Hypothesis testing and OLS Regression

NIPFP

14 and 15 October 2008
Overview

Introduction

Assumptions of OLS regression

Gauss-Markov Theorem

Interpreting the coefficients

Some useful numbers

A Monte-Carlo simulation

Model Specification
The OLS estimator continued

- As we discussed yesterday, the OLS estimator is a means of obtaining good estimates of $\beta_1$ and $\beta_2$, for the relationship $Y = \beta_1 + \beta_2 X_1 + \epsilon$

- Let us now move towards drawing inferences about the true $\beta_1$ and $\beta_2$, given our estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. This requires making some valid assumptions about $X_i$ and $\epsilon$. These assumptions also evoke certain useful statistical properties of OLS, as constrained with the purely numerical properties which we saw yesterday.
Assumptions of OLS regression

- Assumption 1: The regression model is linear in the parameters. \( Y = \beta_1 + \beta_2 X_i + u_i \). This does not mean that \( Y \) and \( X \) are linear, but rather that \( \beta_1 \) and \( \beta_2 \) are linear.
Assumptions of OLS regression

- Assumption 1: The regression model is linear in the parameters. \( Y = \beta_1 + \beta_2 X_i + u_i \). This does not mean that \( Y \) and \( X \) are linear, but rather that \( \beta_1 \) and \( \beta_2 \) are linear.
- Assumption 2: \( X \) values are fixed in repeated sampling.
Assumptions of OLS regression

- Assumption 1: The regression model is linear in the parameters. $Y = \beta_1 + \beta_2 X_i + u_i$. This does not mean that $Y$ and $X$ are linear, but rather that $\beta_1$ and $\beta_2$ are linear.
- Assumption 2: $X$ values are fixed in repeated sampling.
- Assumption 3: The expectation of the disturbance $u_i$ is zero. Thus, the distribution of $u_i$ given a value of $X_i$ (in the population) is symmetric around its mean. (Show figure).
• Assumption 4: The variance of $u_i$ is the same for all observations, i.e. in the above distribution, the distribution of $u_i$ given each value of $X_i$ has the same variance. This is an important property called homoskedasticity.
• Assumption 4: The variance of $u_i$ is the same for all observations, i.e. in the above distribution, the distribution of $u_i$ given each value of $X_i$ has the same variance. This is an important property called **homoskedasticity**.

• Assumption 5: There is no correlation between the $u_i$ (disturbances) of different observations. This is called **auto-correlation** or **serial-correlation**. It is seen more in time series analysis than cross-sectional analysis.
• Assumption 4: The variance of \( u_i \) is the same for all observations, i.e. in the above distribution, the distribution of \( u_i \) given each value of \( X_i \) has the same variance. This is an important property called **homoskedasticity**.

• Assumption 5: There is no correlation between the \( u_i \) (disturbances) of different observations. This is called **auto-correlation** or **serial-correlation**. It is seen more in time series analysis than cross-sectional analysis.

• Assumption 6: The covariance between \( u_i \) and \( X_i \) is zero. Intuitively, since we express \( Y \) as a sum of \( X_i \) and \( U_i \), if these two are correlated, then we must include a covariance term in the summation. So, by assumption, the covariance = 0.
Assumptions of OLS regression

- Assumption 7: The number of sample observations is greater than the number of parameters to be estimated.
Assumptions of OLS regression

- Assumption 7: The number of sample observations is greater than the number of parameters to be estimated.
- Assumption 8: The var(X) must be finite: The X values in a given sample must not all be the same.
Assumptions of OLS regression

- Assumption 7: The number of sample observations is greater than the number of parameters to be estimated.
- Assumption 8: The var(X) must be finite: The X values in a given sample must not all be the same.
- Assumption 9: The regression model is correctly specified. There is no specification error, there is no bias.
Assumptions of OLS regression

- Assumption 7: The number of sample observations is greater than the number of parameters to be estimated.
- Assumption 8: The var(X) must be finite: The X values in a given sample must not all be the same.
- Assumption 9: The regression model is correctly specified. There is no specification error, there is no bias.
- Assumption 10: There is no perfect multicollinearity, no two X; values can be expressed as a perfect linear combination of each other.
Statistical properties that emerge from the assumptions

**Theorem (Gauss Markov Theorem)**

In a linear model in which the errors have expectation zero and are uncorrelated and have equal variances, a best linear unbiased estimator (BLUE) of the coefficients is given by the least-squares estimator.

**BLUE estimator**

- Linear: It is a linear function of a random variable.
- Unbiased: The average or expected value of $\hat{\beta}_2 = \beta_2$.
- Efficient: It has minimum variance among all other estimators.
- However, not all ten classical assumptions have to hold for the OLS estimator to be B, L or U.
Interpreting an OLS coefficient/hypothesis testing

Call:
\texttt{lm(formula = y ~ x)}

Residuals:
\begin{tabular}{cccccc}
Min & 1Q & Median & 3Q & Max \\
-2.77652 & -0.77009 & 0.06778 & 0.60591 & 3.44186 \\
\end{tabular}

Coefficients:
\begin{tabular}{cccccc}
Estimate & Std. Error & t value & Pr(>|t|) \\
(Intercept) & 1.7816 & 0.2132 & 8.355 & 4.41e-13 \\
x & 3.0457 & 0.0398 & 76.531 & < 2e-16 \\
\end{tabular}

Residual standard error: 1.087 on 98 degrees of freedom
Multiple R-squared: 0.9835, Adjusted R-squared: 0.9834
F-statistic: 5857 on 1 and 98 DF, p-value: < 2.2e-16
Interpreting an OLS coefficient/hypothesis testing

```r
density.default(x = r)
N = 100   Bandwidth = 0.2773
```

Number of obs: 1000, R-squared: 0.997
Algebraic notation of the coefficient/estimator

- The least squares result is obtained by minimising 
  \((y - \beta_1 X)'(y - \beta_1 X)\)
- Expanding, 
  \(y'y - \beta_1' X'y - y' X \beta_1 + \beta_1' X' X \beta_1\)
- Differentiating with respect to \(\beta_1\), we get 
  \(-2X'y + 2X'X \beta_1 = 0\)
- Or \(X'X \beta_1 = X'y\)
- Or \(\beta_1 = (XX')^{-1} X'y\)
Testing a hypothesis about the estimator

We know that:

\[ \hat{\beta} = (X'X)^{-1}X'Y \]
\[ = (X'X)^{-1}X'(X\beta + \epsilon) \]
\[ = \beta + (X'X)^{-1}X'\epsilon \]

And now take the expectation:

\[ E[\hat{\beta}] = \beta + (X'X)^{-1}X'E[\epsilon] \]
\[ = \beta + 0 \]
\[ = \beta \]
• So far, we have not used the normality of residual assumption to derive any of our results.
• This assumption, however, is useful to test a hypothesis about an estimator.
• This allows us to test a hypothesis about \( \hat{\beta} \).

**Theorem**

\[
\hat{\beta} \sim \mathcal{N}(\beta, \frac{\sigma^2(X'X)^{-1}}{n})
\]

**Proof.**

• Either with the assumption that \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \)
• Or asymptotically by TCL
Some useful numbers: $R^2$

- $R^2$, or the coefficient of goodness-of-fit of a regression, measures the extent of overlap between the variables $Y$ and $X$. (Show Venn diagram). Since it is a ratio variable, it lies between 0 and 1.

- Technically, it can be expressed as:
  - $\sum Y_i - \bar{Y}^2 = \beta_2^2 \sum X_i - \bar{X}^2 + \sum u_i^2$, or
  - $TSS = ESS + RSS$
  - $R^2 = ESS/TSS$

- This is a useful number, but it must be kept in mind that it is not the best/only indicator of how “good” the regression is.

- Spurious regression: Two numbers that are statistically, but not causally related.

- As you add more variables to the regression, the $R^2$ only increases!
An example with R: Dangers of $R^2$

Call:
```
lm(formula = x ~ y)
```

Residuals:
```
          Min     1Q Median     3Q    Max
-4.8300 -2.6357 -0.1053  2.7757  5.3684
```

Coefficients:
```
                         Estimate Std. Error  t value  Pr(>|t|)
(Intercept)            4.6446     0.3189   14.567 <2e-16
             y          -0.1890     0.3432    -0.551    0.583
```

Residual standard error: 3.024 on 98 degrees of freedom
Multiple R-squared: 0.003084, Adjusted R-squared: -0.007089
F-statistic: 0.3032 on 1 and 98 DF, p-value: 0.5832
An example with R: Dangers of $R^2$

Call:
`lm(formula = x ~ y + m)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4.8994</td>
<td>-2.7182</td>
<td>-0.2155</td>
<td>2.8353</td>
<td>5.5601</td>
</tr>
</tbody>
</table>

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|---------|
| (Intercept)      | 4.5328   | 0.3218     | 14.084  | <2e-16  |
| y                | -0.1355  | 0.3409     | -0.397  | 0.6919  |
| m                | -0.5234  | 0.2976     | -1.759  | 0.0817  |

Residual standard error: 2.992 on 97 degrees of freedom
Multiple R-squared: 0.0339, Adjusted R-squared: 0.01398
F-statistic: 1.702 on 2 and 97 DF, p-value: 0.1878
An example with R: Dangers of $R^2$

Call:
\[
\text{lm(formula = x ~ y + m + z)}
\]

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.9964</td>
<td>-2.4296</td>
<td>-0.3385</td>
<td>2.6638</td>
<td>5.7291</td>
</tr>
</tbody>
</table>

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|)  |
|------------|----------|------------|---------|-----------|
| (Intercept)| 4.5316   | 0.3225     | 14.052  | <2e-16    |
| y          | -0.1402  | 0.3417     | -0.410  | 0.683     |
| m          | -0.4979  | 0.2999     | -1.660  | 0.100     |
| z          | -0.2285  | 0.2904     | -0.787  | 0.433     |

Residual standard error: 2.998 on 96 degrees of freedom
Multiple R-squared: 0.04009, Adjusted R-squared: 0.01009
F-statistic: 1.336 on 3 and 96 DF, p-value: 0.2671
Some useful numbers: Adjusted $R^2$

- This helps reduce the danger of $R^2$, as it adjusts the value of $R^2$ to the number of independent variables in the model.
- $\overline{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$
- But it is still related to $R^2$
Some useful numbers: Akaike Information Criterion

- Another way of measuring goodness of fit, adjusted for the number of variables

\[ \text{AIC} = e^{2k/n} \frac{RSS}{n} \]

- Lower AIC is better, and \( 2k/n \) can be interpreted as the “penalty factor”.
A Monte-Carlo simulation

density.default(x = c)

N = 1000   Bandwidth = 0.044
Some issues in model specification

- Scaling and units of measurement: Interpreting $\hat{\beta}_1$ and $\hat{\beta}_2$ when $X$ is expressed in different ways
- Standardised coefficients
- Various functional forms: Linear, log-linear, lin-log etc
Thank you.