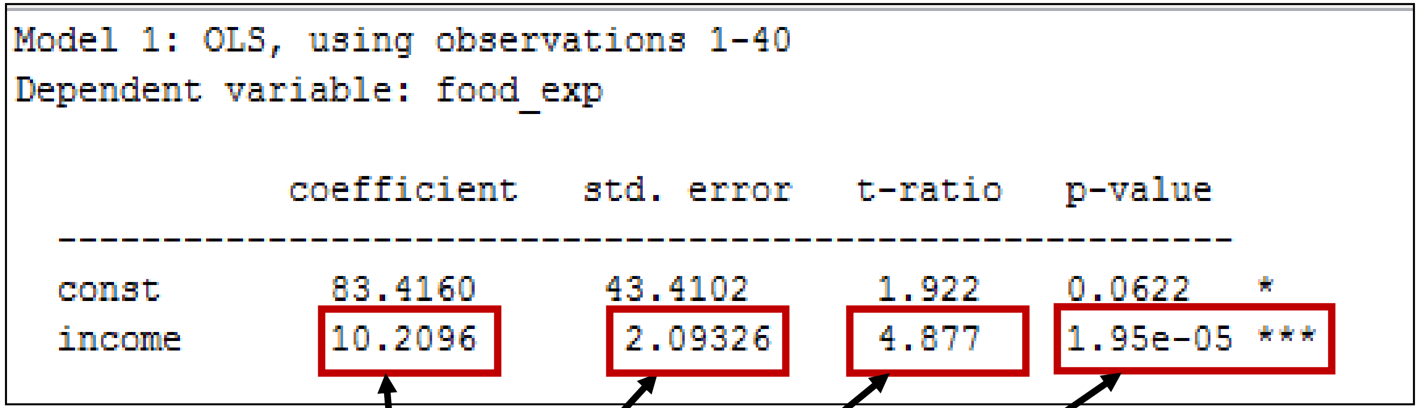
```
Model 1: OLS, using observations 1-40
Dependent variable: food_exp

             coefficient   std. error
  ------------------------------------
  const        83.4160      43.4102
  income       10.2096       2.09326
```

t-distribution table



		Significance level					
Degrees of freedom	*Two-tailed test:*	10%	5%	2%	1%	0.2%	0.1%
	One-tailed test:	5%	2.5%	1%	0.5%	0.1%	0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538

Testing Hypotheses (gretl regression output)



```
Model 1: OLS, using observations 1-40
Dependent variable: food_exp

             coefficient   std. error   t-ratio   p-value
  ----------------------------------------------------------
  const        83.4160      43.4102      1.922    0.0622    *
  income       10.2096       2.09326     4.877    1.95e-05 ***
```

P-value = Prob(Sample Results | H_o true)

$$t = \frac{b_i - c}{s.e(b_i)} = \frac{10.21 - 0}{2.09} = 4.88$$

$$p - value = 1.95\text{e}{-}05 = 1.95 \times 10^{-5} = \frac{1.95}{100000} = 0.0000195$$

Decision: p-value = 0.0000195/2 <0.01; reject the H_o; *** indicates that the coefficient is statistically significant at 1% level.

Note: p-value in GRETL result are for two-tailed test. For one-tailed test divide p-value by 2.

Statistical Inference (Hypothesis testing) Notes

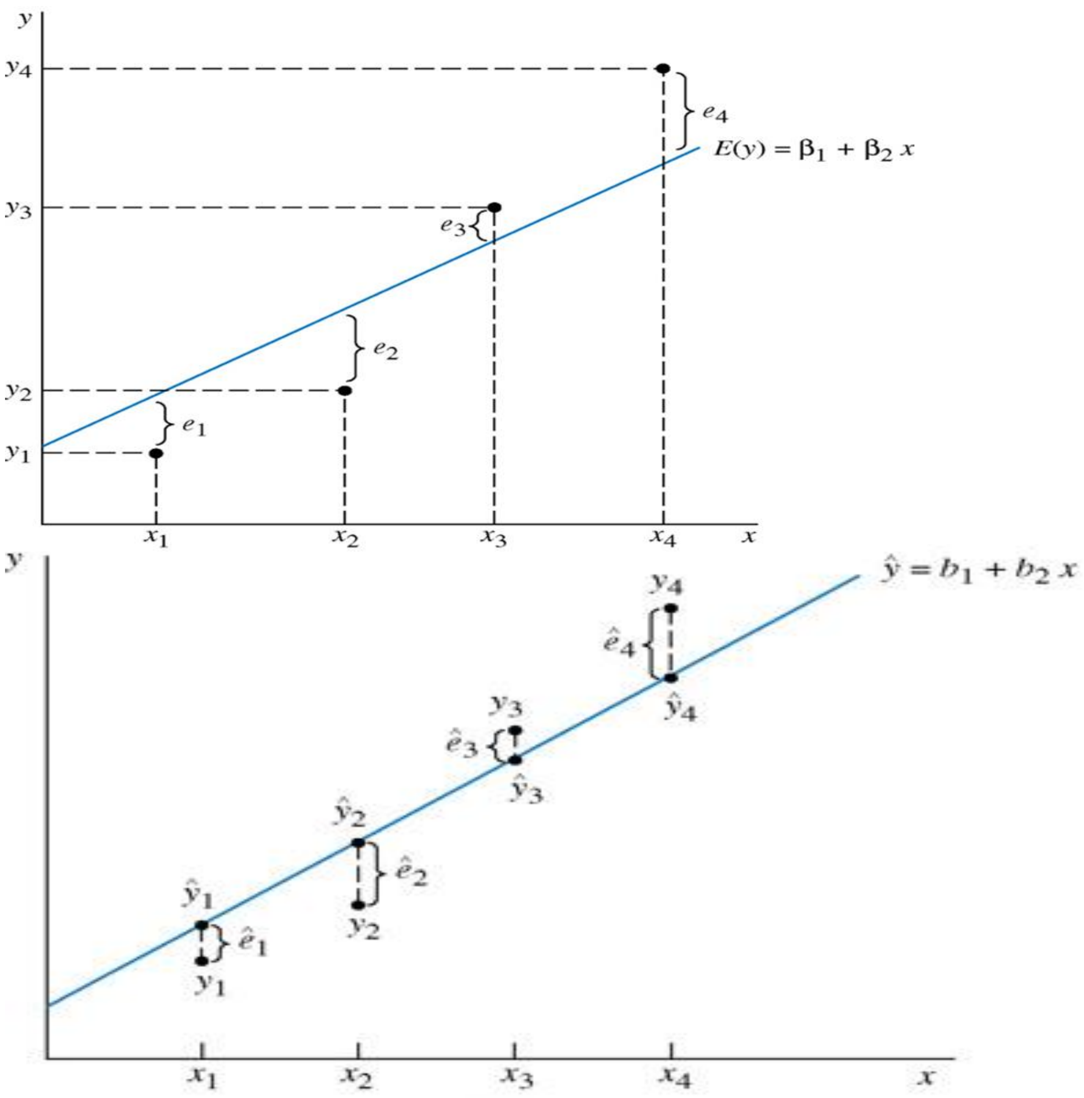
- In Regression analysis, the default null hypothesis is to test that the population parameter equals to zero (H_0:$\beta_i = 0$), i.e., no influence of the independent variable (X) on the dependent variable (Y).
- Decide one-tailed or two-tailed test (based on research hypothesis/question) and set null and alternative hypotheses accordingly
- If *p*-value is not given (in exam), calculate *t*-statistic, compare with appropriate critical value and make decision
- If p-value is given, use the following decision rules **(Note: p-value in GRETL or SAS result are for 2-tailed test. For 1-tailed test, divide p-value in GRETL/SAS result by 2)**
  - p-value<0.01➔Reject H_0 at 1% significance level (or the test is significant at 1% level)
  - 0.01≤p-value<0.05 ➔ Reject H_0 at 5% significance level (or the test is significant at 5% level)
  - 0.05≤p-value<0.10 ➔ Reject H_0 at 10% significance level (or the test is significant at 10% level)
  - p-value≥0.1 ➔Do not reject the null hypothesis (The test is not significant)

Measuring Goodness-of-fit

- An objective of econometric analysis is to use x_i to explain as much of the variation in the dependent variable y_i as possible.
- To develop a measure of the variation in y_i that is explained by the model, separating y_i into its explainable and unexplainable components.

$$y_i = E(y_i) + e_i$$

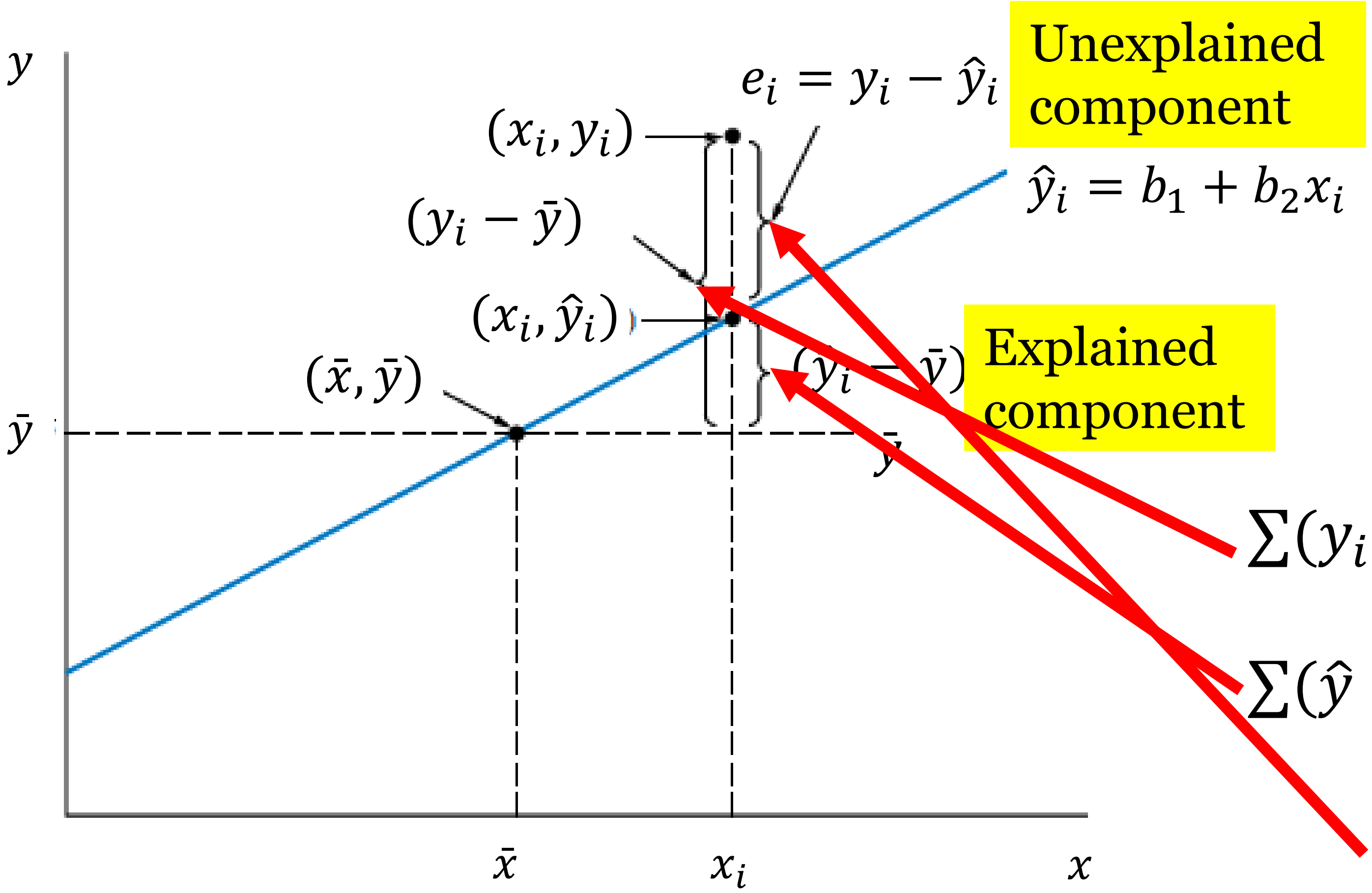
- we can rewrite $y_i = \hat{y}_i + \hat{e}_i$
- Subtracting the sample mean from both sides: $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i$



Measuring Goodness-of-fit

Graphically:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i$$



Unexplained component

Explained component

Squaring and summing both sides we get:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum\hat{e}_i^2$$

(see Hill et al. (2012) Appendix B4 for proof)

$\sum(y_i - \bar{y})^2 =$ Total Sum of Squares **(SST)** or (TSS)

$\sum(\hat{y} - \bar{y})^2 =$ Regression (or) Explained Sum of Squares **(SSR)** or (ESS)

$\sum\hat{e}_i^2 =$ Error (or) Residual Sum of Squares **(SSE)** or (RSS)

Measuring Goodness-of-fit

$\sum(y_i - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum\hat{e}_i^2$ can be written as

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

The **coefficient of determination**, or R^2 , as the proportion of variation in y explained by x within the regression model:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (0 \leq R^2 \leq 1)$$

Measuring Goodness-of-fit

For the food expenditure example, the sums of squares are:

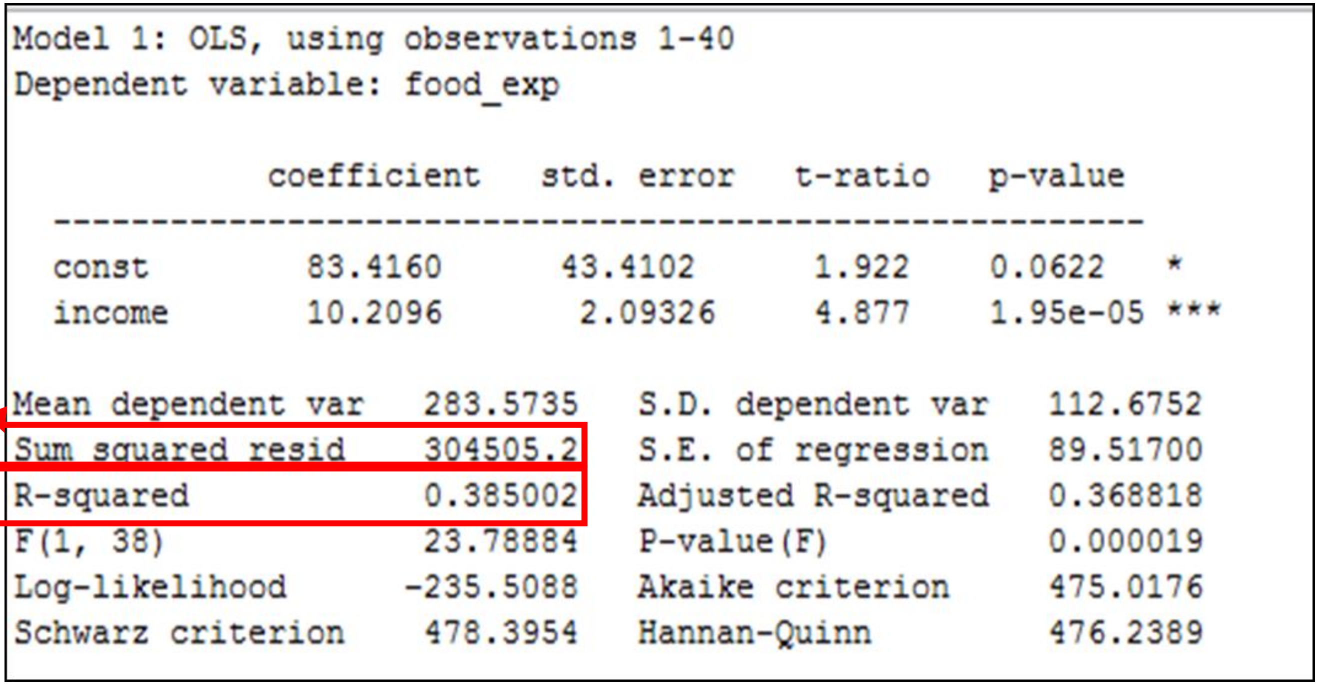
$$SST = \sum (y_i - \bar{y})^2 = 495132.160$$

$$SSE = \sum (y_i - \hat{y})^2 = \sum \hat{e}_i^2 = 304505.176$$

Therefore: $R^2 = 1 - \frac{SSE}{SST}$

$$= 1 - \frac{304505.176}{495132.160}$$

$$= 0.385$$

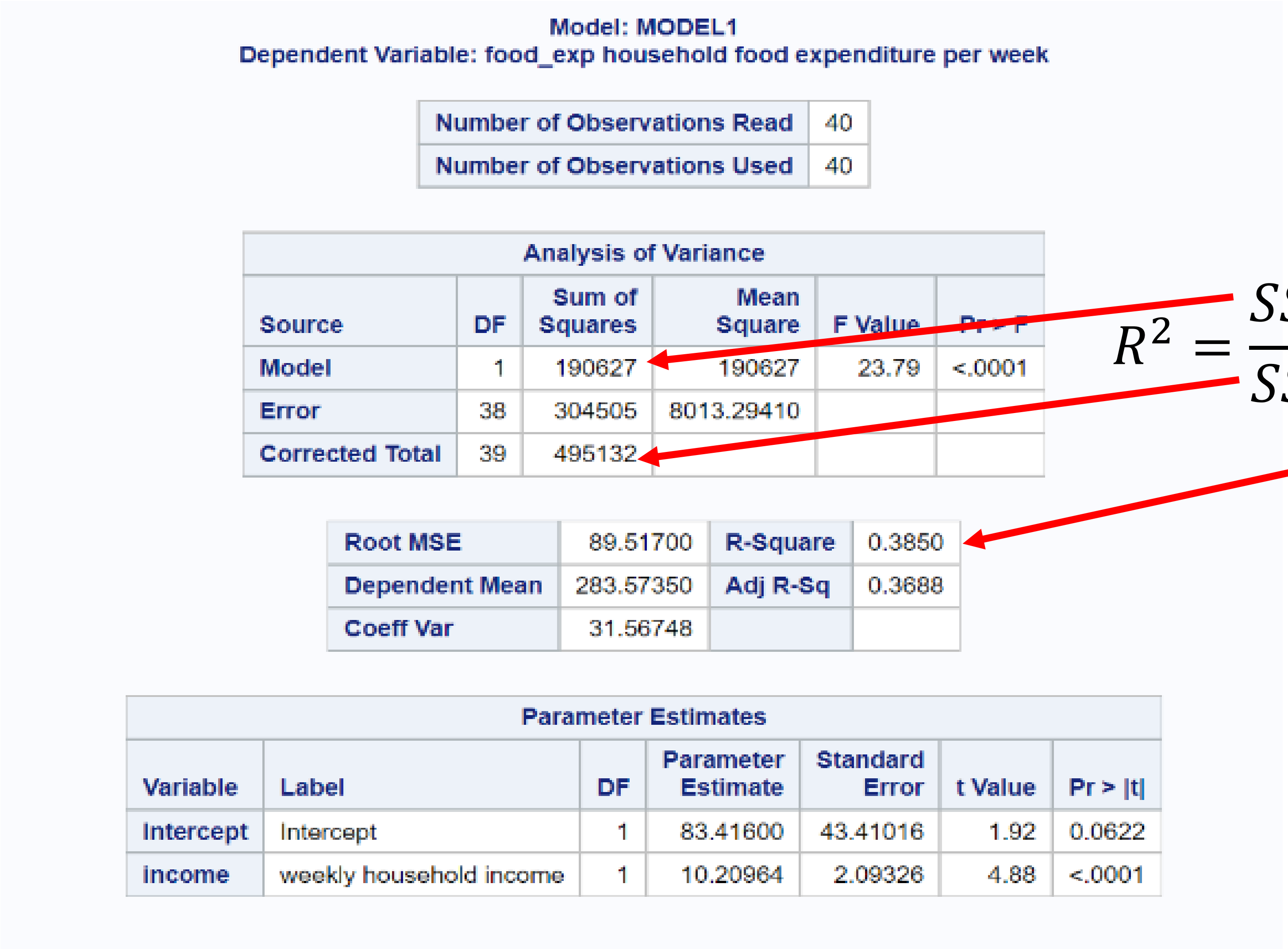


```
Model 1: OLS, using observations 1-40
Dependent variable: food_exp

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const        83.4160      43.4102      1.922    0.0622    *
  income       10.2096       2.09326     4.877    1.95e-05  ***

Mean dependent var   283.5735   S.D. dependent var   112.6752
Sum squared resid    304505.2   S.E. of regression   89.51700
R-squared            0.385002   Adjusted R-squared   0.368818
F(1, 38)             23.78884   P-value(F)           0.000019
Log-likelihood      -235.5088   Akaike criterion     475.0176
Schwarz criterion    478.3954   Hannan-Quinn         476.2389
```

We conclude that 38.5% of the variation in food expenditure (about its sample mean) is explained by our regression model, which uses only income as an explanatory variable

Note: for a simple regression $r_{xy}^2 = 0.62^2 = 0.385 = R^2$

Measuring Goodness-of-fit



Model: MODEL1
Dependent Variable: food_exp household food expenditure per week

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	190627	190627	23.79	<.0001
Error	38	304505	8013.29410		
Corrected Total	39	495132			

Root MSE	89.51700	R-Square	0.3850
Dependent Mean	283.57350	Adj R-Sq	0.3688
Coeff Var	31.56748		

Parameter Estimates

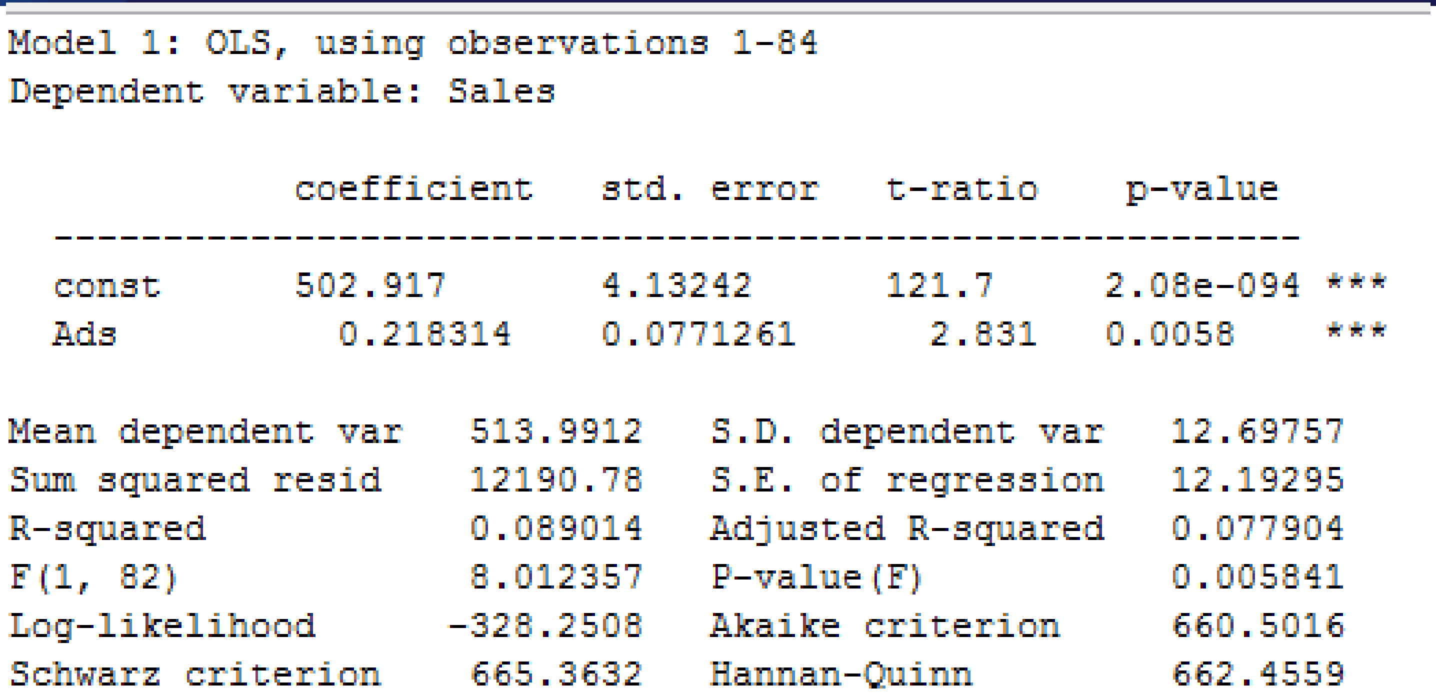
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	83.41600	43.41016	1.92	0.0622
income	weekly household income	1	10.20964	2.09326	4.88	<.0001

$$R^2 = \frac{SSR}{SST} = \frac{190627}{495132} = 0.385$$

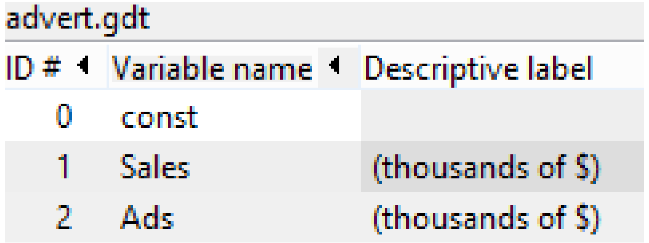
Interpreting the regression results

- The key ingredients in a regression results are:
  1. the coefficient estimates
  2. the standard errors (or *t*-values)
  3. an indication of statistical significance (p-values)
  4. R^2

Interpreting the regression results (Homework)



```
Model 1: OLS, using observations 1-84
Dependent variable: Sales

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const      502.917       4.13242      121.7     2.08e-094 ***
  Ads          0.218314    0.0771261      2.831   0.0058    ***

Mean dependent var   513.9912   S.D. dependent var   12.69757
Sum squared resid    12190.78   S.E. of regression   12.19295
R-squared            0.089014   Adjusted R-squared   0.077904
F(1, 82)             8.012357   P-value(F)           0.005841
Log-likelihood      -328.2508   Akaike criterion     660.5016
Schwarz criterion    665.3632   Hannan-Quinn         662.4559
```



advert.gdt

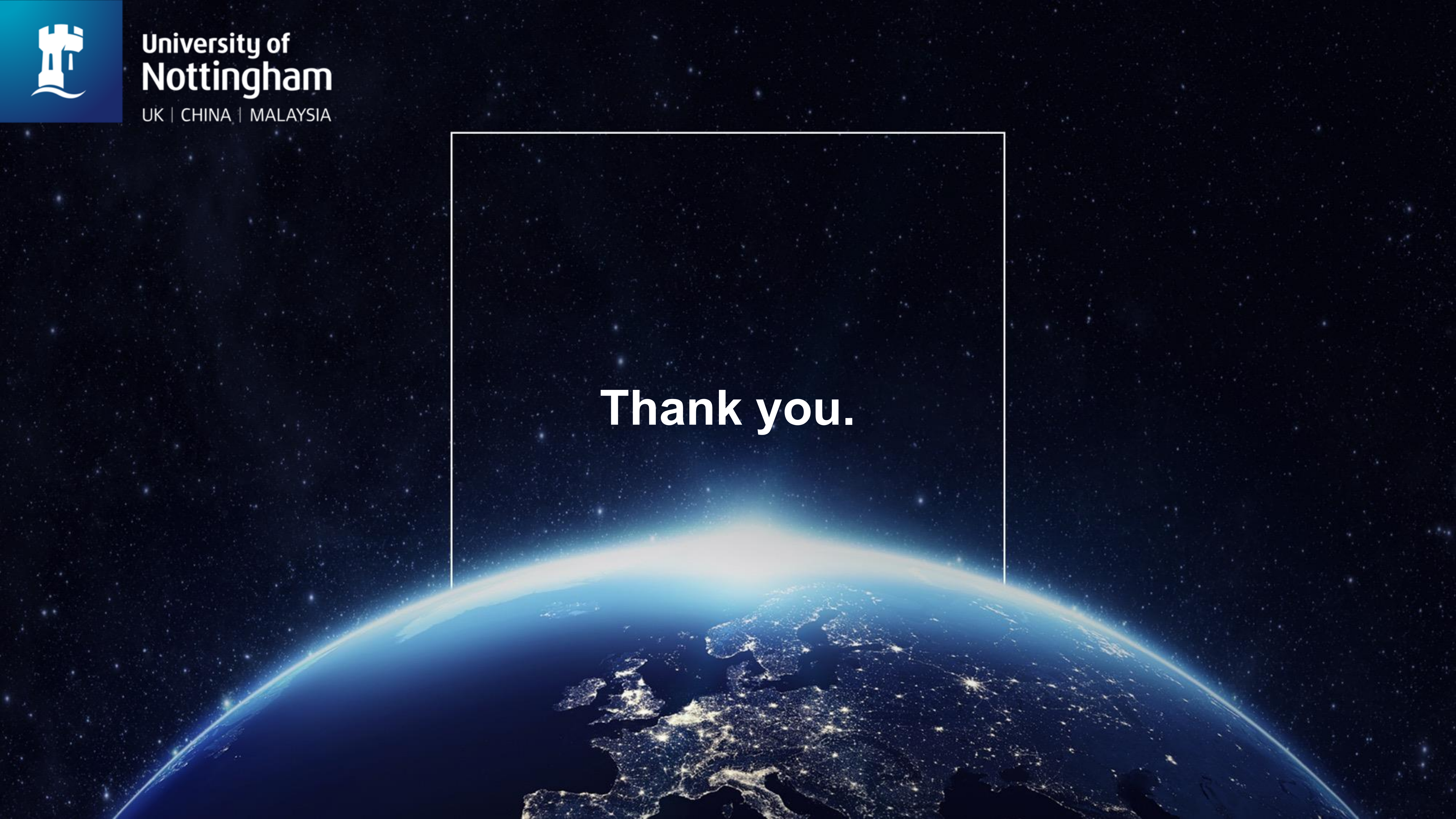
ID #	Variable name	Descriptive label
0	const	
1	Sales	(thousands of $)
2	Ads	(thousands of $)

- What is the hypothesis we should test?
- Write down the regression equation.
- Interpret the estimated coefficient and R².
- Can you reject the hypothesis at 1% significance level? What is P-value of the test.

Summary

After this lecture, you should be able to:

- identify and explain the keys assumption of simple linear regression
- properties of least square estimators
- understand the concept of the variation of OLS estimator
- estimate the regression parameters and interpret the results
- conduct a hypothesis test on the significance of a regression coefficient
- understand and interpret the measure of goodness-of-fit
- report the regression results



Thank you.