Regression Workshop I

Baobao Zhang
bzhang22@syr.edu
Feb. 24, 2025

Download notes and R code using QR code
https://github.com/baobaofzhang/campbell-workshop

Roadmap for this review

① Ordinary least squares (OLS) regression
- univariate regression
- multiple regression
- robust standard errors
- how to output results
- interpreting regression results
  - causal vs. correlational

② Logistic regression
- why we use it for binary outcomes
- how it works
- interpreting results

Starting concepts:

- conditional expectation function: calculates the expected value of a random variable Y given a specific value of another random variable X
- We use regression to estimate the conditional expectation function
  - ↳ we are doing our best to approximate it

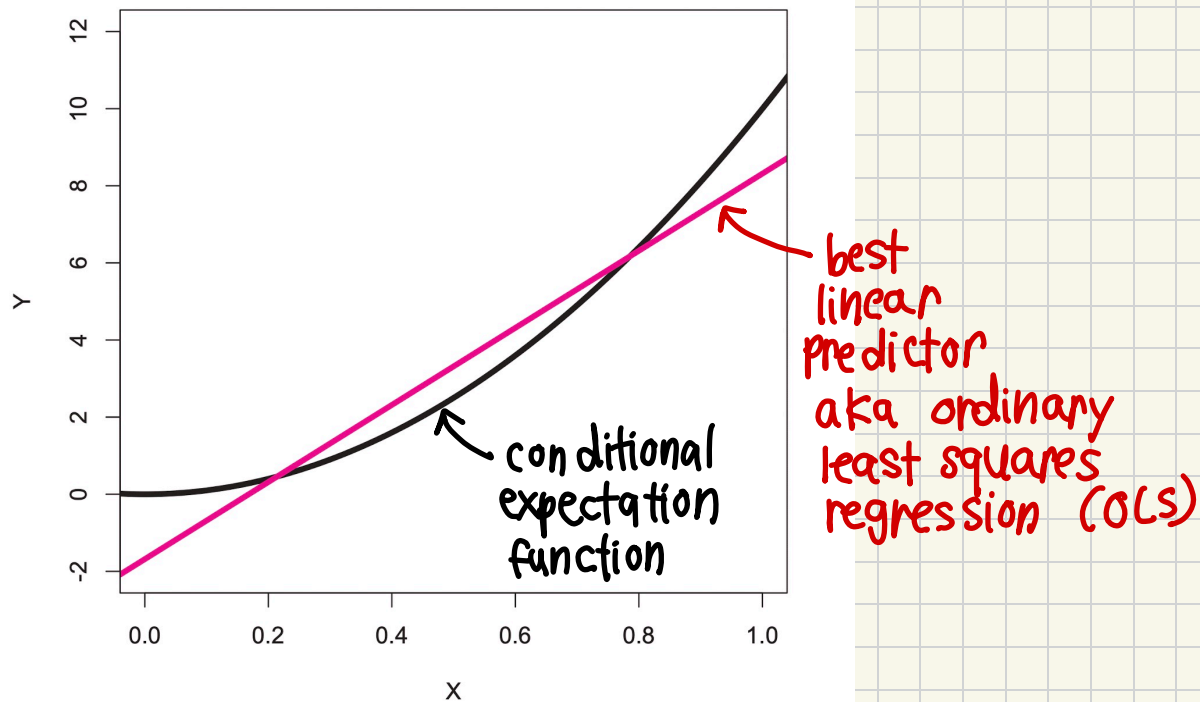


FIGURE 2.2.1. *Plotting the CEF and BLP*

Why do we use regression?

① To estimate the causal effect of some treatment variable D on outcome variable Y

② To examine the correlation between predictor variables X's on outcome variable Y

③ To predict the outcome Y given some values of our predictor variables X's

Univariate Ordinary Least Squares Regression

$$E(Y|X) = \beta_0 + \beta_1 X$$

This is the general form

outcome variable

predictor variable if X predicts Y
treatment variable if X causes Y

sometimes called explanatory variable or independent variable but this language is not precise

β_0 and β_1 are coefficients that we need to estimate

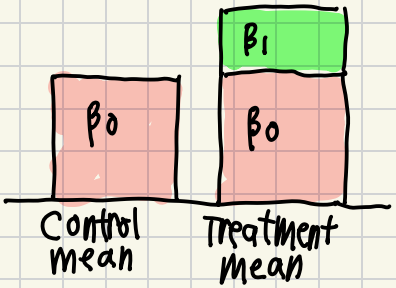
Example 1

After running a field experiment where you randomly assigned half of the university students to a "get-out-the vote" text message, you used OLS regression to estimate the effect of the treatment on whether students voted or not.

control mean

average treatment effect

$$E(Y|D) = \beta_0 + \beta_1 D$$



outcome variable

$Y_i = 0$ if not voted

$Y_i = 1$ if voted

treatment variable

$D_i = 0$ if in control

$D_i = 1$ if in treatment

Let's say the analysis of our experiment gave us the following result:

$$E(Y|D) = 0.6 + 0.05D$$

60% of students in the control group voted

The text message treatment increased voting by 5 percentage points

Example 2

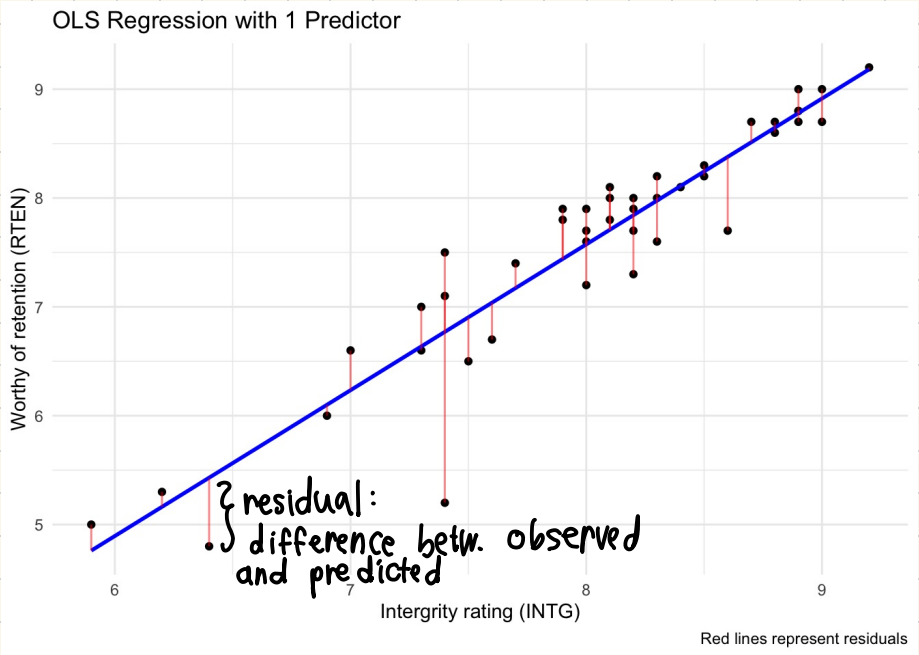
You use a survey dataset of lawyers' evaluation of judges to see if there is a correlation between their evaluation of the judges' intergrity and whether they think the judges are worthy of being retented.

$$E(Y|X) = \beta_0 + \beta_1 X$$

outcome variable: worthy of retention

predictor variable: intergrity

OLS regression minimizes the sum of the squares of the differences between the observed outcome variables (the black dots •) and the predicted come of the regression model (the blue line /)



Let's take a look using R (statistical programming language)

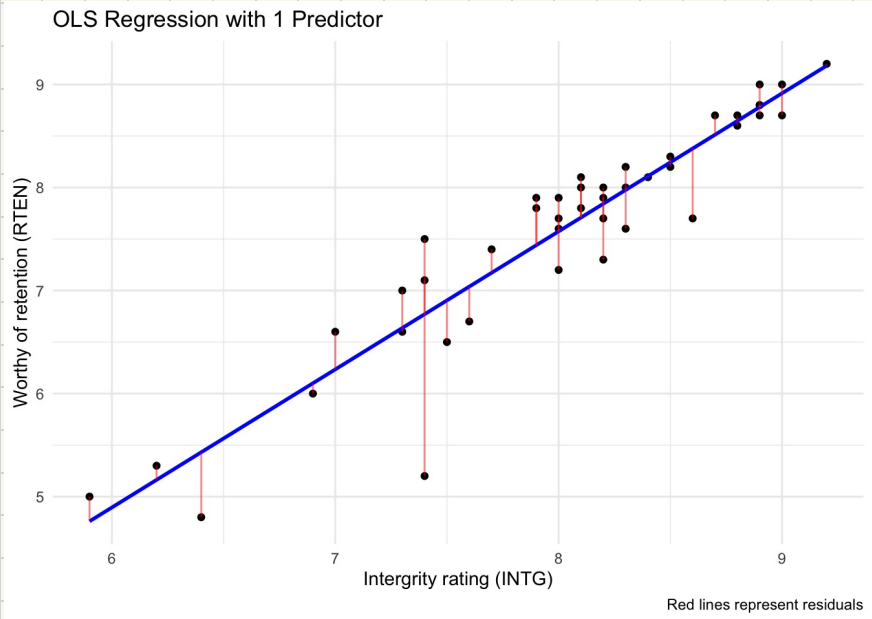
β_0: Value of Y when X = 0

Description: df [2 × 9]

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>	conf.low <dbl>	conf.high <dbl>	df <dbl>
(Intercept)	-3.144739	0.6367794	-4.938505	1.366694e-05	-4.430741	-1.858736	41
INTG	1.339878	0.0747349	17.928405	5.196614e-21	1.188947	1.490808	41

2 rows | 1-8 of 9 columns

β_1: slope of the blue line



Each 1 pt. increase in intergrity rating is correlated with 1.33 in the worthy of retention score

Multiple OLS regression

- 1 outcome variable
- more than 1 predictor or treatment variables
- not multivariate regression → that means regression with more than 1 outcome variable

The simplest example of multiple OLS regression:

$$E(\underbrace{Y}_{\text{outcome}} | X_1, X_2) = \beta_0 + \beta_1 \underbrace{X_1}_{\text{predictor variable 1}} + \beta_2 \underbrace{X_2}_{\text{predictor variable 2}}$$

Example 3

Using the dataset in Example 2, you try to predict judges' worthiness of retention score using

① intergrity score

② their perceived familiarity with the law

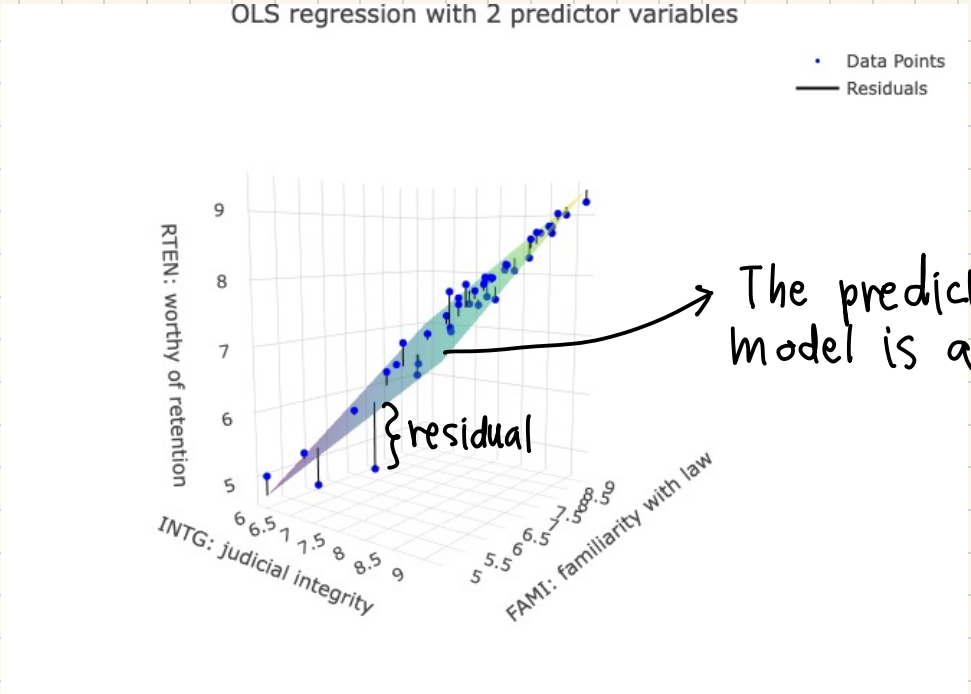
$$E(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Y → worthy of rentention

X_1 → intergrity

X_2 → familiarity with the law

Same idea as before: OLS regression will create a plane (instead of a line) that minimizes the sum of the residuals squared



The predicted model is a plane

Let's take a look in R

Value of Y when INTG and FAMI are set to 0

Description: df [3 × 9]

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>	conf.low <dbl>	conf.high <dbl>	df <dbl>
(Intercept)	-2.4824922	0.4934926	-5.030455	1.074436e-05	-3.4798779	-1.4851066	40
INTG	0.6947431	0.1220622	5.691711	1.281708e-06	0.4480461	0.9414401	40
FAMI	0.6025785	0.1237743	4.868364	1.799417e-05	0.3524212	0.8527358	40

3 rows

Now you try...

Holding FAMI constant,
each 1 point increase in INTG
is correlated with 1 point increase in Y

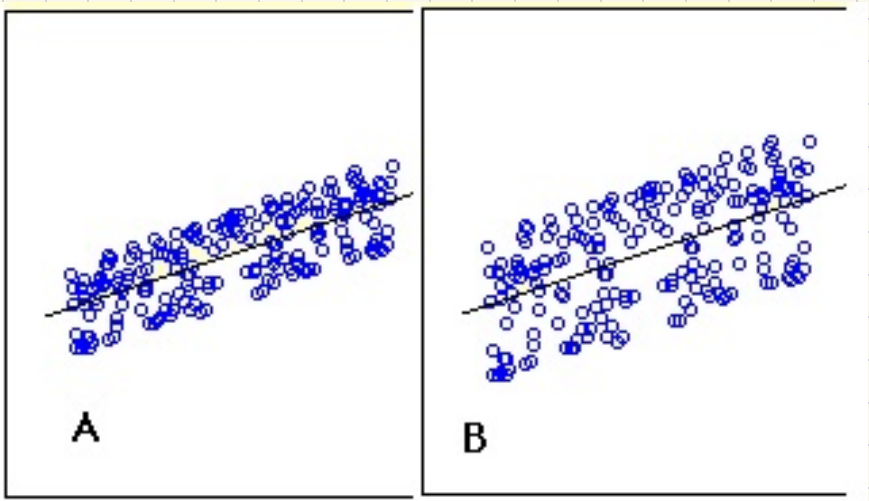
Harder to visualize the OLS regression model in higher dimensions but the same principle applies...

Standard error of a coefficient

- measures how precisely the model estimates the coefficient's value
- always positive
- tends to decrease when you have more units of observations in your dataset and less noisy outcome variables

In the univariate regression case:

$$SE(B_1) = \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$



Which of the datasets would produce the smaller standard error?

Why we need to use robust Standard errors...

Dealing with heteroskedasticity

For single-variate linear regression, we have

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

coefficients are parameters we want to estimate

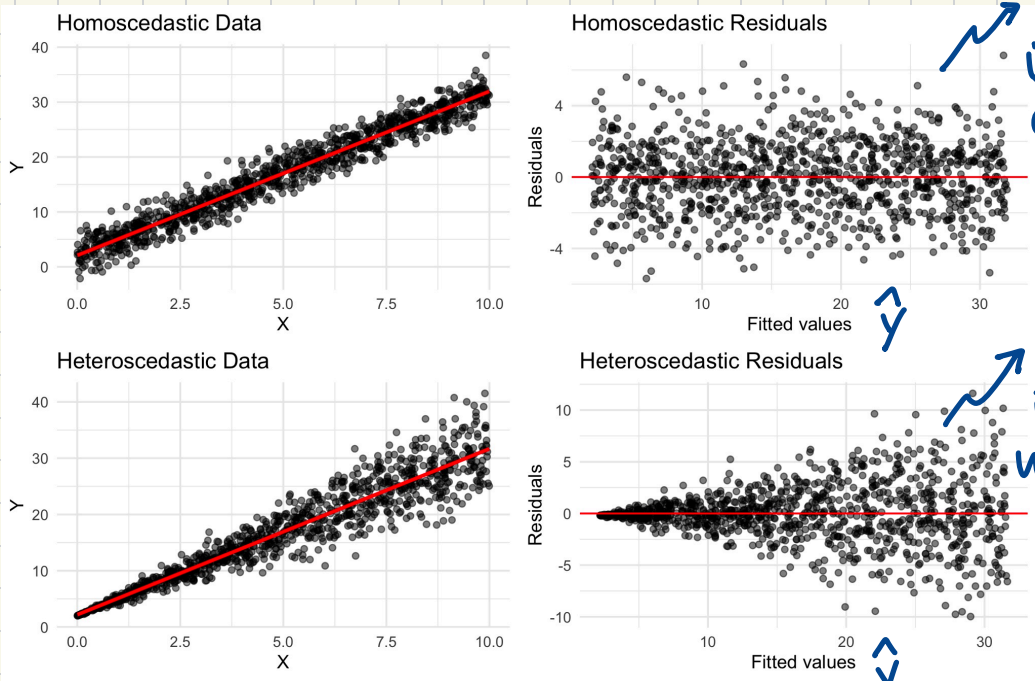
We assume ϵ (the error term) is normally distributed with mean 0 and variance σ^2

$$\epsilon \sim N(0, \sigma^2)$$

This means we assume a constant variance σ^2 but the assumption is likely to be false in many real-world settings!

Homoscedastic = having the same variance

Heteroscedastic = <u>not</u> having the same variance



The variance is constant across the fitted values

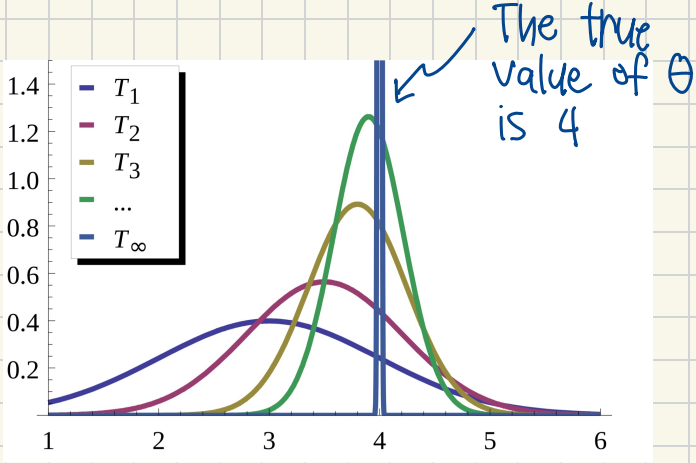
The variance increases with the fitted values

Why is heteroscedasticity a problem?

- <u>Not</u> a problem with regards to our OLS estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots$, etc.
- But a problem with our standard error estimates
  ↳ when the residuals are heteroscedastic, our OLS standard error estimates are inconsistent

What is consistency?

an estimator $\hat{\theta}$ is consistent when it converges to the true value of θ as the number of data points increases



What it could mean for your regression analysis

↳ If your data has heteroskedatic residuals, using normal standard errors will often <u>underestimate</u> the true variability of your coefficient estimate

$\Downarrow$

Falsely rejecting the null hypothesis when you do a t-test or Z-test

Comparing OLS standard errors and "robust" standard errors

OLS standard errors

$$\mathrm{Var}(\beta) = \underbrace{(X^T X)^{-1}}_{\text{bread}} \underbrace{X^T \theta X}_{\text{meat}} \underbrace{(X^T X)^{-1}}_{\text{bread}}$$

X $\qquad\qquad$ X^T

2	4	-1
-10	5	11
18	-7	6

$\Rightarrow$

2	-10	18
4	5	-7
-1	11	6

X is the matrix of predictor variables

X^T is the transpose of X

$\theta = \sigma^2 I_i$

where

$$I_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$n \times n$ matrix with 1 on the diagonals

and

$$\theta = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

σ^2 is the variance of the the residuals

Robust standard errors are robust to heteroscedasticity

↳ There are various types

HC1 Robust SEs

$$Var(\beta) = (X^T X)^{-1} X^T \Theta_{HC1} X (X^T X)^{-1}$$

where for each i

$$HC1 = \frac{n}{n-K} e_i^2$$

n → number of observations

e_i^2 → the squared residual

K → number of coefficients

So, this is what Θ_{HC1} looks like:

$$\Theta_{HC1} = \begin{bmatrix} \frac{n}{n-K} e_1^2 & 0 & \cdots & 0 \\ 0 & \frac{n}{n-K} e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{n}{n-K} e_n^2 \end{bmatrix}$$

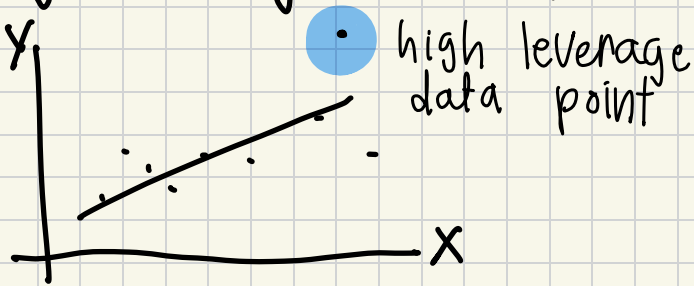
The default robust standard errors is HC3

$$HC3 = \frac{e_i^2}{(1-h_i)^2}$$

the squared residual

the leverage value
a number btw. 0 and 1
higher value means the data point has more leverage (e.g., outliers)

We prefer HC3 because it downweights the influence of high-leverage data points (e.g., outliers)



USE ROBUST SE IN ALMOST ALL CASES PLEASE

How do we use standard errors in hypothesis testing?

We typically assume the following when hypothesis testing:

H_0 (null hypothesis): our coefficient estimate is equal to 0

H_1 (alternative hypothesis): our coefficient estimate is not equal to 0

p-value: what's the probability that we get an observed result as extreme as the one we observe (from our data) assuming the null hypothesis

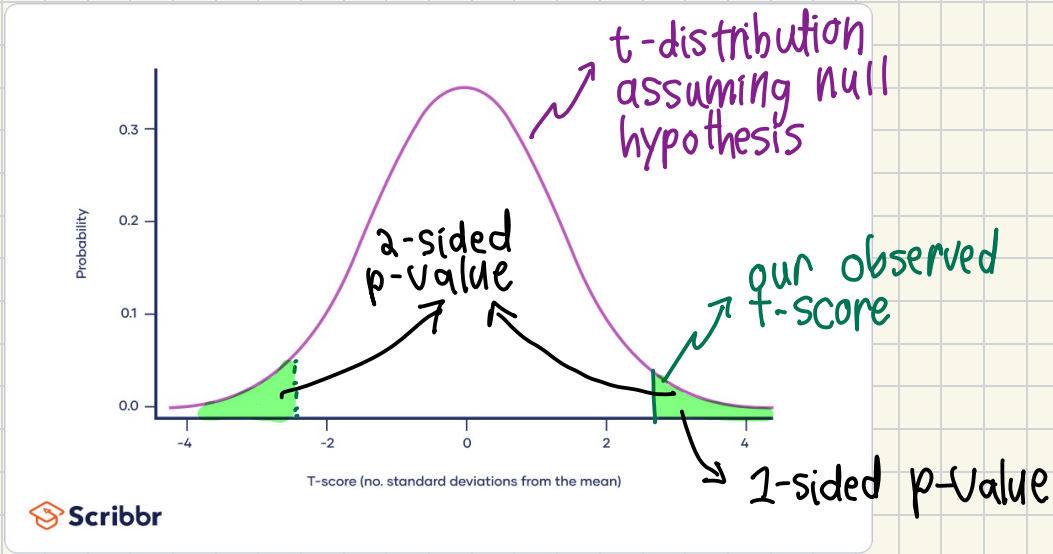
One-sided test

$H_{one\text{-}sided}$: our coefficient $\beta > 0$

$H_{two\text{-}sided}$: our coefficient $\beta \neq 0$

we generate t-score to do our hypothesis test

$$t_{\hat{\beta}} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$



Confidence intervals

interval that captures the true value of the coefficient with a set probability

95% confidence interval for $\hat{\beta}$ means that with 95% probability, the interval will contain the true value of β

How to estimate 95% confidence interval when you have large amounts of data ($n > 50$):

Lower: $\hat{\beta} - 1.96 \cdot SE(\hat{\beta})$

Upper: $\hat{\beta} + 1.96 \cdot SE(\hat{\beta})$

Showing regression results: table

BAN P, HILL SJ. Efficacy of Congressional Oversight. *American Political Science Review*. Published online 2025:1-19. doi:10.1017/S0003055424001242

TABLE 2. Dynamic Effect of Hearings on Improper Payment Rate

different model specifications

	Agency-level			Program-level		
	1	2	3	4	5	6
Years since last agency hearing	−0.098***	−0.10***		−0.17***	−0.17***	
	(0.035)	(0.038)		(0.054)	(0.060)	
Years since last unified govt. hearing		0.0048			0.18	
		(0.072)			(0.14)	
Years since last House agency hearing			−0.077**			−0.051
			(0.032)			(0.042)
Years since last Senate agency hearing			−0.091			−0.26**
			(0.064)			(0.12)
Agency hearing	0.59	0.65		1.14	0.25	
	(0.49)	(0.56)		(1.03)	(1.08)	
Agency hearing × unified government		−0.16			3.40	
		(0.79)			(2.34)	
House agency hearing			0.95			1.10
			(0.53)			(1.10)
Senate agency hearing			−0.63			0.40
			(0.60)			(0.70)
Fiscal year FE	Yes	Yes	Yes	Yes	Yes	Yes
Agency/program FE	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	304	304	304	1,289	1,289	1,289

$\hat{\beta}$

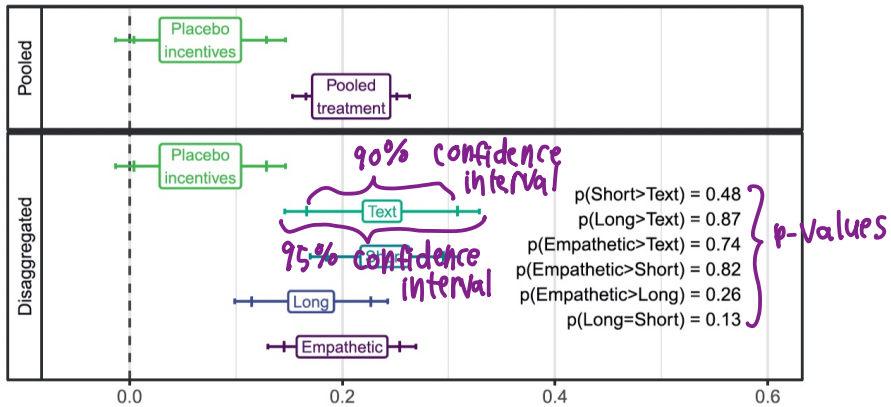
$SE(\hat{\beta})$

stars for p-values

Note: Dependent variable is improper payment rate as a percent of all payments. Standard errors in parentheses; clustered on agency-year for program-level analysis. $^{**}p < 0.05$, $^{***}p < 0.01$.

Showing regression results: coefficient plots

BOWLES J, CROKE K, LARREGUY H, LIU S, MARSHALL J. Sustaining Exposure to Fact-Checks: Misinformation Discernment, Media Consumption, and Its Political Implications. *American Political Science Review*. Published online 2025:1-24. doi:10.1017/S0003055424001394



(c) Intended future take-up

Note: All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation 1. Top panel excludes *Text* from *Pooled treatment* since they were not sent podcasts; *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals. Tables F1–F3 in the Supplementary Material report regression results for each index and its components; Tables F14 and F15 in the Supplementary Material further include LASSO-selected covariates.

When is an coefficient estimate causal

① When the explanatory variable X is randomly assigned or as-if randomly assigned

Example: randomly assigning students to a get-out-the-vote text message

Example: winning a lottery to get $1 million

② controlling for variables that happened before X (pre-treatment), X is as-if randomly assigned

Example: most regression results from an observation study that tries to estimate the effect of X on Y

BOWLES J, CROKE K, LARREGUY H, LIU S, MARSHALL J. Sustaining Exposure to Fact-Checks: Misinformation Discernment, Media Consumption, and Its Political Implications. American Political Science Review. Published online 2025:1-24. doi:10.1017/S0003055424001394

Research Question: Does encouraging people in South Africa to consume fact-check information protect them against believing misinformation?

Estimation

We estimate intent-to-treat (ITT) effects of different combinations of treatment arms relative to our control group. Specifically, we estimate the following pre-specified OLS regressions:

$$Y_{ib} = \alpha_b + \beta Y_{ib}^{pre} + \boldsymbol{\gamma} \mathrm{X}_{ib}^{pre} + \boldsymbol{\tau} \mathrm{T}_{ib} + \varepsilon_{ib}, \tag{1}$$

where Y_{ib} is an outcome for respondent *i* from block *b*, T_{ib} is the vector of individual treatment assignments, α_b are randomization block fixed effects, Y_{ib}^{pre} is the baseline analog of the outcome (where feasible), and X_{ib}^{pre} is a vector of predetermined baseline covariates selected separately for each outcome variable via cross-validated LASSO. The vector $\boldsymbol{\tau}$ captures the ITT effect of each treatment condition. Reflecting the individual-level randomization, robust standard errors are used throughout.

BAN P, HILL SJ. Efficacy of Congressional Oversight. *American Political Science Review*. Published online 2025:1-19. doi:10.1017/S0003055424001242

Research Design and Identification

To estimate the effect of improper payment hearings on improper payment rates, we use a two-way fixed effects identification strategy. Because we are uncertain about how long it might take for agencies to enact reforms to improve their improper payment problems, we estimate a linear trend in improper payments in the years following a hearing. Our specification is

$$\begin{aligned}\{Improper\ payment\ rate\}_{it} &= \alpha_i + \delta_t + \beta Hearing_{it} \\ &+ \gamma\{Years\ since\ last\ hearing\}_{it} + \varepsilon_{it}\end{aligned} \quad (1)$$

where $\{Improper\ payment\ rate\}_{it}$ is the sample-estimated improper payment rate of agency i in fiscal year t, $Hearing_{it}$ takes the value 1 if a witness from agency i testified in at least one oversight hearing on improper payments in year t, $\{Years\ since\ last\ hearing\}_{it}$ counts the number of fiscal years since a witness from that agency was most recently called to testify (year of hearing equals zero), and α and δ are agency and fiscal year fixed effects. These fixed effects capture time-invariant features of agencies (policy domain, relative budget size, etc.) and factors that affect all agencies in each fiscal year (new statutes or executive orders, budget environment, economic conditions, focus of presidential administration, GAO involvement in hearings, etc.).

Research Question:

Does congressional hearings decrease improper payment rates to contractors and clients?

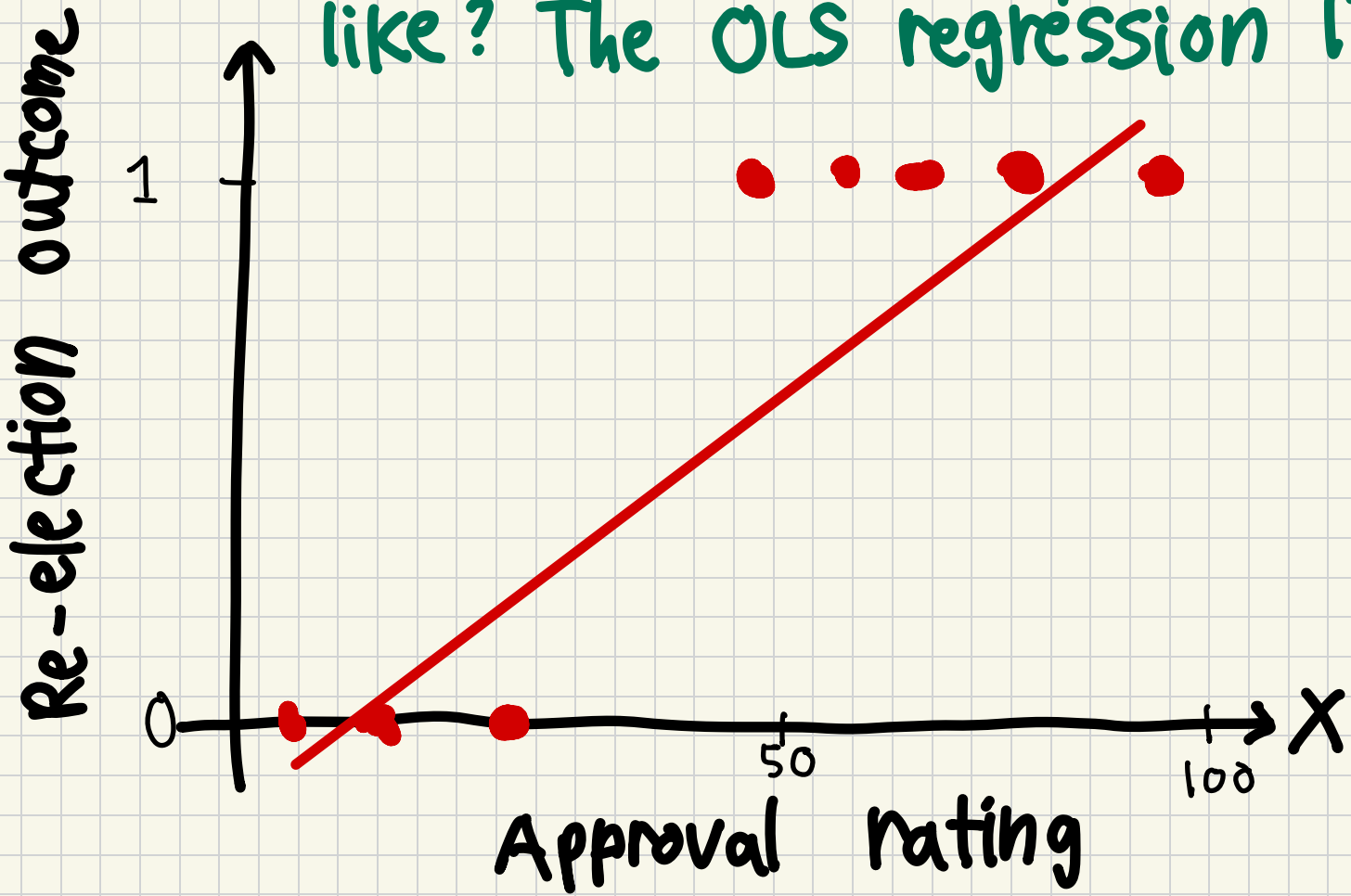
Logistic regression: why do we need it

Answer: to predict the probability of a binary outcome (i.e., got elected vs not) or to classify data into one of two categories

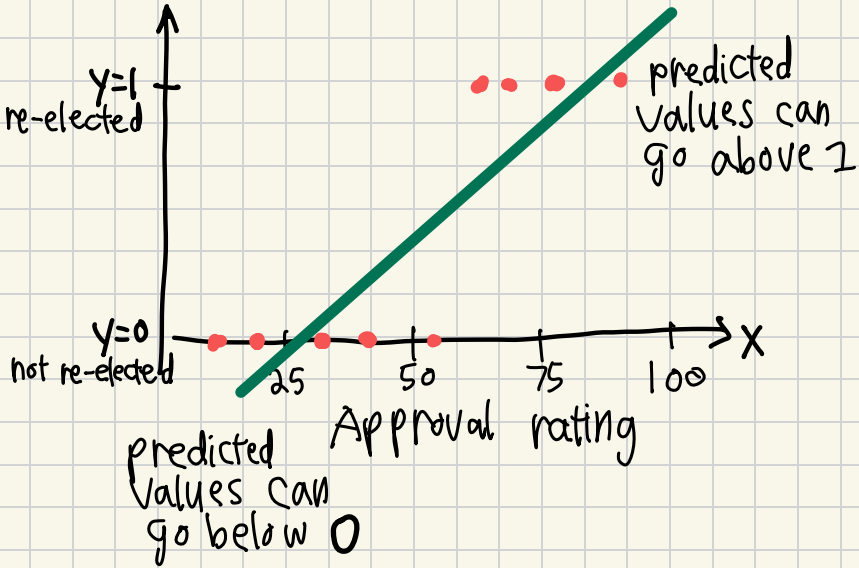
Secondary answer: to analyze the relationship between predictor variables and an outcome variable

Q: Suppose you want to predict the likelihood of a senator getting re-elected based on their approval rating (from 0 to 100). Why would using OLS regression cause a problem?

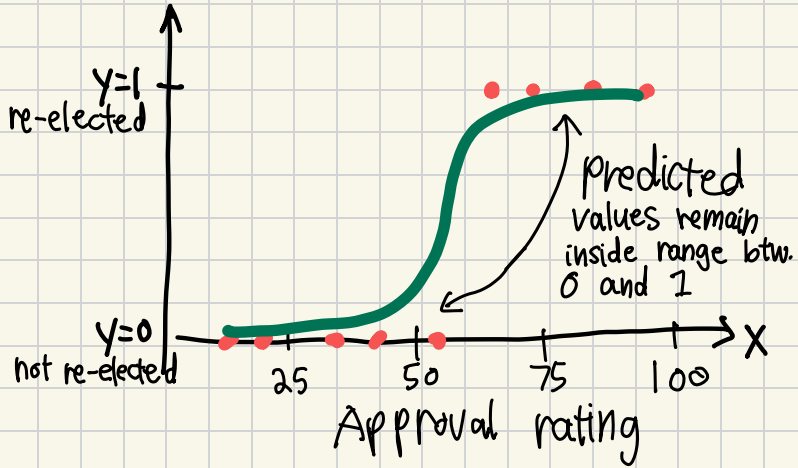
Q: What would the data points look like? The OLS regression line?



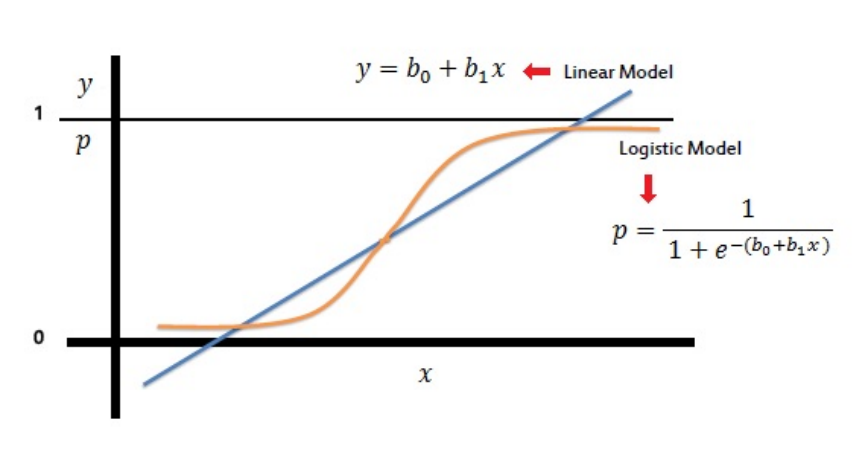
OLS Regression



Logistic Regression



We use the logistic function to fit the data so the predicted outcome is between 0 and 1



Interpting logistic regression

Suppose we use logistic regression to predict whether a member of congress will be re-elected using

- approval rating (0 to 100)
- unemployment rate in their district
- whether they are the same party as the president

We get the following output from R:

How do we interpret the coefficients?

```
z test of coefficients:

                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)     -1.1411819  0.7203479 -1.5842    0.1131
approval_rating  0.0570767  0.0094241  6.0564 1.392e-09 ***
unemployment    -0.4181664  0.1015419 -4.1182 3.819e-05 ***
pres_party       0.3413943  0.3404133  1.0029    0.3159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

X_1 approval_rating, X_2 unemployment, X_3 pres_party

We can write the regression model as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

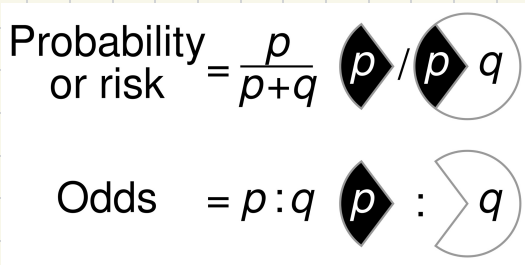
$$= -1.14 + 0.06 X_1 - 0.42 X_2 + 0.34 X_3$$

This is called the log odds

Side-note: what are log-odds?

What are odds?

$$\frac{\text{probability of success}}{\text{probability of failure}} = \frac{p}{1-p}$$



Example: What are the odds of the Eagles winning the Super Bowl if their probability of winning is $1/3$?

$p = 1/3 \quad q = 1-p = 2/3$

$$\frac{p}{1-p} = \frac{1/3}{2/3} = \frac{1}{2}$$

1 to 2 odds

Log odds means we take the natural log of the odds

Example: If the odds of the Eagles winning the Super Bowl is $\frac{1}{2}$, then the log odds is $\ln\left(\frac{1}{2}\right)$

```
z test of coefficients:

                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)     -1.1411819  0.7203479 -1.5842    0.1131
approval_rating  0.0570767  0.0094241  6.0564 1.392e-09 ***
unemployment    -0.4181664  0.1015419 -4.1182 3.819e-05 ***
pres_party       0.3413943  0.3404133  1.0029    0.3159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can write the regression model as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$= -1.14 + 0.06 X_1 - 0.42 X_2 + 0.34 X_3$$

This is called the log odds

β_0 (the intercept): the log odds when all the predictor variables $(X_1, X_2, \ldots X_k)$ are zero

$\beta_1, \beta_2, \ldots, \beta_k$: the amount the log odds change for each unit increase in the predictor variable

Example:

$\beta_1 = 0.06$

$\ln\left(\frac{\text{pr. of re-elected}}{\text{pr. of not re-elected}}\right)$ increases by 0.06 when the candidate's approval rating increases by 1 percentage pt

but we don't typically think in log odds

We are more comfortable thinking in odds

If $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$, we can $\exp\left[\ln\left(\frac{p}{1-p}\right)\right]$ to get $\frac{p}{1-p}$

Let's say that $\text{logit}(p) = \beta_0 + \beta_1 X$

$$\ln\left(\frac{1}{1-p}\right) = \beta_0 + \beta_1 X$$

$$\exp\left[\ln\left(\frac{1}{1-p}\right)\right] = \exp(\beta_0 + \beta_1 X)$$

We can take the exponents of our coefficients to get the odds ratio (typically written as OR)

Results reported as odds ratio (OR)

Term	Odds Ratio	Std.Error	z value	p value	CI Lower	CI Upper
(Intercept)	0.319	0.688	-1.659	0.097	0.081	1.217
approval_rating	1.059	0.010	5.933	0.000	1.040	1.080
unemployment	0.658	0.097	-4.294	0.000	0.539	0.791
pres_party	1.407	0.341	1.002	0.316	0.724	2.763

OR greater than 1: the event is more likely to occur as the predictor variable increases

<u>example</u>: the candidate is more likely to get re-elected as approval rating increases

OR = 1: no association between the predictor variable and the outcome variable

OR less than 1: the event is less likely to occur as the predictor variable increases

Term	Odds Ratio	Std.Error	z value	p value	CI Lower	CI Upper
(Intercept)	0.319	0.688	-1.659	0.097	0.081	1.217
approval_rating	1.059	0.010	5.933	0.000	1.040	1.080
unemployment	0.658	0.097	-4.294	0.000	0.539	0.791
pres_party	1.407	0.341	1.002	0.316	0.724	2.763

Q: Given the table above, is getting re-elected more likely to occur if the unemployment in their district increases?

Q: Is getting re-elected more likely to occur if the candidate is in the same party as the president?

Term	Odds Ratio	Std.Error	z value	p value	CI Lower	CI Upper
(Intercept)	0.319	0.688	-1.659	0.097	0.081	1.217
approval_rating	1.059	0.010	5.933	0.000	1.040	1.080
unemployment	0.658	0.097	-4.294	0.000	0.539	0.791
pres_party	1.407	0.341	1.002	0.316	0.724	2.763

- **Approval rating**: A candidate with an approval rating that's 1 percentage point higher have 1.059 times higher odds of getting re-elected (p-value < 0.001)
- **President's party**: A candidate who is in the same party as the president has 1.407 times higher odds of getting re-elected compared with some who is not in the same party as the president but this association is not statistically significant (p = 0.316)

keeping all the other variables constant :-)