ECN 102, Summer 2020

Week 4 Recap
Multiple Regression

Dummy Variables

- Used to capture categories, binary designation
- Let *floss* $= 1$ if a person flosses every day, *floss* $= 0$ if not
- Use flossing to explain cavities: $cavities = \beta_1 + \beta_2 floss + \epsilon$
- Two categories but only one dummy variable. If m categories, include $m - 1$ dummies to avoid dummy variable trap.
- Coefficients relative to omitted **reference category**
- Those who floss? $\widehat{cavities} = b_1 + b_2$
- Those who don't floss? $\widehat{cavities} = b_1$
- So b_2 tells you how many more (or fewer if negative) cavities a flosser has relative to a non-flosser, on average

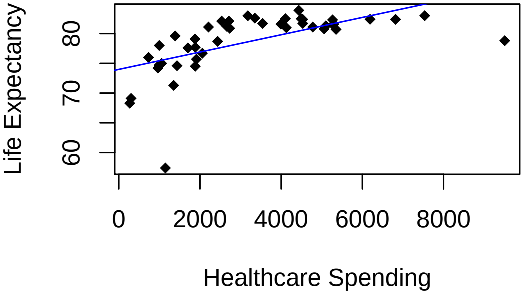
Dummy Variable Trap

- Suppose we have dummy *floss* and another dummy *notfloss*
- If a person flosses: $\textit{floss} = 1$ and $\textit{notfloss} = 0$
- If a person does not floss: $\textit{floss} = 0$ and $\textit{notfloss} = 1$
- In both cases, $\textit{floss} + \textit{notfloss} = 1$
- Can therefore write $\textit{floss} = 1 - \textit{notfloss}$
- **Perfect multicollinearity** means one regressor can be expressed as a perfect linear function of other regressors. OLS explodes.
- For categories, solution is to just drop one of the categories from the regression and make that the reference category.

Logarithms 1

- Can use logarithms (and really any other function) of variables. Still linear in parameters

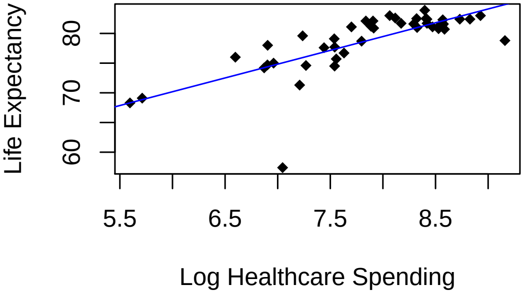
- For example, we might want to explain life expectancy with health care expenditure



- Relationship looks logarithmic to me

Logarithms 2

- Use regression $le = \beta_1 + \beta_2 \log(hcspending) + \epsilon$



- No one ever talks about "changes in log healthcare spending"
- $\widehat{le} = 43.30 + 4.65 \times \log(hcspending)$
- 1% higher healthcare spending is associated with, on average, an increase in life expectancy of about $b_2/100 = 0.0465$ years.

Logarithms 3

Model	Dependent Variable	Regressor	Interpretation of β_2
linear	y	x	$\Delta y = \beta_2 \times \Delta x$
linear-log	y	$\log(x)$	$\Delta y \approx \frac{\beta_2}{100} \times \%\Delta x$
log-linear (semi-elasticity)	$\log(y)$	x	$\%\Delta y \approx 100\beta_2 \times \Delta x$
log-log (elasticity)	$\log(y)$	$\log(x)$	$\%\Delta y \approx \beta_2 \times \%\Delta x$

Table: Interpret Δ as "difference in" rather than "change in" to avoid unintentional causal interpretation. For example, log-linear regression says we expect the percentage difference in y to be $100\beta_2$ times the difference in x. More concretely, when we consider a value of x that is larger by 1 unit, we expect to see a value of y that is larger by $100\beta_2$ percent.

CNLRM 1: Model Specification

- The true model is of form

$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon.$$

- It's linear
- It includes the correct regressors (which might not be linear, e.g. x_2 might be a log of something)
- Disturbances are additive
- The intuition: estimating $y = b_1 + b_2 x_2 + \ldots + b_k x_k + e$ if the true model looks different can't be right

CNLRM 2: Exogenous Explanatory Variables (Zero Conditional Mean)

- $E[\epsilon | x_2, \ldots, x_k] = 0$
- Remember, ϵ is like the "mistake" of the regression line
- When plugging in x, regression line is correct on average
- Wouldn't want to use regression line that's wrong on average
- Equivalent to disturbance term uncorrelated with any regressors, and

$$\hat{y} = b_1 + b_2 x_2 + \ldots + b_k x_k$$

- Then **partial effect** of x_2 on y is

$$\frac{\partial \hat{y}}{\partial x_2} = b_2,$$

where all other x_j are being controlled for by definition of partial derivative

CNLRM $1+2$

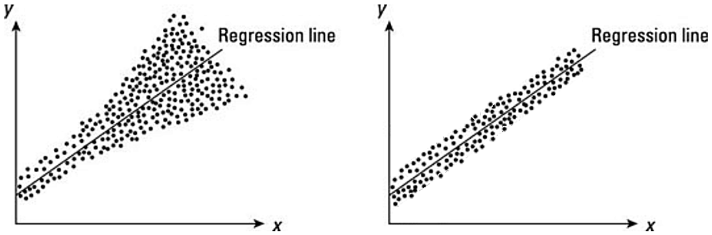
- An estimator is **unbiased** if it is correct, on average
- For example, sample mean $\overline{X}$ is unbiased because $E[\overline{X}] = \mu$
- CNLRM assumption 1 and 2 imply that OLS gives unbiased estimates

$$E[b_j] = \beta_j \quad \text{for all } j$$

- So we expect OLS to give the correct coefficients, on average

CNLRM 3: Homoskedasticity

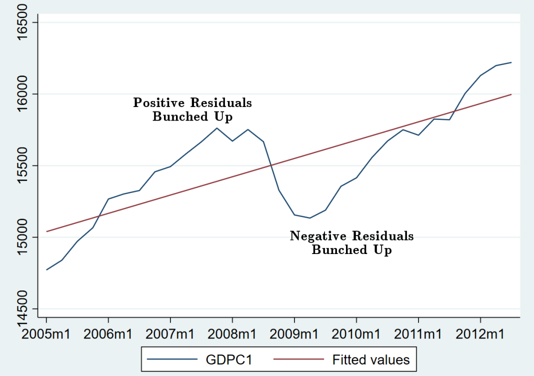
- $\text{Var}(\epsilon \mid X_2, \ldots, X_k) = \sigma_\epsilon^2 < \infty$
- The variance of the disturbance doesn't depend on x: it is constant (and finite)
- The left: heteroskedasticity. The right: homoskedasticity.



- Mostly a technical assumption, makes the math nicer

CNLRM 4: Independent Disturbances

- ϵ_i and ϵ_j are independent for $i \neq j$
- Often fails in time series and panel data
- called **autocorrelation** for time series, called **serial correlation** for panel data



CNLRM 1-4: Consistency and BLUE

- An estimator is **consistent** if it gets closer and closer (probabilistically) to the thing it's trying to estimate as the sample size gets bigger (i.e. law of large numbers is satisfied)
- When you get more and more observations, $\bar{X} \xrightarrow{p} \mu$
- Likewise when CNLRM 1-4 hold, more and more observations implies $b_j \xrightarrow{p} \beta_j$ for all j
- CNLRM 1-4 also imply that OLS gives the **b**est **l**inear **u**nbiased **e**stimates, or **BLUE**
- Here, "best" means the most efficient, i.e. the smallest standard errors

CNLRM 5: Normal Disturbances

- $\epsilon \sim N(0, \sigma^2)$
- Needed for hypothesis testing on small samples
- CNLRM 1-5 imply that OLS gives the **b**est **u**nbiased **e**stimates, or **BUE**
- So OLS is best when compared to both linear and nonlinear models

CNLRM 6: Degrees of Freedom, Variation, Perfect Multicollinearity

- Need to have $n > k$ because $T(0)$ isn't a thing: implies infinitely large standard error
- Need to have variation in x because otherwise the line of best fit is essentially vertical: can't have estimates exploding to infinity
- **Perfect multicollinearity** means that one regressor can be written as a perfect linear function of other regressors: becomes impossible to untangle the coefficients for each regressor (like having fewer equations than unknowns)
- OLS fails catastrophically if any of these conditions are violated