Data Analysis in R

Assumptions & Limits of OLS

Ken Stiller

13th January 2024

Syllabus: Data Analysis in R

1. Introduction
2. Causality & Basics of Statistics
3. Sampling & Measurement
4. Prediction
5. Multivariate Regression
6. Probability & Uncertainty
7. Hypothesis Testing
8. **Assumptions & Limits of OLS**
9. Interactions & Non-Linear Effects

Overview

- Hypothesis Testing & OLS
  - Hypothesis testing in the linear regression framework
  - Interpretation of regression tables and statistical significance
- Limits & Assumptions of OLS
  - Do's and dont's of statistical significance
  - OLS assumptions

Table of Contents

Overview

Hypothesis Testing & OLS

Interpretation

Limitations

OLS Assumptions

Wrap Up

Hypothesis Testing in the Context of Regression

- Typically your hypotheses will look something like this:
  - **Null hypothesis:** There is no association between X and Y
  - **Alternative hypothesis:** There is an association between X and Y

- Note that this hypothesis tells us nothing about the size (or 'magnitude') of the association between X and Y.

- It just tells us we suspect there is some association (i.e., non-zero).

- It's imperative to always consider statistical significance **and** coefficient size

So When do we Reject the Null?

- Here, we are looking at the significance of coefficient estimates ($\hat{\beta}$)
- Yet, this is very similar to what we discussed last week.
- We are interested in the coefficient (β) that describes association between X and Y
- How accurate is our estimate ($\hat{\beta}$)?
- β will also have a sampling distribution. What do you think does this mean?
- We could extract multiple samples from that population, run our regression, and look at the distribution of $\hat{\beta}$

So When do we Reject the Null?

- Here, we are looking at the significance of coefficient estimates ($\hat{\beta}$)
- Yet, this is very similar to what we discussed last week.
- We are interested in the coefficient (β) that describes association between X and Y
- How accurate is our estimate ($\hat{\beta}$)?
- β will also have a sampling distribution. What do you think does this mean?
- We could extract multiple samples from that population, run our regression, and look at the distribution of $\hat{\beta}$

Sampling Distribution - Again

The Idea

Just like statistics such as sample means, medians etc, regression coefficients are also estimators and thus have sampling distributions: Under repeated sampling, we would have variation in $\hat{\beta}$, and we will need to take this variation into consideration while performing hypothesis testing.

- For instance, we can illustrate the joint sampling distribution of $\hat{\beta}_0$ & $\hat{\beta}_1$ by generating simulated data.
- Holding the X values from 25 observations constant, we simulate data from a linear model with the following parameters:

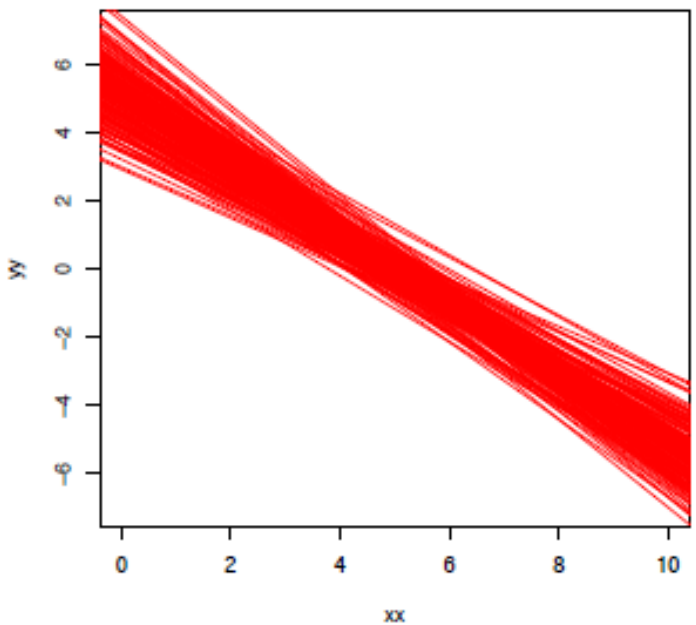
$$\sigma^2 = 4$$

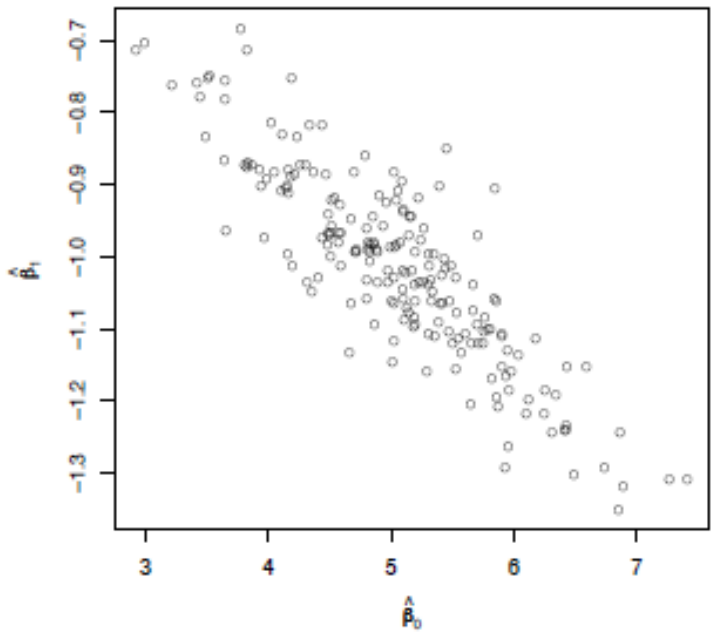
$$\beta_0 = 5$$

$$\beta_1 = -1$$

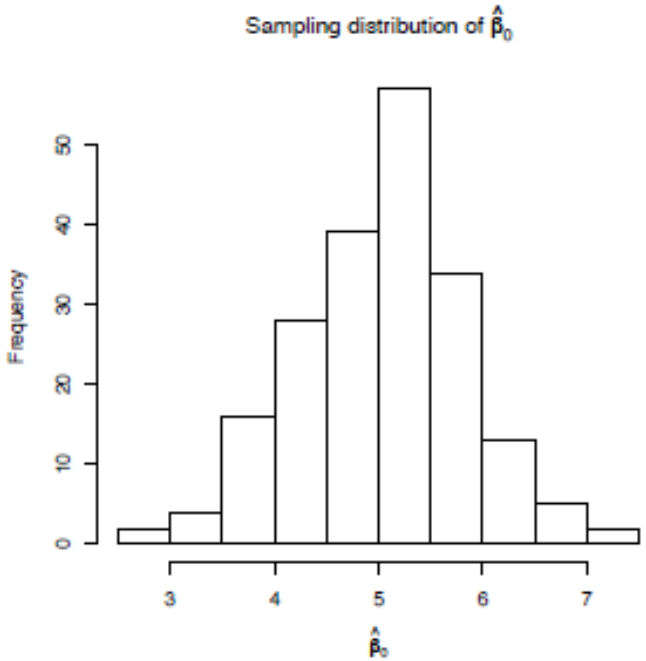
- We then estimate intercepts and slopes for each simulated data set

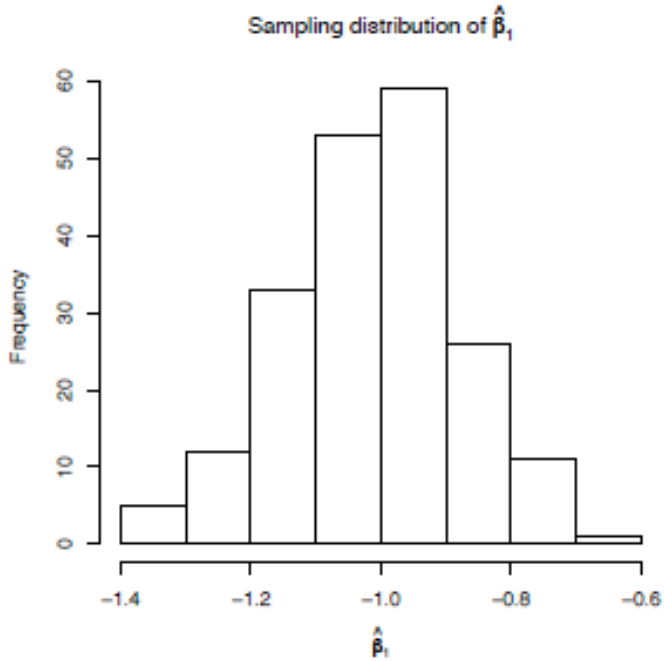
Sampling Distribution - Again II





Sampling Distribution - Again III





Hypothesis Testing with Regression

The (same) logic

We have a hypothesis about how X and Y are associated. Our research hypothesis, H_a, is that this relationship is either positive ($\beta > 0$) or negative ($\beta < 0$). The null hypothesis, (H_0) then, is simply: $H_0 : \beta = 0$.

- Using data, we get some $\hat{\beta}$. How are we going to test H_0?
- The idea is the same as with all other tests we have seen so far
- We need a statistic, which will give us a score that can be mapped into some distribution telling us the probability of observing this score based on a sample under the assumption that the null, (H_0), is true.
- How are we going to get this statistic in linear regression?
- We apply the same logic:

$$\text{statistic-score} = \frac{\hat{\beta} - \beta(\text{ Under the Null })}{\text{Standard Error of } \hat{\beta}} = \frac{\hat{\beta} - c}{\text{Standard Error of } \hat{\beta}}$$

Test of Statistical Significance: t-test

$$t_{n-2} = \frac{\text{Estimate of parameter} - \text{null hypothesis value}}{\text{Standard error of estimate}}$$

$$t_{n-2} = \frac{\hat{\beta}_1 - \beta_{H_0}}{SE\left[\hat{\beta}_1\right]}$$

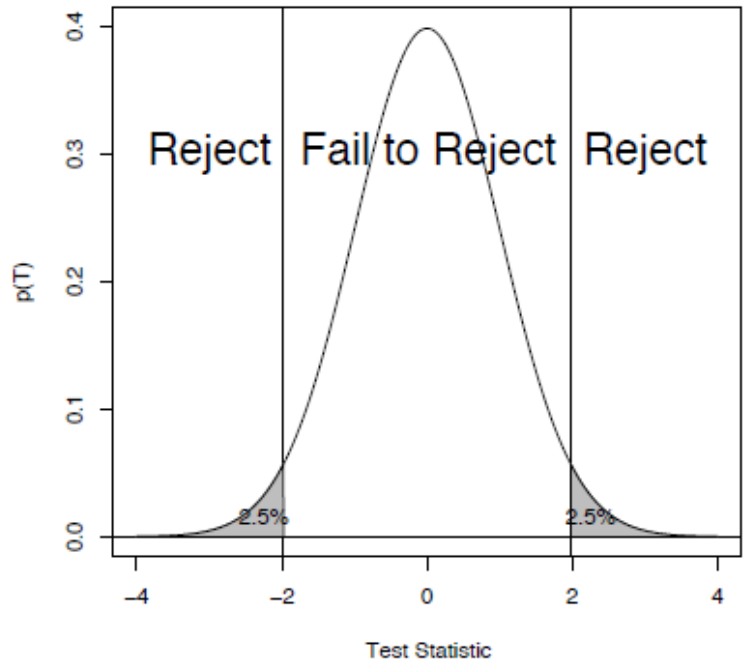
$$\text{since } \beta_{H_0} = 0,$$

$$\text{we get: } t_{n-2} = \frac{\hat{\beta}_1}{SE\left[\hat{\beta}_1\right]}$$

Note: While the null hypothesis often states that the parameter is 0, this logic is not restricted to this special case. This framework allows you to vary the value of the null. How would that change our formula? What would we learn from this?

Statistical Inference with Linear Regression

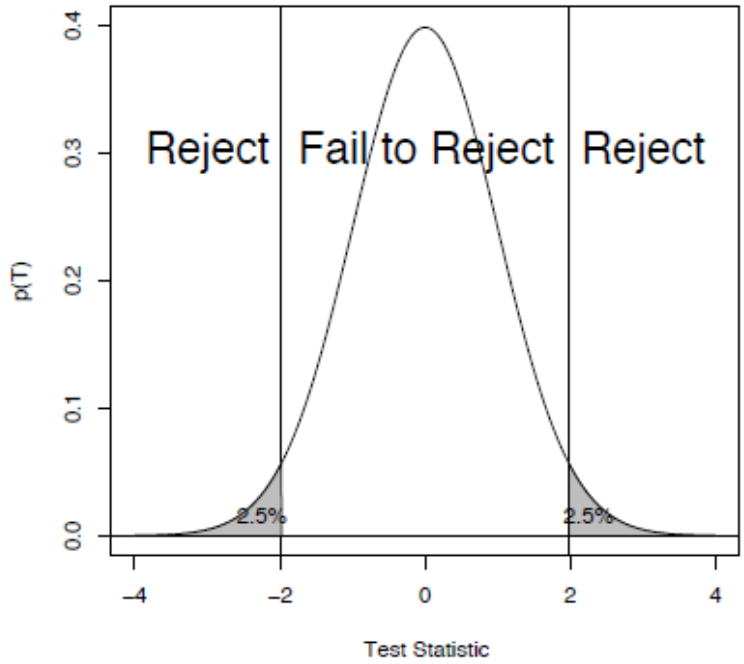
- We have specified the null and alternative hypotheses
- We set $\alpha = 0.05$
- We have seen the null distribution of the test statistic
- We now need to determine the rejection region for $\hat{\beta}_1$
- At $\alpha = 0.05$, what is the critical value for the t_{n-2} distribution?
- If n is reasonably large, 1.96
- So, we reject the null if $|t_{n-2}| > 1.96$



	$\hat{\beta}$	Std. Error	t-statistic
x	−.570	.039	−14.44
Int.	70.38	3.37	20.86

Statistical Inference with Linear Regression II

- Same logic as before
- We determine the $\hat{\beta}_1$ rejection region
- Compute the t-value using $(\hat{\beta}_1 - c)/SE(\hat{\beta}_1)$ and compare to critical value
- Here, $\hat{\beta}_1$ is -.570 and $\widehat{SE(\hat{\beta}_1)}$ is .039
- So, $t = (-.570 - 0)/0.039 = -14.44$
- What do we conclude?
- We reject H_0 at the 0.05 level



	$\hat{\beta}$	Std. Error	t-statistic
x	−.570	.039	−14.44
Int.	70.38	3.37	20.86

P-value Revisited

- Keep in mind: the p-value summarises our evidence against H_0 - it's the probability of observing a t-value at least as extreme as one we observe assuming H_0 is true
- So, a p-value is $P(\frac{\hat{\beta}_1 - c}{\widehat{SE(\hat{\beta}_1)}} > |t_{obs}|)$ or the probability the t-statistic falls outside the observed t-value under H_0
- If a p-value is $\leq \alpha$, we reject H_0.
- Thus, α is the largest p-value, for which we will reject H_0.
- In this case, our t-value was -14.44. Will the p-value be small or large?

	$\hat{\beta}$	Std. Error	t-statistic	$P > \|t\|$	Cl
x	$-.570$	.039	-14.44	0.000	$[-0.648 - -0.492]$
Int.	70.38	3.37	20.86	0.000	$[63.73 - 77.018]$

P-value Revisited

- Keep in mind: the p-value summarises our evidence against H_0 - it's the probability of observing a t-value at least as extreme as one we observe assuming H_0 is true
- So, a p-value is $P(\frac{\hat{\beta}_1 - c}{\widehat{SE(\hat{\beta}_1)}} > |t_{obs}|)$ or the probability the t-statistic falls outside the observed t-value under H_0
- If a p-value is $\leq \alpha$, we reject H_0.
- Thus, α is the largest p-value, for which we will reject H_0.
- In this case, our t-value was -14.44. Will the p-value be small or large?

	$\hat{\beta}$	Std. Error	t-statistic	$P > \|t\|$	Cl
x	$-.570$	.039	-14.44	0.000	$[-0.648 - -0.492]$
Int.	70.38	3.37	20.86	0.000	$[63.73 - 77.018]$

Confidence Intervals with Linear Regression

Same logic here

Suppose - rather than a point estimate - we want to be more conservative and specify an interval as our estimate for the location of the true slope. This is easy now since we know the sampling distribution for our t-value

- We construct the CI using the same formula and logic as in the univariate case that we have already seen:

$$\hat{\beta}_1 \pm t_{\alpha/2} * \widehat{SE(\hat{\beta}_1)}$$

- Typically, with $n > 50$ and $\alpha = 0.05$, $t_{\alpha/2} = 1.96$
- In our example: $.570 \pm 1.96 * .039 = [-.648, -.492]$
- Interpretation: Under repeated sampling, the computed interval will 95 out of 100 times include the true β
- We therefore reject H_0 if the CI doesn't contain c.

Statistical Significance: Example

Table: Effect of university education on turnout

	(1) Voted	(2) Voted
University Education	0.141	0.148
	(0.0176)	(0.0174)
Age		0.166
		(0.0244)
Constant	0.732	0.587
	(0.0113)	(0.0241)
Observations	2,152	2,152

Standard errors in parentheses.
Source: British election survey 2017

- Which estimates are statistically significant?

Statistical Significance: Example II

Table: Effect of university education on turnout

	(1) Voted	(2) Voted
University Education	0.141***	0.148***
	(0.0176)	(0.0174)
Age		0.166***
		(0.0244)
Constant	0.732***	0.587***
	(0.0113)	(0.0241)
Observations	2,152	2,152

Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
Source: British election survey 2017

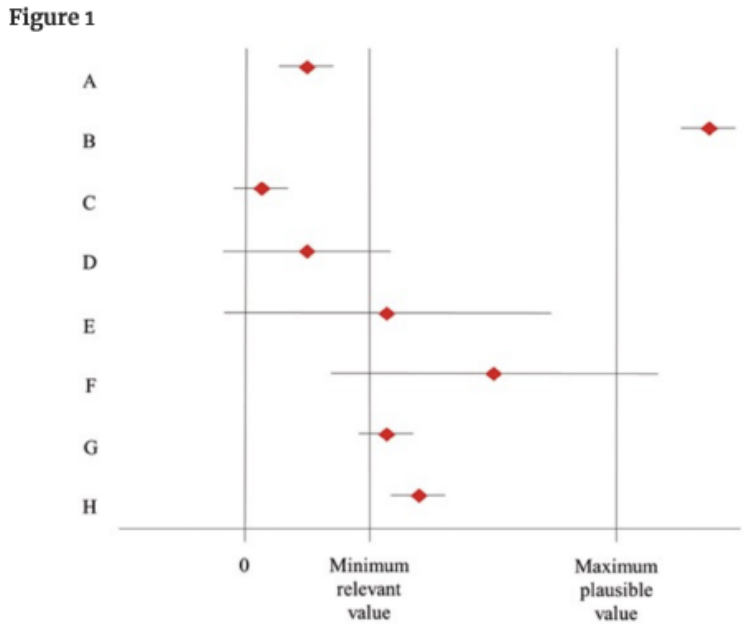
- Which estimates are statistically significant?

Caution with P-Values & Significance!

- p-values are ubiquitous in published science
- They are informative but being too strict on them can lead to problems such as p-hacking and replication bias
- If all we care about is having $p < 0.05$, this generates incentives to tweak models so that we get precisely that
- p-values have been under fire and many different proposals have been put forward (dropping them, reducing α, pre-analysis plans, etc).
- **Statistically significant results may be absolutely meaningless**

P-Values: Example I

- In addition to statistical significance, we care about coefficient magnitude. Is it:
  1. Large enough for us to care?
  2. Credible?



- *Source: Bernardi et al. 2017, "Sing me a Song with Social Significance", ESR*

Caution with P-Values & Significance II

- A statistically significant effect may well be too small [large] to be meaningful
- Remember that p-values are affected by sample size: the larger the sample, the smaller the p-value (else equal)
- With very large samples, associations become more easily significant
- The question then is: are they large enough for us to care?
- Make sure to present CI and substantive estimate of effect

Bias and Efficiency

1. **Bias** refers to the expected difference between the estimate (e.g., sample mean) and the parameter (e.g., population mean)
2. An estimator is more **efficient** when it has a lower variance than another. Lower variance means that under repeated sampling, the estimates are likely to be similar

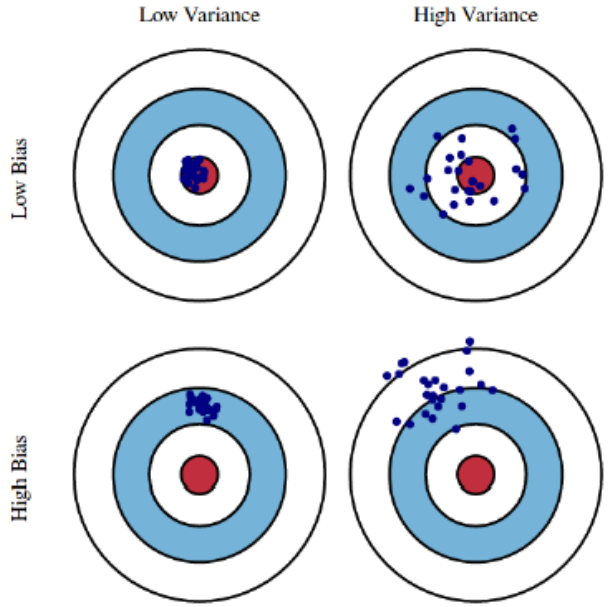


Fig. 1 Graphical illustration of bias and variance.

Source: Blog "Towards Data Science"

Properties of OLS

- Under certain assumptions, OLS is the most efficient estimator among all unbiased estimators (within the family of estimators we want to work with)
- In other words, OLS is the estimator that yields lowest variance among those that have no expected difference from the parameter
- You probably are familiar with the Gauss-Markov assumptions
- Some violations of these assumptions introduce bias, others make OLS less efficient
- There are *some potential* remedies to some violations of these assumptions → see script for more

OLS Assumptions

- Linearity: the population regression model is linear in its parameters.
- Random sampling: the observed data represent a random sample from the population described in the model
- No perfect collinearity: there is variation in the explanatory variables
- Zero conditional mean: the expected value of the error term is zero conditional on the explanatory variables.
- Heteroskedasticity: the error term has constant variance, i.e. the same variance regardless of the value of X
- Normality of error term: the error term follows a normal distribution (after conditioning on independent variables)

2. Random Sampling

- Non-random sampling can lead to bias that may lead a researcher to draw the wrong conclusions about the population from their sample.

- Non-random sampling can happen in different ways:
  1. You sample from a universe that does not represent the whole population (e.g., University students, Facebook users, etc)
  2. You condition on availability of a measure that exists only for a given part of the population
     - e.g., death toll in ongoing conflict. Only places with lower conflict can be sampled

4. Zero Conditional Mean

- A linear regression model is a *linear* approximation of the relationship between explanatory variables X_i and a dependent variable Y

$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{v}_{\text{error term}}$$

  - Y: dependent/outcome variable
  - X: independent/explanatory variable
  - v: unobserved error/disturbance term

- As you know, β is the regression coefficient of X, which describes the association between X and Y

4. Zero Conditional Mean II

- *For any given value of X, the average of unobservables is the same and therefore must equal the average value of v in the population (which we can trivially recode to 0). Combining the two assumptions we get the zero conditional mean assumption.*

- This is by far the most important assumption of all OLS assumptions. This is crucial if you want your coefficients to have a causal interpretation.

- Note that we assume that we assume that v and X are independent, i.e. that the average value of v does not depend on X, which means that $E[v|X]$ is the same for all X (constant).

4. Zero Conditional Mean III

- In order for $\hat{\beta}$ to be unbiased we need the following zero condition mean condition ot hold: $E[v|X] = 0$
  - In observational studies, X is likely to be determined by omitted variables in v, which could be also related to Y
  - thus, $E[\hat{\beta}] \neq \beta$
  - This we know as omitted variable bias

- We learnt that a common practice that aims to account for omitted variable bias is to use C controls in the analysis:

$$Y = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope1}} X + \underbrace{\underbrace{\beta_2}_{\text{slope 2}} C + \underbrace{\epsilon}_{\text{error term}}}_{\text{previous error term}}$$

- Where: $cov(C, Y) \neq 0$ and $cov(C, X) \neq 0$
- In doing so, we aim at $E[v|X] = 0$

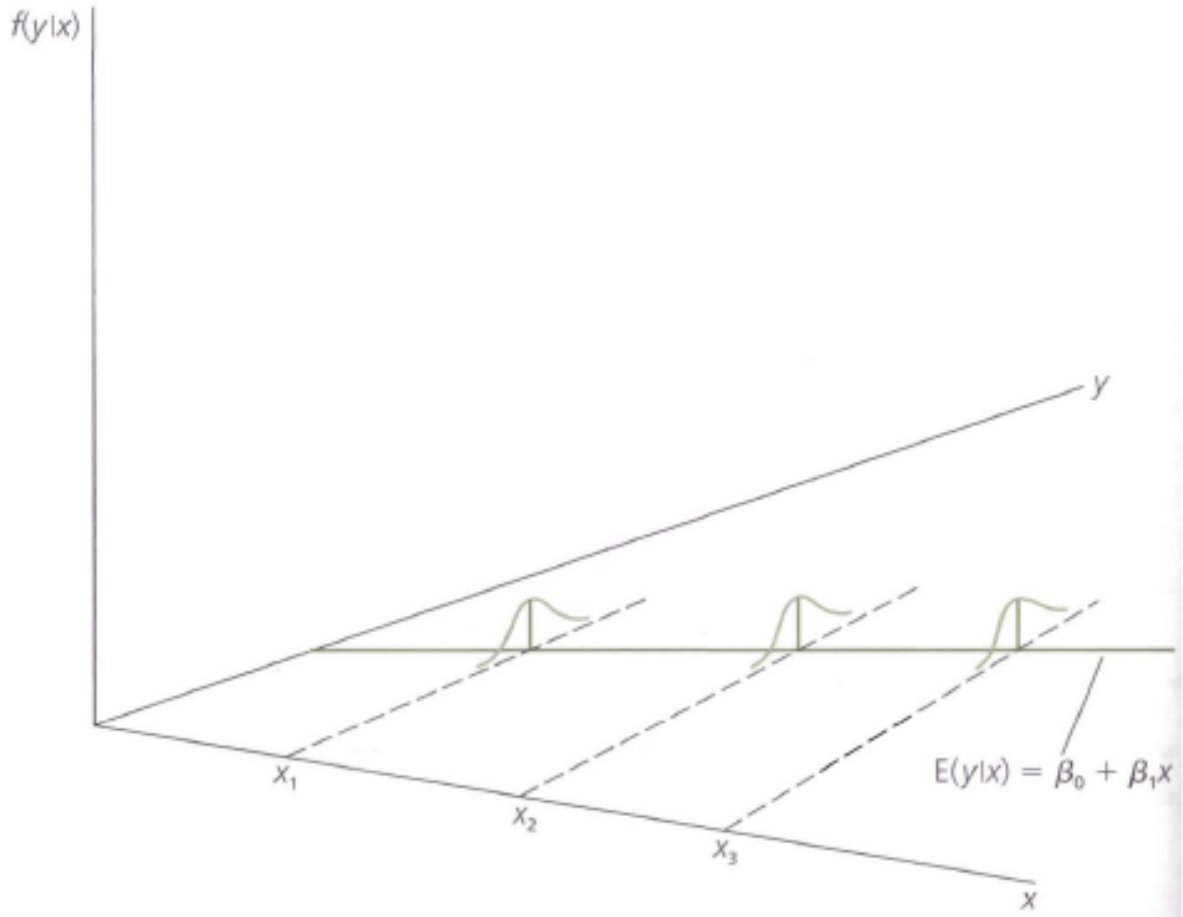
5. Variance of OLS Estimators

- Suppose we have an unbiased estimator and know the mean of the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$. Now, if we want to test hypotheses, we need to know how far $\hat{\beta}_0$ and $\hat{\beta}_1$ are from β_0 and β_1 on average
- In other words, we don't only need an unbiased point estimate, but also an unbiased estimate of the variance and, subsequently, standard errors.
- Thus, we introduce assumption 5: Homoskedasticity

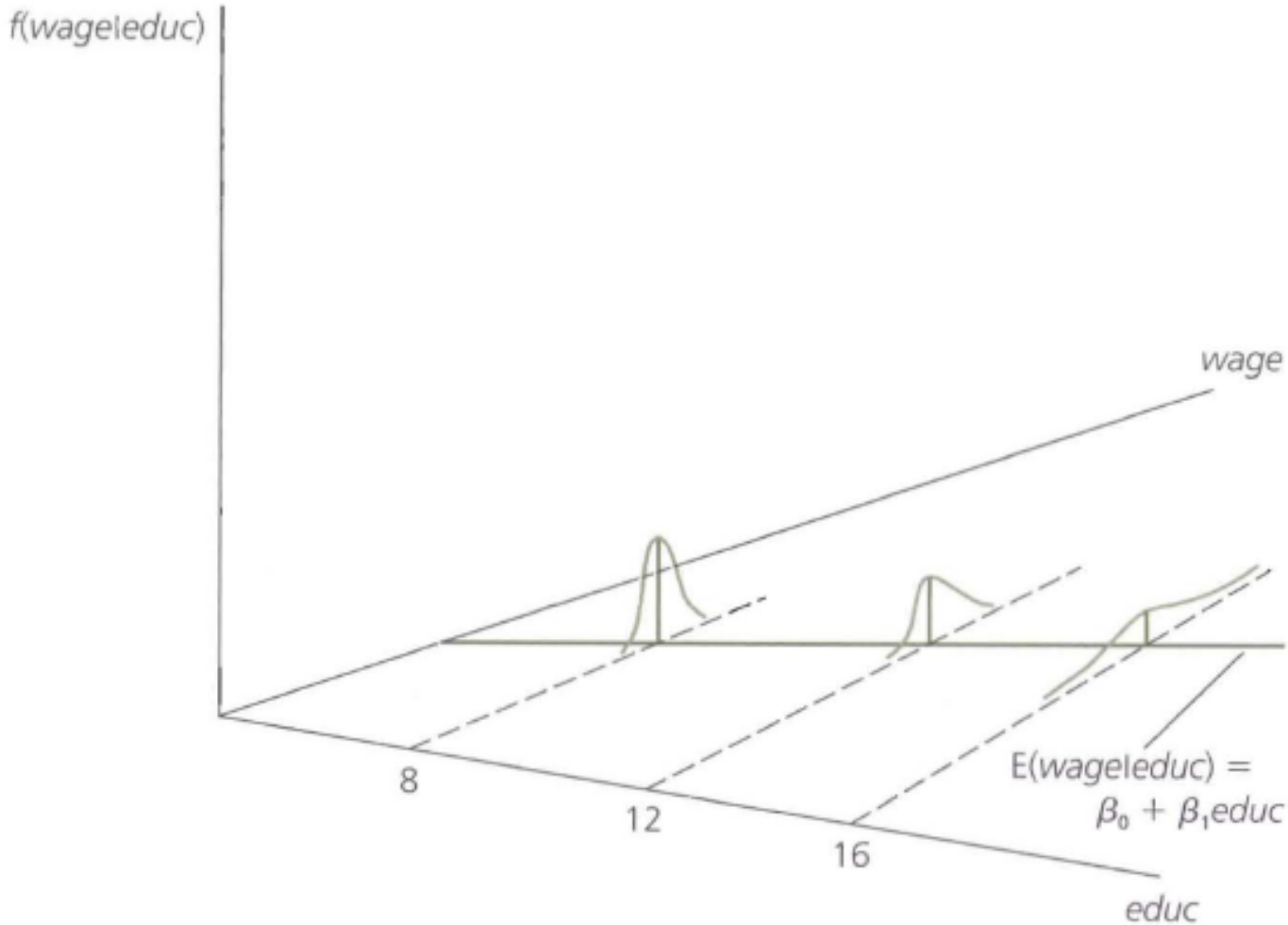
- This assumption holds that the error term has the same (conditional) variance given any value of the explanatory variable, that is, constant variance: $V[v|X] = \sigma^2$

- This assumption is violated if, for instance, $V[v|X = 1] \neq V[v|X = 0]$
- This phenomenon is called heteroskedasticity

Homoskedasticity: $V[v|X] = \sigma^2$



Heteroskedasticity: $V[v|X] \neq \sigma^2$



Heteroskedasticity and Hypothesis Testing

We would like to have: $t_{n-2} = \frac{\hat{\beta}_1}{SE[\hat{\beta}_1]}$

Yet we get: $t_{n-2} = \frac{\hat{\beta}_1}{SE[\hat{\beta}_1|X]}$

- **Potential Solution:**
- We can use heteroskedasticity-robust standard errors (straightforward in `R`)
- This accounts for potential violations of the assumption
- If we fail to do this, measures that depend on estimated variance, such as standard errors could otherwise be biased
- If we fail to account for heteroskedasticity, our inferences might be wrong

Another Violation: Clustering

- For OLS, we assume all our observations are independent
- Is this always the case? Consider the following example: we ask people how often they watch TV; but the sample includes several families, with members watching TV together.
- Are these observations independent?
- Other examples: students in a classroom, municipalities in states

Potential Solution

Cluster your standard errors (i.e., calculate them in a way that accounts for nested nature of the data)

Remedies to Violations of Assumptions

- Linearity: take non-linear transformation of the independent variable (e.g., logs) and check if linearity is met. Often the case if you have exponential distributions (e.g., income)
- Random sampling: affects interpretation of the findings: you can only extrapolate to the kind of sample you draw from.
- No perfect collinearity: check for strong collinearity and remove collinear variables. Make sure to prevent accounting for closely related measures
- Zero conditional mean: violations often caused by omitted variable bias. Include ommitted variables
- Heteroskedasticity: `R` allows you to include "robust" standard errors that are unbiased even under heteroskedasticity; use clustered standard errors if data is clustered
- Normality of error term: Non-normal errors are usually the result of linearity assumption not holding. If you fix that, it's usually fine

Wrap Up

- OLS is robust to violations of most assumptions; there often is some way around them
- This is something that you probably remember from statistics class and it is important to have some understanding about this and keep this in mind - but don't worry too much about it

- Most of the time, you probably will have to think of the following:

1. Clustering (very often; you should always cluster if it makes sense)
2. Random sampling (often), but we knew about this already...
3. Linearity (it may make sense to include a model with raw data and another logged model if one of the variables has an exponential distribution)
4. Potentially also collinearity; but the model won't work if you have perfect collinearity

All In All: Recommendations for Applied Work

- Be aware of the assumptions as a matter of statistical literacy
- Cluster standard errors if your data structure suggests it
- Transform variables if you have non-normal distributions (e.g., exponentials)
- Be aware of your sampling procedures and implications (bias, scope conditions of findings)
- In practice, there often is a trade-off between biased and variance **- don't get lost in minor aspects**