Econometrics

TA Session 3

Lucia Sauer

2025-10-10

Overview

- Review of A1
- Regression Analysis
- Standard Error Calculation
- Introduction to Stata
- Hypothesis Testing
- Export Regression output to LaTeX

How I Grade (and Why I’m Tough)

Correction Principles

When grading, I focus on three things:
1. Rigor:
Be precise with **notation** and **concepts**.
2. Clarity:
Each line should have a clear purpose. Avoid algebraic noise.

> “If your reader has to guess what you mean — rewrite it.”

3. Conciseness:
Prove what’s asked, no more, no less.

> “If you can prove something in 3 clean steps, don’t use 10.”

Common Mistakes (1)

> "Expectations are over populations, not individuals."

$$E(y_i|d_i = 1)$$

captures the expected variable_name for individuals with treatment_name = 1.

It is **not**:

- the expectation of an individual outcome, nor
- the average of the observed y_i's in the sample.

Common Mistakes (2)

Best linear approximation can be written as:

$$E^*(Y|X) = \beta_1 + \beta_2 X$$

or just

$$\beta_1 + \beta_2 X$$

Wrong	Why
$E(Y\|X) \approx \beta_1 + \beta_2 X$	Suggests population expectation equals linear approximation
$Y = \beta_1 + \beta_2 X + u$	Treats the model as exact, not an approximation
$\hat{Y} = \beta_1 + \beta_2 X$	Mixes estimator notation with population relation

Common Mistakes (3)

When deriving β_1 and β_2 from the OLS minimization problem, we are deriving the **population parameters**.

- Step 1:

$$\frac{\partial(.)}{\partial b_1} : \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)$$

- Step 2: once I set to 0, the solution is β_1 and β_2:

$$\frac{\partial(.)}{\partial b_1} : \sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i) = 0$$

End of Review of A1…

you are doing great, keep it up!

Birthweight Data

The `bwght` dataset contains information on 1,388 births in the United States.

It was collected to study the determinants of infant birth weight, particularly the effects of maternal smoking and socioeconomic factors during pregnancy.

```python
import pandas as pd
import wooldridge as woo
df = woo.data('bwght')
df = df.dropna()
df.head(5)
```

	faminc	cigtax	cigprice	bwght	fatheduc	motheduc	parity	male	white	cigs	lbwght	bw
0	13.5	16.5	122.300003	109	12.0	12.0	1	1	1	0	4.691348	6.8
1	7.5	16.5	122.300003	133	6.0	12.0	2	1	0	0	4.890349	8.3
3	15.5	16.5	122.300003	126	12.0	12.0	2	1	0	0	4.836282	7.8

	faminc	cigtax	cigprice	bwght	fatheduc	motheduc	parity	male	white	cigs	lbwght	bw
4	27.5	16.5	122.300003	134	14.0	12.0	2	1	1	0	4.897840	8.3
5	7.5	16.5	122.300003	118	12.0	14.0	6	1	0	0	4.770685	7.3

Regression Analysis

$$\texttt{bwght}_i = \beta_1 + \beta_2\texttt{cigs}_i + \beta_3\texttt{parity}_i + \beta_4\texttt{faminc}_i + \beta_5\texttt{motheduc}_i + \beta_6\texttt{fatheduc}_i + \epsilon_i$$

where:

- `bwght`: birth weight, in pounds;
- `cigs`: average number of cigarettes the mother smoked per day during pregnancy;
- `parity`: birth order of the child;
- `faminc`: annual family income;
- `motheduc`: years of schooling of the mother;
- `fatheduc`: years of schooling of the father.

Results in Python

```python
import statsmodels.api as sm

X = sm.add_constant(df[['cigs', 'parity', 'faminc', 'motheduc', 'fatheduc']])
model = sm.OLS(df['bwght'], X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  bwght   R-squared:                       0.039
Model:                            OLS   Adj. R-squared:                  0.035
Method:                 Least Squares   F-statistic:                     9.553
Date:                Fri, 10 Oct 2025   Prob (F-statistic):           5.99e-09
Time:                        11:27:38   Log-Likelihood:                -5242.2
No. Observations:                1191   AIC:                         1.050e+04
Df Residuals:                    1185   BIC:                         1.053e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        114.5243      3.728     30.716      0.000     107.209     121.839
cigs          -0.5959      0.110     -5.401      0.000      -0.812      -0.379
parity         1.7876      0.659      2.711      0.007       0.494       3.081
faminc         0.0560      0.037      1.533      0.126      -0.016       0.128
motheduc      -0.3705      0.320     -1.158      0.247      -0.998       0.257
fatheduc       0.4724      0.283      1.671      0.095      -0.082       1.027
==============================================================================
Omnibus:                      120.762   Durbin-Watson:                   1.938
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              838.114
Skew:                          -0.119   Prob(JB):                    1.01e-182
Kurtosis:                       7.103   Cond. No.                         266.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Standard Error Calculation

How does the $se(\hat{\beta}_2)$ is calculated?

$$se(\hat{\beta}_2) = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{22}}$$

where:

- $\hat{\sigma}^2 = \frac{1}{n-K}\hat{\epsilon}'\hat{\epsilon}$ is the estimator of the variance of the error term.
- $(\mathbf{X}'\mathbf{X})^{-1}_{22}$ is the second diagonal element of the inverse of the matrix $\mathbf{X}'\mathbf{X}$, which captures the variance of `cigs` conditional on all other regressors.

To calculate $\hat{\sigma}^2$

```
sse = model.ssr
sse
```

```
np.float64(464041.13513001957)
```

$$\frac{SSE}{n-K} = \frac{464041.14}{1191-6}$$

To calculate $(\mathbf{X}'\mathbf{X})^{-1}_{22}$:

```
import numpy as np
XtX = np.dot(X.T, X)
XtX_inv = np.linalg.inv(XtX)
round(XtX_inv[1,1],4) # This is (X'X)^{-1}_{22}
```

```
np.float64(0.0)
```

Now let's calculate the standard error:

```
n_k = model.df_resid  # n - K
se_beta2 = np.sqrt((sse/n_k) * XtX_inv[1,1])
round(se_beta2,4)
```

```
np.float64(0.1103)
```

Introduction to Stata

Stata is a statistical software widely used in econometrics.

- dta. is the format for Stata datasets, but you can also import and export data in various formats like CSV, Excel, and more.
- do. files are the native file formats of Stata where you can write and execute commands and share scripts with others.
- Stata's interface is made up of various windows that can be moved around or closed to optimize the user's experience.

To follow best practices, I found this tutorial very useful: Stata Tutorial

---

Useful commands:

- `use` to load a dataset
- `describe` to get a summary of the dataset
- `regress` to run a regression
- `summarize` to get summary statistics
- `gen` to create new variables
- `browse` to view the dataset in a spreadsheet-like format
- `outreg` to export regression results to LaTeX

Example in Stata

Let's load the `bwght` dataset.

```
**********************************************************
* 2. Load Dataset
**********************************************************

* Install bcuse if not already installed
ssc install bcuse

* Load Wooldridge dataset
bcuse bwght, clear

* Inspect data
br in 1/10
```

Run the regression:

```
**********************************************************
* 3. Regression Analysis
**********************************************************

* Model:
* bwght_i =  1 +  2*cigs_i +  3*parity_i +  4*faminc_i +  5*motheduc_i +  6*fatheduc_i +  _i
```

```
regress bwght cigs parity faminc motheduc fatheduc
```

Hypothesis Testing

We want to test:

1. Whether `cigs` is statistically significant.
2. Whether `motheduc` and `fatheduc` are jointly statistically significant.

> Exercise
>
> 1. Indicate **null and alternative hypotheses**.
> 2. Write the expression of the **t-test statistic** or **F-test statistic** used for this test, and its assumed distribution.
> 3. Indicate the decision rule.
> 4. Draw the **acceptance and rejection regions** associated with this test. Properly label the axes. Include any relevant value to clearly identify both regions.
> 5. At what **significance level** would the regressors be statistically significant? Draw the **p-value** associated to the test performed. Properly label the axes. What is the minimum significance level that would make this regressor be significant?

Regression Output with stargazer

stargazer is a package that creates well-formatted regression tables and exports them to LaTeX, HTML, or text formats.

> Caution
>
> You will need to:
>
> ```
> uv sync
> ```
>
> your env because I have installed it Stargazer for this session.

Example usage

```python
from stargazer.stargazer import Stargazer
stargazer = Stargazer([model])
with open("regression_output.tex", "w") as f:
    f.write(stargazer.render_latex())

#Print the LaTeX code
print(stargazer.render_latex())
```

```
\begin{table}[!htbp] \centering
\begin{tabular}{@{\extracolsep{5pt}}lc}
\\[-1.8ex]\hline
\hline \\[-1.8ex]
& \multicolumn{1}{c}{\textit{Dependent variable: bwght}} \
\cr \cline{2-2}
\\[-1.8ex] & (1) \\
\hline \\[-1.8ex]
 cigs & -0.596$^{***}$ \\
& (0.110) \\
 const & 114.524$^{***}$ \\
& (3.728) \\
 faminc & 0.056$^{}$ \\
& (0.037) \\
 fatheduc & 0.472$^{*}$ \\
& (0.283) \\
 motheduc & -0.370$^{}$ \\
& (0.320) \\
 parity & 1.788$^{***}$ \\
& (0.659) \\
\hline \\[-1.8ex]
 Observations & 1191 \\
 $R^2$ & 0.039 \\
 Adjusted $R^2$ & 0.035 \\
 Residual Std. Error & 19.789 (df=1185) \\
 F Statistic & 9.553$^{***}$ (df=5; 1185) \\
\hline
\hline \\[-1.8ex]
\textit{Note:} & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
\end{tabular}
\end{table}
```

How to attach tables in LaTeX

Two versions:

1. You can upload the .tex file to Overleaf and call it in your main .tex file with:

```
\input{regression_output.tex}
```

2. You can copy the code from the .tex file and paste it directly in your main .tex file.