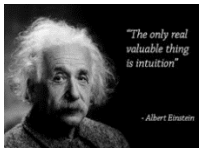
OLS Regression Assumptions



OLS Assumptions: Intuition

- How does **OLS** find a regression line that **minimizes the SSE**?
  →The coefficient vector is obtained with this **matrix** operation
  $\beta = (X'X)^{-1}X'Y$
- It can be shown **mathematically** that this matrix operation yields the OLS line
- This matrix operation may seem complicated, but it can be computed easily with computational **software** that handles matrix operations, like **SAS**, **R**, etc.
- But this formula is derived from a complex mathematical **proof** that simplifies substantially when some **assumptions are met**.
- If some of these assumptions are **not met**, the above formula does not necessarily produce coefficients for a regression line that minimizes the SSE.
- And **other methods** may be more appropriate

It's Good to be BLUE

When the OLS assumptions are met, the OLS estimators are said to be **BLUE**

- ✓ **Best** → Estimators have the lowest variance
- ✓ **Linear** → The regression is a linear combination of variables and coefficients, however many regression models use transformations (e.g., log, quadratic), but they are still considered linear models.
- ✓ **Unbiased** → the estimated β coefficients represent the true effect (remember that some models yield biased coefficients, for example is important variables are omitted from the model.

OLS Main Assumptions

- **(YC) Y is continuous** – if Y is a dummy, categorical, discrete or truncated, other methods are needed (e.g., Logistic, Probit, etc.)
- **(YN) Y is normally distributed** – not critical if the errors are normally distributed; there are acceptable transformations [e.g., log(Y), Box-Cox, 1/Y, quadratic, etc.] when Y is not normally distributed
- **(XI) X's are independent** (uncorrelated) – some correlation in the X's are tolerable if multicollinearity is not severe. With high multicollinearity, other methods are more appropriate (e.g., structural equation models)
- **(LI) Y and X's have linear relationship** – if not, some X's can be transformed to create a linear model (e.g., $Y = \beta_0 + \beta_1 X + \beta_1 X^2 + \varepsilon$)
- **(OI) Observations are independent** – if one observation is influenced by another (e.g., the temperatures hour by hour, survey responses by people who are related), methods that correct for serial correlation are more appropriate.
- **(EI) Errors are independent** – if errors are correlated it is an indication that there is a missing variable in the model.
- **(EA) The error average is 0** (+/– errors average out) – OLS takes care of this
- **(EV) The error variance is constant** – uneven residuals cause observations with large errors to pull the regression line making it biased. Methods like Weighted Least Squares (WLS) correct this problem

Two Reasons for not Using OLS:

1. **OLS assumptions are not met**
(you really can't use OLS); or
2. **Another method has higher predictive accuracy**
(you can use OLS, but another method is better)

KOGOD SCHOOL
of
BUSINESS