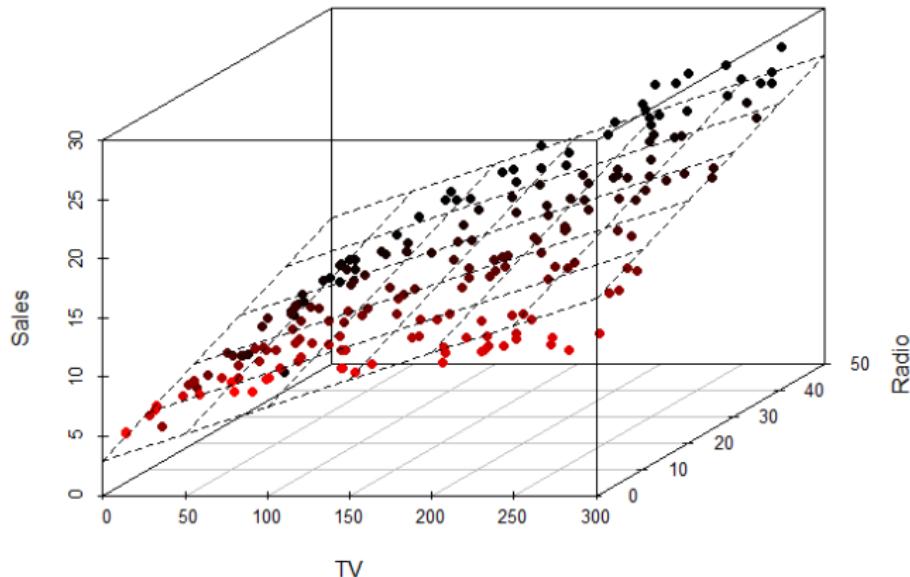


Linear Regression and Logistic Regression



Introduction to R Programming
District Data Labs

Running example: IMDB scores and film budgets

Do movies with higher budgets receive higher IMDB ratings?

Are other factors driving this relationship?

The `lm()` function

Linear regression notation

The “Best Fit” Line

Drawing a best-fit line through a scatterplot

How to find the best fit line?

Calculating the sum of squared errors/residuals (SSE/SSR)

Non-best-fit lines can be closer to many observations

Interpreting coefficients for continuous X variables

The very famous, classic sentence

Interpreting the constant

Some common mistakes

Interpreting coefficients for categorical X variables

Managing categorical variables

Recoding/labeling values

Combining categories

Reordering categories

Binary (Dummy) Variables

Unordered Categorical Variables

Ordered Categorical Variables

Making inferences

Standard errors

Confidence Intervals

t -statistics and p -values

Logistic (logit) regression

What these models do

Running logit models in R

Odds ratios

Expressing results as probabilities

The running example:

IMDB is the Internet Movie Data Base. It lists every movie, and every individual credited on a movie, and provides more information than anyone would ever need.

The running example:

IMDB is the Internet Movie Data Base. It lists every movie, and every individual credited on a movie, and provides more information than anyone would ever need.

Every movie has an IMDB score, the average rating of the movie as voted on by users of this website.

The running example:

IMDB is the Internet Movie Data Base. It lists every movie, and every individual credited on a movie, and provides more information than anyone would ever need.

Every movie has an IMDB score, the average rating of the movie as voted on by users of this website.

Question: do movies with higher budgets receive higher IMDB ratings?

The running example:

IMDB is the Internet Movie Data Base. It lists every movie, and every individual credited on a movie, and provides more information than anyone would ever need.

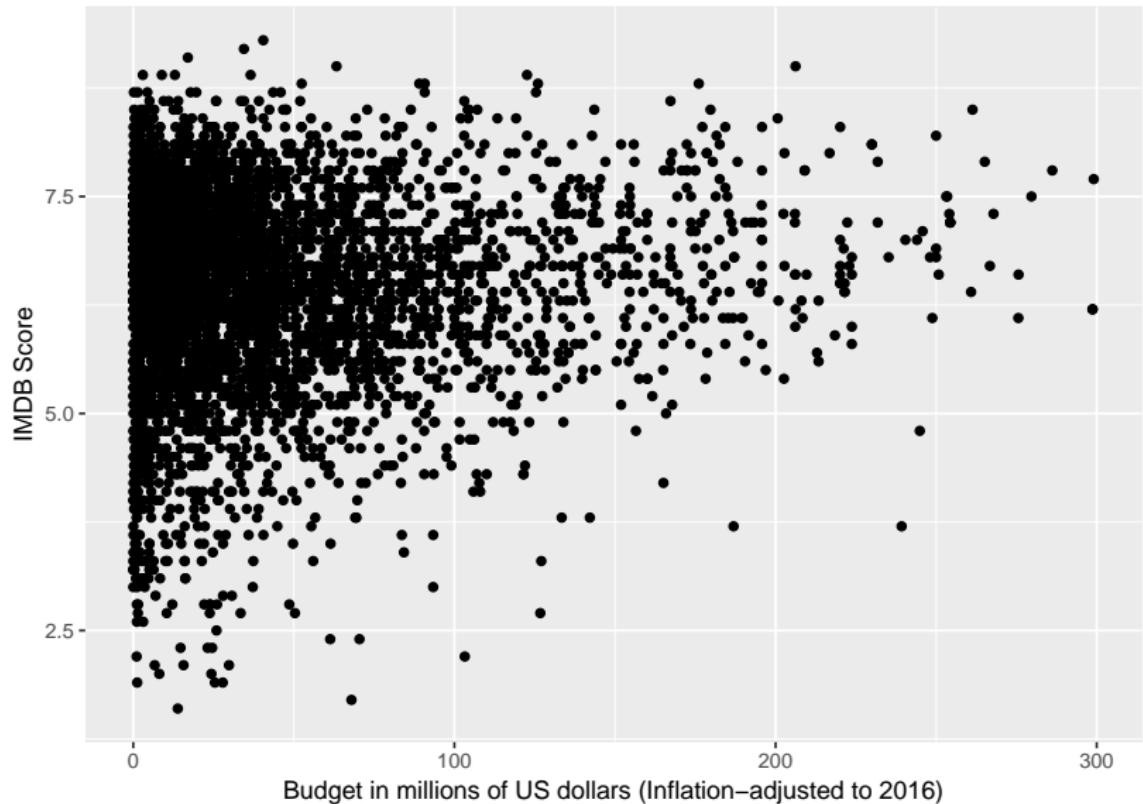
Every movie has an IMDB score, the average rating of the movie as voted on by users of this website.

Question: do movies with higher budgets receive higher IMDB ratings?

From the scatterplot and single regression alone, it appears that movies with higher budgets get slightly higher IMDB scores, on average.

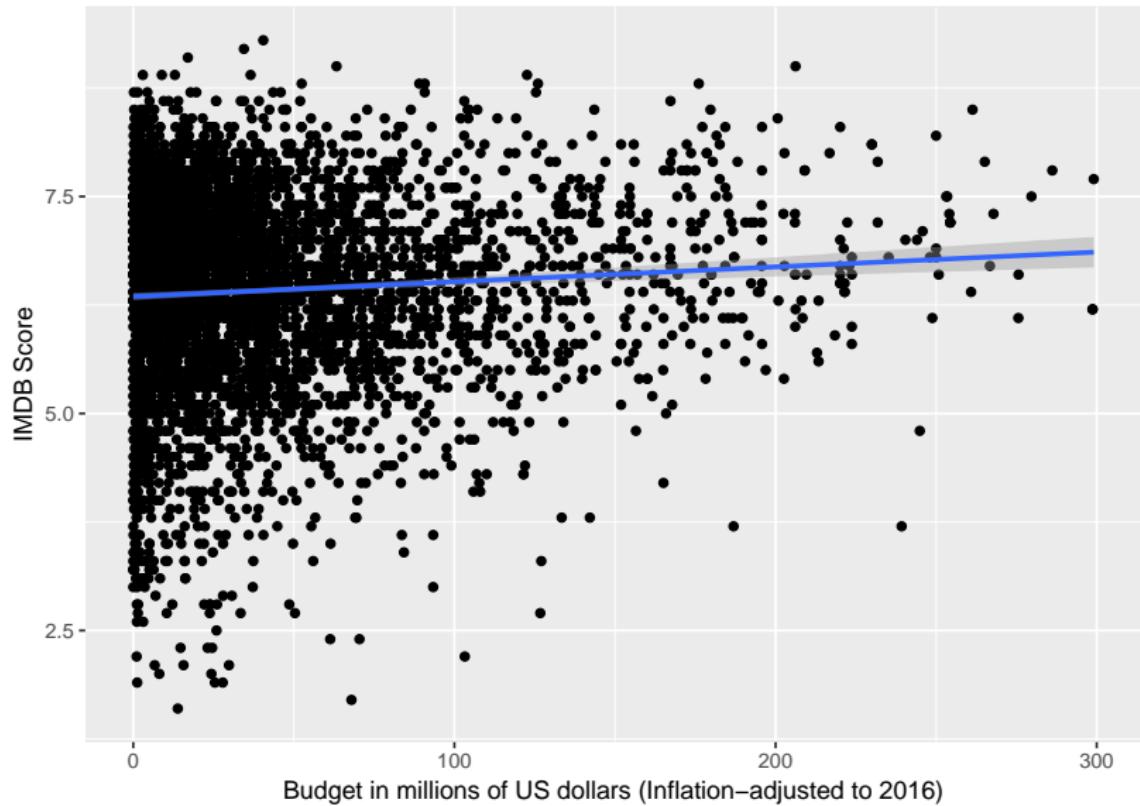
The running example:

Do movies with higher budgets receive higher IMDB ratings?



The running example:

Do movies with higher budgets receive higher IMDB ratings?



The summary table from the lm() command

```
> reg <- lm(imdb_score ~ budget, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7941	-0.6237	0.1122	0.7916	2.8832

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.3822859	0.0191259	333.698	< 2e-16 ***
budget	0.0008521	0.0002128	4.004	6.33e-05 ***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 1.112 on 4537 degrees of freedom
(517 observations deleted due to missingness)

Multiple R-squared: 0.003521, Adjusted R-squared: 0.003301
F-statistic: 16.03 on 1 and 4537 DF, p-value: 6.328e-05

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I observe: sales of sunscreen increase with sales of ice cream

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

I need to control for:

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

I **need to control for**: warm weather!

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

I **need to control for**: warm weather!

I **observe**: kids who enjoy **building clocks** go on to be scientists

and engineers

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

I **need to control for**: warm weather!

I **observe**: kids who enjoy **building clocks** go on to be scientists and engineers

I **claim**: we must make clock-building part of the curriculum!

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

I **need to control for**: warm weather!

I **observe**: kids who enjoy **building clocks** go on to be scientists and engineers

I **claim**: we must make clock-building part of the curriculum!

I **need to control for**:

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

I **need to control for**: warm weather!

I **observe**: kids who enjoy **building clocks** go on to be scientists and engineers

I **claim**: we must make clock-building part of the curriculum!

I **need to control for**: how much the students like science!

Are other factors driving this relationship?

If there's an **omitted variable** that may be related to BOTH variables in the regression, include it as a control.

I **observe**: sales of sunscreen increase with sales of ice cream

I **claim**: a chemical in sunscreen induces ice cream cravings

I **need to control for**: warm weather!

I **observe**: kids who enjoy **building clocks** go on to be scientists and engineers

I **claim**: we must make clock-building part of the curriculum!

I **need to control for**: how much the students like science!

Controls account for **alternative explanations**. If we include controls, and we still see a relationship between the variables, we are more certain the relationship is **real**.

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. Year:

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation;

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**:

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive;

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.
3. **Popularity of the cast**:

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.
3. **Popularity of the cast**: famous actors are more expensive;

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.
3. **Popularity of the cast**: famous actors are more expensive; people like famous actors!

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.
3. **Popularity of the cast**: famous actors are more expensive; people like famous actors!
4. **Marketing the cast**:

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.
3. **Popularity of the cast**: famous actors are more expensive; people like famous actors!
4. **Marketing the cast**: big marketing campaigns featuring the cast are expensive;

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.
3. **Popularity of the cast**: famous actors are more expensive; people like famous actors!
4. **Marketing the cast**: big marketing campaigns featuring the cast are expensive; but the campaign may make people like the movie more.

The running example:

The concern: there might be another variable that is **related BOTH to the dependent variable (IMDB score) and the independent variable (budget)**. For example:

1. **Year**: later years have more expensive movies, even after adjusting for inflation; people rate recent movies more highly.
2. **Duration**: longer movies are more expensive; longer movies are usually better.
3. **Popularity of the cast**: famous actors are more expensive; people like famous actors!
4. **Marketing the cast**: big marketing campaigns featuring the cast are expensive; but the campaign may make people like the movie more.

We can account for these factors by including additional (control) variables in the regression.

The running example:

We therefore include these controls in the analysis:

The running example:

We therefore include these controls in the analysis:

1. year — the **number of years** since 1916 (the year of the oldest movie in the data)

The running example:

We therefore include these controls in the analysis:

1. year — the **number of years** since 1916 (the year of the oldest movie in the data)
2. duration — length of the movie in **minutes**

The running example:

We therefore include these controls in the analysis:

1. year — the **number of years** since 1916 (the year of the oldest movie in the data)
2. duration — length of the movie in **minutes**
3. cast_total_facebook_likes — **total number of facebook likes in 1000s** for every member of the cast

The running example:

We therefore include these controls in the analysis:

1. year — the **number of years** since 1916 (the year of the oldest movie in the data)
2. duration — length of the movie in **minutes**
3. cast_total_facebook_likes — **total number of facebook likes in 1000s** for every member of the cast
4. facenumber_in_poster — the **number of human faces** that appear on the official move poster. (A crude measure of how the marketing focuses on the cast.)

The running example:

We therefore include these controls in the analysis:

1. year — the **number of years** since 1916 (the year of the oldest movie in the data)
2. duration — length of the movie in **minutes**
3. cast_total_facebook_likes — **total number of facebook likes in 1000s** for every member of the cast
4. facenumber_in_poster — the **number of human faces** that appear on the official move poster. (A crude measure of how the marketing focuses on the cast.)

Our plan: we will go over the meaning, calculation, and interpretation of every number in the output from R's `lm()` function.

The lm() function

```
reg <- lm(imdb_score ~ budget + duration + year +  
           facenumber_in_poster +  
           cast_total_facebook_likes, data = imdb)  
summary(reg)
```

The lm() function

```
reg <- lm(imdb_score ~ budget + duration + year +  
           facenumber_in_poster +  
           cast_total_facebook_likes, data = imdb)  
summary(reg)
```

The lm() command stands for **linear model**.

The lm() function

```
reg <- lm(imdb_score ~ budget + duration + year +  
           facenumber_in_poster +  
           cast_total_facebook_likes, data = imdb)  
summary(reg)
```

The lm() command stands for **linear model**.

This command takes **two arguments**:

1. **The formula:** First type the name of the DV, then a tilde \sim , then the IVs separated by plus signs $+$.

(There are ways to put other cool things in the formula, but we'll get to that later).

The lm() function

```
reg <- lm(imdb_score ~ budget + duration + year +  
           facenumber_in_poster +  
           cast_total_facebook_likes, data = imdb)  
summary(reg)
```

The lm() command stands for **linear model**.

This command takes **two arguments**:

1. **The formula:** First type the name of the DV, then a tilde \sim , then the IVs separated by plus signs $+$.

(There are ways to put other cool things in the formula, but we'll get to that later).

2. **The data:** The name of the data frame that the variables are stored in.

The summary table from the R lm() command

```
> reg <- lm(imdb_score ~ budget + duration + year + facenumber_in_poster +
+           cast_total_facebook_likes, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget + duration + year + facenumber_in_poster +
    cast_total_facebook_likes, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7664	-0.5586	0.0990	0.6909	3.0156

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0297394	0.1395064	43.222	< 2e-16 ***
budget	-0.0005112	0.0002037	-2.509	0.0121 *
duration	0.0158795	0.0007005	22.667	< 2e-16 ***
year	-0.0152222	0.0012434	-12.242	< 2e-16 ***
facenumber_in_poster	-0.0401398	0.0075809	-5.295	1.25e-07 ***
cast_total_facebook_likes	0.0054011	0.0008239	6.556	6.16e-11 ***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 1.021 on 4517 degrees of freedom

(533 observations deleted due to missingness)

Multiple R-squared: 0.161, Adjusted R-squared: 0.1601

F-statistic: 173.4 on 5 and 4517 DF, p-value: < 2.2e-16

Linear regression notation

The simplest linear regression uses a model that looks like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Linear regression notation

The simplest linear regression uses a model that looks like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Know the notation:

Linear regression notation

The simplest linear regression uses a model that looks like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Know the notation:

y is the **dependent variable**, or the **outcome**. This is the variable that represents the phenomenon you are trying to **predict or explain**.

Linear regression notation

The simplest linear regression uses a model that looks like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Know the notation:

y is the **dependent variable**, or the **outcome**. This is the variable that represents the phenomenon you are trying to predict or explain.

x refers to the **independent variable**. Also called: regressor, predictor, explanatory variable, or just your X. This is the variable that you believe has an effect on the outcome.

Linear regression notation

The simplest linear regression uses a model that looks like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Know the notation:

y is the **dependent variable**, or the **outcome**. This is the variable that represents the phenomenon you are trying to predict or explain.

x refers to the **independent variable**. Also called: regressor, predictor, explanatory variable, or just your X . This is the variable that you believe has an effect on the outcome.

i refers to one generic observation in the data. The total number of observations is called a **sample size**, and is often denoted N .

Linear regression notation

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

α , β , and anything else that the regression estimates, is called a **parameter**. α is the **constant** or **intercept**. Sometimes it is denoted β_0 .

Linear regression notation

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

α , β , and anything else that the regression estimates, is called a **parameter**. α is the **constant** or **intercept**. Sometimes it is denoted β_0 .

β is a **coefficient**, but it is also called an effect, slope, or a “beta.” This is the **most important quantity in regression models**

Linear regression notation

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

α , β , and anything else that the regression estimates, is called a **parameter**. α is the **constant** or **intercept**. Sometimes it is denoted β_0 .

β is a **coefficient**, but it is also called an effect, slope, or a “beta.” This is the **most important quantity in regression models**

ε represents the **errors**, also called **residuals**. It shows the difference between the predicted and observed values of y .

Linear regression notation

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

α , β , and anything else that the regression estimates, is called a **parameter**. α is the **constant** or **intercept**. Sometimes it is denoted β_0 .

β is a **coefficient**, but it is also called an effect, slope, or a “beta.” This is the **most important quantity in regression models**

ε represents the **errors**, also called **residuals**. It shows the difference between the predicted and observed values of y .

Our goal: use data to find the **best estimates** for α and β . Once we have found these values, then $y = \alpha + \beta x$ is a **best fit line** for values in a scatterplot.

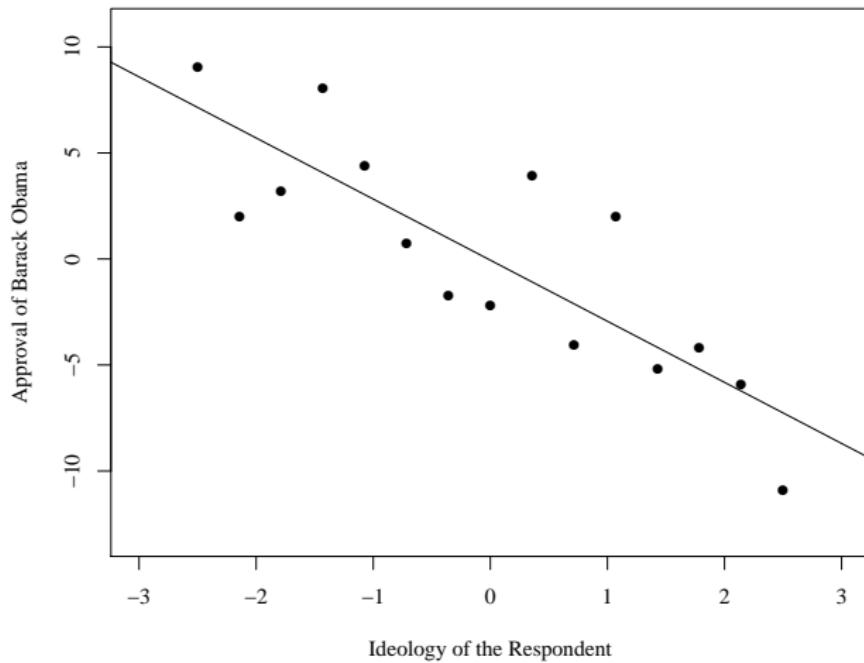
Drawing a line through a scatterplot

Table: Example Data

Obs. Number	Left-Right Ideology	Obama Approval	Obs. Number	Left-Right Ideology	Obama Approval
1	-2.50	9.04	9	0.36	2.94
2	-2.14	1.97	10	0.71	-4.06
3	-1.79	3.21	11	1.07	1.99
4	-1.43	8.07	12	1.43	-5.18
5	-1.07	4.38	13	1.79	-4.22
6	-0.71	0.74	14	2.14	-5.92
7	-0.36	-1.71	15	2.50	-10.89
8	0.00	-2.20			

Drawing a line through a scatterplot

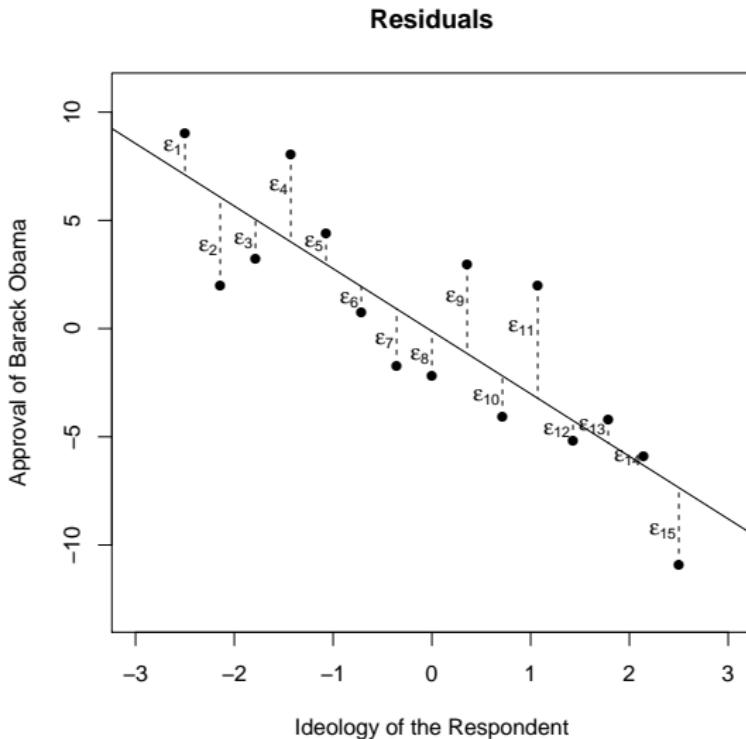
Best Fit Line



What makes this line a “best fit” anyway?

Drawing a line through a scatterplot

ε is the **vertical** distance between each point and the **best-fit line**.
The line is drawn to make these regression errors **as small as possible**.



How to find the best fit line?

It so happens that the best fit line must have a slope of -2.89 and an intercept of -0.12. The linear regression model is

$$(\text{Obama Approval})_i = -0.12 - 2.89(\text{Left-Right Ideology})_i + \varepsilon_i.$$

How to find the best fit line?

It so happens that the best fit line must have a slope of -2.89 and an intercept of -0.12. The linear regression model is

$$(\text{Obama Approval})_i = -0.12 - 2.89(\text{Left-Right Ideology})_i + \varepsilon_i.$$

How did we get these numbers?

How to find the best fit line?

It so happens that the best fit line must have a slope of -2.89 and an intercept of -0.12. The linear regression model is

$$(\text{Obama Approval})_i = -0.12 - 2.89(\text{Left-Right Ideology})_i + \varepsilon_i.$$

How did we get these numbers? These are the values of the slope and intercept make the **predicted values of y** as close as possible to the **observed values of y** .

How to find the best fit line?

It so happens that the best fit line must have a slope of -2.89 and an intercept of -0.12. The linear regression model is

$$(\text{Obama Approval})_i = -0.12 - 2.89(\text{Left-Right Ideology})_i + \varepsilon_i.$$

How did we get these numbers? These are the values of the slope and intercept make the **predicted values of y** as close as possible to the **observed values of y** .

Example: someone has a left-right ideology score of -1.43. What do we predict their level of Obama approval would be?

$$(\text{Obama Approval}) = -0.12 - 2.89(-1.43) + \varepsilon,$$

$$(\text{Obama Approval}) = 4.01 + \varepsilon.$$

How to find the best fit line?

It so happens that the best fit line must have a slope of -2.89 and an intercept of -0.12. The linear regression model is

$$(\text{Obama Approval})_i = -0.12 - 2.89(\text{Left-Right Ideology})_i + \varepsilon_i.$$

How did we get these numbers? These are the values of the slope and intercept make the **predicted values of y** as close as possible to the **observed values of y** .

Example: someone has a left-right ideology score of -1.43. What do we predict their level of Obama approval would be?

$$(\text{Obama Approval}) = -0.12 - 2.89(-1.43) + \varepsilon,$$

$$(\text{Obama Approval}) = 4.01 + \varepsilon.$$

In the data, we **observe** that the person with ideology at -1.43 approves of Obama at 8.07. So the **error/residual** for this observation is

$$\varepsilon = 8.07 - 4.01 = 4.06$$

How to find the best fit line?

Step 1: choose **candidate** values for α and β .

How to find the best fit line?

Step 1: choose **candidate** values for α and β .

Step 2: for every observation, use these values to **calculate the error**:

$$\varepsilon_i = y_i - (\alpha - \beta x_i)$$

How to find the best fit line?

Step 1: choose **candidate** values for α and β .

Step 2: for every observation, use these values to **calculate the error**:

$$\varepsilon_i = y_i - (\alpha - \beta x_i)$$

Step 3: calculate the **sum of squared errors/residuals** (SSE or SSR):

$$\sum_{i=1}^N \varepsilon_i^2$$

How to find the best fit line?

Step 1: choose **candidate** values for α and β .

Step 2: for every observation, use these values to **calculate the error**:

$$\varepsilon_i = y_i - (\alpha - \beta x_i)$$

Step 3: calculate the **sum of squared errors/residuals** (SSE or SSR):

$$\sum_{i=1}^N \varepsilon_i^2$$

If our choices for α and β represent the **best fit line**, then the SSR will be LOWER than the SSR we'd get with any other choice of α and β .

Calculating the SSR

Obs. Number	Left-Right Ideology	Obama Approval	Linear Prediction	Residual (ε)
1	-2.50	9.04	7.10	1.94
2	-2.14	1.97	6.07	-4.11
3	-1.79	3.21	5.04	-1.83
4	-1.43	8.07	4.01	4.06
5	-1.07	4.38	2.97	1.40
6	-0.71	0.74	1.94	-1.20
7	-0.36	-1.71	0.91	-2.62
8	0.00	-2.20	-0.12	-2.08
9	0.36	2.94	-1.15	4.09
10	0.71	-4.06	-2.19	-1.87
11	1.07	1.99	-3.22	5.21
12	1.43	-5.18	-4.25	-0.93
13	1.79	-4.22	-5.28	1.07
14	2.14	-5.92	-6.32	0.40
15	2.50	-10.89	-7.35	-3.54

Calculating the SSR

For the best fit line in the example, we calculate SSR as:

$$\begin{aligned} & (1.94)^2 + (-4.11)^2 + (-1.83)^2 + (4.06)^2 \\ & + (1.40)^2 + (-1.20)^2 + (-2.62)^2 + (-2.08)^2 \\ & + (4.09)^2 + (-1.87)^2 + (5.21)^2 + (-0.93)^2 \\ & + (1.07)^2 + (0.40)^2 + (-3.54)^2 = \textcolor{red}{117.15}. \end{aligned}$$

Calculating the SSR

For the best fit line in the example, we calculate SSR as:

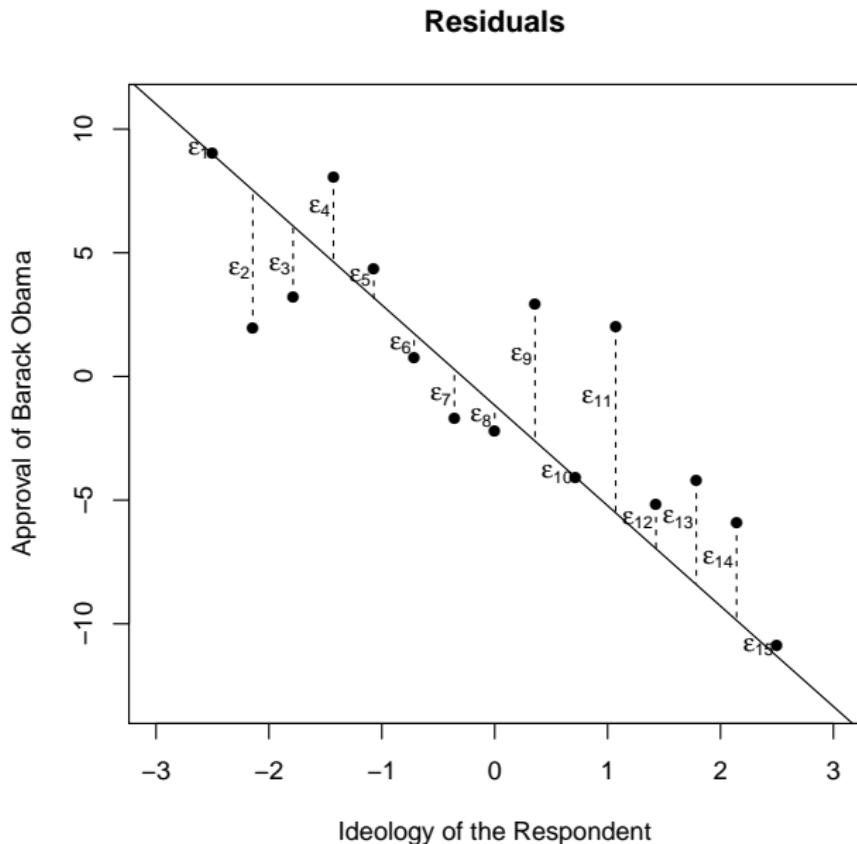
$$\begin{aligned} & (1.94)^2 + (-4.11)^2 + (-1.83)^2 + (4.06)^2 \\ & + (1.40)^2 + (-1.20)^2 + (-2.62)^2 + (-2.08)^2 \\ & + (4.09)^2 + (-1.87)^2 + (5.21)^2 + (-0.93)^2 \\ & + (1.07)^2 + (0.40)^2 + (-3.54)^2 = \textcolor{red}{117.15}. \end{aligned}$$

Now consider a second linear model that has a slope of -4 and an intercept of -1.

A line that's not the best fit

Obs. Number	Left-Right Ideology	Obama Approval	Linear Prediction	Residual (ε)
1	-2.50	9.04	8.98	0.06
2	-2.14	1.97	7.53	-5.56
3	-1.79	3.21	6.08	-2.87
4	-1.43	8.07	4.63	3.44
5	-1.07	4.38	3.18	1.19
6	-0.71	0.74	1.73	-0.99
7	-0.36	-1.71	0.28	-1.99
8	0.00	-2.20	-1.17	-1.03
9	0.36	2.94	-2.62	5.56
10	0.71	-4.06	-4.07	0.01
11	1.07	1.99	-5.52	7.51
12	1.43	-5.18	-6.96	1.79
13	1.79	-4.22	-8.41	4.20
14	2.14	-5.92	-9.86	3.95
15	2.50	-10.89	-11.31	0.42

A line that's not the best fit



A line that's not the best fit

Note: this second regression line is **better for several datapoints!** In particular, observations 1, 10, and 15.

A line that's not the best fit

Note: this second regression line is **better for several datapoints!** In particular, observations 1, 10, and 15.

But, other residuals are a lot bigger now, like observations 2, 9, and 11.

A line that's not the best fit

Note: this second regression line is **better for several datapoints!** In particular, observations 1, 10, and 15.

But, other residuals are a lot bigger now, like observations 2, 9, and 11.

If we calculate the new SSR, we find that the bad outweighs the good:

$$\begin{aligned} & (0.06)^2 + (-5.56)^2 + (-2.87)^2 + (3.44)^2 \\ & + (1.19)^2 + (-0.99)^2 + (-1.99)^2 + (-1.03)^2 \\ & + (5.56)^2 + (0.01)^2 + (7.51)^2 + (1.79)^2 \\ & + (4.20)^2 + (3.95)^2 + (0.42)^2 = \textcolor{red}{182.22}. \end{aligned}$$

Interpreting coefficients

```
> reg <- lm(imdb_score ~ budget + duration + year + facenumber_in_poster +
+           cast_total_facebook_likes, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget + duration + year + facenumber_in_poster +
    cast_total_facebook_likes, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7664	-0.5586	0.0990	0.6909	3.0156

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0297394	0.1395064	43.222	< 2e-16 ***
budget	-0.0005112	0.0002037	-2.509	0.0121 *
duration	0.0158795	0.0007005	22.667	< 2e-16 ***
year	-0.0152222	0.0012434	-12.242	< 2e-16 ***
facenumber_in_poster	-0.0401398	0.0075809	-5.295	1.25e-07 ***
cast_total_facebook_likes	0.0054011	0.0008239	6.556	6.16e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.021 on 4517 degrees of freedom

(533 observations deleted due to missingness)

Multiple R-squared: 0.161, Adjusted R-squared: 0.1601

F-statistic: 173.4 on 5 and 4517 DF, p-value: < 2.2e-16

Interpreting coefficients

There is a very famous, classic sentence that every regression student learns for interpreting β s:

Interpreting coefficients

There is a very famous, classic sentence that every regression student learns for interpreting β s:

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Interpreting coefficients

There is a very famous, classic sentence that every regression student learns for interpreting β s:

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

This sentence:

1. is useful,

Interpreting coefficients

There is a very famous, classic sentence that every regression student learns for interpreting β s:

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

This sentence:

1. is useful,
2. really, really sucks.

Interpreting coefficients

There is a very famous, classic sentence that every regression student learns for interpreting β s:

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

This sentence:

1. is useful,
2. really, really sucks.

You may find it useful to say this sentence to yourself when you first see results. But **never write this sentence** in a paper. It lacks style!

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

How to use this sentence: there are 5 parts to plug in

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

How to use this sentence: there are **5 parts** to plug in

1. the unit of X ,

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

How to use this sentence: there are **5 parts** to plug in

1. the unit of X ,
2. the name of X ,

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

How to use this sentence: there are **5 parts** to plug in

1. the unit of X ,
2. the name of X ,
3. the value of β in terms of units of Y ,

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

How to use this sentence: there are **5 parts** to plug in

1. the unit of X ,
2. the name of X ,
3. the value of β in terms of units of Y ,
4. the name of Y ,

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

How to use this sentence: there are **5 parts** to plug in

1. the unit of X ,
2. the name of X ,
3. the value of β in terms of units of Y ,
4. the name of Y ,
5. the names of the other X variables in the model.

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X :

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X :

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X : a movie's budget

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X : a movie's budget
3. the value of β in terms of units of Y :

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X : a movie's budget
3. the value of β in terms of units of Y : -0.0005 points

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X : a movie's budget
3. the value of β in terms of units of Y : -0.0005 points
4. the name of Y :

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X : a movie's budget
3. the value of β in terms of units of Y : -0.0005 points
4. the name of Y : the IMDB score

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X : a movie's budget
3. the value of β in terms of units of Y : -0.0005 points
4. the name of Y : the IMDB score
5. the names of the other X variables in the model:

Interpreting coefficients

Example: interpret the coefficient on budget.

1. the unit of X : \$1 million, inflation-adjusted to 2016
2. the name of X : a movie's budget
3. the value of β in terms of units of Y : -0.0005 points
4. the name of Y : the IMDB score
5. the names of the other X variables in the model: the release year, the duration of the movie, the number facebook likes for the cast, and the number of faces on the official poster.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Interpreting coefficients

- A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.
- A \$1 million (inflation-adjusted to 2016) increase

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget is associated with a 0.0005 point decrease

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget is associated with a 0.0005 point decrease in the movie's IMDB score, on average, after controlling for

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget is associated with a 0.0005 point decrease in the movie's IMDB score, on average, after controlling for the release year, the duration of the movie, the number facebook likes for the cast, and the number of faces on the official poster.

Interpreting coefficients

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget is associated with a 0.0005 point decrease in the movie's IMDB score, on average, after controlling for the release year, the duration of the movie, the number facebook likes for the cast, and the number of faces on the official poster.

Interpreting coefficients

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget is associated with a 0.0005 point decrease in the movie's IMDB score, on average, after controlling for the release year, the duration of the movie, the number facebook likes for the cast, and the number of faces on the official poster. **bleep bloop bloop death to all humans**

Interpreting coefficients

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget is associated with a 0.0005 point decrease in the movie's IMDB score, on average, after controlling for the release year, the duration of the movie, the number facebook likes for the cast, and the number of faces on the official poster. **bleep bloop bloop death to all humans**

Now that you understand what the numbers are telling you, can you [state the result with more style](#)? There's no one right answer. Here's one attempt:

Interpreting coefficients

A \$1 million (inflation-adjusted to 2016) increase in a movie's budget is associated with a 0.0005 point decrease in the movie's IMDB score, on average, after controlling for the release year, the duration of the movie, the number facebook likes for the cast, and the number of faces on the official poster. **bleep bloop bloop death to all humans**

Now that you understand what the numbers are telling you, can you [state the result with more style](#)? There's no one right answer. Here's one attempt:

After taking into account the year in which the movie was released, the length of the film, and the popularity and marketing of the cast, a increase in the film's budget of \$1 million (inflation-adjusted to 2016) does not appear to raise the movie's IMDB score. In fact, raising the budget is associated with a slight, .0005 point decrease in the IMDB score, on average.

Try interpreting the other coefficients

```
> reg <- lm(imdb_score ~ budget + duration + year + facenumber_in_poster +
+           cast_total_facebook_likes, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget + duration + year + facenumber_in_poster +
    cast_total_facebook_likes, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7664	-0.5586	0.0990	0.6909	3.0156

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0297394	0.1395064	43.222	< 2e-16 ***
budget	-0.0005112	0.0002037	-2.509	0.0121 *
duration	0.0158795	0.0007005	22.667	< 2e-16 ***
year	-0.0152222	0.0012434	-12.242	< 2e-16 ***
facenumber_in_poster	-0.0401398	0.0075809	-5.295	1.25e-07 ***
cast_total_facebook_likes	0.0054011	0.0008239	6.556	6.16e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.021 on 4517 degrees of freedom

(533 observations deleted due to missingness)

Multiple R-squared: 0.161, Adjusted R-squared: 0.1601

F-statistic: 173.4 on 5 and 4517 DF, p-value: < 2.2e-16

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Duration:

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Duration:

A 1 minute increase

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Duration:

A 1 minute increase in the movie's length

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Duration:

A 1 minute increase in the movie's length is associated with a 0.016 point increase

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Duration:

A 1 minute increase in the movie's length is associated with a 0.016 point increase in the movie's IMDB score, on average, after controlling for

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Duration:

A 1 minute increase in the movie's length is associated with a 0.016 point increase in the movie's IMDB score, on average, after controlling for the movie's budget, the release year, the number facebook likes for the cast, and the number of faces on the official poster.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Duration: (if one minute \rightarrow 0.016 point increase, then 30 minutes \rightarrow 0.48 increase)

Consider two movies with the same budget, release year, cast, and marketing. If one movie is 30 minutes longer than the other, it will have an IMDB score that is 0.48 higher, on average.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Year:

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Year:

A 1 year increase

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Year:

A 1 year increase in the movie's release year

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Year:

A 1 year increase in the movie's release year is associated with a 0.015 point decrease

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Year:

A 1 year increase in the movie's release year is associated with a 0.015 point decrease in the movie's IMDB score, on average, after controlling for

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Year:

A 1 year increase in the movie's release year is associated with a 0.015 point decrease in the movie's IMDB score, on average, after controlling for the movie's budget and duration, the number facebook likes for the cast, and the number of faces on the official poster.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Year:

Holding constant a movie's budget, cast, duration, and marketing, if the film is released one year later, its IMDB score will be 0.015 points lower, on average.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Number of faces on the poster:

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Number of faces on the poster:

An additional face on the movie poster

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Number of faces on the poster:

An additional face on the movie poster is associated with a 0.04 point decrease

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Number of faces on the poster:

An additional face on the movie poster is associated with a 0.04 point decrease in the movie's IMDB score, on average, after controlling for

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Number of faces on the poster:

An additional face on the movie poster is associated with a 0.04 point decrease in the movie's IMDB score, on average, after controlling for the movie's budget, duration, and release year, and the number of facebook likes for the cast.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Number of faces on the poster:

If a movie's budget, release year, duration, and casting is held constant, placing an additional face on the movie poster results in an average decrease in the IMDB score of .04 points.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Facebook likes for the cast (in thousands):

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Facebook likes for the cast (in thousands):

1000 additional facebook likes for the cast

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Facebook likes for the cast (in thousands):

1000 additional facebook likes for the cast is associated with a 0.005 point increase

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Facebook likes for the cast (in thousands):

1000 additional facebook likes for the cast is associated with a 0.005 point increase in the movie's IMDB score, on average, after controlling for

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Facebook likes for the cast (in thousands):

1000 additional facebook likes for the cast is associated with a 0.005 point increase in the movie's IMDB score, on average, after controlling for the movie's budget, duration, and release year, and the number of faces on the official poster.

Interpreting coefficients

A one-unit increase in X is associated with a β change in Y , on average, after controlling for the other X variables in the model.

Facebook likes for the cast (in thousands):

Given two films with the same marketing, release year, budget, and duration, if one film's cast has 1000 more facebook likes than the other, it will obtain an IMDB score that is on average .005 points higher.

Interpreting the constant

The constant is a strange thing. It has to be there to make the slopes more meaningful. But it has a mostly useless interpretation:

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

Interpreting the constant

The constant is a strange thing. It has to be there to make the slopes more meaningful. But it has a mostly useless interpretation:

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

In order to use this sentence, plug in:

Interpreting the constant

The constant is a strange thing. It has to be there to make the slopes more meaningful. But it has a mostly useless interpretation:

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

In order to use this sentence, plug in:

- ▶ the names of the X variables,

Interpreting the constant

The constant is a strange thing. It has to be there to make the slopes more meaningful. But it has a mostly useless interpretation:

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

In order to use this sentence, plug in:

- ▶ the names of the X variables,
- ▶ the name of Y ,

Interpreting the constant

The constant is a strange thing. It has to be there to make the slopes more meaningful. But it has a mostly useless interpretation:

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

In order to use this sentence, plug in:

- ▶ the names of the X variables,
- ▶ the name of Y ,
- ▶ and the value of the constant α .

Interpreting the constant

The constant is a strange thing. It has to be there to make the slopes more meaningful. But it has a mostly useless interpretation:

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

In order to use this sentence, plug in:

- ▶ the names of the X variables,
- ▶ the name of Y ,
- ▶ and the value of the constant α .

There's **no good way** to phrase this with style if the situation where **all X s are simultaneously equal to zero is nonsensical**.

Interpreting the constant

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

Interpreting the constant

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

Example: $\alpha = 6.03$.

Interpreting the constant

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

Example: $\alpha = 6.03$.

When budget, duration, years since 1916, the number of facebook likes for the cast, and the number of faces on the poster are all simultaneously equal to 0, the movie's IMDB score is equal to 6.03, on average.

Interpreting the constant

When all of the X variables are simultaneously equal to 0, Y is equal to α , on average.

Example: $\alpha = 6.03$.

When budget, duration, years since 1916, the number of facebook likes for the cast, and the number of faces on the poster are all simultaneously equal to 0, the movie's IMDB score is equal to 6.03, on average.

A movie with no budget, lasting 0 minutes, released in 1916, with no facebook likes for the cast and no faces on the official poster will receive an average IMDB score of 6.03.

Some notes about interpreting coefficients

If a variable is included **only as a control**, don't bother interpreting the coefficient. Concentrate on the x variable of theoretical interest.

Some notes about interpreting coefficients

If a variable is included **only as a control**, don't bother interpreting the coefficient. Concentrate on the x variable of theoretical interest.

At this point, you should be **ignoring the p -values**. Whether or not these coefficients are “significant” does not change their interpretations.

Smallest p -value \neq Most influential variable

Some notes about interpreting coefficients

If a variable is included **only as a control**, don't bother interpreting the coefficient. Concentrate on the x variable of theoretical interest.

At this point, you should be **ignoring the p -values**. Whether or not these coefficients are “significant” does not change their interpretations.

Smallest p -value \neq Most influential variable

Instead of “change” in the template, you can say “**increase**” for positive coefficients and “**decrease**” for negative coefficients (writing β without the negative sign)

Some notes about interpreting coefficients

Changing the units of X will **change the coefficient too!**

Some notes about interpreting coefficients

Changing the units of X will **change the coefficient too!**

So suppose we code Facebook likes so that 1 means 10,000 likes instead of 1000 likes. Then instead of $\beta = 0.005$, the coefficient would be $\beta = 0.05$. **This is now the largest coefficient.**

Some notes about interpreting coefficients

Changing the units of X will **change the coefficient too!**

So suppose we code Facebook likes so that 1 means 10,000 likes instead of 1000 likes. Then instead of $\beta = 0.005$, the coefficient would be $\beta = 0.05$. **This is now the largest coefficient.**

That is, 10,000 additional Facebook likes for the cast is associated with a .05 increase in the IMDB score, on average, controlling for

...

Some notes about interpreting coefficients

Changing the units of X will **change the coefficient too!**

So suppose we code Facebook likes so that 1 means 10,000 likes instead of 1000 likes. Then instead of $\beta = 0.005$, the coefficient would be $\beta = 0.05$. **This is now the largest coefficient.**

That is, 10,000 additional Facebook likes for the cast is associated with a .05 increase in the IMDB score, on average, controlling for

...

As a result, coefficients are **not comparable** unless they have **exactly the same unit**. In general

Largest coefficient \neq Most influential variable

Recoding values

Recoding values means **replacing many values** of a categorical variable with new values, **simultaneously**.

There are a few reasons why you may want to recode:

1. **Change the labels** of values. Make it easier to see and remember what each value means. Better to code as "Male", "Female" than 1, 2.
2. **Replace missing codes** with NA.
3. **Change the order** of categories for display in graphs, or in case you want to treat the variable as ordinal and categories are out of order.

Labeling categorical values

Suppose you have a categorical variable with values 1, 2, 3, 4, **8** and **9**.

You look in the data's **codebook** (hopefully it has one) and see that

Category	Meaning
1	I speak Spanish primarily
2	I speak both Spanish and English equally
3	I speak English primarily but can speak Spanish
4	I can not speak Spanish
8	refused
9	skipped

We can replace the numeric values with their **written meanings**, without changing the way the categorical data is treated in R.

Using fct_recode()

There are many ways to replace numeric categories with their **written meanings**. The easiest method uses the `fct_recode()` function from the `forcats` package (one of the `tidyverse`).

Here's an example of how to use `fct_recode()` :

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),
                                              "I speak Spanish primarily" = "1",
                                              "I speak both Spanish and English equally" = "2",
                                              "I speak English primarily but can speak Spanish" = "3",
                                              "I can not speak Spanish" = "4",
                                              NULL = "8",
                                              NULL = "9"))
```

Let's break down the elements of this code:

Using fct_recode()

This function must be only ever used **inside** the `mutate()` function.

Type `mutate()`, then the data frame, then the name of the categorical variable you are editing, then an equal sign.

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),  
                                              "I speak Spanish primarily" = "1",  
                                              "I speak both Spanish and English equally" = "2",  
                                              "I speak English primarily but can speak Spanish" = "3",  
                                              "I can not speak Spanish" = "4",  
                                              NULL = "8",  
                                              NULL = "9"))
```

Parentheses will appear automatically and will indent correctly when you push enter. **Leave the closing parentheses alone.**

Using fct_recode()

Then type `fct_recode()`

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),  
                                              "I speak Spanish primarily" = "1",  
                                              "I speak both Spanish and English equally" = "2",  
                                              "I speak English primarily but can speak Spanish" = "3",  
                                              "I can not speak Spanish" = "4",  
                                              NULL = "8",  
                                              NULL = "9"))
```

The first argument of `fct_recode()` is the categorical variable, which must be of the **factor class**. If it is not (here it is numeric), use `as.factor()` to coerce the variable:

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),  
                                              "I speak Spanish primarily" = "1",  
                                              "I speak both Spanish and English equally" = "2",  
                                              "I speak English primarily but can speak Spanish" = "3",  
                                              "I can not speak Spanish" = "4",  
                                              NULL = "8",  
                                              NULL = "9"))
```

Using fct_recode()

The press enter, and on each new line write the **new categorical text label**, in quotes, equal to the **old categorical label**, also in quotes.

Remember, as with the `rename()` function: **new first, then old**.

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),
  "I speak Spanish primarily" = "1",
  "I speak both Spanish and English equally" = "2",
  "I speak English primarily but can speak Spanish" = "3",
  "I can not speak Spanish" = "4",
  NULL = "8",
  NULL = "9"))
```

This code works whether the old labels are **numbers or text**.

Using fct_recode()

Finally, for the categories you want to set to be **missing**, write the new category labels as `NULL`, with **no quotes**:

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),
                                              "I speak Spanish primarily" = "1",
                                              "I speak both Spanish and English equally" = "2",
                                              "I speak English primarily but can speak Spanish" = "3",
                                              "I can not speak Spanish" = "4",
                                              NULL = "8",
                                              NULL = "9")
```

This code is more **space-consuming** than other approaches. But the advantage is that we can more easily keep track of the new and old categories, minimizing the risk of **confusing which label goes with which number**.

Using `fct_recode()`

One more nice thing that `fct_recode()` can do: [combine categories](#).

To combine two old categories into the same new category, just use the [same new category label](#) for multiple old categories.

For example, suppose we want to group every category in which a person knows at least some Spanish as **Yes**, and the category in which a person knows no Spanish as **No**. We can write:

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),  
                                              "Yes" = "1",  
                                              "Yes" = "2",  
                                              "Yes" = "3",  
                                              "No" = "4",  
                                              NULL = "8",  
                                              NULL = "9"))
```

Reordering categories

The categories of a factor variable have a built-in **order**. The order controls a few things:

1. The order of the categories appear in any **table**
2. The order the categories appear left-to-right in any **graph**
3. The meaning of the variable when used in a **regression model**

Sometimes the categories have a *natural* ordering: for example, we can arrange categories in order of how much Spanish a person speaks.

Sometimes the categories **don't have a natural ordering**, but it makes sense to choose a particular order because it makes a **table or graph looks better**, or to change the base category in a regression.

Reordering categories

To change the order, use the `fct_relevel()` function. It works a lot like `fct_recode()`, only instead of writing old categories, simply write the existing categories in the order you want.

For example:

```
anes <- mutate(anes, speakspanish = fct_relevel(speakspanish,  
                                                 "I can not speak Spanish",  
                                                 "I speak English primarily but can speak Spanish",  
                                                 "I speak both Spanish and English equally",  
                                                 "I speak Spanish primarily"))
```

We'll talk more about **why it matters to change the order** in more detail over the next few weeks.

Reordering categories

Also, note that both the `fct_recode()` and `fct_relevel()` functions can be called within the **same** call to `mutate()`:

```
anes <- mutate(anes, speakspanish = fct_recode(as.factor(speakspanish),
                                              "I speak Spanish primarily" = "1",
                                              "I speak both Spanish and English equally" = "2",
                                              "I speak English primarily but can speak Spanish" = "3",
                                              "I can not speak Spanish" = "4",
                                              NULL = "8",
                                              NULL = "9"),
               speakspanish = fct_relevel(speakspanish,
                                            "I can not speak Spanish",
                                            "I speak English primarily but can speak Spanish",
                                            "I speak both Spanish and English equally",
                                            "I speak Spanish primarily"))
```

If you have **multiple categorical variables** to edit in this way, you can place all the calls to `fct_recode()` and `fct_relevel()` in the **same** `mutate()` command.

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i1} is gender: 0 for men, 1 for women

x_{i2} is region: Northeast, Midwest, Mountain, West Coast,
Southwest, Southeast.

x_{i3} is interest in politics, 7-point scale from “not at all” to
“extremely”

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i1} is gender: 0 for men, 1 for women

x_{i2} is region: Northeast, Midwest, Mountain, West Coast, Southwest, Southeast.

x_{i3} is interest in politics, 7-point scale from “not at all” to “extremely”

The following regression equation

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

is incorrect. We will correct it in a little bit.

Binary (Dummy) Variables

A binary variable is a variable that can take on **only 2 values**.

Binary (Dummy) Variables

A binary variable is a variable that can take on **only 2 values**.

Usually these values are **0 and 1**. In R, they might be TRUE and FALSE, but R converts these logical values to 0 and 1 for the math of the regression.

Binary (Dummy) Variables

A binary variable is a variable that can take on **only 2 values**.

Usually these values are **0 and 1**. In R, they might be TRUE and FALSE, but R converts these logical values to 0 and 1 for the math of the regression.

Sometimes these values are coded as 1 and 2. That **messes** with the constant (since the constant is the mean of y when all x variables are 0, including binary ones). But it **won't change any coefficients** to keep a binary variable coded this way.

Binary (Dummy) Variables

A binary variable is a variable that can take on **only 2 values**.

Usually these values are **0 and 1**. In R, they might be TRUE and FALSE, but R converts these logical values to 0 and 1 for the math of the regression.

Sometimes these values are coded as 1 and 2. That **messes** with the constant (since the constant is the mean of y when all x variables are 0, including binary ones). But it **won't change any coefficients** to keep a binary variable coded this way.

Recoding binary variables as 0/1 is the best practice, however.

Binary (Dummy) Variables

Binary variables can mean **several things**:

Binary (Dummy) Variables

Binary variables can mean **several things**:

- ▶ Is something **true**?

$0 = \text{No}$, $1 = \text{Yes}$

Binary (Dummy) Variables

Binary variables can mean **several things**:

- ▶ Is something **true**?

$0 = \text{No}$, $1 = \text{Yes}$

- ▶ There are **two groups**.

$0 = \text{one group}$, $1 = \text{the other group}$

Binary (Dummy) Variables

Binary variables can mean several things:

- ▶ Is something true?

$0 = \text{No}$, $1 = \text{Yes}$

- ▶ There are two groups.

$0 = \text{one group}$, $1 = \text{the other group}$

- ▶ There's a special subset of the observations that behave differently.

$0 = \text{not in the subset}$, $1 = \text{in the subset}$

Binary (Dummy) Variables

In a linear regression, binary variables require special treatment.

If a binary variable is the dependent variable, then a special kind of regression called **logistic regression** can be used. We'll cover that later today.

Binary (Dummy) Variables

In a linear regression, binary variables require special treatment.

If a binary variable is the dependent variable, then a special kind of regression called **logistic regression** can be used. We'll cover that later today.

If a binary variable is an independent variable, then pay special attention to what **unit** means in the coefficient sentence:

A one-unit increase in X	=	A comparison of the 1 group against the 0 group
----------------------------------	---	---

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i1} is gender: 0 for men, 1 for women

The (incorrect for now) regression equation

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

Interpret $\beta_1 = -30$.

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i1} is gender: 0 for men, 1 for women

The (incorrect for now) regression equation

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

Interpret $\beta_1 = -30$.

“Women, compared to men, watch 30 fewer minutes of TV news on average, after accounting for respondents' region and interest in politics.”

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i1} is gender: 0 for men, 1 for women

The (incorrect for now) regression equation

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

Interpret $\beta_1 = -30$.

“Women, compared to men, watch 30 fewer minutes of TV news on average, after accounting for respondents' region and interest in politics.”

Do not start talking about a “one-unit increase in gender.”

Unordered Categorical Variables

An **unordered categorial** (a.k.a. nominal) variable

- ▶ is non-continuous (a small number of discrete values),
- ▶ has **more than 2 categorical values**,
- ▶ but has no natural way to put these categories in an **order**.

Unordered Categorical Variables

An **unordered categorial** (a.k.a. nominal) variable

- ▶ is non-continuous (a small number of discrete values),
- ▶ has **more than 2 categorical values**,
- ▶ but has no natural way to put these categories in an **order**.

Some common unordered categorical variables in political science:

- ▶ race,
- ▶ religion,
- ▶ region,
- ▶ party affiliation (multiparty systems),
- ▶ actions available to a particular actor,
- ▶ answers to an ideological question that cannot be placed neatly on a L/R scale.

Unordered Categorical Variables

If an unordered categorical variable is the dependent variable, then a special kind of regression called **multinomial logistic regression** can be used. We don't cover that here, but it's another big topic in statistics.

Unordered Categorical Variables

If an unordered categorical variable is the dependent variable, then a special kind of regression called **multinomial logistic regression** can be used. We don't cover that here, but it's another big topic in statistics.

If an unordered categorical variable is an independent variable, then we have to do something to address the fact that

A one-unit increase in X

makes no sense!

Unordered Categorical Variables

The problem: The actual numbers that are applied to the values of an unordered categorical variable are **arbitrary and have no meaning**, other than distinguishing the categories.

So adding 1 is a meaningless thing to do.

Unordered Categorical Variables

The problem: The actual numbers that are applied to the values of an unordered categorical variable are arbitrary and have no meaning, other than distinguishing the categories.

So adding 1 is a meaningless thing to do.

The solution: break the unordered categorical variable up into binary (dummy) variables for each category!

Each dummy variable means “1 if x is the category in question, 0 if x is not.”

Unordered Categorical Variables

The problem: The actual numbers that are applied to the values of an unordered categorical variable are **arbitrary and have no meaning**, other than distinguishing the categories.

So adding 1 is a meaningless thing to do.

The solution: **break the unordered categorical variable up** into binary (dummy) variables for each category!

Each dummy variable means “**1 if x is the category in question, 0 if x is not.**”

The leave out ONE of these dummy variables. The one you decide to leave out is called the **base category**, and **all other categories are compared to this one.**

Unordered Categorical Variables

R, by default, breaks up factor variables for you, and leaves out the **first category** that appears when using the `levels()` or `table()` functions on the variable.

Unordered Categorical Variables

R, by default, breaks up factor variables for you, and leaves out the first category that appears when using the `levels()` or `table()` functions on the variable.

I prefer to leave out the largest category, since that makes the comparisons more meaningful.

Unordered Categorical Variables

R, by default, breaks up factor variables for you, and leaves out the **first category** that appears when using the `levels()` or `table()` functions on the variable.

I prefer to leave out the largest category, since that makes the comparisons more meaningful.

To **change the base category**, use `fct_relevel()` within `mutate()` prior to the regression to move the intended base to be first among the categories.

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i2} is region: Northeast, Midwest, Mountain, West Coast,
Southwest, Southeast.

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i2} is region: Northeast, Midwest, Mountain, West Coast, Southwest, Southeast.

The following regression equation

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

is incorrect because we didn't break x_{i2} into dummies for each category.

Example

Suppose we treat **Northeast as the base category** (and leave out that dummy variable D_{i1}). Then we create the following variables:

- ▶ D_{i2} : 1 if x_{i2} is **Midwest**, 0 otherwise.
- ▶ D_{i3} : 1 if x_{i2} is **Mountain**, 0 otherwise.
- ▶ D_{i4} : 1 if x_{i2} is **West Coast**, 0 otherwise.
- ▶ D_{i5} : 1 if x_{i2} is **Southwest**, 0 otherwise.
- ▶ D_{i6} : 1 if x_{i2} is **Southeast**, 0 otherwise.

Example

Suppose we treat **Northeast as the base category** (and leave out that dummy variable D_{i1}). Then we create the following variables:

- ▶ D_{i2} : 1 if x_{i2} is **Midwest**, 0 otherwise.
- ▶ D_{i3} : 1 if x_{i2} is **Mountain**, 0 otherwise.
- ▶ D_{i4} : 1 if x_{i2} is **West Coast**, 0 otherwise.
- ▶ D_{i5} : 1 if x_{i2} is **Southwest**, 0 otherwise.
- ▶ D_{i6} : 1 if x_{i2} is **Southeast**, 0 otherwise.

The correct regression equation is

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

where the δ terms are just more coefficients.

Example

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Suppose that

- ▶ $\delta_2 = 20$
- ▶ $\delta_4 = -15$
- ▶ $\delta_6 = 40$
- ▶ $\delta_3 = 35$
- ▶ $\delta_5 = -10$

How are these coefficients interpreted?

Example

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Suppose that

- ▶ $\delta_2 = 20$
- ▶ $\delta_4 = -15$
- ▶ $\delta_6 = 40$
- ▶ $\delta_3 = 35$
- ▶ $\delta_5 = -10$

How are these coefficients interpreted?

δ_2 : “Compared to people from the Northeast, people from the Midwest watch 20 more minutes of TV news on average, after accounting for gender and interest in politics.”

Example

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Suppose that

- ▶ $\delta_2 = 20$
- ▶ $\delta_4 = -15$
- ▶ $\delta_6 = 40$
- ▶ $\delta_3 = 35$
- ▶ $\delta_5 = -10$

How are these coefficients interpreted?

δ_3 : “Compared to people from the Northeast, people from the Mountain region watch 35 more minutes of TV news on average, after accounting for gender and interest in politics.”

Example

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Suppose that

- ▶ $\delta_2 = 20$
- ▶ $\delta_4 = -15$
- ▶ $\delta_6 = 40$
- ▶ $\delta_3 = 35$
- ▶ $\delta_5 = -10$

How are these coefficients interpreted?

δ_4 : “Compared to people from the Northeast, people from the West Coast watch 15 fewer minutes of TV news on average, after accounting for gender and interest in politics.”

Example

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Suppose that

- ▶ $\delta_2 = 20$
- ▶ $\delta_4 = -15$
- ▶ $\delta_6 = 40$
- ▶ $\delta_3 = 35$
- ▶ $\delta_5 = -10$

How are these coefficients interpreted?

δ_5 : “Compared to people from the Northeast, people from the Southwest watch 10 fewer minutes of TV news on average, after accounting for gender and interest in politics.”

Example

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Suppose that

- ▶ $\delta_2 = 20$
- ▶ $\delta_4 = -15$
- ▶ $\delta_6 = 40$
- ▶ $\delta_3 = 35$
- ▶ $\delta_5 = -10$

How are these coefficients interpreted?

δ_6 : “Compared to people from the Northeast, people from the Southeast watch 40 more minutes of TV news on average, after accounting for gender and interest in politics.”

Ordered Categorical Variables

An ordered categorial variable

Ordered Categorical Variables

An ordered categorial variable

- ▶ is non-continuous (a small number of discrete values),

Ordered Categorical Variables

An ordered categorial variable

- ▶ is non-continuous (a small number of discrete values),
- ▶ has more than 2 categorical values,

Ordered Categorical Variables

An **ordered categorial variable**

- ▶ is non-continuous (a small number of discrete values),
- ▶ has more than 2 categorical values,
- ▶ but **has a natural way to put these categories in an order.**

Ordered Categorical Variables

An **ordered categorial variable**

- ▶ is non-continuous (a small number of discrete values),
- ▶ has more than 2 categorical values,
- ▶ but **has a natural way to put these categories in an order.**

Some common ordered categorical variables in political science:

- ▶ income level,
- ▶ response to a left-right ideology question,
- ▶ measures of the degree to which a case exhibits particular characteristics (e.g. democratic quality, strong federalism, ethnic tension).

Ordered Categorical Variables

If an ordered categorical variable is the dependent variable, then a special kind of regression called **ordered logistic regression** can be used. We don't cover that here, but it's another big topic in statistics.

Ordered Categorical Variables

If an ordered categorical variable is the dependent variable, then a special kind of regression called **ordered logistic regression** can be used. We don't cover that here, but it's another big topic in statistics.

If an ordered categorical variable is an independent variable, then we have a choice to make. Do we

Ordered Categorical Variables

If an ordered categorical variable is the dependent variable, then a special kind of regression called **ordered logistic regression** can be used. We don't cover that here, but it's another big topic in statistics.

If an ordered categorical variable is an independent variable, then we have a choice to make. Do we

- ▶ treat the ordinal variable like a **continuous variable**?

Ordered Categorical Variables

If an ordered categorical variable is the dependent variable, then a special kind of regression called **ordered logistic regression** can be used. We don't cover that here, but it's another big topic in statistics.

If an ordered categorical variable is an independent variable, then we have a choice to make. Do we

- ▶ treat the ordinal variable like a **continuous variable**?
- ▶ or treat the ordinal variable like an **unordered categorical variable**?

Ordered Categorical Variables

If an ordered categorical variable is the dependent variable, then a special kind of regression called **ordered logistic regression** can be used. We don't cover that here, but it's another big topic in statistics.

If an ordered categorical variable is an independent variable, then we have a choice to make. Do we

- ▶ treat the ordinal variable like a **continuous variable**?
- ▶ or treat the ordinal variable like an **unordered categorical variable**?

Each choice has pros and cons.

Ordered Categorical Variables

Choice 1: treat as continuous.

Ordered Categorical Variables

Choice 1: treat as continuous.

Pro: no special transformations are necessary. A “one-unit increase” refers to a one-level increase on the ordinal scale.

Ordered Categorical Variables

Choice 1: treat as continuous.

Pro: no special transformations are necessary. A “one-unit increase” refers to a one-level increase on the ordinal scale.

Con: because you derive **only one effect for the whole variable**, you assume that every level has the same effect on y . That is:

- ▶ moving from “extremely conservative” to “very conservative” has the same effect on y as moving from “very liberal” to “extremely liberal”

Ordered Categorical Variables

Choice 1: treat as continuous.

Pro: no special transformations are necessary. A “one-unit increase” refers to a one-level increase on the ordinal scale.

Con: because you derive **only one effect for the whole variable**, you assume that every level has the same effect on y . That is:

- ▶ moving from “extremely conservative” to “very conservative” has the same effect on y as moving from “very liberal” to “extremely liberal”

There are situations in which this assumption is probably okay.

There are others when it is a very poor assumption. Be aware of what works best for your particular analysis.

Ordered Categorical Variables

Choice 2: treat as unordered categorical.

Ordered Categorical Variables

Choice 2: treat as unordered categorical.

Pro: does not assume that every level increase has the same effect on y , regardless of where on the scale you happen to be.

Ordered Categorical Variables

Choice 2: treat as unordered categorical.

Pro: does not assume that every level increase has the same effect on y , regardless of where on the scale you happen to be.

Con: if there are **a lot of categories** or a lot of other unordered categorical variables, you might burn all of your degrees of freedom or make the results unwieldy.

Also: *more nuance than you might need.* Makes interpretation more difficult.

Ordered Categorical Variables

Choice 2: treat as unordered categorical.

Pro: does not assume that every level increase has the same effect on y , regardless of where on the scale you happen to be.

Con: if there are **a lot of categories** or a lot of other unordered categorical variables, you might burn all of your degrees of freedom or make the results unwieldy.

Also: *more nuance than you might need.* Makes interpretation more difficult.

My recommendation – break up an ordinal variable **if it is a real IV of interest**. But don't bother if it is just a control.

Ordered Categorical Variables

If a variable is a **factor**, and you want to treat it as **continuous**:

- ▶ Use `levels()` to confirm that the categories are in a meaningful order. If not, use `fct_relevel()` within `mutate()` to put them in order.
- ▶ Use `as.numeric()` around the variable in the regression command.

Ordered Categorical Variables

If a variable is a **factor**, and you want to treat it as **continuous**:

- ▶ Use `levels()` to confirm that the categories are in a meaningful order. If not, use `fct_relevel()` within `mutate()` to put them in order.
- ▶ Use `as.numeric()` around the variable in the regression command.

If a variable is **numeric**, and you want to treat it as **unordered categorical**, use `factor()` to coerce the variable to the factor class, labeling the categories (like we do when we clean data).

Ordered Categorical Variables

If a variable is a **factor**, and you want to treat it as **continuous**:

- ▶ Use `levels()` to confirm that the categories are in a meaningful order. If not, use `fct_relevel()` within `mutate()` to put them in order.
- ▶ Use `as.numeric()` around the variable in the regression command.

If a variable is **numeric**, and you want to treat it as **unordered categorical**, use `factor()` to coerce the variable to the factor class, labeling the categories (like we do when we clean data).

Don't assign the ordered class to any variable. By default, R performs a method called **polynomial contrasts** with ordered class variables. It's not easy or useful to interpret.

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i3} is interest in politics, 7-point scale from “not at all” to “extremely”

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i3} is interest in politics, 7-point scale from “not at all” to “extremely”

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Interpret $\beta_3 = 10$.

Example

Level of analysis: random sample from adult US population

y_i is minutes spent watching TV news

x_{i3} is interest in politics, 7-point scale from “not at all” to “extremely”

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} + \beta_3 x_{i3} + \varepsilon_i$$

Interpret $\beta_3 = 10$.

“As a person increasing one-level on the seven-point interest in politics scale, becoming more interested, they watch 10 additional minutes of cable news on average, after accounting for gender and region.”

Example

Or, we could break up the interest variable too, yielding a regression equation like:

$$y_i = \alpha + \beta_1 x_{i1} + \delta_2 D_{i2} + \delta_3 D_{i3} + \delta_4 D_{i4} + \delta_5 D_{i5} + \delta_6 D_{i6} \\ + \gamma_2 G_{i2} + \gamma_3 G_{i3} + \gamma_4 G_{i4} + \gamma_5 G_{i5} + \gamma_6 G_{i6} + \gamma_7 G_{i7} + \varepsilon_i,$$

and interpret the new coefficients in the same way we interpret the coefficients for the unordered categorical variable.

Standard errors

```
> reg <- lm(imdb_score ~ budget + duration + year + facenumber_in_poster +
+           cast_total_facebook_likes, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget + duration + year + facenumber_in_poster +
    cast_total_facebook_likes, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7664	-0.5586	0.0990	0.6909	3.0156

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.0297394	0.1395064	43.222	< 2e-16	***
budget	-0.0005112	0.0002037	-2.509	0.0121	*
duration	0.0158795	0.0007005	22.667	< 2e-16	***
year	-0.0152222	0.0012434	-12.242	< 2e-16	***
facenumber_in_poster	-0.0401398	0.0075809	-5.295	1.25e-07	***
cast_total_facebook_likes	0.0054011	0.0008239	6.556	6.16e-11	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.021 on 4517 degrees of freedom

(533 observations deleted due to missingness)

Multiple R-squared: 0.161, Adjusted R-squared: 0.1601

F-statistic: 173.4 on 5 and 4517 DF, p-value: < 2.2e-16

Standard errors

Statistics is primarily the study of **uncertainty**. Any statement we make **must** be accompanied by a *second* statement about how certain we are about the first statement.

Standard errors

Statistics is primarily the study of **uncertainty**. Any statement we make **must** be accompanied by a *second* statement about how certain we are about the first statement.

A high degree of certainty \neq a large effect.

We can be very certain that an effect is small, or we can find a very large effect and be very uncertain about its true value.

Standard errors

Statistics is primarily the study of **uncertainty**. Any statement we make **must** be accompanied by a *second* statement about how certain we are about the first statement.

A high degree of certainty \neq a large effect.

We can be very certain that an effect is small, or we can find a very large effect and be very uncertain about its true value.

A **standard error** is a measurement of our uncertainty about a result. The smaller the standard error, the more **certain** we are about the likely true values.

Standard errors

Statistics is primarily the study of **uncertainty**. Any statement we make **must** be accompanied by a *second* statement about how certain we are about the first statement.

A high degree of certainty \neq a large effect.

We can be very certain that an effect is small, or we can find a very large effect and be very uncertain about its true value.

A **standard error** is a measurement of our uncertainty about a result. The smaller the standard error, the more **certain** we are about the likely true values.

Standard errors also have no interpretation. They can only be used to construct two meaningful, interpretable quantities: **confidence intervals** and **p-values**.

The 95% confidence interval

Suppose you draw one value x from a general normal distribution

$$N(\mu, \sigma^2).$$

There is only a **2.5% chance** that x is less than the following quantity:

$$\mu - 1.96\sigma,$$

And there is a **97.5% chance** that x is less than

$$\mu + 1.96\sigma.$$

The 95% confidence interval

Suppose you draw one value x from a general normal distribution

$$N(\mu, \sigma^2).$$

There is only a **2.5% chance** that x is less than the following quantity:

$$\mu - 1.96\sigma,$$

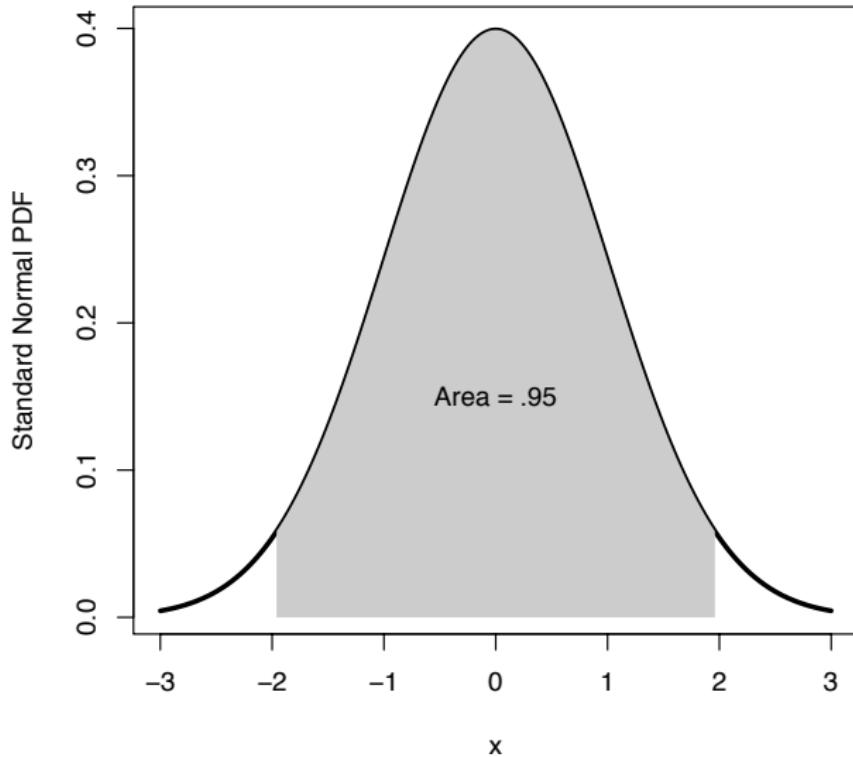
And there is a **97.5% chance** that x is less than

$$\mu + 1.96\sigma.$$

So there's a **95% chance** that x is in the interval

$$[\mu - 1.96\sigma, \mu + 1.96\sigma].$$

The 95% confidence interval



The 95% confidence interval

The same math applies to regressions. We are interested in the true population/DGP value of β . We have a sample estimate $\hat{\beta}$ and a standard error σ_{β} . The relative likelihood of true values of β are represented by

$$N(\hat{\beta}, \sigma_{\beta}^2).$$

The 95% confidence interval

The same math applies to regressions. We are interested in the true population/DGP value of β . We have a sample estimate $\hat{\beta}$ and a standard error σ_β . The relative likelihood of true values of β are represented by

$$N(\hat{\beta}, \sigma_\beta^2).$$

There is only a 2.5% chance that the true β is less than the following quantity:

$$\hat{\beta} - 1.96\sigma_\beta,$$

The 95% confidence interval

The same math applies to regressions. We are interested in the true population/DGP value of β . We have a sample estimate $\hat{\beta}$ and a standard error σ_β . The relative likelihood of true values of β are represented by

$$N(\hat{\beta}, \sigma_\beta^2).$$

There is only a 2.5% chance that the true β is less than the following quantity:

$$\hat{\beta} - 1.96\sigma_\beta,$$

And there is a 97.5% chance that the true β is less than

$$\hat{\beta} + 1.96\sigma_\beta.$$

The 95% confidence interval

The same math applies to regressions. We are interested in the true population/DGP value of β . We have a sample estimate $\hat{\beta}$ and a standard error σ_β . The relative likelihood of true values of β are represented by

$$N(\hat{\beta}, \sigma_\beta^2).$$

There is only a 2.5% chance that the true β is less than the following quantity:

$$\hat{\beta} - 1.96\sigma_\beta,$$

And there is a 97.5% chance that the true β is less than

$$\hat{\beta} + 1.96\sigma_\beta.$$

So there's a 95% chance that the true β is in the interval

$$[\hat{\beta} - 1.96\sigma_\beta, \hat{\beta} + 1.96\sigma_\beta].$$

The 95% confidence interval

Example: $\beta = 2.67$, $\sigma_\beta = .28$.

The 95% confidence interval

Example: $\beta = 2.67$, $\sigma_\beta = .28$.

95% lower bound:

$$2.67 - 1.96(.28) = \textcolor{red}{2.12}$$

The 95% confidence interval

Example: $\beta = 2.67$, $\sigma_\beta = .28$.

95% lower bound:

$$2.67 - 1.96(.28) = \textcolor{red}{2.12}$$

95% upper bound:

$$2.67 + 1.96(.28) = \textcolor{red}{3.22}$$

The 95% confidence intervals

R doesn't display the confidence intervals automatically. Instead, use the `confint()` command.

```
> confint(reg)
```

	2.5 %	97.5 %
(Intercept)	5.7562385478	6.303240231
budget	-0.0009105205	-0.000111797
duration	0.0145061203	0.017252935
year	-0.0176599057	-0.012784443
facenumber_in_poster	-0.0550021788	-0.025277448
cast_total_facebook_likes	0.0037859224	0.007016326

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

For example, `confint(reg, level = .9)` gives the 90% CI.

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

For example, `confint(reg, level = .9)` gives the 90% CI.

If we decrease the level, what happens to the interval?

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

For example, `confint(reg, level = .9)` gives the 90% CI.

If we decrease the level, what happens to the interval? **It gets smaller.** Why?

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

For example, `confint(reg, level = .9)` gives the 90% CI.

If we decrease the level, what happens to the interval? **It gets smaller.** Why?

A smaller percent means

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

For example, `confint(reg, level = .9)` gives the 90% CI.

If we decrease the level, what happens to the interval? **It gets smaller.** Why?

A smaller percent means **less confidence**.

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

For example, `confint(reg, level = .9)` gives the 90% CI.

If we decrease the level, what happens to the interval? **It gets smaller.** Why?

A smaller percent means **less confidence**. So we can report a closer interval around the coefficient, but we are less confident in that interval.

Other confidence intervals

You can specify a **different confidence interval** (that uses a different value from 1.96) by using the level option for the `confint()` command in R.

For example, `confint(reg, level = .9)` gives the 90% CI.

If we decrease the level, what happens to the interval? **It gets smaller.** Why?

A smaller percent means **less confidence**. So we can report a closer interval around the coefficient, but we are less confident in that interval.

On the other hand, a higher percent means **more confidence**, but a **wider interval**.

How to interpret a confidence interval

Confidence intervals are the second most meaningful statistic in the regression table, just after the coefficients. You might say

“We believe with 95% confidence that the true coefficient is between 2.13 and 3.22.”

Warning: while most people will describe confidence intervals this way, the most accurate interpretation is *weird*:

How to interpret a confidence interval

Confidence intervals are the second most meaningful statistic in the regression table, just after the coefficients. You might say

“We believe with 95% confidence that the true coefficient is between 2.13 and 3.22.”

Warning: while most people will describe confidence intervals this way, the most accurate interpretation is *weird*:

If we were to

- ▶ take repeated samples from the same population,
- ▶ (or repeatedly run the same data generating process)

and **run the same regression on each dataset we draw**,

How to interpret a confidence interval

Confidence intervals are the second most meaningful statistic in the regression table, just after the coefficients. You might say

“We believe with 95% confidence that the true coefficient is between 2.13 and 3.22.”

Warning: while most people will describe confidence intervals this way, the most accurate interpretation is *weird*:

If we were to

- ▶ take repeated samples from the same population,
- ▶ (or repeatedly run the same data generating process)

and **run the same regression on each dataset we draw**, then the true population parameter would be covered by **95% of these models' 95% confidence intervals**.

t-statistics

```
> reg <- lm(imdb_score ~ budget + duration + year + facenumber_in_poster +
+           cast_total_facebook_likes, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget + duration + year + facenumber_in_poster +
    cast_total_facebook_likes, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7664	-0.5586	0.0990	0.6909	3.0156

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0297394	0.1395064	43.222	< 2e-16 ***
budget	-0.0005112	0.0002037	-2.509	0.0121 *
duration	0.0158795	0.0007005	22.667	< 2e-16 ***
year	-0.0152222	0.0012434	-12.242	< 2e-16 ***
facenumber_in_poster	-0.0401398	0.0075809	-5.295	1.25e-07 ***
cast_total_facebook_likes	0.0054011	0.0008239	6.556	6.16e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.021 on 4517 degrees of freedom

(533 observations deleted due to missingness)

Multiple R-squared: 0.161, Adjusted R-squared: 0.1601

F-statistic: 173.4 on 5 and 4517 DF, p-value: < 2.2e-16

p-values

```
> reg <- lm(imdb_score ~ budget + duration + year + facenumber_in_poster +
+           cast_total_facebook_likes, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget + duration + year + facenumber_in_poster +
    cast_total_facebook_likes, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7664	-0.5586	0.0990	0.6909	3.0156

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0297394	0.1395064	43.222	< 2e-16 ***
budget	-0.0005112	0.0002037	-2.509	0.0121 *
duration	0.0158795	0.0007005	22.667	< 2e-16 ***
year	-0.0152222	0.0012434	-12.242	< 2e-16 ***
facenumber_in_poster	-0.0401398	0.0075809	-5.295	1.25e-07 ***
cast_total_facebook_likes	0.0054011	0.0008239	6.556	6.16e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.021 on 4517 degrees of freedom
(533 observations deleted due to missingness)

Multiple R-squared: 0.161, Adjusted R-squared: 0.1601
F-statistic: 173.4 on 5 and 4517 DF, p-value: < 2.2e-16

t statistics and *p* values

These quantities are the most **widely abused** results in regression analysis.

It is very important for you to understand exactly what these values mean. It is even more important for you to understand what these values **don't mean**.

t statistics and *p* values

These quantities are the most **widely abused** results in regression analysis.

It is very important for you to understand exactly what these values mean. It is even more important for you to understand what these values **don't mean**.

A coefficient of 0 means

- ▶ a **flat sloped** best-fit line – perfectly horizontal
- ▶ the value of y is **independent** of the value of x
- ▶ “a one-unit increase in x produces **no change** in y , on average, controlling . . .”

t statistics and *p* values

These quantities are the most **widely abused** results in regression analysis.

It is very important for you to understand exactly what these values mean. It is even more important for you to understand what these values **don't mean**.

A coefficient of 0 means

- ▶ a **flat sloped** best-fit line – perfectly horizontal
- ▶ the value of y is **independent** of the value of x
- ▶ “a one-unit increase in x produces **no change** in y , on average, controlling . . .”

All of which are ways to state that x has **no effect** on y .

t statistics and *p* values

These quantities are the most **widely abused** results in regression analysis.

It is very important for you to understand exactly what these values mean. It is even more important for you to understand what these values **don't mean**.

A coefficient of 0 means

- ▶ a **flat sloped** best-fit line – perfectly horizontal
- ▶ the value of y is **independent** of the value of x
- ▶ “a one-unit increase in x produces **no change** in y , on average, controlling ...”

All of which are ways to state that x has **no effect** on y .

Goal: to test whether sufficient evidence exists to **reject the null hypothesis that a coefficient β is 0**.

Interpretation of p -values

In the most technical sense, the p -value is

The probability that a random sample could have had a t value
with this distance from 0, assuming the truth is that $\beta=0$

Interpretation of p -values

In the most technical sense, the p -value is

The probability that a random sample could have had a t value with this distance from 0, assuming the truth is that $\beta=0$

If the probability is really low, then one of two things must be true

Interpretation of p -values

In the most technical sense, the p -value is

The probability that a random sample could have had a t value with this distance from 0, assuming the truth is that $\beta=0$

If the probability is really low, then one of two things must be true

1. the sample was **really, really extraordinary and unlikely**,

Interpretation of p -values

In the most technical sense, the p -value is

The probability that a random sample could have had a t value with this distance from 0, assuming the truth is that $\beta=0$

If the probability is really low, then one of two things must be true

1. the sample was **really, really extraordinary and unlikely**,
2. or **the assumption that $\beta = 0$ is wrong**.

Interpretation of p -values

In the most technical sense, the p -value is

The probability that a random sample could have had a t value with this distance from 0, assuming the truth is that $\beta=0$

If the probability is really low, then one of two things must be true

1. the sample was **really, really extraordinary and unlikely**,
2. or **the assumption that $\beta = 0$ is wrong**.

For very small values of p , we **reject** the possibility of the first option and go with the second.

Interpretation of p -values

In the most technical sense, the p -value is

The probability that a random sample could have had a t value with this distance from 0, assuming the truth is that $\beta=0$

If the probability is really low, then one of two things must be true

1. the sample was **really, really extraordinary and unlikely**,
2. or **the assumption that $\beta = 0$ is wrong**.

For very small values of p , we **reject** the possibility of the first option and go with the second.

Then we conclude that $\beta \neq 0$.

Interpretation of p -values

In the most technical sense, the p -value is

The probability that a random sample could have had a t value with this distance from 0, assuming the truth is that $\beta=0$

If the probability is really low, then one of two things must be true

1. the sample was **really, really extraordinary and unlikely**,
2. or **the assumption that $\beta = 0$ is wrong**.

For very small values of p , we **reject** the possibility of the first option and go with the second.

Then we conclude that $\beta \neq 0$.

Which we understand to mean that **x has some non-zero effect on y** .

Interpretation of p -values

Some common standards for concluding that $\beta \neq 0$:

Interpretation of p -values

Some common standards for concluding that $\beta \neq 0$:

- ▶ $p < .05$ – the most common standard in social science

Interpretation of p -values

Some common standards for concluding that $\beta \neq 0$:

- ▶ $p < .05$ – the most common standard in social science
- ▶ $p < .01$ – a more conservative standard for concluding that x has an effect on y

Interpretation of p -values

Some common standards for concluding that $\beta \neq 0$:

- ▶ $p < .05$ – the most common standard in social science
- ▶ $p < .01$ – a more conservative standard for concluding that x has an effect on y
- ▶ $p < .1$ – a less conservative standard

Interpretation of p -values

Some common standards for concluding that $\beta \neq 0$:

- ▶ $p < .05$ – the most common standard in social science
- ▶ $p < .01$ – a more conservative standard for concluding that x has an effect on y
- ▶ $p < .1$ – a less conservative standard

If the standard is met, we call a coefficient “statistically significantly different from 0” although many researchers just say “significant.”

Interpretation of p -values

Some common standards for concluding that $\beta \neq 0$:

- ▶ $p < .05$ – the most common standard in social science
- ▶ $p < .01$ – a more conservative standard for concluding that x has an effect on y
- ▶ $p < .1$ – a less conservative standard

If the standard is met, we call a coefficient “statistically significantly different from 0” although many researchers just say “significant.”

Important: choose a standard before running any tests and stick with it. **Bending the standard to favor a conclusion is academically dishonest.**

MISinterpretation of *p*-values

Mistake 1: “**type 1 error**” — concluding that a hypothesis is false even though it is actually true.

MISinterpretation of *p*-values

Mistake 1: “**type 1 error**” — concluding that a hypothesis is false even though it is actually true.

$p = .05$ means there’s only a **1/20 chance** that your t could have been as big as it is, assuming that the true $\beta = 0$.

MISinterpretation of *p*-values

Mistake 1: “**type 1 error**” — concluding that a hypothesis is false even though it is actually true.

$p = .05$ means there’s only a **1/20 chance** that your t could have been as big as it is, assuming that the true $\beta = 0$.

But a 1/20 chance means that if you do **20 tests**, you **WOULD** expect one on average to be unusual!

MISinterpretation of *p*-values

Mistake 1: “**type 1 error**” — concluding that a hypothesis is false even though it is actually true.

$p = .05$ means there’s only a **1/20 chance** that your t could have been as big as it is, assuming that the true $\beta = 0$.

But a 1/20 chance means that if you do **20 tests**, you **WOULD** expect one on average to be unusual!

Some researchers do *test after test after test after test*. That will eventually lead you to claim that a **null relationship is significant**. See: <http://xkcd.com/882/>.

MISinterpretation of *p*-values

Mistake 2: “**type 2 error**” — concluding that a hypothesis is true even though it is actually false.

MISinterpretation of *p*-values

Mistake 2: “**type 2 error**” — concluding that a hypothesis is true even though it is actually false.

$p = .35$ means we cannot reject the null of independence between x and y . But that **does NOT mean that x and y actually are independent!**

MISinterpretation of *p*-values

Mistake 2: “**type 2 error**” — concluding that a hypothesis is true even though it is actually false.

$p = .35$ means we cannot reject the null of independence between x and y . But that **does NOT mean that x and y actually are independent!**

A null finding is not “**no effect**” but rather a **lack of enough evidence** to meet an arbitrary standard of $p < .05$. Don’t write “has no effect” in your papers.

MISinterpretation of *p*-values

Mistake 2: “**type 2 error**” — concluding that a hypothesis is true even though it is actually false.

$p = .35$ means we cannot reject the null of independence between x and y . But that **does NOT mean that x and y actually are independent!**

A null finding is not “**no effect**” but rather a **lack of enough evidence** to meet an arbitrary standard of $p < .05$. Don’t write “has no effect” in your papers.

Some researchers go so far as to **delete** the x variables whose coefficients fail to achieve significance. That’s a **really stupid** thing to do, especially if some of these variables are theoretically important controls.

MISinterpretation of p -values

Mistake 3: interpreting the size of the p -values

MISinterpretation of *p*-values

Mistake 3: interpreting the size of the *p*-values

The size of a *p*-value says NOTHING about the size of the effect of *x* on *y*. **That's the coefficient!**

MISinterpretation of *p*-values

Mistake 3: interpreting the size of the *p*-values

The size of a *p*-value says NOTHING about the size of the effect of *x* on *y*. **That's the coefficient!**

It is possible for strong effects to have high *p*-values and it is possible for small effects to have low *p*-values.

MISinterpretation of *p*-values

Mistake 3: interpreting the size of the *p*-values

The size of a *p*-value says NOTHING about the size of the effect of *x* on *y*. **That's the coefficient!**

It is possible for strong effects to have high *p*-values and it is possible for small effects to have low *p*-values.

Don't fall into the **traps** of saying that one variable is

MISinterpretation of *p*-values

Mistake 3: interpreting the size of the *p*-values

The size of a *p*-value says NOTHING about the size of the effect of *x* on *y*. **That's the coefficient!**

It is possible for strong effects to have high *p*-values and it is possible for small effects to have low *p*-values.

Don't fall into the **traps** of saying that one variable is

- ▶ “more significant” than another,

MISinterpretation of *p*-values

Mistake 3: interpreting the size of the *p*-values

The size of a *p*-value says NOTHING about the size of the effect of *x* on *y*. **That's the coefficient!**

It is possible for strong effects to have high *p*-values and it is possible for small effects to have low *p*-values.

Don't fall into the **traps** of saying that one variable is

- ▶ “more significant” than another,
- ▶ “strongly” or “marginally” significant,

MISinterpretation of *p*-values

Mistake 3: interpreting the size of the *p*-values

The size of a *p*-value says NOTHING about the size of the effect of *x* on *y*. **That's the coefficient!**

It is possible for strong effects to have high *p*-values and it is possible for small effects to have low *p*-values.

Don't fall into the **traps** of saying that one variable is

- ▶ “more significant” than another,
- ▶ “strongly” or “marginally” significant,
- ▶ “increasingly significant” when we add controls.

MISinterpretation of *p*-values

Mistake 3: interpreting the size of the *p*-values

The size of a *p*-value says NOTHING about the size of the effect of *x* on *y*. **That's the coefficient!**

It is possible for strong effects to have high *p*-values and it is possible for small effects to have low *p*-values.

Don't fall into the **traps** of saying that one variable is

- ▶ “more significant” than another,
- ▶ “strongly” or “marginally” significant,
- ▶ “increasingly significant” when we add controls.

A coefficient is **significant** or **not**. Don't interpret the size of *p*.

Logistic (logit) regression

Linear regression models, what we've focused on to this point, are for studying **continuous dependent variables**.

Logistic (logit) regression

Linear regression models, what we've focused on to this point, are for studying **continuous dependent variables**.

If the dependent variable is **binary**, you can study it with a method called **logistic regression** (also called “logit models”).

Logistic (logit) regression

Linear regression models, what we've focused on to this point, are for studying **continuous dependent variables**.

If the dependent variable is **binary**, you can study it with a method called **logistic regression** (also called “logit models”).

There are some **similarities** and some important **differences** between linear regression and logit.

Logistic (logit) regression

How logit is similar to linear regression:

Logistic (logit) regression

How logit is similar to linear regression:

- (1) Both use the **linear model**:

$$\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$$

where

- ▶ α is an intercept (or constant) we will estimate
- ▶ each β is a **coefficient** we will estimate
- ▶ x variables can be of theoretical interest, or controls

Logistic (logit) regression

How logit is similar to linear regression:

- (1) Both use the **linear model**:

$$\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$$

where

- ▶ α is an intercept (or constant) we will estimate
- ▶ each β is a **coefficient** we will estimate
- ▶ x variables can be of theoretical interest, or controls

- (2) Coefficients for **binary** x variables still compare the 1s (TRUE) group to the 0s (FALSE) group. Coefficients for **unordered categorical** variables must still be broken into **dummies**.

Logistic (logit) regression

How logit is similar to linear regression:

- (3) Positive coefficients mean that x makes the outcome **more likely**.

Logistic (logit) regression

How logit is similar to linear regression:

- (3) Positive coefficients mean that x makes the outcome **more likely**.
- (4) Negative coefficients mean that x makes the outcome **less likely**.

Logistic (logit) regression

How logit is similar to linear regression:

- (3) Positive coefficients mean that x makes the outcome **more likely**.
- (4) Negative coefficients mean that x makes the outcome **less likely**.
- (5) Coefficients of 0 mean that x has **no effect** on the outcome.
 p -values test whether each coefficient is **significantly different than zero**.

Logistic (logit) regression

How logit is different from linear regression:

Logistic (logit) regression

How logit is different from linear regression:

- (1) While we can interpret the sign and significance of coefficients, the coefficients themselves **don't have a useful interpretation**.

Logistic (logit) regression

How logit is different from linear regression:

- (1) While we can interpret the sign and significance of coefficients, the coefficients themselves **don't have a useful interpretation**.
- (2) Instead of interpreting coefficients, we can report **odds ratios**, **probabilities**, or **changes in probability**.

Logistic (logit) regression

How logit is different from linear regression:

- (1) While we can interpret that sign and significance of coefficients, the coefficients themselves **don't have a useful interpretation**.
- (2) Instead of interpreting coefficients, we can report **odds ratios**, **probabilities**, or **changes in probability**.
- (3) Logit is an example of a **generalized linear model** (GLM). GLMs don't minimize the sum of squared errors.

Logistic (logit) regression

How logit is different from linear regression:

- (1) While we can interpret that sign and significance of coefficients, the coefficients themselves **don't have a useful interpretation**.
- (2) Instead of interpreting coefficients, we can report **odds ratios**, **probabilities**, or **changes in probability**.
- (3) Logit is an example of a **generalized linear model** (GLM). GLMs don't minimize the sum of squared errors.
- (4) Instead they calculate a likelihood function that expresses how likely different parameters are given the data we observe. We estimate coefficients to **maximize the likelihood function**.

Running logit models in R

To run a logit model in R, use the `glm()` function:

```
logit <- glm(vote ~ age + marital + education +
              union + race + gender,
              data = anes, family=binomial(link="logit"))
```

Running logit models in R

To run a logit model in R, use the `glm()` function:

```
logit <- glm(vote ~ age + marital + education +
              union + race + gender,
              data = anes, family=binomial(link="logit"))
```

This function is similar to the `lm()` function:

- ▶ First, type the dependent variable, then a tilde ~
- ▶ Then type the **independent variables and controls** separated by plus signs
- ▶ Then use the data argument to specify the data frame that contains these variables
- ▶ `summary(logit)` displays the results

Running logit models in R

To run a logit model in R, use the `glm()` function:

```
logit <- glm(vote ~ age + marital + education +
              union + race + gender,
              data = anes, family=binomial(link="logit"))
```

This function is similar to the `lm()` function:

- ▶ First, type the dependent variable, then a tilde ~
- ▶ Then type the **independent variables and controls** separated by plus signs
- ▶ Then use the data argument to specify the data frame that contains these variables
- ▶ `summary(logit)` displays the results

The difference is also including

`family=binomial(link="logit")`, which tells R to run a logit, not another kind of GLM.

Running logit models in R

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.054219	0.267143	-3.946	7.94e-05 ***
age	-0.002611	0.002858	-0.914	0.360916
maritalNo longer married	0.218070	0.107966	2.020	0.043403 *
maritalNever married	0.804424	0.123571	6.510	7.52e-11 ***
educationSome college	0.038824	0.204404	0.190	0.849359
educationCollege degree	0.620889	0.210886	2.944	0.003238 **
educationGraduate degree	1.374357	0.219475	6.262	3.80e-10 ***
educationHS diploma	0.110228	0.217442	0.507	0.612203
union	0.462263	0.120015	3.852	0.000117 ***
raceBlack	3.953488	0.364617	< 2e-16	***
raceOther	0.717218	0.160059	4.481	7.43e-06 ***
raceHispanic	1.763333	0.169434	10.407	< 2e-16 ***
genderFemale	0.384729	0.086809	4.432	9.34e-06 ***
genderOther	13.267704	225.817853	0.059	0.953148

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 , , 1			

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3864.0 on 2799 degrees of freedom

Residual deviance: 3183.5 on 2786 degrees of freedom

(1471 observations deleted due to missingness)

AIC: 3211.5

Odds

Odds: The probability **for** an event divided by the probability **against** the event.

Example: probability of $y = 1$ is .6. Odds are

Odds

Odds: The probability **for** an event divided by the probability **against** the event.

Example: probability of $y = 1$ is .6. Odds are

$$\frac{.6}{.4} = \frac{6}{4} = \frac{3}{2} = \text{"3 to 2 for"}$$

Odds

Odds: The probability **for** an event divided by the probability **against** the event.

Example: probability of $y = 1$ is .6. Odds are

$$\frac{.6}{.4} = \frac{6}{4} = \frac{3}{2} = \text{"3 to 2 for"}$$

If we flip the fraction, it's called **"2 to 3 against"**.

Odds

Odds: The probability **for** an event divided by the probability **against** the event.

Example: probability of $y = 1$ is .6. Odds are

$$\frac{.6}{.4} = \frac{6}{4} = \frac{3}{2} = \text{"3 to 2 for"}$$

If we flip the fraction, it's called **"2 to 3 against"**.

Usually, when a gambler talks about odds, it's the odds **against**.
Odds are useful in gambling because they are interpreted as payouts. **The second number is the bet required to get the first number in profit.**

Odds

Odds: The probability **for** an event divided by the probability **against** the event.

Example: probability of $y = 1$ is .6. Odds are

$$\frac{.6}{.4} = \frac{6}{4} = \frac{3}{2} = \text{"3 to 2 for"}$$

If we flip the fraction, it's called **"2 to 3 against"**.

Usually, when a gambler talks about odds, it's the odds **against**. Odds are useful in gambling because they are interpreted as payouts. **The second number is the bet required to get the first number in profit.**

So in the above example, if I bet \$3 on an event and it happens, I get \$5: my \$3 returned and **\$2 in profit**.

Odds ratios from a logit model

In a logit model, the **odds for** an outcome are given by the whole linear model taken as the power of e :

$$\text{Odds} = e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}.$$

Odds ratios from a logit model

In a logit model, the **odds for** an outcome are given by the whole linear model taken as the power of e :

$$\text{Odds} = e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}.$$

Odds Ratio: How do the odds change (**multiplicatively**) for a one-unit increase in x ?

Odds ratios from a logit model

In a logit model, the **odds for** an outcome are given by the whole linear model taken as the power of e :

$$\text{Odds} = e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}.$$

Odds Ratio: How do the odds change (**multiplicatively**) for a **one-unit increase in x** ? For logit, the odds ratio is

$$\begin{aligned}\frac{\text{Odds with } (x_{i1} + 1)}{\text{Odds with } (x_{i1})} &= \frac{e^{\alpha + \beta_1(\cancel{x_{i1}} + 1) + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{e^{\alpha + \beta_1 \cancel{x_{i1}} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} \\ &= \frac{e^{\alpha + \beta_1 \cancel{x_{i1}} + \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{e^{\alpha + \beta_1 \cancel{x_{i1}} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} = \frac{e^{\beta_1} e^{\alpha + \beta_1 \cancel{x_{i1}} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{e^{\alpha + \beta_1 \cancel{x_{i1}} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} = e^{\beta_1}.\end{aligned}$$

Odds ratios from a logit model

In a logit model, the **odds for** an outcome are given by the whole linear model taken as the power of e:

$$\text{Odds} = e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}.$$

Odds Ratio: How do the odds change (**multiplicatively**) for a **one-unit increase in x** ? For logit, the odds ratio is

$$\begin{aligned}\frac{\text{Odds with } (x_{i1} + 1)}{\text{Odds with } (x_{i1})} &= \frac{e^{\alpha + \beta_1(x_{i1}+1) + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} \\ &= \frac{e^{\alpha + \beta_1 x_{i1} + \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} = \frac{e^{\beta_1} e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} = e^{\beta_1}.\end{aligned}$$

Odds ratio interpretation of a logit coefficient: “a one-unit increase in x_{ik} is associated with **multiplying the odds** by e^{β_k} , on average, after controlling for the other x variables.”

Odds ratios from a logit model

Example: logit coefficient is $\beta = 0.25$.

Odds ratios from a logit model

Example: logit coefficient is $\beta = 0.25$.

A one-unit increase in x_i is associated with multiplying the odds by $e^{0.25} = 1.28$, on average, after controlling for the other x variables.

Odds ratios from a logit model

Example: logit coefficient is $\beta = 0.25$.

A one-unit increase in x_i is associated with multiplying the odds by $e^{0.25} = 1.28$, on average, after controlling for the other x variables.

A one-unit increase in x_i is associated with the outcome becoming 28% “more likely”, on average, after controlling for the other x variables.

Odds ratios from a logit model

Example: logit coefficient is $\beta = 0.25$.

A one-unit increase in x_i is associated with multiplying the odds by $e^{0.25} = 1.28$, on average, after controlling for the other x variables.

A one-unit increase in x_i is associated with the outcome becoming 28% “more likely”, on average, after controlling for the other x variables.

Example: logit coefficient is $\beta = -0.5$.

Odds ratios from a logit model

Example: logit coefficient is $\beta = 0.25$.

A one-unit increase in x_i is associated with multiplying the odds by $e^{.25} = 1.28$, on average, after controlling for the other x variables.

A one-unit increase in x_i is associated with the outcome becoming 28% “more likely”, on average, after controlling for the other x variables.

Example: logit coefficient is $\beta = -0.5$.

A one-unit increase in x_i is associated with multiplying the odds by $e^{-0.5} = .61$, on average, after controlling for the other x variables.

Odds ratios from a logit model

Example: logit coefficient is $\beta = 0.25$.

A one-unit increase in x_i is associated with multiplying the odds by $e^{0.25} = 1.28$, on average, after controlling for the other x variables.

A one-unit increase in x_i is associated with the outcome becoming 28% “more likely”, on average, after controlling for the other x variables.

Example: logit coefficient is $\beta = -0.5$.

A one-unit increase in x_i is associated with multiplying the odds by $e^{-0.5} = .61$, on average, after controlling for the other x variables.

A one-unit increase in x_i is associated with the outcome becoming 39% “less likely”, on average, after controlling for the other x variables.

How to calculate odds ratios in R

After running a logit model and saving it as an object (named `logit`, for example), display the odds ratios by typing

```
exp(coef(logit)) :
```

```
> exp(coef(logit)) ## odds ratios
   (Intercept) age maritalNo longer married
3.484645e-01 9.973924e-01 1.243674e+00
maritalNever married educationSome college educationCollege degree
2.235408e+00 1.039587e+00 1.860582e+00
educationGraduate degree educationHS diploma union
3.952535e+00 1.116533e+00 1.587662e+00
raceBlack raceOther raceHispanic
5.211685e+01 2.048726e+00 5.831840e+00
genderFemale genderOther
1.469215e+00 5.782166e+05
```

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Odds are used in things like horse racing, but social scientists **NEVER** talk about probability in this weird way. Which makes more sense?

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Odds are used in things like horse racing, but social scientists **NEVER** talk about probability in this weird way. Which makes more sense?

.25 probability

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Odds are used in things like horse racing, but social scientists **NEVER** talk about probability in this weird way. Which makes more sense?

.25 probability or

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Odds are used in things like horse racing, but social scientists **NEVER** talk about probability in this weird way. Which makes more sense?

$$\text{.25 probability} \quad \text{or} \quad \frac{p}{1-p} = \frac{.25}{.75} = \text{3 to 1 odds against}$$

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Odds are used in things like horse racing, but social scientists **NEVER** talk about probability in this weird way. Which makes more sense?

.25 probability or $\frac{p}{1-p} = \frac{.25}{.75} = 3 \text{ to 1 odds against}$

.8 probability

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Odds are used in things like horse racing, but social scientists **NEVER** talk about probability in this weird way. Which makes more sense?

.25 probability or $\frac{p}{1-p} = \frac{.25}{.75} = 3 \text{ to 1 odds against}$

.8 probability or

Why odds ratios are going out of style

We use odds because they are **easy to calculate**. Before **powerful computers** they were the only viable option for interpreting logit results.

We still use them! I'm not entirely sure why. Tradition? Training?
It's troubling because odds ratios have some strange properties:

The odds are odd!

Odds are used in things like horse racing, but social scientists **NEVER** talk about probability in this weird way. Which makes more sense?

.25 probability or $\frac{p}{1-p} = \frac{.25}{.75} = 3 \text{ to 1 odds against}$

.8 probability or $\frac{p}{1-p} = \frac{.8}{.2} = 4 \text{ to 1 odds for}$

Why odds ratios are going out of style

Odds ratios can be misleading!

Why odds ratios are going out of style

Odds ratios can be misleading!

Example: 45% of white voters chose Obama

$$\text{Odds of voting for Obama for white voters} = \frac{.45}{.55} = 0.82$$

Why odds ratios are going out of style

Odds ratios can be misleading!

Example: 45% of white voters chose Obama

$$\text{Odds of voting for Obama for white voters} = \frac{.45}{.55} = 0.82$$

95% of African-American voters chose Obama

$$\text{Odds of voting for Obama for African-American voters} = \frac{.95}{.05} = 19$$

Why odds ratios are going out of style

Odds ratios can be misleading!

Example: 45% of white voters chose Obama

$$\text{Odds of voting for Obama for white voters} = \frac{.45}{.55} = 0.82$$

95% of African-American voters chose Obama

$$\text{Odds of voting for Obama for African-American voters} = \frac{.95}{.05} = 19$$

The odds ratio for being African-American, relative to white, is

$$\text{Odds ratio} = \frac{19}{0.82} = 23.2.$$

Why odds ratios are going out of style

Odds ratios can be misleading!

Example: 45% of white voters chose Obama

$$\text{Odds of voting for Obama for white voters} = \frac{.45}{.55} = 0.82$$

95% of African-American voters chose Obama

$$\text{Odds of voting for Obama for African-American voters} = \frac{.95}{.05} = 19$$

The odds ratio for being African-American, relative to white, is

$$\text{Odds ratio} = \frac{19}{0.82} = 23.2.$$

But it's **hyperbolic** to say: African-American voters are 23.2 times (or 2,220%) more likely to vote for Obama than white voters.

Why odds ratios are going out of style

Odds are easy to misunderstand and incorrectly interpret!

Why odds ratios are going out of style

Odds are easy to misunderstand and incorrectly interpret!

Odds ratios are the multiplicative change in the **odds for** an event.

Why odds ratios are going out of style

Odds are easy to misunderstand and incorrectly interpret!

Odds ratios are the multiplicative change in the **odds for** an event.

They are **NOT** multiplicative changes in probability.

Why odds ratios are going out of style

Odds are easy to misunderstand and incorrectly interpret!

Odds ratios are the multiplicative change in the **odds for** an event.

They are **NOT** multiplicative changes in probability.

$$\text{Mult. change in odds} = \frac{.25/.75}{.15/.85} = \mathbf{1.89}$$

Why odds ratios are going out of style

Odds are easy to misunderstand and incorrectly interpret!

Odds ratios are the multiplicative change in the **odds for** an event.

They are **NOT** multiplicative changes in probability.

$$\text{Mult. change in odds} = \frac{.25/.75}{.15/.85} = \mathbf{1.89}$$

$$\text{Mult. change in probability} = \frac{.25}{.15} = \mathbf{1.67}$$

Why odds ratios are going out of style

Odds are easy to misunderstand and incorrectly interpret!

Odds ratios are the multiplicative change in the **odds for** an event.

They are **NOT** multiplicative changes in probability.

$$\text{Mult. change in odds} = \frac{.25/.75}{.15/.85} = \mathbf{1.89}$$

$$\text{Mult. change in probability} = \frac{.25}{.15} = \mathbf{1.67}$$

When a researcher says “more likely” it’s not clear if that refers to **odds or probability**.

Why odds ratios are going out of style

Odds are easy to misunderstand and incorrectly interpret!

Odds ratios are the multiplicative change in the **odds for** an event.

They are **NOT** multiplicative changes in probability.

$$\text{Mult. change in odds} = \frac{.25/.75}{.15/.85} = \mathbf{1.89}$$

$$\text{Mult. change in probability} = \frac{.25}{.15} = \mathbf{1.67}$$

When a researcher says “more likely” it’s not clear if that refers to **odds or probability**.

Instead, let’s compute **predicted probability** and **marginal changes in probability**.

Predicted probability

The best way to express the results of logit models is as the **probability that $y = 1$** or **changes in probability** with respect to an x variable.

Predicted probability

The best way to express the results of logit models is as the **probability that $y = 1$** or **changes in probability** with respect to an x variable.

For logit, the formula for predicted probability is

$$P(y_i = 1) = \frac{1}{1 + e^{-(\hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik})}}$$

where $\hat{\cdot}$ indicates the estimate for the parameter.

Predicted probability

The best way to express the results of logit models is as the **probability that $y = 1$** or **changes in probability** with respect to an x variable.

For logit, the formula for predicted probability is

$$P(y_i = 1) = \frac{1}{1 + e^{-(\hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik})}}$$

where $\hat{\cdot}$ indicates the estimate for the parameter.

These probabilities are **different** for every observation, depending on the covariates.

Predicted probability

The best way to express the results of logit models is as the **probability that $y = 1$** or **changes in probability** with respect to an x variable.

For logit, the formula for predicted probability is

$$P(y_i = 1) = \frac{1}{1 + e^{-(\hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik})}}$$

where $\hat{\cdot}$ indicates the estimate for the parameter.

These probabilities are **different** for every observation, depending on the covariates.

Strategy: plot the predicted probabilities over the x of interest to show what's really going on with the model.

The OLS Estimator

The **formula for a β coefficient** in a single linear regression is

$$\beta = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2},$$

where \bar{x} and \bar{y} are the **sample means of x and y** .

The OLS Estimator

The **formula for a β coefficient** in a single linear regression is

$$\beta = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2},$$

where \bar{x} and \bar{y} are the **sample means of x and y** .

And the **formula for the constant α** is

$$\alpha = \bar{y} - \beta \bar{x}.$$

The OLS Estimator

The **formula for a β coefficient** in a single linear regression is

$$\beta = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2},$$

where \bar{x} and \bar{y} are the **sample means of x and y** .

And the **formula for the constant α** is

$$\alpha = \bar{y} - \beta \bar{x}.$$

We can **prove** that these are the correct OLS formulas for simple regression using calculus. However, the formulas are more complicated when the model includes **control variables**.

Multiple linear regression

When there are controls, we can use the **OLS estimator** to directly get coefficients for this regression:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

How would the math for the OLS estimator change?

Multiple linear regression

When there are controls, we can use the **OLS estimator** to directly get coefficients for this regression:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

How would the math for the OLS estimator change?

In this case, the formulas get more complicated:

Multiple linear regression

When there are controls, we can use the **OLS estimator** to directly get coefficients for this regression:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

How would the math for the OLS estimator change?

In this case, the formulas get more complicated:

$$\hat{\alpha} = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left(\sum_{i=1}^N (x_{1i} - \bar{x}_1)^2 \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2},$$

and the formula for $\hat{\beta}_2$ is the same as the formula for $\hat{\beta}_1$ with the subscripts 1 and 2 interchanged.

Multiple linear regression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left(\sum_{i=1}^N (x_{1i} - \bar{x}_1)^2 \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2},$$

The reason why this formula for β_1 is so much more complicated than the formula for single regression is that it **controls** for x_2 . Notice all the products of x_1 and x_2 in this equation.

Multiple linear regression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left(\sum_{i=1}^N (x_{1i} - \bar{x}_1)^2 \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2},$$

The reason why this formula for β_1 is so much more complicated than the formula for single regression is that it **controls** for x_2 . Notice all the products of x_1 and x_2 in this equation.

Can you imagine what the formula with 3 X variables looks like? How about 7? 15? Are these formulas just too big to ever write down or understand?

Multiple linear regression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \times \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left(\sum_{i=1}^N (x_{1i} - \bar{x}_1)^2 \times \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2},$$

The reason why this formula for β_1 is so much more complicated than the formula for single regression is that it **controls** for x_2 . Notice all the products of x_1 and x_2 in this equation.

Can you imagine what the formula with 3 X variables looks like? How about 7? 15? Are these formulas just too big to ever write down or understand?

Not if you use linear algebra!

Multiple linear regression

Y is an $(n \times 1)$ column vector containing the values of the dependent variable.

Multiple linear regression

Y is an $(n \times 1)$ column vector containing the values of the dependent variable.

X is an $(n \times [k + 1])$ matrix that contains each of the k independent variables in a column along with one column of all 1s to represent the intercept.

Multiple linear regression

Y is an $(n \times 1)$ column vector containing the values of the dependent variable.

X is an $(n \times [k + 1])$ matrix that contains each of the k independent variables in a column along with one column of all 1s to represent the intercept.

\hat{B} is a $([k + 1] \times 1)$ column vector of coefficient estimates from a multiple regression.

Multiple linear regression

Y is an $(n \times 1)$ column vector containing the values of the dependent variable.

X is an $(n \times [k + 1])$ matrix that contains each of the k independent variables in a column along with one column of all 1s to represent the intercept.

\hat{B} is a $([k + 1] \times 1)$ column vector of coefficient estimates from a multiple regression.

All of the parameters in \hat{B} are estimated simultaneously using the following formula:

$$\hat{B} = (X'X)^{-1}X'Y$$

This formula is called the **ordinary least squares (OLS)** estimator for a multiple regression.

How to calculate standard errors

The `lm()` function calculates standard errors for you. It follows four steps.

How to calculate standard errors

The `lm()` function calculates standard errors for you. It follows **four steps**.

- (1) After calculating the coefficients, calculate the **sum of squared errors / residuals** (SSE / SSR).

How to calculate standard errors

The `lm()` function calculates standard errors for you. It follows **four steps**.

(1) After calculating the coefficients, calculate the **sum of squared errors / residuals** (SSE / SSR).

(2) Calculate the **degrees of freedom** — a measure of the amount of statistical power in the data — with this formula:

$$DF = N - (K + 1) = N - K - 1$$

where

- ▶ N is the **sample size** in the regression
- ▶ K is the **number of covariates**, and
- ▶ 1 represents the constant.

How to calculate standard errors

(Remember that the formula for the coefficients is best expressed with **matrix algebra**,

$$\hat{B} = (X'X)^{-1}X'Y,$$

where Y is an $(n \times 1)$ column vector containing the values of the dependent variable,

and X is an $(n \times [k + 1])$ matrix that contains each of the k independent variables in a column along with one column of all 1s to represent the intercept.)

How to calculate standard errors

(Remember that the formula for the coefficients is best expressed with **matrix algebra**,

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

where Y is an $(n \times 1)$ column vector containing the values of the dependent variable,

and X is an $(n \times [k + 1])$ matrix that contains each of the k independent variables in a column along with one column of all 1s to represent the intercept.)

(3) Calculate a **variance-covariance matrix** for the coefficients with this formula:

$$V(\beta) = \sqrt{\frac{\text{SSE}}{\text{DF}}} (X'X)^{-1} = \sqrt{\frac{\sum_{i=1}^N \varepsilon_i^2}{N - K - 1}} (X'X)^{-1}$$

How to calculate standard errors

To see this matrix in R, use the `vcov()` command.

	(Intercept)	budget	duration	year
(Intercept)	1.946204e-02	3.044016e-06	-6.325672e-05	-1.463362e-04
budget	3.044016e-06	4.149578e-08	-3.572508e-08	-1.127333e-08
duration	-6.325672e-05	-3.572508e-08	4.907607e-07	1.443062e-07
year	-1.463362e-04	-1.127333e-08	1.443062e-07	1.546119e-06
facenumber_in_poster	-7.611410e-06	6.993496e-08	-2.131794e-07	-5.230102e-07
cast_total_facebook_likes	1.245103e-05	-1.645036e-08	-6.015909e-08	-1.324895e-07
	facenumber_in_poster	cast_total_facebook_likes		
(Intercept)		-7.611410e-06		1.245103e-05
budget		6.993496e-08		-1.645036e-08
duration		-2.131794e-07		-6.015909e-08
year		-5.230102e-07		-1.324895e-07
facenumber_in_poster		5.747075e-05		-5.534960e-07
cast_total_facebook_likes		-5.534960e-07		6.787731e-07

- (4) The **square roots of the diagonal elements** (highlighted) are the **standard errors** for each coefficient.

How to calculate a p -value

Goal: to test whether sufficient evidence exists to **reject the null hypothesis that a coefficient β is 0.**

If we have enough evidence to “reject” this null hypothesis, we can conclude that β really is different from 0, and that x has a real effect on y (assuming everything important is controlled).

How to calculate a *p*-value

Goal: to test whether sufficient evidence exists to **reject the null hypothesis that a coefficient β is 0.**

If we have enough evidence to “reject” this null hypothesis, we can conclude that β really is different from 0, and that x has a real effect on y (assuming everything important is controlled).

Step 1: compute the t statistic for a coefficient β

$$t = \frac{\beta}{\sigma_\beta}$$

t is just the coefficient divided by its standard error.

How to calculate a *p*-value

Goal: to test whether sufficient evidence exists to **reject the null hypothesis that a coefficient β is 0.**

If we have enough evidence to “reject” this null hypothesis, we can conclude that β really is different from 0, and that x has a real effect on y (assuming everything important is controlled).

Step 1: compute the t statistic for a coefficient β

$$t = \frac{\beta}{\sigma_\beta}$$

t is just the coefficient divided by its standard error.

Example:

$$t_{\text{budget}} = \frac{\beta_{\text{budget}}}{\sigma_{\text{budget}}} = \frac{-0.00051}{.0002} = -2.51.$$

How to calculate a *p*-value

We are testing whether this t statistic is close to or far away from 0.

- 0. t is close to 0 when either

How to calculate a *p*-value

We are testing whether this *t* statistic is **close to or far away from**

0. *t* is close to 0 when either

- ▶ the **coefficient is close to 0**, or

How to calculate a *p*-value

We are testing whether this *t* statistic is close to or far away from

0. *t* is close to 0 when either

- ▶ the **coefficient is close to 0**, or
- ▶ the **standard error is very large**.

How to calculate a *p*-value

We are testing whether this *t* statistic is **close to or far away from**

0. *t* is close to 0 when either

- ▶ the **coefficient is close to 0**, or
- ▶ the **standard error is very large**.

So even very large coefficients can fail to be significantly different from 0 if the standard error is too large.

How to calculate a *p*-value

We are testing whether this t statistic is close to or far away from

0. t is close to 0 when either

- ▶ the coefficient is close to 0, or
- ▶ the standard error is very large.

So even very large coefficients can fail to be significantly different from 0 if the standard error is too large.

Step 2: Suppose for a moment that β (and therefore t) were really = 0 in the population / DGP. When you take a random sample, however, β and t will be nonzero, but for purely random reasons.

How to calculate a *p*-value

We are testing whether this t statistic is close to or far away from

0. t is close to 0 when either

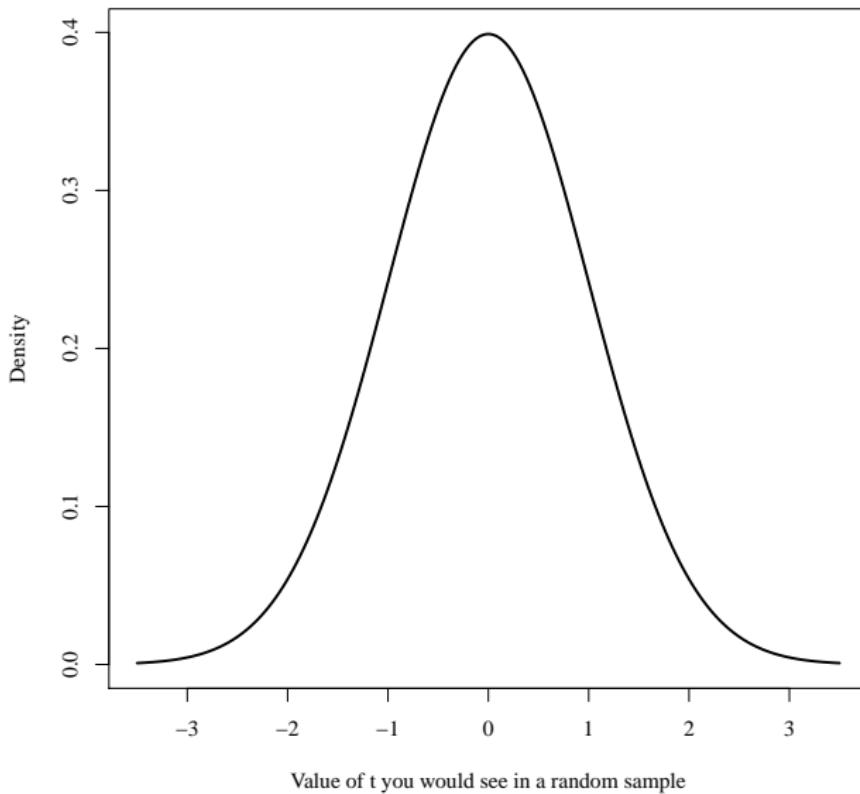
- ▶ the coefficient is close to 0, or
- ▶ the standard error is very large.

So even very large coefficients can fail to be significantly different from 0 if the standard error is too large.

Step 2: Suppose for a moment that β (and therefore t) were really = 0 in the population / DGP. When you take a random sample, however, β and t will be nonzero, but for purely random reasons.

If you take more and more random samples, the t values for a particular coefficient from all of these samples follow (very close to) a standard normal distribution:

Supposing that t=0 in the Population / DGP



How to calculate a *p*-value

Let's say for example you perform a regression on your sample and find that

$$t = \frac{\beta}{\sigma_\beta} = -1.96.$$

(Remember: we're in the hypothetical world in which **the true value of $t=0$.**)

How to calculate a *p*-value

Let's say for example you perform a regression on your sample and find that

$$t = \frac{\beta}{\sigma_\beta} = -1.96.$$

(Remember: we're in the hypothetical world in which **the true value of $t=0$.**)

Step 3: Take the absolute value of t , since we are only interested in the **distance** between t and 0. So in our example $|t| = 1.96$.

How to calculate a *p*-value

Let's say for example you perform a regression on your sample and find that

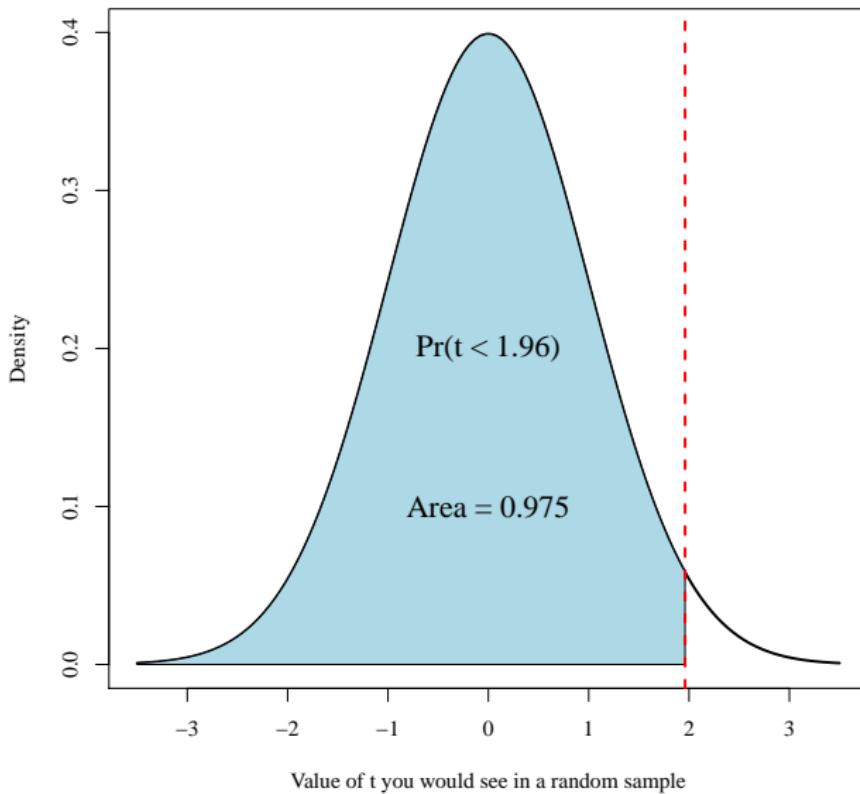
$$t = \frac{\beta}{\sigma_\beta} = -1.96.$$

(Remember: we're in the hypothetical world in which **the true value of $t=0$.**)

Step 3: Take the absolute value of t , since we are only interested in the **distance** between t and 0. So in our example $|t| = 1.96$.

Step 4: Use the standard normal distribution to compute the probability that a sample (in the hypothetical $t=0$ world) would have a t value less than $|t|$:

Supposing that t=0 in the Population / DGP



How to calculate a *p*-value

The probability that the random variable is less than a specified value is the **area under the curve** of the distribution from that value all the way towards the left.

How to calculate a *p*-value

The probability that the random variable is less than a specified value is the **area under the curve** of the distribution from that value all the way towards the left.

You can calculate this probability by plugging the value of $|t|$ into the CDF. For the standard normal CDF we write

$$\Phi(1.96) = .975.$$

How to calculate a *p*-value

The probability that the random variable is less than a specified value is the **area under the curve** of the distribution from that value all the way towards the left.

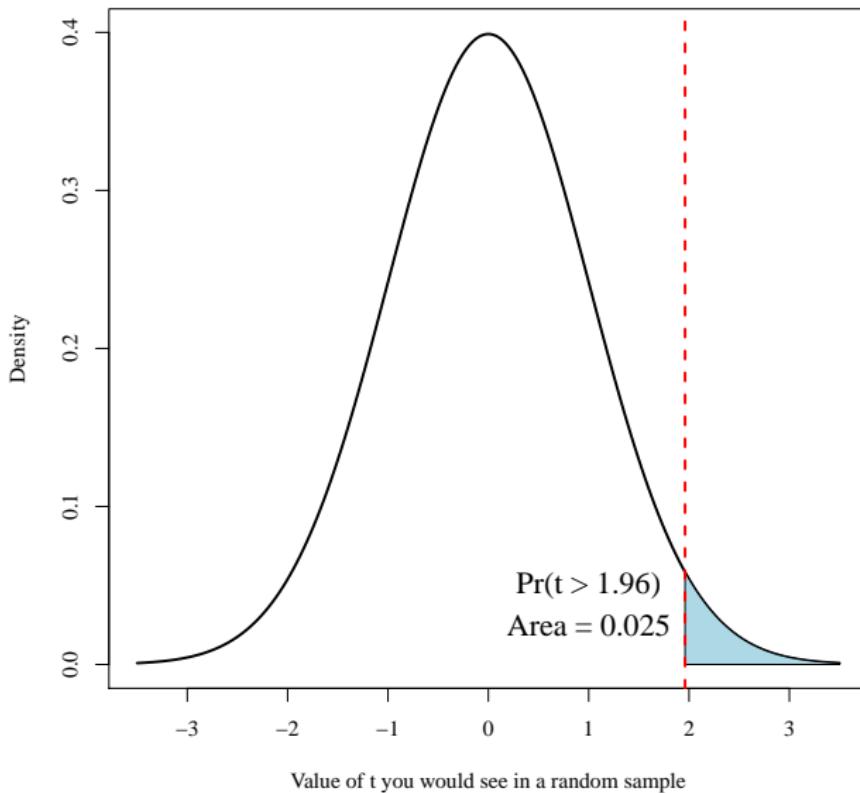
You can calculate this probability by plugging the value of $|t|$ into the CDF. For the standard normal CDF we write

$$\Phi(1.96) = .975.$$

Step 5: Now calculate the probability that a sample (in the hypothetical $t=0$ world) would have a t value **greater** than $|t|$. That's just

$$1 - \Phi(|t|).$$

Supposing that t=0 in the Population / DGP



How to calculate a *p*-value

Step 6: multiply this area by 2.

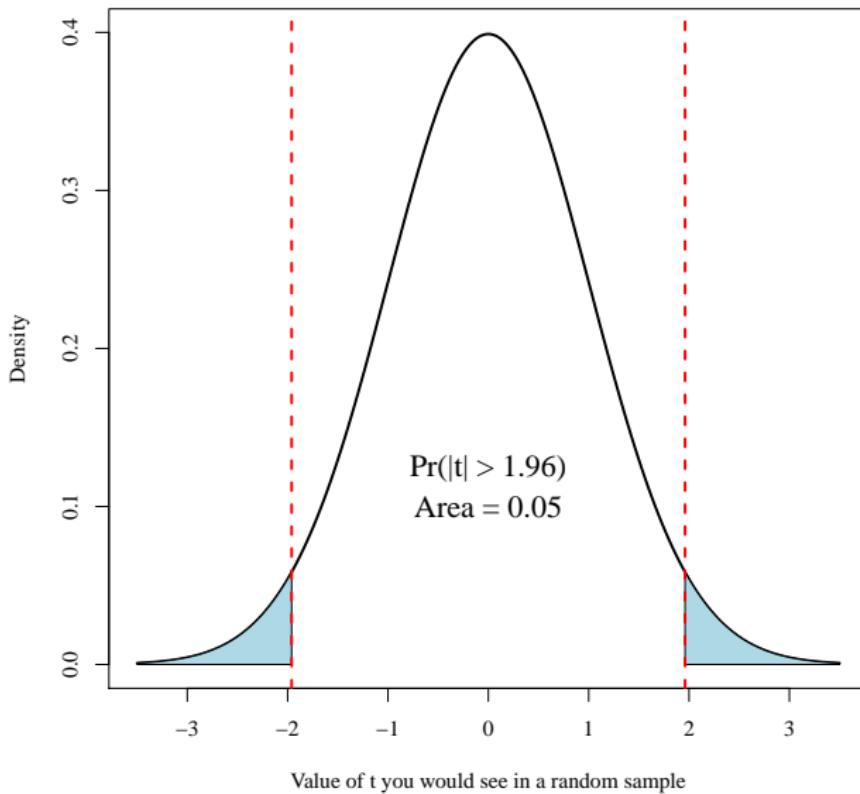
How to calculate a *p*-value

Step 6: multiply this area by 2.

Since we took the absolute value of t , we can achieve the same distance from 0 in either the left or right side of the distribution. So we consider the **other tail** of the distribution and add these areas together.

This result is the *p-value* of the coefficient.

Supposing that t=0 in the Population / DGP



p-values (R)

```
> reg <- lm(imdb_score ~ budget + duration + year + facenumber_in_poster +
+           cast_total_facebook_likes, data = imdb)
> summary(reg)
```

Call:

```
lm(formula = imdb_score ~ budget + duration + year + facenumber_in_poster +
    cast_total_facebook_likes, data = imdb)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7664	-0.5586	0.0990	0.6909	3.0156

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0297394	0.1395064	43.222	< 2e-16 ***
budget	-0.0005112	0.0002037	-2.509	0.0121 *
duration	0.0158795	0.0007005	22.667	< 2e-16 ***
year	-0.0152222	0.0012434	-12.242	< 2e-16 ***
facenumber_in_poster	-0.0401398	0.0075809	-5.295	1.25e-07 ***
cast_total_facebook_likes	0.0054011	0.0008239	6.556	6.16e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.021 on 4517 degrees of freedom
(533 observations deleted due to missingness)

Multiple R-squared: 0.161, Adjusted R-squared: 0.1601
F-statistic: 173.4 on 5 and 4517 DF, p-value: < 2.2e-16

Calculating p -values by hand

Example: $\beta_{\text{budget}} = -.00051$, $\sigma_{\text{budget}} = .0002$. What is the p -value?

Calculating p -values by hand

Example: $\beta_{\text{budget}} = -.00051$, $\sigma_{\text{budget}} = .0002$. What is the p -value?

Step 1: Compute the t statistic

$$t_{\text{budget}} = \frac{\beta_{\text{budget}}}{\sigma_{\text{budget}}} = \frac{-0.00051}{.0002} = -2.51.$$

Calculating p -values by hand

Example: $\beta_{\text{budget}} = -.00051$, $\sigma_{\text{budget}} = .0002$. What is the p -value?

Step 1: Compute the t statistic

$$t_{\text{budget}} = \frac{\beta_{\text{budget}}}{\sigma_{\text{budget}}} = \frac{-0.00051}{0.0002} = -2.51.$$

Steps 2 and 3: suppose this t comes from a standard normal distribution. Take the **absolute value**:

$$|t_{\text{budget}}| = 2.51.$$

Calculating p -values by hand

Example: $\beta_{\text{budget}} = -.00051$, $\sigma_{\text{budget}} = .0002$. What is the p -value?

Step 1: Compute the t statistic

$$t_{\text{budget}} = \frac{\beta_{\text{budget}}}{\sigma_{\text{budget}}} = \frac{-0.00051}{0.0002} = -2.51.$$

Steps 2 and 3: suppose this t comes from a standard normal distribution. Take the **absolute value**:

$$|t_{\text{budget}}| = 2.51.$$

Step 4: Compute the area under the standard normal curve up to $|t|$ (in R, use the `pnorm()` command):

```
> pnorm(t.budget)
      budget
0.9939516
```

Calculating p -values by hand

Step 5: Compute the area under the standard normal curve for values greater than $|t|$:

```
> 1 - pnorm(t.budget)
      budget
0.006048412
```

Calculating p -values by hand

Step 5: Compute the area under the standard normal curve for values greater than $|t|$:

```
> 1 - pnorm(t.budget)
      budget
0.006048412
```

Step 6: multiply by 2

```
> 2*(1 - pnorm(t.budget))
      budget
0.01209682
```

Calculating p -values by hand

Step 5: Compute the area under the standard normal curve for values greater than $|t|$:

```
> 1 - pnorm(t.budget)
      budget
0.006048412
```

Step 6: multiply by 2

```
> 2*(1 - pnorm(t.budget))
      budget
0.01209682
```

So the p -value is 0.012.