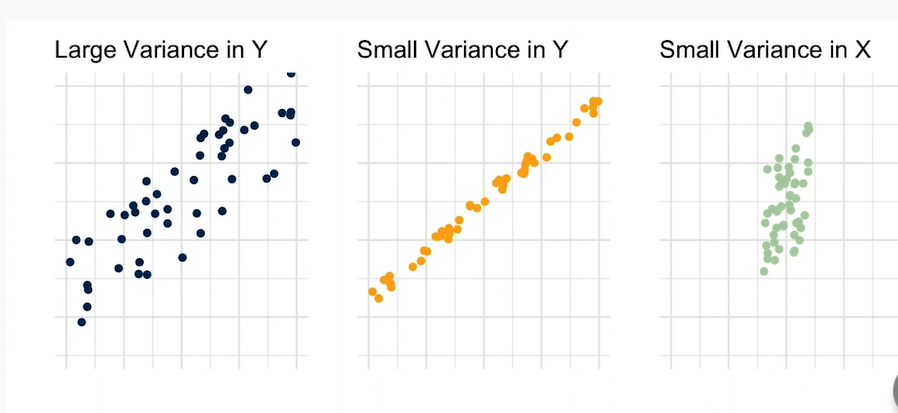
Regression and Multi-factor Experiments

1. Omitted Variable Bias (OVB) and Attenuation Bias (when X is mismeasured, our estimated treatment effect will be biased toward 0)
   1. Example of OVB
      1. Returns to Schooling example: this is observational data about whether further education would increase your wages.
         1. When work experience is included as a covariate, the explanatory coefficient of education drops (it is originally an omitted variable in the short. Schooling and experience are negatively correlated with each other. Schooling and experience are positively correlated with earnings.
         2. Ability is also an omitted variable. When IQ is included as a covariate, the estimated returns to schooling fell, so the OVB caused an overestimate of returns to schooling.
         3. Still have more OVB. We can never know if we have the right set of covariates without an experiment.
   2. Bad Controls: when you include other outcome variables as covariates on the right-hand side of the regression.
      1. Example: your occupation can be an outcome of going to college rather than it being just a covariate affecting your income.
      2. Think about the effect X has on the covariate. If you’re including the covariate which is a bad control, you’re having a second effect of X on the outcome.
   3. Robust standard errors and confidence intervals in regression
      1. When running an experiment, adding covariates cannot decrease precision because the treatment should be uncorrelated with all covariates.
      2. Rule of thumb about standard errors
         1. Standard errors are larger when variance in Y is larger.
         2. Standard errors are smaller when variance in X is larger.



* + 1. Standard errors in regression output
       1. Treatment effect (with 95% confidence interval) is about slope coefficient ±2 standard errors
       2. OLS standard errors assume ach observation’s idiosyncratic component, ε is i.i.d.
       3. Independence is sensible in a randomized experiment
    2. Heteroskedasticity (different observations have different error variances) and homoscedastic (different observations have the same error variances)
       1. Homoskedasticity is the default assumption under which OLS standard errors are usually computed. Default is just to use robust standard error even if homoskedasticity is true.
       2. Vertical error variance causes uncertainty about line’s true slope.
       3. Fanned out data means many lines could be fit depending on the sample
       4. More accurate plot with accuracy distant from grand mean: end points anchor slope.
          1. Leverage: data points nearer ends of regression line influence slope more.

When estimating the variance of , take the weighted sum of squared residuals, divide by the total variance in X. The weights in weighted sum correspond to leverage of each observation. Squared residuals, , weighted by squared horizontal deviations from the mean.

* + - 1. Robust standard errors/heteroskedasticity-robust standard errors/Eicher-White standard errors /Huber-White standard errors: estimate accurate confidence intervals, even when error variance varies with X.
         1. Do not require knowledge of shape of heteroskedasticity or which x-variables correlate with variance.
  1. If treatment assignment is clustered, must use clustered standard errors to avoid unintentionally overstating precision of estimates.
  2. Multifactor Design
     1. Regression Analysis of multifactor experiments
        1. Estimate regressions with interaction terms to estimate how much more one treatment matters when the other treatment is turned on.
     2. Example: The Visible Hand. Three dimensions: Race and class, ad quality, and asking price. Do Black sellers hurt their sales more by using bad grammar than White sellers?
     3. Example: Hispanic and Grammar.
        1. estimates the average response in the omitted category (Colin, Good Grammar)
        2. estimates the effect of ethnicity when there are no grammar errors,
        3. estimates the effect of grammar, when the ethnicity signal is Colin, .
        4. , the interaction coefficient, estimates how much more the grammar errors matter for Jose than for Clin.
        5. Perform a Wald-test on the regression coefficient:
     4. Regression gracefully handles non-binary categorical variables (where an F-test would be required).

1. Big Picture in Estimating Causal Effects
   1. Questions to ask
      1. What is the causal relationship of interest?
      2. What is the ideal experiment to measure this?
         1. Even if you’re doing observational research, ask this question!
         2. Can you redefine your question?
      3. What is your identification strategy?
         1. Where does the variation come from?
         2. Why is this variation independent of potential outcomes?
      4. How are you computing the confidence intervals?