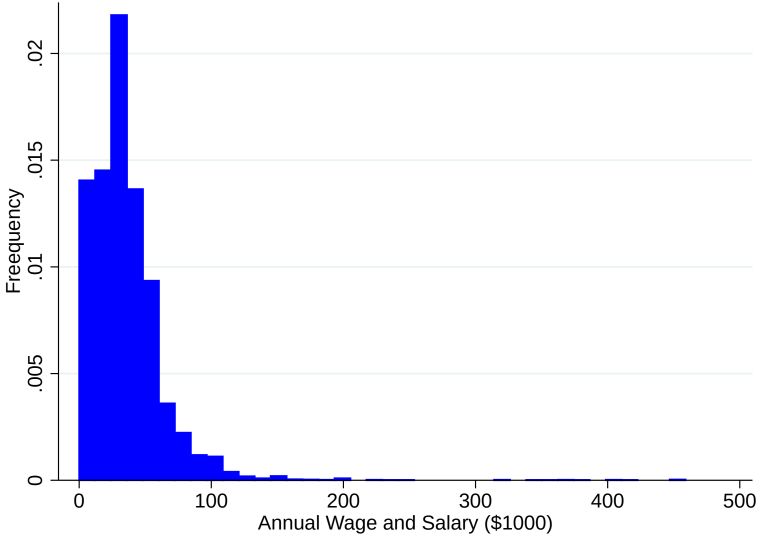
Recitation 1: Regression Review

Christina Patterson

Outline For Recitation

1. Statistics
   - Bias, sampling variance and hypothesis testing.
   - Two important statistical theorems: Law of large numbers (LLN) and Central Limit Theorem (CLT)
2. Simple Regression
   - The mechanics of Ordinary Least Squares (OLS) regression: coefficients, and standard errors
3. Multiple Regression
   - The mechanics of multiple regression: interpretations and dummy variables.

Statistics: Distribution of Wage and Salary Earnings



Statistics: Basic Definitions

- Population mean: $E[x_i]$
  - We cannot observe this
  - this is called a *parameter*
- Sample mean: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - We can calculate this: $\$34,945$
  - this is called a *sample statistic* and it is an *estimator* of $E[x_i]$
- A statistic is *unbiased* when $E[\bar{x}] = E[x_i]$

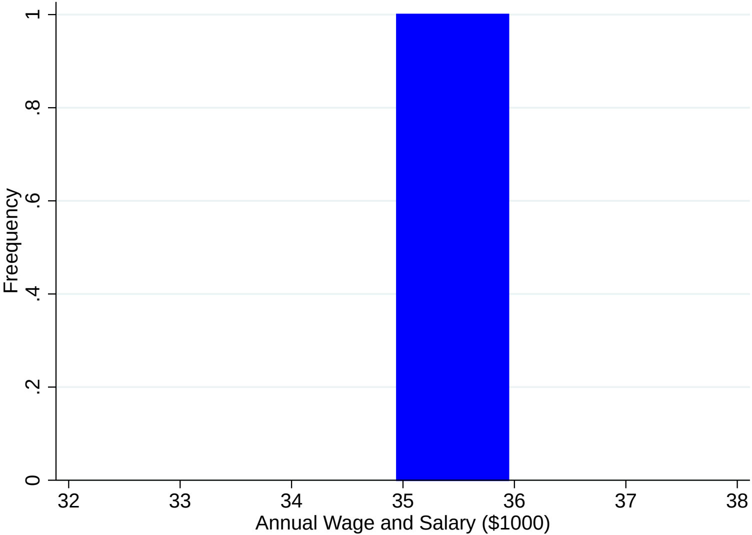
Statistics: Variance

- It's important to distinguish between two types of variance:
  1. *Population Variance:* Underlying variation in earnings in the population
  $$Var(x_i) = E\left[(x_i - E[x_i])^2\right] = \sigma_x^2$$
  2. *Sampling Variance:* The variance of the sample mean (i.e. if I kept drawing samples, how much would the sample mean vary?)
  $$Var(\bar{x}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{\sigma_x^2}{n}$$
- Note that as $n \to \infty$, the sampling variance goes to 0
- This is the result of the *Law of Large Numbers*, which is a theorem that states that in sufficiently large samples, the sample average converges to the expected value.

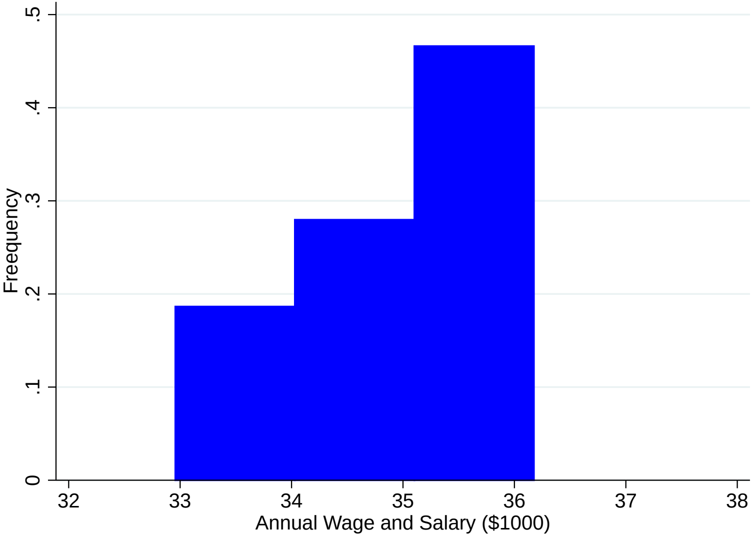
Statistics: Hypothesis Testing

- Suppose you want to test the hypothesis that the average earnings in the population are $40,000$, i.e. $E[x] = 40,000$.
- The idea is that if you had thousands of different samples of individual and you calculated the mean earnings in each of those samples, those means would be normally distributed even if the distribution of incomes in the population is very non-normal (which in fact it is)

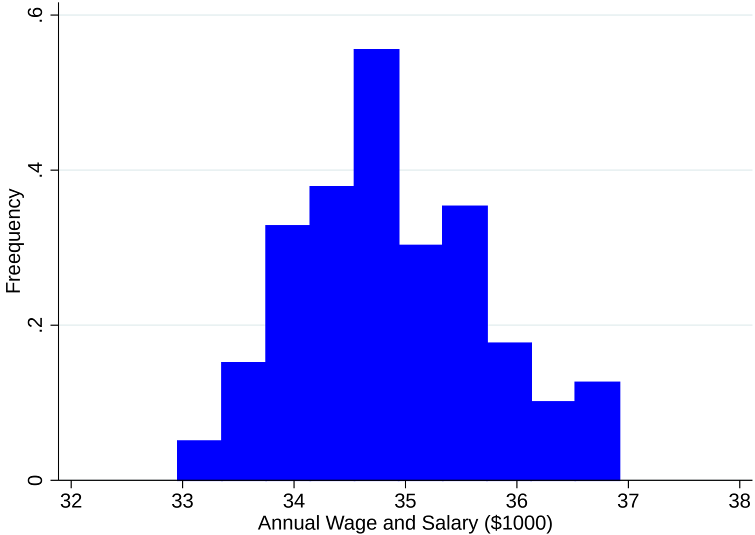
The Power of the Central Limit Theorem (1 draw)



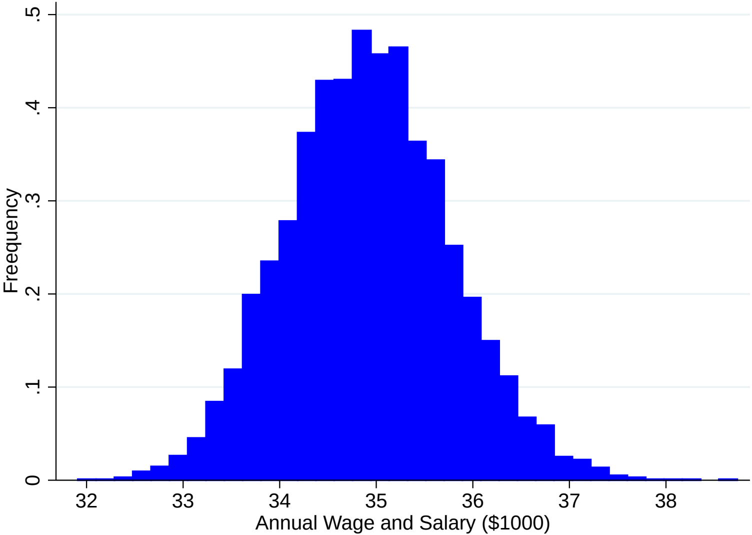
The Power of the Central Limit Theorem (10 draws)



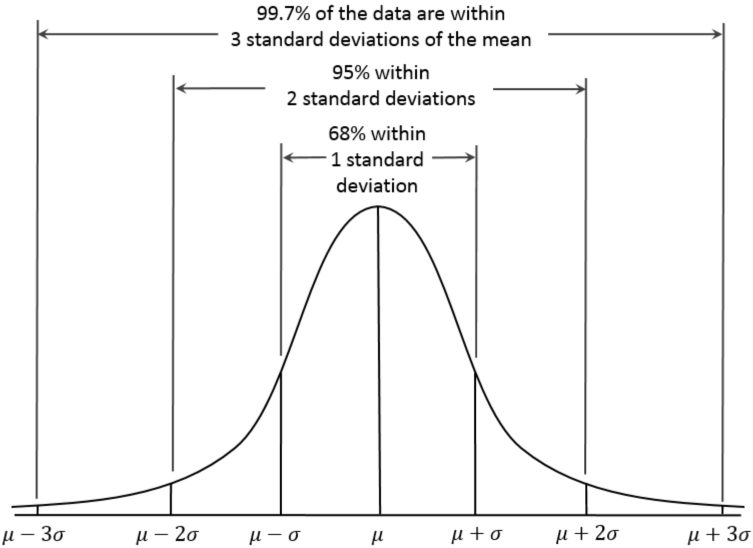
The Power of the Central Limit Theorem (100 draws)



The Power of the Central Limit Theorem (5000 draws)



ASIDE: The Normal Distribution



Statistics: The Power of the Central Limit Theorem

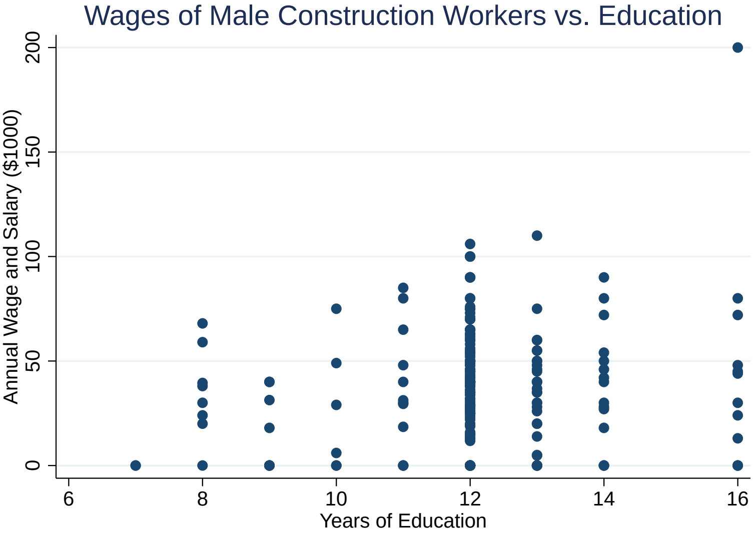
- In practice, we construct a *t-statistic*:

$$t_{40,000} = \frac{\bar{x} - 40,000}{\sqrt{Var(\bar{x})}}$$

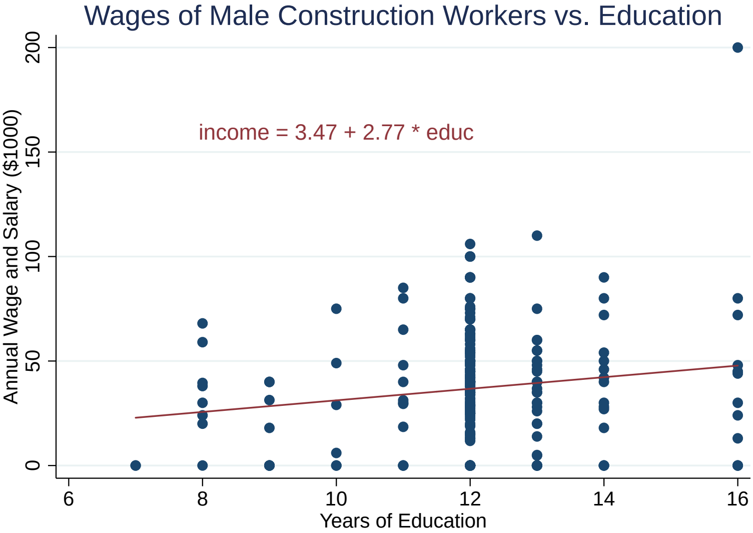
- The t-statistic is far above 2 or below -2 when the observed average is sufficiently far from our hypothesized value.
- In this case:

$$t_{40,000} = \frac{34.945 - 40}{.356} = -14.2$$

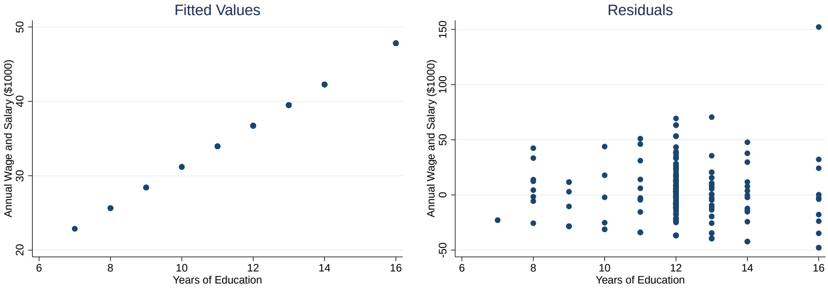
Simple Regression



Two Variables: Mechanics of OLS Regression



Two Variables: Mechanics of OLS Regression



- Fitted values (left plot): $\widehat{income}_i = \hat{a} + \hat{b} educ_i$
- Residuals (right plot): $e_i = income_i - \widehat{income}_i$

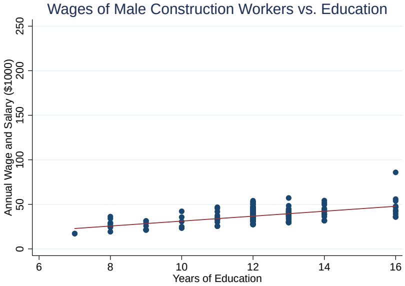
Two Variables: OLS standard errors

- Suppose I want to test the hypothesis that income increases with education (i.e. $\beta > 0$)
- This involves calculating the standard error of the estimated coefficient $\hat{\beta}$:

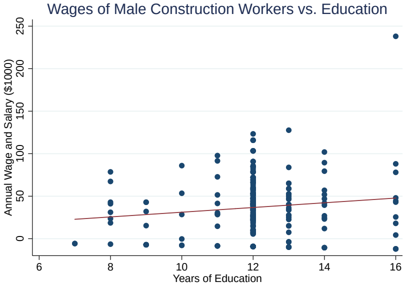
$$S.e.(\hat{\beta}) = \frac{\sigma_e}{\sqrt{n}} \frac{1}{\sigma_x}$$

  - Increasing in the variance of the residuals (σ_e).
  - Decreasing in sample size (n)
  - Decreasing in the variance of the regressor (σ_x).

OLS standard errors are increasing in the varaince of the residuals

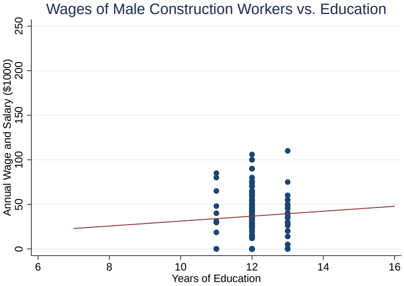


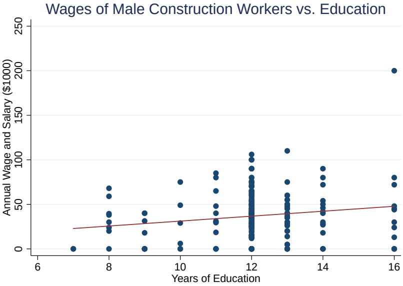
Wages of Male Construction Workers vs. Education



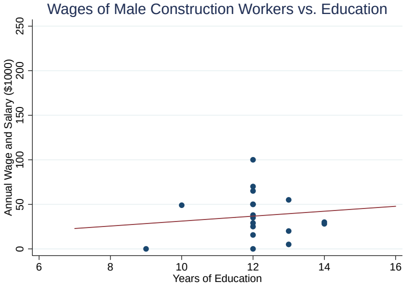
Wages of Male Construction Workers vs. Education

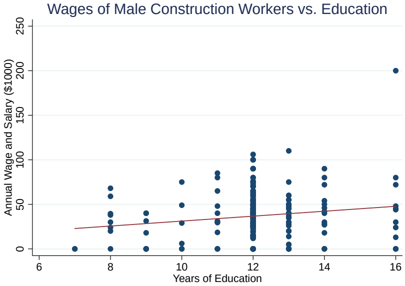
OLS standard errors are decreasing in the variance of the regressor


Wages of Male Construction Workers vs. Education


Wages of Male Construction Workers vs. Education

OLS standard errors are decreasing in the number of observations





Hypothesis Testing

- Like in the one variable case, we can construct a t-statistic:

$$t_0 = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

	(1) wage
Years of School (*b*)	2.772**
	(1.135)
Constant (*a*)	3.469
	(13.80)
Observations	245
R-squared	0.024

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Multivariate Regression

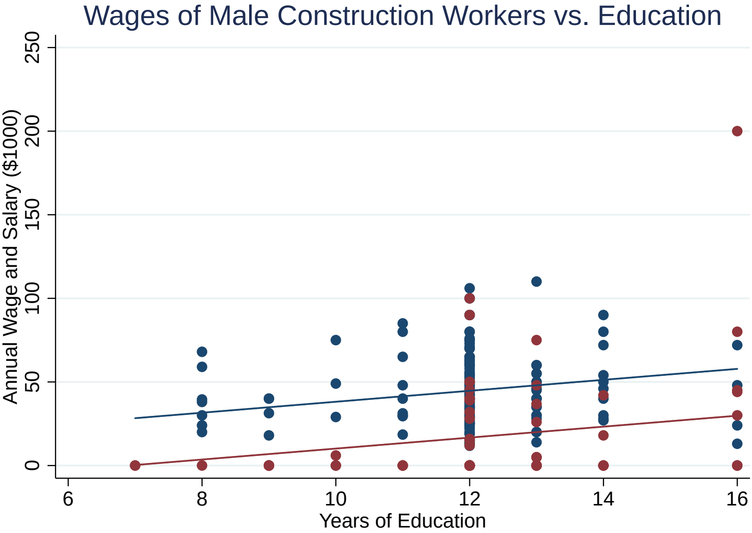
- Supposed we want to add another variable – whether the worker is self-employed.

	(1)
	wage
Years of School (b_1)	3.278***
	(1.009)
Self Employed (b_2)	-28.01***
	(3.426)
Constant (a)	5.364
	(12.25)
Observations	245
R-squared	0.235

Standard errors in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Multivariate Regression



14.03 / 14.003 Microeconomic Theory and Public Policy
Fall 2016