Statistical Reasoning

Week 10

Sciences Po - Louis de Charsonville

Spring 2018

Outline

Research Paper

Research Paper

Timeline

2^{nd} **draft**	**10 April**
Week 11	17 April
Final draft	**24 April**

Writing

Explore associations

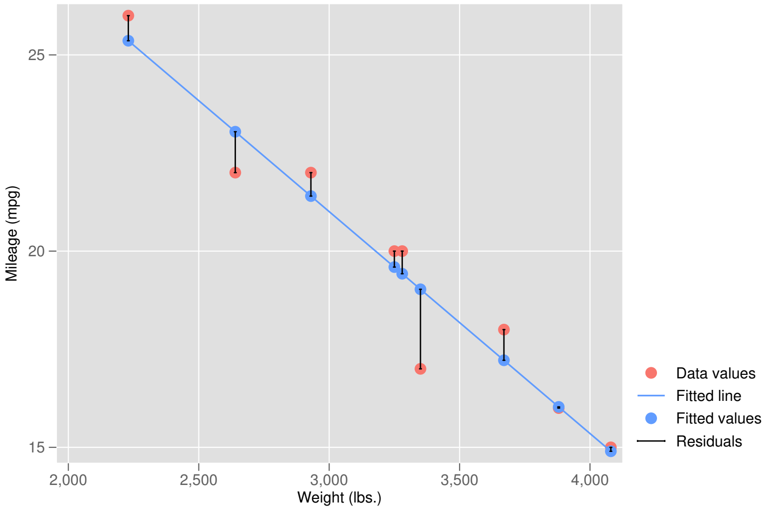
- Stata Guide, Sec. 10 `ttest`, `prtest`, `tab`, `chi2`, `pwcorr`
- Stata Guide, Sec. 11 `sc`, `lowess`, `pwcorr`, `reg`, `rfvplot`

Write up **substantive results** as sentences ; cite significance tests and other statistics in brackets : ($\rho = .7$) ($p < .05$).

Go through editing

- Remove technical content
- Rewrite until concision
- Keep your message clear

Regression


25
20
15
Mileage (mpg)
2,000
2,500
3,000
3,500
4,000
Weight (lbs.)
Data values
Fitted line
Fitted values
Residuals

Regression

Mathematical Form

$$Y = \alpha + \beta X + \sum_i \gamma_i C_i + \epsilon \tag{1}$$

Key ingredients

- Dependent variable, or outcome variable Y
- Treatment variable X
- Control variables C_i

Key outcomes

- Intercept α
- Effect of treatment β
- Effect of controls γ_i

Assumptions of the regression

Simple Linear Regression

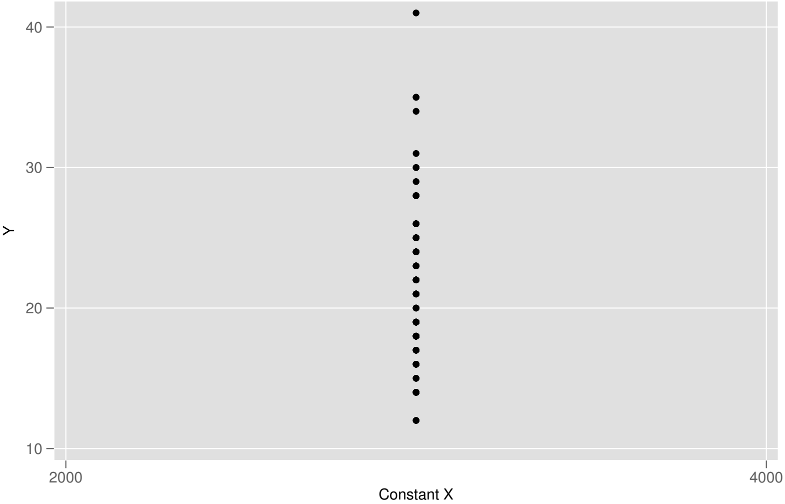
$$Y = \alpha + \beta X + \epsilon \qquad (2)$$

β and α are correctly estimated under the following **assumptions** :

1. H_1 : Sample variation in X
2. H_2 :Random sampling : $\{Y_i, X_i\}$ are independent and indentically distributed (i.i.d.)
3. H_3 : Zero Conditional mean : $E(\epsilon|X) = 0$ or in *plain English* : "values of the residuals, ϵ, does not depend on X.
4. H_4 : Linear in *parameters*
5. H_5 : Heteroscedasticity : $Var(\epsilon|X) = Var(\epsilon)$. Variance of the residuals does not depend of X

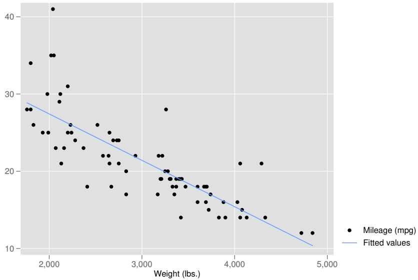
Break H_1 : No sample variation in X

H_1 not true $\Rightarrow$ X is constant

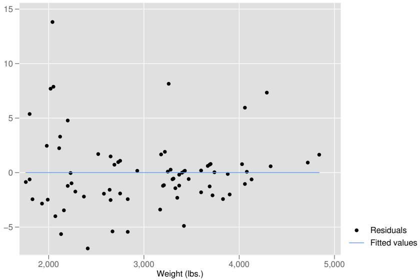


H_3 : ϵ does not depend of X

H_3 : $E(\epsilon|X) = 0$



`tw (sc mpg weight) (lfit mpg weight)`



`tw (sc epsilon weight) (lfit epsilon weight)`

H_4 : Linear in *parameters*

H_4 not true $\Rightarrow$ non-linearity in parameters

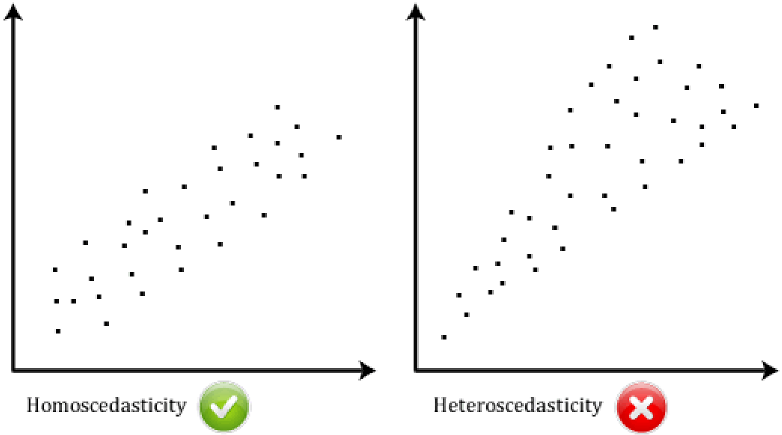
Example

$$Y = \alpha + \beta^2 X + \epsilon$$

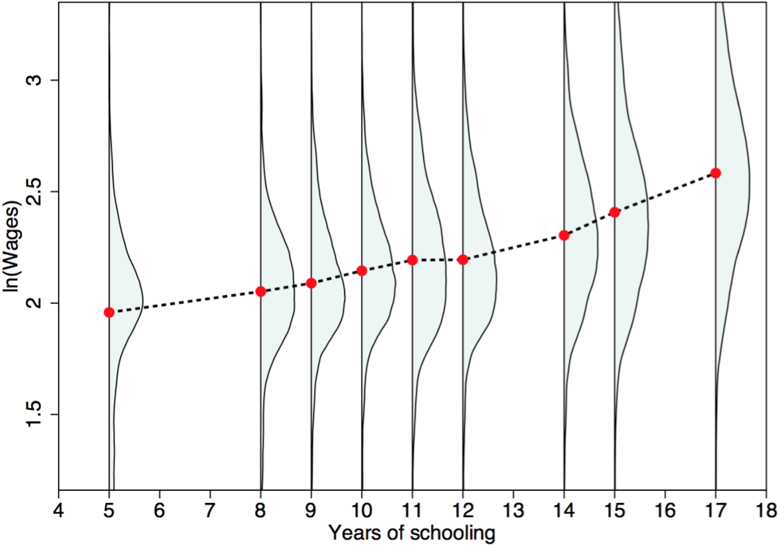
$$Y = \alpha + e^{\beta} X + \epsilon$$

H_5 : Homoskedasticity

H_5 : Homoskedasticity, variance residuals should be independent of X, e.g. $Var(\epsilon|X) = Var(\epsilon)$



Example of heteroskedasticity



Interpretation of R^2

- R^2 measures the fraction of the sample variance in Y explained by the regressors, X.
- **Low** R^2 does **not** say anything about whether we estimate **causal effects**.
- **Low** R^2 says that the model is **not useful for prediction**.

Interpretation of coefficients

Model

$$Y = \alpha + \beta X + \epsilon$$

- An increase in one unit of X is associated with an increase of β units of Y.

Standardization

- Each variable can be normalized to fit $\mathcal{N}(0,1)$ so that their **standardized coefficients** have comparable standard deviations units.
- Interpret unstandardized coefficients
- Use standardization for comparisons

Why controls ? 1/4

Do hospitals make people healthier ?

Do hospitals make people healthier ?

Group	Sample size	Mean Health Status	Std.Error
Hospital	7774	2.74	0.014
No Hospital	900049	2.07	0.003

NHIS data

- Mean difference : 0.71 (t-stat : 58.9)
- People who have been hospitalised in the past 12 months declare a significantly lower health status.

Do hospitals make people healthier ?

Let's denote

- Y_i health status of obs i
- $D_i = \{0, 1\}$ a binary variable for hospitalisation.
- Rephrase question with notation : "Is Y_i affected by hospital care ?"

Two potential outcomes :

- Y_{1i} if $D_i = 1$ (individual status if he goes to hospital)
- Y_{0i} if $D_i = 0$ (individual status had he not gone to hospital, *irrespective of whether he actually went*)

We would like to know $Y_{1i} - Y_{0i}$.

Why controls ? 3/4

Naive comparison of average

Average difference in average health =

+ Average health of hospitalised people

− Average health of non-hospitalised people

Decomposition in 4 terms :

Average difference in average health =

+ Average health of hospitalised people

− Average health of HP had they not gone to hospital

+ Average health of HP had they not gone to hospital

− Average health of non-hospitalised people

Average treatment effect on the treated

+ Average health of hospitalised people

− Average health of HP had they not gone to hospital

Selection bias

+ Average health of HP had they not gone to hospital

− Average health of non-hospitalised people

Notations

- $E[Y_i|D_i = 1]$ = Average health of hospitalised people
- $E[Y_i|D_i = 0]$ = Average health of non-hospitalised people
- $E[Y_{0i}|D_i = 1]$ = Average health of hospitalised people had they not gone to hospital
- $E[Y_{0i}|D_i = 0]$ = Average health of unhospitalised people

Mathematically

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

Observed difference in averages = average treatment effect on the treated + selection bias

Controls and causality

- There is causality if the variable of interest is independent of potential outcomes.
- In randomized controlled trials, this is typically the case.
- In non-random assignment, we need to assume that after controlling for C_i, both the treated and non-treated groups are equivalent in their remaining characteristics

Conditional Independence Assumption

The dependent variable (or outcome) is independent of the independent variable of interest (or treatment), conditionals on control.
Formally :

$$Y \perp X | C$$

Omitted variable bias

Let's assume the underlying true model is :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{3}$$

We estimate :

$$Y = \tilde{\alpha} + \tilde{\beta}_1 X_1 + \epsilon \tag{4}$$

How different is $\tilde{\beta}_1$ from β_1 ?

Omitted variable bias

Let's assume the underlying true model is :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{3}$$

We estimate :

$$Y = \tilde{\alpha} + \tilde{\beta_1} X_1 + \epsilon \tag{4}$$

How different is $\tilde{\beta_1}$ from β_1 ?

Bias on $\tilde{\beta_1}$ depends on the correlation between X_1 and X_2 :

	$Corr(X_1, X_2) > 0$	$Corr(X_1, X_2) < 0$
$\tilde{\beta_1} > 0$	+	-
$\tilde{\beta_1} < 0$	-	+

Bad controls

- More controls are not always better
- Bad controls : variables that are themselves outcome variables in the experiment
- Good control : have been fixed at the time the dependent variable was determined.
- Timing uncertain / Unknown ? → Explicit assumptions about what happened first, or assumption that none of the control variables are themselves caused by the regressor of interest.

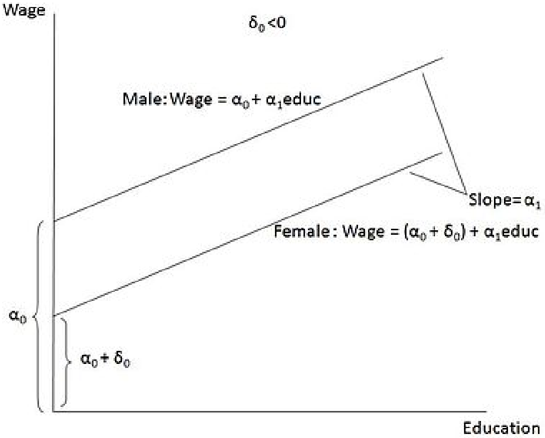
Single coefficient of dummy X_3

$$Y = \alpha_0 + \alpha_1 X_1 + \delta_0 * 0 + \epsilon$$
$$Y = \alpha_0 + \alpha_1 X_1 + \delta_0 * 1 + \epsilon$$

The omitted category $X_3 = 0$ is called the **reference category** and is part of the baseline model $Y = \alpha$, for which all coefficients are null

Example

$$Income = \alpha_0 + \alpha_1 * education + 0 * female + \epsilon$$



Practice

- Rerun week9.do and analyze residuals.