Econometrics Cheat Sheet

by Tyler Ransom, University of Oklahoma
@tyleransom

Data & Causality

Basics about data types and causality.

Types of data

Experimental	Data from randomized experiment
Observational	Data collected passively
Cross-sectional	Multiple units, one point in time
Time series	Single unit, multiple points in time
Longitudinal (or Panel)	Multiple units followed over multiple time periods

Experimental data

- Correlation $\implies$ Causality
- Very rare in Social Sciences

Statistics basics

We examine a **random sample** of data to learn about pop.

Random sample	Representative of population
Parameter (θ)	Some number describing population
Estimator of θ	Rule assigning value of θ to sample e.g. Sample average, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$
Estimate of θ	What the estimator spits out for a particular sample ($\hat{\theta}$)
Sampling distribution	Distribution of estimates across all possible samples
Bias of estimator W	$E(W) - \theta$
Efficiency	W efficient if $Var(W) < Var(\widetilde{W})$
Consistency	W consistent if $\hat{\theta} \to \theta$ as $N \to \infty$

Hypothesis testing

The way we answer yes/no questions about our population using a sample of data. e.g. "Does increasing public school spending increase student achievement?"

null hypothesis (H_0)	Typically, $H_0 : \theta = 0$
alt. hypothesis (H_a)	Typically, $H_0 : \theta \neq 0$
significance level (α)	Tolerance for making Type I error; (e.g. 10%, 5%, or 1%)
test statistic (T)	Some function of the sample of data
critical value (c)	Value of T such that reject H_0 if $\lvert T\rvert > c$; c depends on α; c depends on if 1- or 2-sided test
p-value	Largest α at which fail to reject H_0; reject H_0 if $p < \alpha$

Simple Regression Model

Regression is useful because we can estimate a *ceteris paribus* relationship between some variable x and our outcome y

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

We want to estimate $\hat{\beta}_1$, which gives us the effect of x on y.

OLS formulas

To estimate $\hat{\beta}_0$ and $\hat{\beta}_1$, we make two assumptions:

1. $E(u) = 0$
2. $E(u|x) = E(u)$ for all x

When these hold, we get the following formulas:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\widehat{Cov}(y, x)}{\widehat{Var}(x)}$$

fitted values ($\hat{y}_i$)	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
residuals ($\hat{u}_i$)	$\hat{u}_i = y_i - \hat{y}_i$
Total Sum of Squares	$SST = \sum_{i=1}^{N} (y_i - \overline{y})^2$
Expl. Sum of Squares	$SSE = \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2$
Resid. Sum of Squares	$SSR = \sum_{i=1}^{N} \hat{u}_i^2$
R-squared (R^2)	$R^2 = \frac{SSE}{SST}$; "frac. of var. in y explained by x"

Algebraic properties of OLS estimates

$\sum_{i=1}^{N} \hat{u}_i = 0$ (mean & sum of residuals is zero)

$\sum_{i=1}^{N} x_i \hat{u}_i = 0$ (zero covariance bet. x and resids.)

The OLS line (SRF) always passes through $(\overline{x}, \overline{y})$

$SSE + SSR = SST$

$0 \leq R^2 \leq 1$

Interpretation and functional form

Our model is restricted to be **linear in parameters**

But not linear in x

Other functional forms can give more realistic model

Model	DV	RHS	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y \approx (\beta_1/100)\,[1\%\Delta x]$
Log-level	$\log(y)$	x	$\%\Delta y \approx (100\beta_1)\,\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y \approx \beta_1 \%\Delta x$
Quadratic	y	$x + x^2$	$\Delta y = (\beta_1 + 2\beta_2 x)\,\Delta x$

Note: DV = dependent variable; RHS = right hand side

Multiple Regression Model

Multiple regression is more useful than simple regression because we can more plausibly estimate *ceteris paribus* relationships (i.e. $E(u|x) = E(u)$ is more plausible)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

$\hat{\beta}_1, \ldots, \hat{\beta}_k$: **partial effect** of each of the x's on y

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}_1 - \cdots - \hat{\beta}_k \overline{x}_k$$

$$\hat{\beta}_j = \frac{\widehat{Cov}(y, \text{residualized } x_j)}{\widehat{Var}(\text{residualized } x_j)}$$

where "residualized x_j" means the residuals from OLS regression of x_j on all other x's (i.e. $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_k$)

Gauss-Markov Assumptions

1. y is a linear function of the β's
2. y and x's are randomly sampled from population
3. No perfect multicollinearity
4. $E(u|x_1, \ldots, x_k) = E(u) = 0$ (Unconfoundedness)
5. $Var(u|x_1, \ldots, x_k) = Var(u) = \sigma^2$ (Homoskedasticity)

When (1)-(4) hold:	OLS is unbiased; i.e. $E(\hat{\beta}_j) = \beta_j$
When (1)-(5) hold:	OLS is Best Linear Unbiased Estimator

Variance of u (a.k.a. "error variance")

$$\hat{\sigma}^2 = \frac{SSR}{N - K - 1} = \frac{1}{N - K - 1}\sum_{i=1}^{N} \hat{u}_i^2$$

Variance and Standard Error of $\hat{\beta}_j$

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \; j = 1, 2, ..., k$$

where

$$SST_j = (N-1)Var(x_j) = \sum_{i=1}^{N}(x_{ij} - \overline{x}_j)$$

$$R_j^2 = R^2 \text{ from a regression of } x_j \text{ on all other } x\text{'s}$$

Standard deviation:	$\sqrt{Var}$
Standard error:	$\sqrt{\widehat{Var}}$

$$se(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)}}, \; j = 1, \ldots, k$$

Classical Linear Model (CLM)

Add a 6th assumption to Gauss-Markov:

6. u is distributed $N\left(0, \sigma^2\right)$

Need this to know what the exact *distribution* of $\hat{\beta}_j$ is

- If A(6) fails, need **asymptotics** to test β's
- Then, interpret distr. of $\hat{\beta}_j$ as asymptotic (not exact)

Testing Hypotheses about the β's

- Under A (1)–(6), can test hypotheses about the β's
- Or, (much more plausible) A (1)–(5) + asymptotics

t-test for simple hypotheses

To test a simple hypothesis like

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

use a t-test:

$$t = \frac{\hat{\beta}_j - 0}{se\left(\hat{\beta}_j\right)}$$

where 0 is the null hypothesized value.

Reject H_0 if $p < \alpha$ or if $|t| > c$ (See: Hypothesis testing)

F-test for joint hypotheses

Can't use a t-test for joint hypotheses, e.g.:

$$H_0 : \beta_3 = 0,\ \beta_4 = 0,\ \beta_5 = 0$$
$$H_a : \beta_3 \neq 0 \text{ OR } \beta_4 \neq 0 \text{ OR } \beta_5 \neq 0$$

Instead, use F statistic:

$$F = \frac{(SSR_r - SSR_{ur})/(df_r - df_{ur})}{SSR_{ur}/df_{ur}} = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(N-k-1)}$$

where

$$SSR_r = SSR \text{ of restricted model (if } H_0 \text{ true)}$$
$$SSR_{ur} = SSR \text{ of unrestricted model (if } H_0 \text{ false)}$$
$$q = \# \text{ of equalities in } H_0$$
$$N-k-1 = \text{Deg. Freedom of unrestricted model}$$

Reject H_0 if $p < \alpha$ or if $F > c$ (See: Hypothesis testing)

Note: $F > 0$, always

Qualitative data

- Can use qualitative data in our model
- Must create a **dummy variable**
- e.g. "Yes" represented by 1 and "No" by 0

dummy variable trap: Perfect collinearity that happens when too many dummy variables are included in the model

$$y = \beta_0 + \beta_1 happy + \beta_2 not\_happy + u$$

The above equation suffers from the dummy variable trap because units can only be "happy" or "not happy," so including both would result in perfect collinearity with the intercept

Interpretation of dummy variables

Interpretation of dummy variable coefficients is always relative to the excluded category (e.g. *not_happy*):

$$y = \beta_0 + \beta_1 happy + \beta_2 age + u$$

β_1: avg. y for those who are happy *compared to* those who are unhappy, holding fixed age

Interaction terms

interaction term: When two x's are multiplied together

$$y = \beta_0 + \beta_1 happy + \beta_2 age + \beta_3 happy \times age + u$$

β_3: difference in age **slope** for those who are happy *compared to* those who are unhappy

Linear Probability Model (LPM)

When y is a dummy variable, e.g.

$$happy = \beta_0 + \beta_1 age + \beta_2 income + u$$

β's are interpreted as *change in probability*:

$$\Delta \Pr(y = 1) = \beta_1 \Delta x$$

By definition, homoskedasticity is violated in the LPM

CLM violations

Omitted Variable Bias

When an important x is excluded: **omitted variable bias**.

Bias depends on two forces:

1. Partial effect of x_2 on y (i.e. β_2)
2. Correlation between x_2 and x_1

Which direction does the bias go?

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive Bias	Negative Bias
$\beta_2 < 0$	Negative Bias	Positive Bias

Note: "Positive bias" means β_1 is too big;
"Negative bias" means β_1 is too small