L.9: Writing an Empirical Report

Econometrics 1: ver. 2024 Fall Semester

Naoki Awaya

How to Write an Empirical Report

Basic Structure

- A typical structure of an empirical paper is as follows:

1. Introduction
2. Data description
3. Econometric models
4. Estimation results
5. Conclusion

+ References, Acknowledgement, Appendix, etc

- For a full-length article, a literature review section should be in between sections 1 and 2.
- Sections 3 – 4 may be integrated in one section.

Section 1. Introduction

Introduction

- The purpose of this section is to present the main research objective of your paper.
  - What questions would you like to answer in this paper?
  - **Research background**: why are these questions important?
- Typical formats of research questions:
  - What are the major determinants of Y?
  - To what extent X is important as a determinant of Y?
  - How the effects of X on Y differ between A and B?
- What are the expected results?

Section 1. Introduction

- Literature review
  - Any academic literature already written on the same topic?
  - In what ways will your research differ and contribute to the existing literature?
- Preview of the findings
  - Describe your data and empirical strategy, and briefly present your main findings to the reader.
- In the final paragraph of this section, present the roadmap for the rest of the paper.

Section 2. Data description

- Describe the data to be used.
  - Data source (e.g., organization, URL, etc)
  - Is the data publicly available or confidential?
- Briefly describe the variables used and their definitions (may be in a table format).

Table 1: Variables and definitions

Variable	Definition
Income	Annual income in million JPY
Education	Education in years
...	

- Before proceeding to the main econometric analysis, describe some basic statistics of the data.

Section 2. Data description: summarytools

- Summary table for the whole dataset.

```
>   library(summarytools)
>   setwd("C:/Users/naway/Dropbox/lecture_materials/Econometrics 1 (EDP)/lecture_no
>   data <- read.csv("apartments.csv")
>   descr(data,
+         stats = c("mean", "med", "sd", "min", "max"),
+         transpose = TRUE)
```

```
## Descriptive Statistics
## data
## N: 500
##
##                Mean   Median   Std.Dev      Min       Max
## ----------- ------- -------- --------- ------- --------
##         area   45.36    43.24     24.24   12.92   127.04
##          com    0.37     0.00      0.48    0.00     1.00
##        floor    7.81     5.00      7.88    1.00    52.00
##          ind    0.21     0.00      0.40    0.00     1.00
##        price   32.78    26.00     23.28    3.28   195.88
##        renov    0.20     0.00      0.40    0.00     1.00
##       stdist    0.47     0.43      0.27    0.03     1.63
```

Section 2. Data description: summarytools

- Create summary tables for the renovated apartments (`renov = 1`) and non-renovated ones (`renov = 0`), separately.

```
>   stby(data,
+        INDICES = data$renov,
+        FUN = descr,
+        stats = c("mean", "med", "sd", "min", "max"),
+        transpose = TRUE)
```

The output in the next page.

Section 2. Data description: summarytools

```
## Descriptive Statistics
## data
## Group: renov = 0
## N: 401
##
##                 Mean   Median   Std.Dev      Min       Max
## ------------ ------- -------- --------- ------- --------
##         area   44.68    39.38     24.51   12.92   127.04
##          com    0.37     0.00      0.48    0.00     1.00
##        floor    8.12     5.00      8.21    1.00    52.00
##          ind    0.21     0.00      0.41    0.00     1.00
##        price   33.90    26.30     24.59    3.90   195.88
##        renov    0.00     0.00      0.00    0.00     0.00
##       stdist    0.48     0.46      0.27    0.03     1.50
##
## Group: renov = 1
## N: 99
##
##                 Mean   Median   Std.Dev      Min       Max
## ------------ ------- -------- --------- ------- --------
##         area   48.12    50.66     23.03   14.44   105.64
##          com    0.35     0.00      0.48    0.00     1.00
##        floor    6.55     5.00      6.29    1.00    45.00
##          ind    0.20     0.00      0.40    0.00     1.00
##        price   28.23    25.50     16.27    3.28    80.50
##        renov    1.00     1.00      0.00    1.00     1.00
##       stdist    0.47     0.39      0.31    0.03     1.63
```

Section 2. Data description: summarytools

- Cross table of (renov, com)

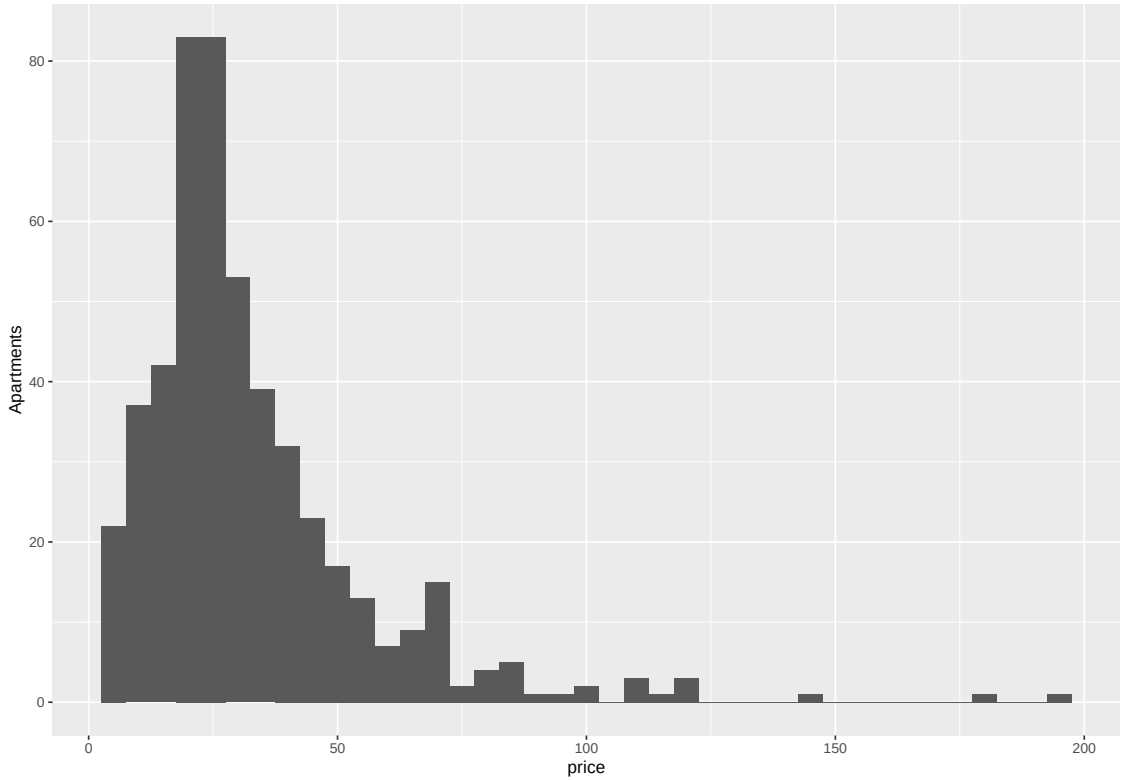
```
> Renovated <- ifelse(data$renov == 1, "Yes", "No")
> Commercial <- ifelse(data$com == 1, "Yes", "No")
> ctable(x = Renovated, y = Commercial, prop = "r")
```

```
## Cross-Tabulation, Row Proportions
## Renovated * Commercial
##
## ---------- ----------- ------------ ------------ -------------
##              Commercial           No          Yes         Total
##   Renovated
##           No               251 (62.6%)  150 (37.4%)  401 (100.0%)
##          Yes                64 (64.6%)   35 (35.4%)   99 (100.0%)
##        Total               315 (63.0%)  185 (37.0%)  500 (100.0%)
## ---------- ----------- ------------ ------------ -------------
```

Section 2. Data description: ggplot

- Histogram of `price`

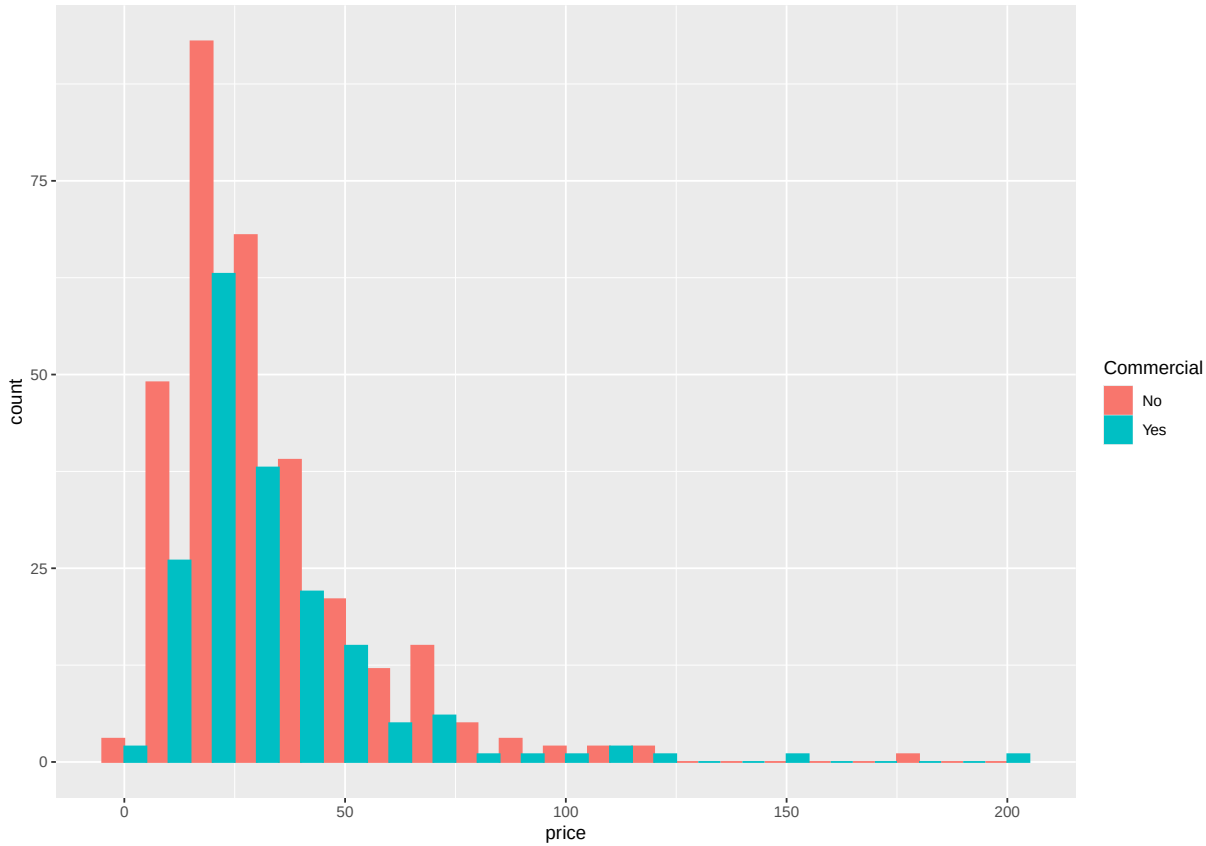
```
> library(tidyverse)
> ggplot(data, aes(price)) +
+   geom_histogram(binwidth = 5) +
+   ylab("Apartments")
```



Section 2. Data description: ggplot

- Histogram of price by groups com = 1 and com = 0

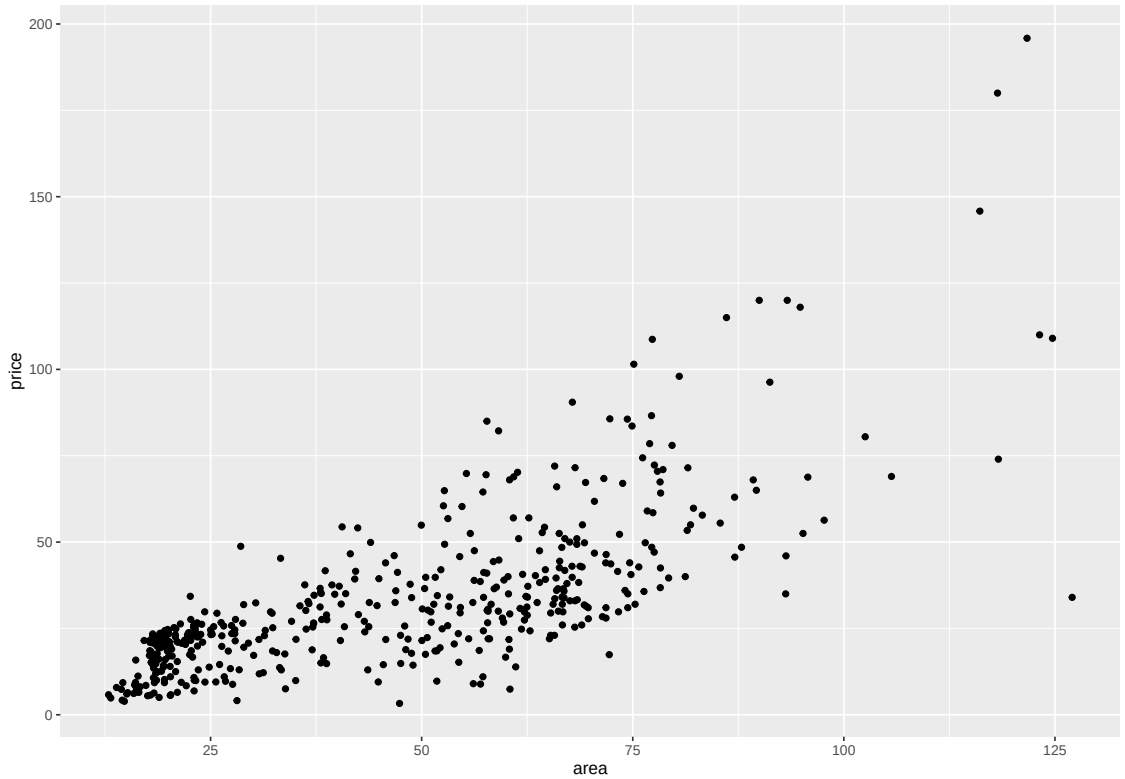
```
> df <- data.frame(price = data$price, Commercial = Commercial)
> ggplot(df, aes(x = price, fill = Commercial, color = Commercial)) +
+   geom_histogram(binwidth = 10, position = "dodge")
```



Section 2. Data description: ggplot

- Scatter plot of `price` vs. `area`

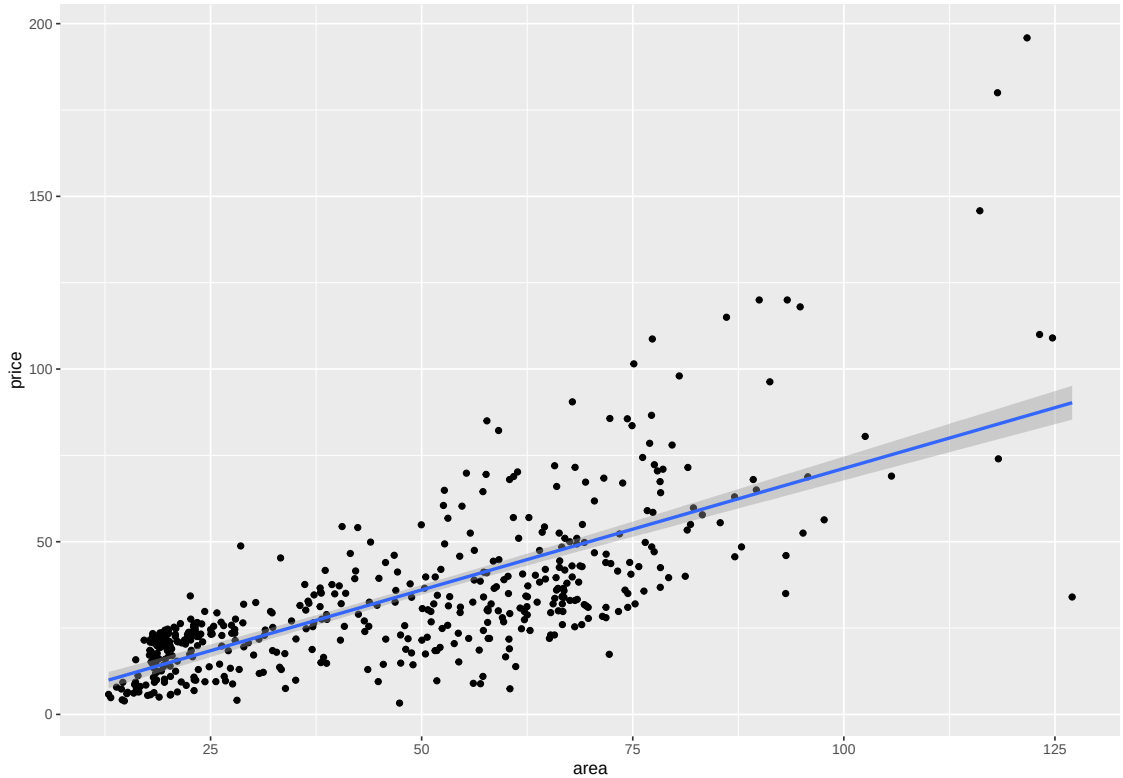
```
> ggplot(data, aes(area, price)) + geom_point()
```



Section 2. Data description: ggplot

- Add a linear regression line.

```
> ggplot(data, aes(area, price)) + geom_point() +
+   geom_smooth(method = lm)
```



The grayed region is the 95% CI.

Section 3. Econometric models

- Present the econometric models to be examined in equation form.

$$\begin{aligned}\text{Model 1.}\quad \texttt{price} &= \beta_0 + \texttt{area}\beta_1 + \texttt{stdist}\beta_2 + \texttt{renov}\beta_3 \\ &\quad + \texttt{com}\beta_4 + \texttt{ind}\beta_5 + \epsilon \\ \text{Model 2.}\quad \texttt{logprice} &= \beta_0 + \texttt{area}\beta_1 + \texttt{stdist}\beta_2 + \texttt{renov}\beta_3 \\ &\quad + \texttt{com}\beta_4 + \texttt{ind}\beta_5 + \epsilon\end{aligned}$$

- If the model contains many regressors, you can suppress some unimportant ones: e.g., $\texttt{Y} = \beta_0 + \texttt{X}_1\beta_1 + \texttt{X}_2\beta_2 + \cdots + \epsilon$
- It is always better to consider multiple models and compare their results.
- If the models are not complicated, this section and the next may be merged into one section.

Section 4. Estimation results

- Report the estimation results in a table.

```
> library(modelsummary)
> data$logprice <- log(data$price)
> m1 <- lm(price ~ area + stdist + renov + com + ind, data)
> m2 <- lm(logprice ~ area + stdist + renov + com + ind, data)
> modelsummary(list(m1,m2),
+   gof_omit = "Log.Lik.|AIC|BIC|F|RMSE",
+   stars = TRUE,
+   notes = "Standard errors in parentheses.",
+   output = "table.docx")
```

Section 4. Estimation results

	(1)	(2)
(Intercept)	7.233***	2.471***
	(2.056)	(0.060)
area	0.746***	0.020***
	(0.028)	(0.001)
stdist	-14.442***	-0.200**
	(2.585)	(0.076)
renov	-8.267***	-0.216***
	(1.685)	(0.049)
com	1.335	0.068
	(1.564)	(0.046)
ind	-1.359	0.002
	(1.806)	(0.053)
Num.Obs.	500	500
R2	0.590	0.556
R2 Adj.	0.586	0.552

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, ***

Section 4. Estimation results

- Clearly state the key findings from the estimation results.
  - What variables have significant/insignificant effects?
  - Are the signs of the coefficients consistent with your hypothesis?
- Compare the results from different models and draw comprehensive conclusions.
- Are the results consistent with prior studies? If not, any possible explanation?
- Ideally, discuss some policy implications and suggestions based on the obtained results.

Section 5. Conclusion

- Briefly review the main contents of the paper in one or two paragraphs.
- Bringing together the results of your analysis and your interpretations, present your policy suggestions to the reader.
- Finally, you may address the potential limitations of your study and remaining future tasks.

Mid-term Paper Assignment

Mid-term Paper Assignment

Mid-term Paper Assignment: write a short empirical paper using linear regression analysis with **R**.

Requirements:

- The paper must be typed in English, 4-5 pages in body length (not including the title and reference page), 11-12 pt. font, and single-authored.
- The paper must have its Title (in a separate page) and Conclusions. The title page contains the title of your paper, your name, and ID.
- Before presenting the estimation results, clearly write the regression model(s) you estimate in equation form.
- The estimation results should be summarized in a table. Do not include a screenshot or a photocopy of your PC screen.
- The script file used in your analysis must be attached to the paper as Appendix (just copy-and-paste the commands from the script file).

Mid-term Paper Assignment

Data

- Data taken from the GSS (General Social Survey) 2018.
- For details of the data, see **data description.pdf**.
- What variables to be included in the model as dependent or explanatory variables are all up to you. The variables may be transformed according to your research interests.

Evaluation

- The purpose of the assignment is to check your ability
  - to perform regression analysis and statistical hypothesis testing with **R**,
  - to correctly interpret the estimation results,
  - and to describe the obtained results in a research-paper style.

Mid-term Paper Assignment

- This assignment accounts for 30% of your grade.
- Submission deadline: To be added in a future version. Submit a **pdf** file of your paper through **Waseda Moodle**.[a]

[a] Using Microsoft Word is not mandatory. You may use Rmarkdown, LaTex, or any text editor you prefer.