Two-Variable Regression Model: The Problem of Estimation

Introducing the Ordinary Least Squares Estimator

Jamie Monogan

University of Georgia

Intermediate Political Methodology
By the end of this meeting, participants should be able to:

- Use ordinary least squares to estimate coefficients of a two-variable linear model by hand and through software.
- Compute standard errors and $r^2$ values by hand and through software.
- Describe the numerical properties of least squares estimators.
- Name the assumptions of the Classical Linear Regression Model.
- Explain the meaning of the Gauss-Markov Theorem.
- Describe the statistical properties of least squares estimators.
The Method of Ordinary Least Squares (OLS)

- Our goal: fit a line that minimizes squared residuals. (Why squared?)
- For the two-variable population regression model, 
  \( Y_i = \beta_1 + \beta_2 X_i + u_i \), we do this by choosing \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) to minimize the sum of squared residuals.
- The sum of squared residuals is:
  \( \sum \hat{u}^2_i = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \).
- With a bit of differential calculus, we can create the normal equations and find a formula for \( \hat{\beta}_1 \) & \( \hat{\beta}_2 \).

Representation of the sum of squared residuals function. The horizontal axis represents possible values of \( \hat{\beta} \). The vertical axis represents values of \( \sum \hat{u}^2_i \).
Calculating the Slope

- The sample regression coefficient, or slope ($\hat{\beta}_2$), can be computed in a number of ways, ordinary least squares (OLS) being just one technique among many.

- In two-variable regression the formula to compute $\hat{\beta}_2$ is:
  $$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

- In other words, the slope is computed by taking the difference of each value of the dependent variable from the mean of the dependent variable multiplied by the difference of each value of independent variable from the mean of the independent variable.

- Then a sum of all those products, divided by the sum of squared deviations of the independent variable.

- Intuition: how much do the two variables go together (covariance) divided by how much does the independent variable vary (variance).
A Word on Notation

- Gujarati & Porter (and several other texts) frequently use “deviation notation.”
- It is essential to recognize when deviation notation is being used.
- In the textbook, lower-case letters refer to the value of a variable minus its mean.
- In other words: \( x_i = X_i - \bar{X} \), and \( y_i = Y_i - \bar{Y} \).
- Hence, we could write: \( \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum (x_i)^2} \)
- Although this is more compact than the equation on the last slide, I personally prefer the long-hand form to avoid confusion.
- Remember this all the way through page 922. The text does not always remind you of what lower-case letters mean.
An Example: 2010 Congressional Races in North Carolina

<table>
<thead>
<tr>
<th>District</th>
<th>Republican (Y)</th>
<th>Obama (X)</th>
<th>((Y - .540) \times (X - .497))</th>
<th>((X - .497)^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.408</td>
<td>.630</td>
<td>-.018</td>
<td>.018</td>
</tr>
<tr>
<td>2</td>
<td>.496</td>
<td>.520</td>
<td>-.001</td>
<td>.001</td>
</tr>
<tr>
<td>3</td>
<td>.720</td>
<td>.380</td>
<td>-.021</td>
<td>.014</td>
</tr>
<tr>
<td>4</td>
<td>.431</td>
<td>.620</td>
<td>-.013</td>
<td>.015</td>
</tr>
<tr>
<td>5</td>
<td>.659</td>
<td>.380</td>
<td>-.014</td>
<td>.014</td>
</tr>
<tr>
<td>6</td>
<td>.753</td>
<td>.360</td>
<td>-.029</td>
<td>.019</td>
</tr>
<tr>
<td>7</td>
<td>.462</td>
<td>.470</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>8</td>
<td>.440</td>
<td>.520</td>
<td>-.002</td>
<td>.001</td>
</tr>
<tr>
<td>9</td>
<td>.690</td>
<td>.450</td>
<td>-.007</td>
<td>.002</td>
</tr>
<tr>
<td>10</td>
<td>.713</td>
<td>.360</td>
<td>-.024</td>
<td>.019</td>
</tr>
<tr>
<td>11</td>
<td>.457</td>
<td>.470</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>12</td>
<td>.342</td>
<td>.710</td>
<td>-.042</td>
<td>.045</td>
</tr>
<tr>
<td>13</td>
<td>.446</td>
<td>.590</td>
<td>-.009</td>
<td>.009</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>.540</strong></td>
<td><strong>.497</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td></td>
<td></td>
<td><strong>-.176</strong></td>
<td><strong>.156</strong></td>
</tr>
</tbody>
</table>

So we determine \( \hat{\beta}_2 = -1.123 \)

Note: We can get some error from rounding. Wait until you calculate the final quantity before rounding. (I show rounded values on the slides along the way, but use full precision at each stage of calculation.)
Calculating the Intercept

From this point we can also compute the sample intercept as well.

\[ \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}, \]

where:
- \( \hat{\beta}_1 \) = sample intercept.
- \( \hat{\beta}_2 \) = sample slope.
- \( \bar{Y} \) = mean of the dependent variable.
- \( \bar{X} \) = mean of the independent variable.

From the previous slide, we can see that .540 - (-1.123 * .497) = 1.098.

So the sample intercept is 1.098.

How would we do this using software?
- R: `lm(y~x)`
- Stata: `reg y x`

We could easily check our work by entering our data into one of these programs.
The OLS estimator can be computed from observable quantities (e.g. \(X\) & \(Y\)).

The OLS estimator yields point estimates of the relevant population parameters.

Computed from sample data, the OLS estimator yields a sample regression line that has the following properties:

- Passes through the point \((\bar{X}, \bar{Y})\). That is, \(\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}\).
- The mean of the predicted values equals the mean of the outcome: \(\bar{\hat{Y}} = \bar{Y}\).
- The mean value of the residuals is zero: \(\bar{\hat{u}} = 0\).
- The residuals \((\hat{u}_i)\) are uncorrelated with the predicted values \((\hat{Y}_i)\).
- The residuals \((\hat{u}_i)\) are uncorrelated with the input \((X_i)\).
Assumptions of the Classical Linear Regression Model

1. Linear in the parameters.
2. Fixed $X$ values or $X$ values independent of the error term. ($X$ is exogenous.) $\text{cov}(X_i, u_i) = 0.$
3. Correct functional form, or the disturbance has a conditional mean of zero. $E(u_i|X_i) = 0, \forall i.$
4. Homoscedasticity or constant variance of $u_i$. $\text{var}(u_i) = \sigma^2.$
5. No autocorrelation between disturbances. $\text{cov}(u_i, u_j) = 0$ for $i \neq j.$
6. The number of observations $n$ must be greater than the number of parameters to be estimated.
7. The $X$ variables cannot be fixed. $\text{var}(X) > 0.$

* We can argue that assumptions 2-5 are all we need to prove the Gauss–Markov Theorem. True, you cannot violate 1, 6, & 7. But I consider 1, 6, & 7 merely to describe the process of specifying a linear model.
Calculating Standard Errors
Always Report a Measure of Uncertainty

- The standard error of estimate or standard error of regression is calculated as:
  \[ \hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n - 2}} \]

- Note, we use \( n - 2 \) because we lose two degrees of freedom in estimating the bivariate regression model.

- From this, we compute the standard error of each sample coefficient:
  \[ se(\hat{\beta}_1) = \hat{\sigma} \sqrt{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}} \]
  \[ se(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}} \]

- Sometimes we need the covariance between two sample estimates. For example: \( cov(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X} \cdot var(\hat{\beta}_2) \).
Example: 2010 Congressional Races in North Carolina

<table>
<thead>
<tr>
<th>Dist</th>
<th>Y</th>
<th>X</th>
<th>(\hat{Y}_i = 1.098 - 1.123X_i)</th>
<th>(\hat{u}_i = Y_i - \hat{Y}_i)</th>
<th>(\hat{u}_i^2)</th>
<th>(X_i^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.408</td>
<td>.630</td>
<td>.390</td>
<td>.018</td>
<td>.0003</td>
<td>.3969</td>
</tr>
<tr>
<td>2</td>
<td>.496</td>
<td>.520</td>
<td>.514</td>
<td>-.018</td>
<td>.0003</td>
<td>.2704</td>
</tr>
<tr>
<td>3</td>
<td>.720</td>
<td>.380</td>
<td>.671</td>
<td>.049</td>
<td>.0024</td>
<td>.1444</td>
</tr>
<tr>
<td>4</td>
<td>.431</td>
<td>.620</td>
<td>.402</td>
<td>.029</td>
<td>.0009</td>
<td>.3844</td>
</tr>
<tr>
<td>5</td>
<td>.659</td>
<td>.380</td>
<td>.671</td>
<td>-.012</td>
<td>.0001</td>
<td>.1444</td>
</tr>
<tr>
<td>6</td>
<td>.753</td>
<td>.360</td>
<td>.694</td>
<td>.059</td>
<td>.0035</td>
<td>.1296</td>
</tr>
<tr>
<td>7</td>
<td>.462</td>
<td>.470</td>
<td>.570</td>
<td>-.108</td>
<td>.0117</td>
<td>.2209</td>
</tr>
<tr>
<td>8</td>
<td>.440</td>
<td>.520</td>
<td>.514</td>
<td>-.074</td>
<td>.0055</td>
<td>.2704</td>
</tr>
<tr>
<td>9</td>
<td>.690</td>
<td>.450</td>
<td>.592</td>
<td>.098</td>
<td>.0095</td>
<td>.2025</td>
</tr>
<tr>
<td>10</td>
<td>.713</td>
<td>.360</td>
<td>.694</td>
<td>.019</td>
<td>.0004</td>
<td>.1296</td>
</tr>
<tr>
<td>11</td>
<td>.457</td>
<td>.470</td>
<td>.570</td>
<td>-.113</td>
<td>.0128</td>
<td>.2209</td>
</tr>
<tr>
<td>12</td>
<td>.342</td>
<td>.710</td>
<td>.300</td>
<td>.042</td>
<td>.0017</td>
<td>.5041</td>
</tr>
<tr>
<td>13</td>
<td>.446</td>
<td>.590</td>
<td>.435</td>
<td>.011</td>
<td>.0001</td>
<td>.3481</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>.0492</strong></td>
<td><strong>3.3666</strong></td>
</tr>
</tbody>
</table>

Hence, we get: \(\hat{\sigma} = \sqrt{.0492/11} = .0669\).

\[
\text{se}(\hat{\beta}_1) = .0669\sqrt{\frac{3.3666}{13*.156}} = .0860.
\]

\[
\text{se}(\hat{\beta}_2) = \frac{.0669}{\sqrt{.156}} = .1691.
\]
Gauss-Markov Theorem

- Gujarati & Porter 2009, 72: “Under the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.”

- I would say: under the “weak set” of assumptions (2-5) OLS is BLUE
  - (B)est (i.e., efficient)
  - (L)inear
  - (U)nbiased
  - (E)stimator

- The “strong set” of assumptions assumes the disturbance terms have Normal or Gaussian distribution (more on this next week).

- What is meant by efficiency? Minimum variance of the sampling distribution.
Recall: Pearson’s $r$, the Correlation Coefficient

- Pearson’s $r$ is a symmetric measure of bivariate association. It shows how well the independent variable predicts the dependent variable. This measure will range between -1 and 1.
- A correlation coefficient of 0 would suggest the absence of any relationship between the two variables. A value of 1 would imply a perfect positive relationship, and a value of -1 would imply a perfect negative relationship. (Do we see perfect linear relationships in the social sciences?)
- The square of a Pearson’s $r$ ($r^2$) calculates the amount of variance explained by the predictor.
- Equation 3.5.13 in the text will jog your memory for calculating $r$. 
Generally speaking it is considered best in the social sciences to have the largest $r^2$ possible (it is bounded at 1).

However, as the number approaches 1, concerns about the data emerge. There are no perfect relationships in the social sciences and so $r^2$ over .90 or so are immediately suspect. (Unless we’re in the time series world.)

Generally, it is best to compare your $r^2$ against scholars doing similar work.

For example, macro research regularly finds $r^2$ values of .75 or greater while political psychology research rarely reaches .25.

Good research asks good questions. So finding a high $r^2$ on a question that has been decided in the literature is less important than research with a low $r^2$ on an interesting question or a brand new idea.
Our Plan for Calculating $r^2$

- The total variance in $Y$ can be split into explained and residual variance: $TSS = ESS + RSS$.
- Compute the proportional reduction in error ($r^2$).

$$r^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- This is the square of the linear coefficient or Pearson’s $r$.
- There are a few formulas we can use for $r^2$ (one predictor) or $R^2$ (multiple predictors), but we generally stick to the sum of squares based formulae when we start estimating multiple regression models.
- For our example NC elections: $\sum (Y_i - \bar{Y})^2 = 0.2466 = TSS$. We also know $\sum \hat{u}_i^2 = .0492 = RSS$.
- Hence $r^2 = 1 - \frac{.0492}{.2466} = .8005$. 
Homework Data: 2010 Congressional Races in Georgia

<table>
<thead>
<tr>
<th>District</th>
<th>Republican (Y)</th>
<th>Obama (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.717</td>
<td>.370</td>
</tr>
<tr>
<td>2</td>
<td>.486</td>
<td>.540</td>
</tr>
<tr>
<td>3</td>
<td>.691</td>
<td>.350</td>
</tr>
<tr>
<td>4</td>
<td>.254</td>
<td>.790</td>
</tr>
<tr>
<td>5</td>
<td>.263</td>
<td>.790</td>
</tr>
<tr>
<td>7</td>
<td>.671</td>
<td>.390</td>
</tr>
<tr>
<td>8</td>
<td>.526</td>
<td>.430</td>
</tr>
<tr>
<td>10</td>
<td>.673</td>
<td>.370</td>
</tr>
<tr>
<td>12</td>
<td>.433</td>
<td>.540</td>
</tr>
<tr>
<td>13</td>
<td>.295</td>
<td>.720</td>
</tr>
</tbody>
</table>
For Next Time

- Read Gujarati & Porter chapter 4 (Classical Normal Linear Regression Model).
- Study the ten observations of Georgia congressional races in 2010 listed on the previous slide.
- Write down a population regression model for Georgia congressional districts in which the share of the vote that went to Obama in 2008 has a linear effect on the Republican share of the vote in 2010.
- By hand, estimate the sample slope and intercept of this regression model.
- By hand, calculate the standard error of the sample slope and intercept.
- By hand, calculate the $r^2$ value for this model.
- Report all of these quantities and show all of your work in a typed document.
- You have the option to check your work using the software of your choice. Show me how you did your hand calculations, though.