The Simple Linear Regression Model: Specification and Estimation

- Theory suggests many relationships between variables
- These relationships suggest that we can:
  - Answer the How Much Questions
  - Forecast or predict the corresponding value of one variable
- The SLRM can be used to accomplish these goals.

Suppose that we are interested in studying the relationship between consumption and income.

We can gather data from different households and then analyze it.

But how do we analyze it?

We know that theory suggests that consumption is a function of income (and other variables that for now we will ignore). Therefore our economic model would be something like this:

\[ C = F(I) \]
The Simple Linear Regression Model: Specification and Estimation

- We are interested in answering questions like: How much will consumption increase if income increases by 10%?

- To answer this question we need to assume a functional form for our consumption-income relationship like:

  \[ C = a + bI \]  

  (1)

Equation (1) represents our economic model, it shows that as income increases consumption increases.

NOTE: Economic models are not used to predict exact relationships among variables, they predict the average relationship between the variables.

We can transform our economic model, equation (1), into an econometric model.

- We can see that theory suggests a linear relation between \( C \) and \( I \).

- We must account for the fact that our theoretical model needs to be modified in order to include a random disturbances.

  Random what?? And why ????
Our econometric model includes the random disturbances and can be expressed as:

\[ C_i = \beta_1 + \beta_2 I_i + \epsilon_i \]

This is a simple linear regression model.

It is important to differentiate the econometric model in equation (2) from two other concepts/analysis that you may be familiar with:

- Correlation
- Causation

Now that we have a clear idea of what the regression model is all about, we must state and clearly explain the assumptions behind it.
The Simple Linear Regression Model: Specification and Estimation

- The Assumptions of the SLRM:
  1. The dependent variable can be expressed as a linear function of the regressors.
  2. The disturbance term has a mean of zero.
  3. The variance of the random error/disturbance is: \( \text{var}(e) = \sigma^2 = \text{var}(\epsilon) \)
  4. The covariance between any pair of random errors is zero.
  5. The variable x (regressor) is not random and must take at least two different values.
  6. (optional) The values of \( e \) are normally distributed about their mean, if the values of \( y \) are normally distributed, and vice versa.

Now the great task at hand is to estimate the values for \( \beta_1 \) and \( \beta_2 \).

We want the estimation to be a rule, something that makes sense and at the same time that is the best way to do it (or at least that follows logical principles).

Let's look at the data.

---

Fig. 14
The Simple Linear Regression Model: Specification and Estimation

- On the previous graph, draw a line that, for you, represents the linear relationship between the two variables (x and y).
- How are we going to conclude which of the lines we drew is the correct one?
- Is there any criteria that we can follow to choose the BEST line?

What is it that we are trying to achieve?

“We want to choose the line that minimizes the errors/residuals”

\[ e = Y - \hat{Y} \]

What do we mean with this statement?

We can use the Least Square Principle.

“To fit a line to the data values we should fit the line so that the sum of the squares of the vertical distances from each point to the line is as small as possible”
The Simple Linear Regression Model: Specification and Estimation

The Rule: Minimize the Sum of Squared Residuals

Min \( S = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} e_i^2 \)

The Simple Linear Regression Model: Specification and Estimation

Min \( S = \sum_{i=1}^{N} e_i^2 \)

Note: for this derivation the estimate of \( \beta_1 \) is \( a \) and the estimate of \( \beta_2 \) is \( b \)

\[ = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \]

\[ = \sum_{i=1}^{N} (y_i - a - b \cdot x_i)^2 \]

The Simple Linear Regression Model: Specification and Estimation

1-FOC:

\[ \frac{\partial \left( \sum_{i=1}^{N} e_i^2 \right)}{\partial a} = \sum_{i=1}^{N} (-2) \cdot (y_i - a - b \cdot x_i) \]

\[ = -2 \cdot \left[ \sum_{i=1}^{N} (y_i - a - b \cdot x_i) \right] = 0 \]
The Simple Linear Regression Model: Specification and Estimation

\[ \sum_{i=1}^{N} (y_i - a - b \cdot x_i) = 0 \]

Or,

\[ \sum_{i=1}^{N} e_i = 0 \]

**Recall:** For OLS the residuals sum to zero

---

The Simple Linear Regression Model: Specification and Estimation

2-FOC:

\[ \frac{\partial}{\partial b} \left( \sum_{i=1}^{N} e_i \right) = \sum_{i=1}^{N} \left( -2 \cdot x_i \right) \cdot (y_i - a - b \cdot x_i) \]

\[ = -2 \left[ \sum_{i=1}^{N} x_i \cdot (y_i - a - b \cdot x_i) \right] = 0 \]

---

The Simple Linear Regression Model: Specification and Estimation

\[ \sum_{i=1}^{N} x_i \cdot (y_i - a - b \cdot x_i) = \]

\[ \sum_{i=1}^{N} x_i \cdot e_i = 0 \]
The Simple Linear Regression Model: Specification and Estimation

After some algebra:

\[ \sum_{i=1}^{N} y_i = N \cdot a + b \cdot \sum_{i=1}^{N} x_i \quad \text{From 1 FOC} \]

and

\[ \sum_{i=1}^{N} x_i \cdot y_i = a \cdot \sum_{i=1}^{N} x_i + b \cdot \sum_{i=1}^{N} x_i^2 \quad \text{From 2 FOC} \]

The Simple Linear Regression Model: Specification and Estimation

\[ \sum_{i=1}^{N} y_i = N \cdot a + b \cdot \sum_{i=1}^{N} x_i \]

Divide by N:

\[ \bar{y} = a + b \cdot \bar{x} \]

Or,

\[ a = \bar{y} - b \cdot \bar{x} \]

Recall: OLS passes through the mean of the data

The Simple Linear Regression Model: Specification and Estimation

\[ \sum_{i=1}^{N} x_i \cdot y_i = a \cdot \sum_{i=1}^{N} x_i + b \cdot \sum_{i=1}^{N} x_i^2 \]

Substitute the solution for \( a \):

\[ \sum_{i=1}^{N} x_i \cdot y_i = (\bar{y} - b \cdot \bar{x}) \sum_{i=1}^{N} x_i + b \cdot \sum_{i=1}^{N} x_i^2 \]

Or,

\[ \sum_{i=1}^{N} x_i \cdot y_i = \bar{y} \sum_{i=1}^{N} x_i - b \cdot \sum_{i=1}^{N} x_i + b \cdot \sum_{i=1}^{N} x_i^2 \]
The Simple Linear Regression Model: Specification and Estimation

Make use of the fact that: \[ \sum x_i = N \cdot \bar{x} \]
\[ \sum x_i y_i = N \cdot \bar{x} \bar{y} - b \cdot N \cdot \bar{x}^2 - b \cdot \sum x_i^2 \]

Rearrange terms:
\[ b \cdot \left[ \sum x_i^2 \right] = N \cdot \bar{x}^2 \]
\[ \sum x_i y_i - N \cdot \bar{x} \bar{y} \]
\[ Or, \]
\[ b = \frac{\sum x_i y_i - N \cdot \bar{x} \bar{y}}{\sum x_i^2 - N \cdot \bar{x}^2} \]

Factoring:
\[ b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]
\[ Or, \]
\[ b = \frac{Cov(x, y)}{Var(x)} \]
Thus: \[ a = \bar{y} - b \cdot \bar{x} \]

and \[ b = \frac{\text{Cov}(x,y)}{\text{Var}(x)} \]

Given the final solution for \( b \), note that it could be computed directly from the data.

The Simple Linear Regression Model: Specification and Estimation

\[ \begin{bmatrix} \partial \sum \varepsilon^2 / \partial \hat{a} - \partial \sum \varepsilon^2 / \partial \hat{b} \\ \partial \sum \varepsilon^2 / \partial \hat{a} - \partial \sum \varepsilon^2 / \partial \hat{b} \end{bmatrix} = \begin{bmatrix} 2N & 2 \sum x_i \\ 2 \sum x_i & 2 \sum x_i^2 \end{bmatrix} \]

If this matrix is positive definite, then we have found a minimum.

The diagonal elements are always positive.
The Simple Linear Regression Model: Specification and Estimation

Must show that the determinant,

\[(4N \cdot \sum x_i^2) - 4(\sum x_i)^2 > 0\]

is positive.

The Simple Linear Regression Model: Specification and Estimation

Again, make use of the fact that: \[\sum x_i = N \cdot \bar{x}\]

Then, \[(4N \cdot \sum x_i^2) - 4(\sum x_i)^2 =\]

\[(4N \cdot \sum x_i^2) - 4N\bar{x}^2 =\]

\[4N(\sum x_i^2 - \bar{x}^2) =\]

\[4N(\sum (x_i - \bar{x})^2) > 0\]

The Simple Linear Regression Model: Specification and Estimation

- We have a way of estimating the intercept and the slope for our model. But we must keep in mind the assumptions of the econometric model.

- We need to understand the properties of the estimators for the slope and the intercept.

- We need to measure the efficiency of our estimations (and predictions) for the betas and the dependent variable.

- And finally we need to know how to draw conclusions and present the results of our estimations.
First let's look at the results for a practical example. (In class)
- The estimation of the betas
- The interpretation of the results
- The functional form matters
- The transformation of the data can be helpful (elasticities example)

The properties of the estimators for the intercept and the slope parameters.
- The estimators are random variables
- The estimates are non-random
- They are unbiased estimators of the parameters
- They are the best estimators

Let's show that they are unbiased
"When the expected value of any estimator of a parameter equals the true parameter value, then the estimator is unbiased"

Then we need to show that the expected values of $a$ and $b$ are equal to the true parameters $\beta_1$ and $\beta_2$.

*This is shown in class, be there!*
The Simple Linear Regression Model: Specification and Estimation

- The OLS estimators are the best.

"When comparing two linear and unbiased estimators, we always want to use the one with the smaller variance, since the estimation rule gives us the higher probability of obtaining an estimate that is close to the true parameter."

- The precision of the estimators is related to their variance so we need to look at the variance of the estimators and the factors that influence their variance.

- This is done in class, be there!

The Simple Linear Regression Model: Specification and Estimation

- The Gauss Markov Theorem

"Under the assumptions of the linear regression model (without the normality assumption), the estimators for the intercept and the slope have the smallest variance of all linear and unbiased estimators of \( \beta_1 \) and \( \beta_2 \). They are the BEST LINEAR UNBIASED ESTIMATORS (BLUE) of \( \beta_1 \) and \( \beta_2 \).

The Simple Linear Regression Model: Specification and Estimation

- What is the GMT really saying:
  - Best compared to what? ???
  - Why are they the best? ???
  - In order for the GMT to hold, all the assumptions of the SLRM should hold, except the normality assumption. ???

- Proof: This is done in class, be there!
The Simple Linear Regression Model: Specification and Estimation

- As it was mentioned before, the least squares estimators are random variables.

- We have derived their expected value (mean) and their variance. These properties do not depend on the normality assumption, but if we assume that they are normally distributed then we can make probability statements since we would have their PDF's.

- If we assume that the random errors/disturbances are normally distributed with mean zero and variance $\sigma^2$, then the probability distribution of the least squares estimators are also normal.

- The normality of OLS estimators is of great importance in many aspects of statistical inference, as we will see in the next slides.

- What if the errors are not normally distributed?

The Variance of the Error Term:

- Biased estimator
- Unbiased estimator
- **This is done in class, be there!**
The Simple Linear Regression Model: Specification and Estimation

- The assumption that the random errors are normally distributed allows us to do the following:
  - Hypothesis testing
  - Interval estimation (confidence intervals)
  - Prediction
  - Interval Prediction

- Note: if the assumption of normality does not hold, then the sample size must be sufficiently large so that the least squares estimator's distributions are approximately normal, in which case the procedures developed here are also approximate.

A deeper look at the Normal Distribution

**Poincaré:** “...everyone believes in the [Normal] law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think is an empirical fact.”
A deeper look at the Normal Distribution

- The 68-95-99.7% Rule—all normal density curves satisfy the following property:
  - 68% of the observations fall within 1 standard deviation of the mean
  - 95% of the observations fall within 2 standard deviations of the mean
  - 99.7% of the observations fall within 3 standard deviations of the mean

Useful Distributions

- The $\chi^2$ distribution
  - The sum of independent standard normal r.v.s
- The $t$ distribution
  - The ratio of a standard normal r.v. to the square root of an independent $\chi^2$ r.v. divided by its d.o.f. (example in class, pg 93)
- The $F$ distribution
  - The ratio of two $\chi^2$ r.v.s divided by their respective d.o.f.

What is gained by assuming normally distributed errors?

- $\varepsilon_i$ dist. as $N(0, \sigma^2)$
- $\hat{\beta}_i$ dist. as $N(\beta_i, \sigma^2 \sum (y - \tau)^2)$
- And we can make probability statements about $\hat{\beta}_i$
Hypothesis testing

- The purpose of a hypothesis test is to determine whether an estimation result is consistent with the theory that gave rise to the econometric model.

- **t-tests**
  - Significance of individual coefficient
  - The equality of two individual coefficients
  - A single linear restriction

- Why are we using a t-test? ???

- Recall that
  \[ Z = \frac{\beta - \hat{\beta}}{\sigma} \sim N(0,1) \]

  and
  \[ T = \sum \frac{\hat{\beta} - \beta}{\sigma} \sim \chi_{(n-1)}^2 \]

- Then we have:
  \[ t = \frac{b_i - \beta_i}{se(b_i)} \sim t_{(T-2)} \]
Hypothesis testing

- The null hypothesis (H₀) and the alternative hypothesis (H₁ or H₂)

- Choose a significance level
  - Typically the .05 (5%) level or the .01 (1%) level is used

- Calculate a test statistic (e.g., t_{test})

---

Hypothesis testing

- Find the critical value of the test statistic (e.g., t_{critical} derived from tables in the book)

- Apply the following decision rule:
  - If |test value| > critical value, then reject H₀
  - e.g., if |t_{test}| > t_{critical}, then reject H₀
  - If |t_{test}| < t_{critical}, then fail to reject H₀

  - WHY? ????

---

Hypothesis testing

- Economic theory suggests that domestic currency depreciation leads to higher inflation rates:
  - Tradable goods effect
  - Pass-through effect
Hypothesis testing

☐ Lets test this theory.

    We will do this example in class!!

☐ Null hypothesis: ???

☐ Alternative Hypothesis: ???

☐ Select a significance level

    90% Significance 99% Significance
Hypothesis testing

Calculate the test statistic: ???

Find the critical value:

Note: the critical value depends on the significance level (for a two-tailed test, the significance level $\alpha$ is divided between the two tails) and the degrees of freedom (number of observations minus number of slope parameters to be estimated)
Apply decision rule—

⇒ reject the null hypothesis?  

Hypothesis testing

When doing hypothesis testing we have to keep in mind the possible errors that we can make:

- Type I Error
  - Rejecting a true null

- Type II Error
  - Failing to reject a false null

The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone, if the null hypothesis $H_0$, is true.

The $p$-value is largest significance level resulting in acceptance of $H_0$.

It is the probability of wrongly rejecting the null hypothesis if it is in fact true.
Hypothesis Testing

- It is equal to the significance level of the test for which we would only just reject the null hypothesis. The p-value is compared with the actual significance level of our test and, if it is smaller, the result is significant. That is, if the null hypothesis were to be rejected at the 5% significance level, this would be reported as "p < 0.05".

- Small p-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis. It indicates the strength of evidence for say, rejecting the null hypothesis \( H_0 \), rather than simply concluding "Reject \( H_0 \)" or "Fail to reject \( H_0 \)"

Discussion about one-tailed tests

A note on How to choose the null hypothesis

Interval Estimation

- Gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

(Note: it is another use of the t-statistic)
The confidence level is the probability value \((1 - \alpha)\) associated with a confidence interval.

It is often expressed as a percentage. For example, say \(\alpha = 0.05\), or 5%, then confidence level = \((1 - 0.05) = 0.95\), that is, a 95% confidence level.

Why is this so? ???

\[
\begin{align*}
\text{Interval Estimation} \\
\quad &\text{Prob}(t \geq t_c) = \text{Prob}(t \leq t_c) = \frac{\alpha}{2} \\
\quad &\text{Prob}(-t_c \leq t \leq t_c) = 1 - \alpha \\
\quad &\text{Prob}\left(-t_c \leq \frac{\beta - \beta_{\text{hypoth}}}{\text{s.e.}(\hat{\beta})} \leq t_c\right) = 1 - \alpha \\
\quad &\text{Prob}\left(\hat{\beta} - \left[t_c \times \text{s.e.}(\hat{\beta})\right] \leq \beta_{\text{hypoth}} \leq \hat{\beta} + \left[t_c \times \text{s.e.}(\hat{\beta})\right]\right) = 1 - \alpha
\end{align*}
\]

Interpretation anyone? ???

The width of the confidence.

Does it matter? ???

Point vs. interval estimator ???
Interval Estimation
- One of the objectives of the linear regression analysis is to predict the values of the dependent variable.
- How can we make a prediction? ???
- Do we have confidence interval for our predictions? ??? (hint: think of another use of the t-statistic)

The Simple Linear Regression Model: the $R^2$
- After we have estimated the linear regression model we would like to be able to make statements about how much of the variation of the dependent variable is explained by the model.
- In other words we want a measure of the proportion of variation in $y$ explained by $x$ within the regression model.

The Simple Linear Regression Model: the $R^2$
- Recall that the SLRM has two components

\[ C_i = \beta_1 + \beta_2 I_i + \epsilon_i \]

Non Random Random/stochastic
The Simple Linear Regression Model: the $R^2$

- Graphically this means:

\[ \sum (y_i - \bar{y})^2 = \text{Total Sum of squares: } \text{SST} \]
\[ \sum (\hat{y}_i - \bar{y})^2 = \text{Explain Sum of squares: } \text{SSR} \]
\[ \sum \hat{e}_i^2 = \text{Error Sum of squares: } \text{SSE} \]

- Then we have:

\[ \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \]

- All we are looking for is a proportion measure.

- With these sums we can get a ratio of the SSR and the SST. This ratio is called the $R^2$ or the coefficient of determination.
The Simple Linear Regression Model: the $R^2$

- Interpretation of the $R^2$. ???

Final notes for the SLRM

- Scaling the data

- The SLRM is more powerful and useful than you think:
  - Choosing the Functional Form
  - The usefulness of the Normal Distribution shows up again!!