Introductory Econometrics Cheat Sheet

by Marten Walk, *University of Halle 2024*

1a Types of Data

- Cross-Sectional: 20 countries in one year
- Time-Series: 1 country over 20 years
- Panel: 20 countries, 20 years

2 Simple Regression Model

useful for simple *ceteribus paribus* relationship

$$y = \beta_0 + \beta_1 x + u$$

- y = dependent / explained / regressand
- x = independent / explanatory / regressor

Assumptions:

1. $E(u) = 0$: avg. of unobserved is zero
2. $E(u|x) = u$: errors are independent

When these hold, coefficients are:

$$\widehat{\beta}_1 = \frac{\mathrm{Cov}_{x,y}}{\mathrm{Var}_x} = \frac{\sum x_i y_i - n\bar{y}\bar{x}}{\sum x^2 - n\bar{x}^2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

fitted values	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$
residuals	$\hat{u} = y_i - \hat{y}_i$
Total Sum Sqares	SST = $\sum (y_i - \bar{y})^2$
Sum of Sq. Regression	SSR = $\sum (\hat{y}_i - \bar{y})^2$
Sum Sq. Residuals/Error	SSE = $\sum (y_i - \hat{y}_i)^2$

$$\boldsymbol{R^2} = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2 - n\bar{y}^2}$$

$$0 \le R^2 \le 1$$

adj $R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)}$

Standard Error: $\hat{\sigma} = \sqrt{\frac{1}{n-2} \cdot \sum \hat{u}_i^2}$

for coefficient: $\mathrm{se}(\beta_1) = \frac{\sigma}{\sqrt{\sum (x-\bar{x})^2}}$

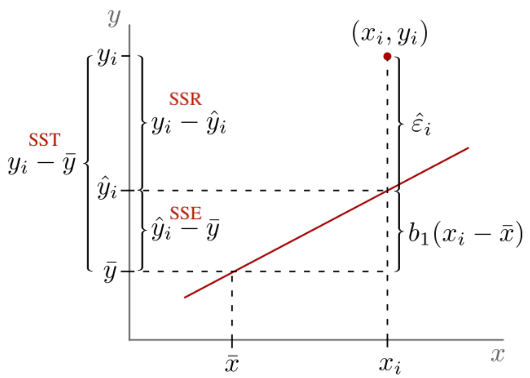
2a Algebraic Properties

$\sum \hat{u}_i = 0$: mean / sum of residuals = zero

$\sum x_i \hat{u}_i = 0$: no covariance x and u

OLS line passes trough $(\bar{x}, \bar{y})$

SST = SSR + SSE



2b How To *Regression*

Step by Step to calcualte a simple regression

i	x_i	y_i	x_i^2	$y_i x_i$	$\hat{y}_i$	$\hat{u}_i$	$\hat{u}_i^2$	y_i^2	SST_x
1	...	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...	...
$\sum$	**...**	**...**	**...**	**...**		**0**	**...**	**...**	**...**
	$\bar{x}, \bar{y}$		β_1, β_0				$\rightarrow R^2$		$\mathrm{se}(\beta_1)$

1. $\bar{x} = \frac{1}{n} \sum x$
2. $\bar{y} = \frac{1}{n} \sum y$
3. β_1: see left
4. β_0 = $\bar{y} - \hat{y}\bar{x}$
5. R^2
6. adj R^2
7. σ
8. $\mathrm{se}(\beta_1)$

2c Interpretation

OLS = linear in parameters, not linear in x!

Model	expl.	indep.	interpretation
level-level	y	x	$\Delta y = \beta_1 \Delta x$
log-level	ln y	x	$\%\Delta y \approx (100 \cdot \beta_1)\Delta x$
level-log	y	ln x	$\Delta y \approx \left(\frac{\beta_1}{100}\right)[\%\Delta x]$
log-log	ln y	ln x	$\%\Delta y \approx \beta_1 \%\Delta x$
Quadratic	y	$x + x^2$	$\Delta y = (\beta_1 + 2\beta_2 x)\Delta x$

2ca Percentage / Percentage Points

$\ln \text{wage} = \beta_0 + 0.05 \text{ unempl.rate}$

- unempl. rate increases by one **% point** (ex: 8⟶9)
- $(0.05 \cdot 100) = 5\%$ wage $\uparrow$

$\ln y = \beta_0 + \beta_1 \ln x$

- rate by one **%** (ex: 8⟶8.08)
- 0.05% wage $\uparrow$

2cb why Logarithmics?

- reduces skewness (example: income)
- extreme values = less influential
- **Note:** not defined for 0 (they just drop out)

exact Interpretation (log-level): *for values > 0.2*

$$\%\Delta y = 100 \cdot [\exp(\beta_2) - 1]$$

---

3 Reminder: Derivates

$$x^a \rightarrow a \cdot x^{1-a}$$

$$\ln x \rightarrow 1/x$$

$$e^x \rightarrow e^x$$

$$\sqrt{x} = x^{\frac{1}{2}} \rightarrow x^{-\frac{1}{2}} = 1/\sqrt{x}$$

$$1/x = x^{-1} \rightarrow x^{-2} = 1/x^2$$

for partial effect calculation, e.g $\frac{\delta y}{\delta x}$

4 Multiple Regression Model

more plausibly estimate the effect of multiple factors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + u$$

4a Gauss-Markov Assumptions

1. **Linear**: y is linear function of $\beta' s$
2. **Random**: y and x are randomly sampled from pop
3. **Non-Collinearity**: regressors arent 100% correlated
4. **Exogenity**: $E(u|x_i, ..., x_k) = E(u) = 0$
   - Regressors arent correlated with error term
5. **Homoscedasticity**:
   - $\mathrm{Var}(u \mid x_i, ..., x_k) = \mathrm{Var}(u) = \sigma^2$
   - Variance of the error is constant
6. **Normality**: u is distributed $N(0, \sigma^2)$

(1)-(4) hold: OLS is unbiased

(1)-(5) hold: OLS is Best Linear Based Estimator

(1)-(6) hold: Classic Linear Model (CLM)
- allows testing hypotheses about β
- if not (6), you need asymptotics
  - not exact interpretations of β

4b Omitted Variable Bias

real model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

est. model: $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$

regression of both regressors: $x_1 = \delta_0 + \delta_1 x_2$

relationship: $\tilde{\beta}_1 = \widehat{\beta_1} + \widehat{\beta_2}\hat{\delta}_1$

5 Asymptotics

Efficient Estimators have

- consistency (variance goes down with N)
- asymptotically normal distribution
- asymptotic variance smaller than other estimators

⇒ OLS = asymptotically efficient!

6 Tests & CI

How to answer yes/no Questions about our models, e.g "Does Cigarette Smoking affect health?"

null hypothesis	$H_0 : \theta = 0$ (typically)
alt. hypothesis	$H_1 : \theta \neq 0$
significance level	$\alpha = \{1\%, 5\%, 10\%\}$
Test Statistic T	sample from our data
critical value c	Value to reject H_0 at given α, if $\lvert T\rvert > c$
p-value	reject H_0 if $p \leq \alpha$
degrees of freedom	$\mathrm{df} = N - k - 1$

6a How to test

1. Define $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
2. Calculate Test Statistic T: $\frac{\beta_1}{\mathrm{se}(\beta_1)}$
3. Find critical value for given α
   - if $\mathrm{df} > 120 \Rightarrow$ normal distribution (**one sided**)
   - else: search in t-distribution table for $t_{29,\alpha/2}$
4. if $|T| > c$, reject H_0

If you want to test for positive / negative influence:

1. define $H_0 : \beta_1 \leq 0$ vs $H_1 : \beta_1 > 0$
3. find critical value (**two sided**)

if test for a value of β

1. define $H_0 : \beta_1 = z$ vs $H_1 : \beta_1 \neq z$
2. test-statistic: $t = \frac{\hat{\beta} - z}{\mathrm{se}\ \hat{\beta}}$

6b Confidence Intervals

$$\left[\hat{\beta}_j \pm c \cdot \mathrm{se}\left(\hat{\beta}_j\right)\right]$$

Reject $H_0 : \beta_1 = z$ if z in [CI]

6c Critical Values

α	one-sided (normal)	two-sided
10%	1.64	1.28
5%	1.96	1.64
1%	2.58	2.33

for t-distribution: look at tables

- one-sided: take $\alpha/2$
- two-sided: take α as given

6d F-statistic

test significance of overall model (incl all predictors)

$$F = \frac{\text{Mean Sum Sq. Regression}}{\text{Mean Sum Sq. Error}}$$

- MSR = SSR $/k$
- MSE = SSE $/N - k - 1$

F-statistic critival value:

- numerator degrees of freedom = k
- demoninator degrees of freedom = N-k-1

alternative: **from** R^2

if restricted model:
- drop the relevant variables
- calculate R^2 of both models

$$F = \frac{\frac{R^2_{\mathrm{ur}} - R^2_r}{q}}{\frac{1 - R^2_{\mathrm{ur}}}{n-k-1}}$$

- q = number of dropped vars
- R^2_{ur} = R-squared from unrestricted regression
- R^2_r = R-squared from restricted / second regression

Note: if no restricted model, just drop the R^2_r and set $q = k$

7 Dummy & Interactions

7a Scaling / Conversions

- y $\cdot\, c \Rightarrow \hat{\beta}_1 \cdot c$
- x $\cdot\, c \Rightarrow \frac{\hat{\beta}_1}{c}$
- $\ln(c \cdot y) \Rightarrow$ no change in slope estimate
  - only intercept: $\hat{\beta}_0^{\text{new}} = \hat{\beta}_1^{\text{old}} + \ln c$
- $\ln(c \cdot x) \Rightarrow$ also no change in slope
- $\hat{\beta}_0^{\text{new}} = \hat{\beta}_0^{\text{old}} - \beta_j \ln c$

7b Dummies / Binaries

to represent qualitative factors in TRUE/FALSE format

$$\text{wage} = \beta_0 + \beta_1 \text{ educ} + \beta_2 \text{ female}$$

if $\beta_2 < 0 \Rightarrow$ women earn less

7c Binary as Dependent Variable

= linear probability model

$$\Pr(y = 1 \mid x) = \beta_0 + \beta_1 x_1 + \dots + u$$

- try to explain the binary outcome
- interpret β as increase in probability
- always heteroskedastic

7d Interactions

to explain an explanatory variable, that depends on another explanatory variable

$$\text{wage} = \beta_0 + \beta_1 \text{ fem} + \beta_2 \text{ married} + \beta_3 \textbf{ fem} \cdot \textbf{married}$$

- single men = base scenario
- married women = include all coefficients

8 R Code

some usefol commands

Variables

```
x <- 1 #Declare a Variale
is.numeric(x) #check type
```

Data Wrangling

```
data <- read.csv("data.csv")
head(data) #show first 5 rows
summary(data) #show summary statistics
describe(data) #show some general info
table(is.na(data)) #inspect missing values
table(is.na(data$col)) #in a column
data %>% select(-c("row1", "row2")) #drop
```

Describe Data

```
nrow(data) #number of rows
ncol(data) #number of columns
sapply(data, class) #types of data in rows
sd(data$col) # standard deviation
```

Dummy Var

```
data$female[data$gender=="Female"]<-1 #create
based on condition
attr(data$female, "label") <- "is Women"
# Convert into Factor
data$female <- factor(data$female,
  levels = c(0, 1),
  labels = c("male", "female"))
```

Plot with ggplot

```
# initialize with data and aesthetics
ggplot(data, aes(x = "xcol", y = "ycol")) +
  geom_scatter() + #scatterplot
  geom_hline(yintercept=1) #horizontal line
```

```
  labs(x = "Sat", y = "Freq") + #labels
  xlim(0, 2000) # limit how far x line goes
```

Regression

```
reg <- lm(y ~ x, data=data)
summary(reg)
```

Fitted Values and Residuals

```
data$fit <- predict(reg) #fitted values
data$res <- residuals(reg) #residuals
sum(reg$Res) #=0 to check
```

How to Read an R Regression Output [*summary(reg)*]

```
## Call:
## lm(formula = lwage ~ educ + exper + tenure, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8545 -0.2327  0.0142  0.2471  1.3162
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.330651   0.114378  46.605  < 2e-16 ***
## educ        0.075357   0.006435  11.711  < 2e-16 ***
## exper       0.014119   0.003338   4.229 2.57e-05 ***
## tenure      0.012755   0.002559   4.984 7.42e-07 ***
## married     0.199171   0.040820   4.879 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.3831 on 930 degrees of freedom
## Multiple R-squared:  0.1762, Adjusted R-squared:  0.1727
## F-statistic: 49.73 on 4 and 930 DF,  p-value: < 2.2e-16
```

1. regression formula
2. estimated coefficients
3. statisticla significance
4. Degrees of Freedom (allows calculation of N)
5. R^2 and adj. R^2