

Linear model 3



COLLABORATIVE PLAYLIST

psych252

<https://tinyurl.com/psych252spotify24>

PLAY ...

02/05/2024

Feedback

Feedback

The correlation champion game was fun!

You sped through explaining how to interpret and present the statistics on a research paper. Perhaps if you could give a more in-depth explaining or a recap of it, and also why you chose to show a particular statistics as opposed to others generated by R. This portion is extremely important for us. Thank you!

Things that came up

n-1 in the sample variance

Instructions for HW3 slightly misleading? #28



Bendix Kemmann
5 days ago in Homework - HW3

PIN

STAR

WATCH

62
VIEWS



3

The sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$ is an unbiased estimator of the population standard deviation. In this exercise, we will run a simulation to compare s with another estimator $s' = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n}}$ and show that s' is biased.

If I'm not mistaken, these instructions in HW3 are strictly speaking inaccurate. What is unbiased is not the sample *standard deviation*, but rather the sample *variance*. One way of seeing that the sample standard deviation is still a biased estimator of the standard deviation of the population is to consider that s_{n-1}^2 is an unbiased estimator for the variance σ_2 of the population. As the square root function is non-linear, the sample standard deviation is actually still biased (the expected value is *not* equal to the true value). In particular, the square root of the sample variance is an underestimate, as the very simulation in this exercise shows (increase the number of simulations to 100 million, say, and the estimate will still be slightly below the stipulated 'ground truth').

Comment Edit Delete Endorse ...

1 Answer



Tobi Gerstenberg STAFF
3 days ago



1

Thank for the post Bendix. You're right -- that's a very good point!! Thanks for spotting this. We will update the homework next time around, and I'll talk about this briefly in the "Things that came up section" on Monday's class.



Comment Edit Delete Endorse ...

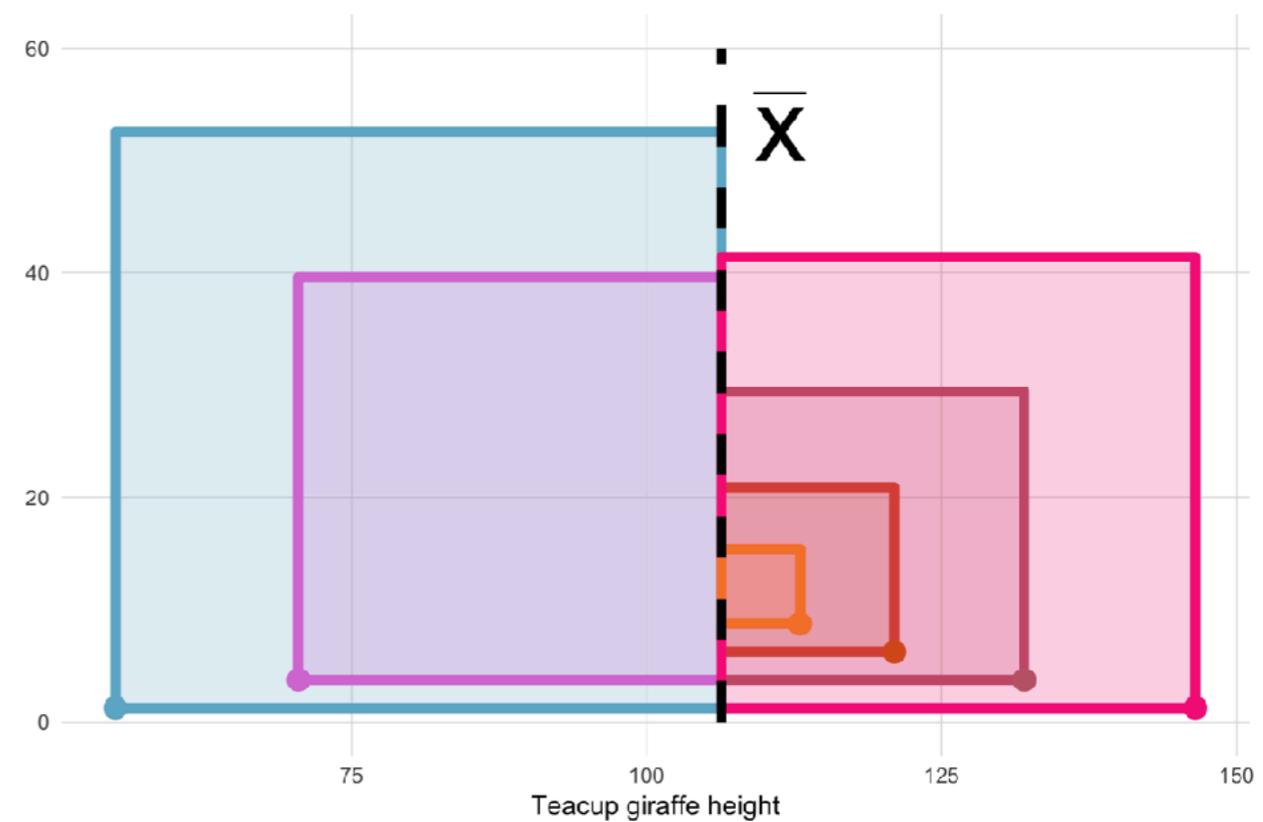
<https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/measuring-spread-quantitative/v/sample-standard-deviation-and-bias>

$n-1$ in sample variance

$$\text{Var}(Y) = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

definition of the sample variance

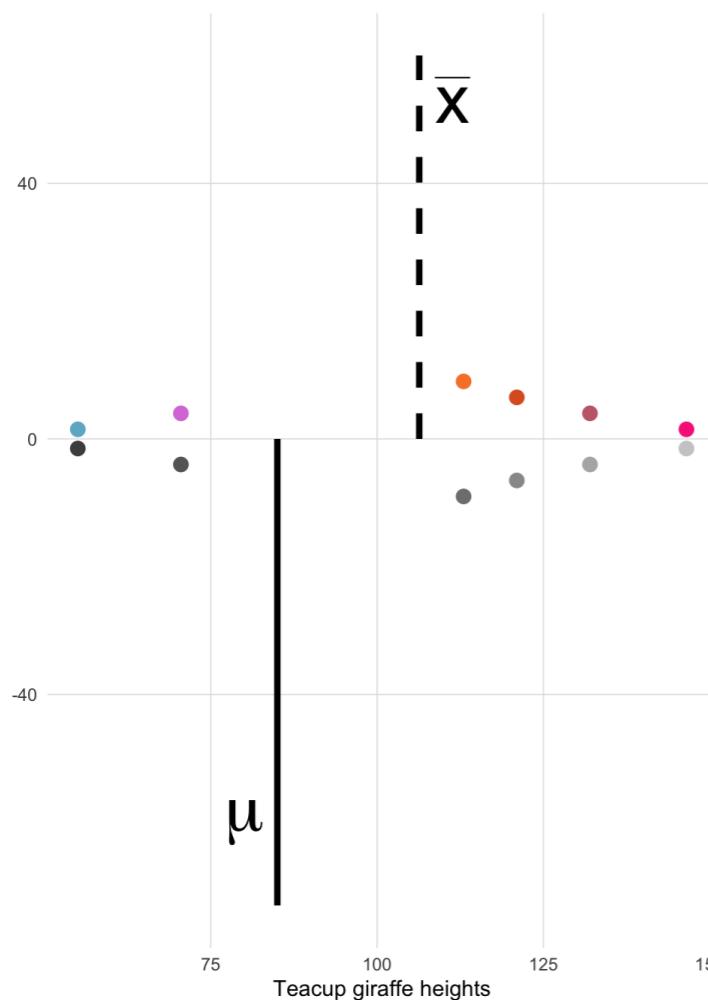
sum of squares =
squared deviation
from the mean



$n-1$ in standard deviation

$$\text{Var}(Y) = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

definition of the sample variance



\bar{x} = sample mean

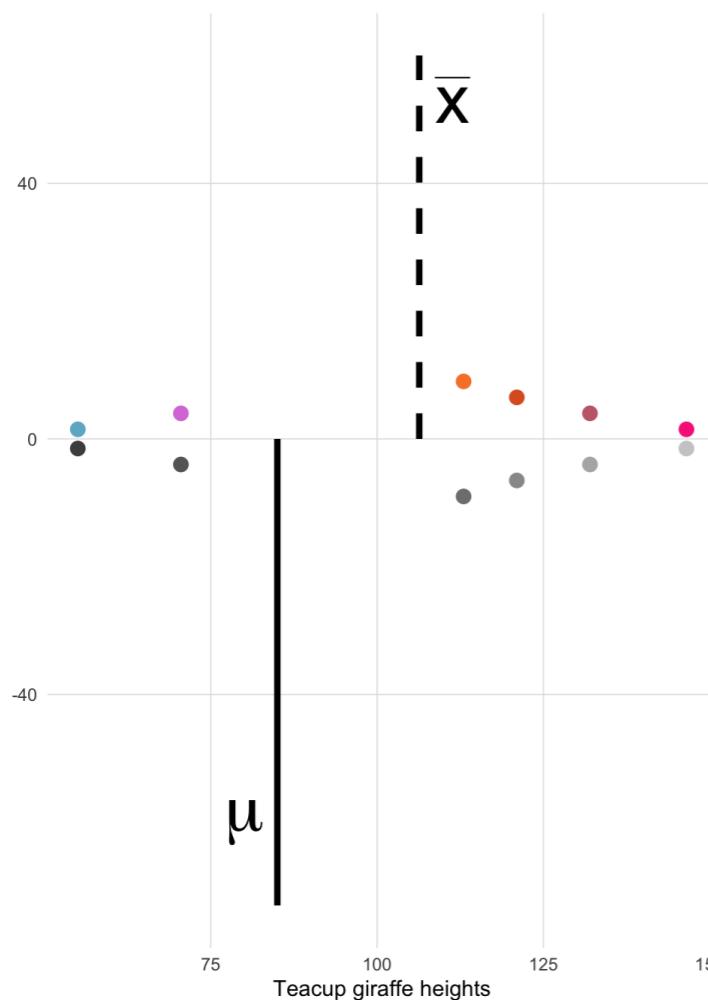
- the sample mean will be somewhat different from the population mean
- this means, we are likely to underestimate the true variance

μ = population mean

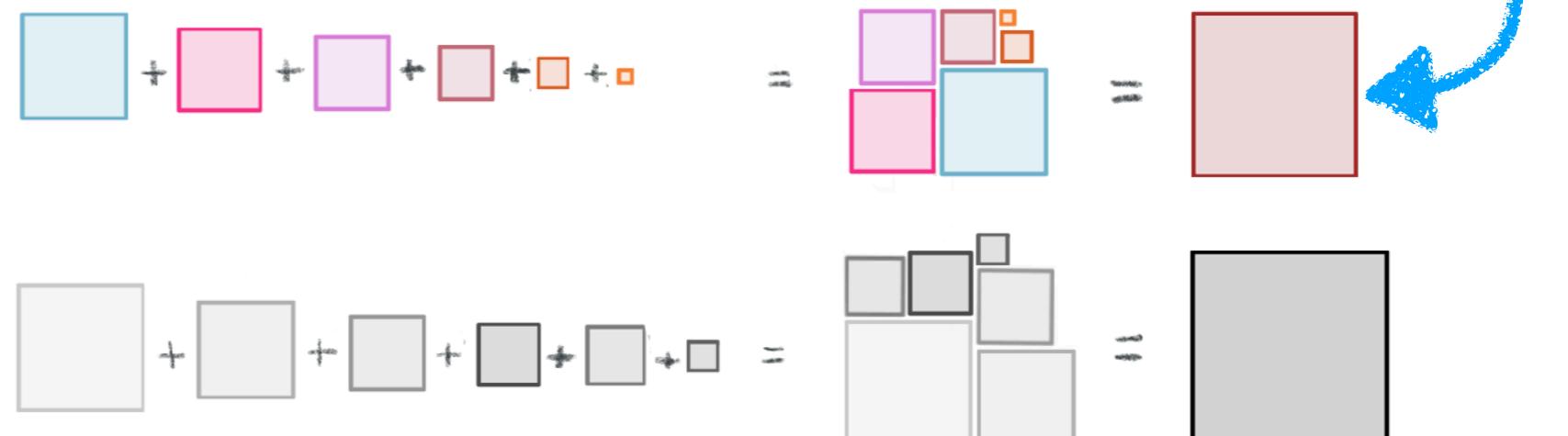
$n-1$ in standard deviation

$$\text{Var}(Y) = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

definition of the sample variance



\bar{x} = sample mean



μ = population mean

this area
is smaller

Plan for today

- Quick recap
- Multiple regression
 - two continuous predictors
 - one categorical predictor
 - one continuous and one categorical predictor
- Interactions
- `lm()` output

Quick recap

Quick recap: Regression (conceptual)

Linear model: Simple regression

$$\text{Data} = \text{Model} + \text{Error}$$

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

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

↑
the model is a linear combination of predictors

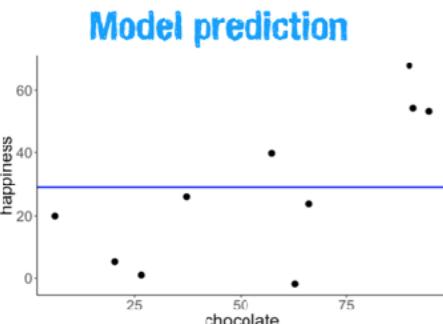
The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

H_0 : Chocolate consumption and happiness are unrelated.

Model C

$$Y_i = \beta_0 + \epsilon_i$$



Fitted model

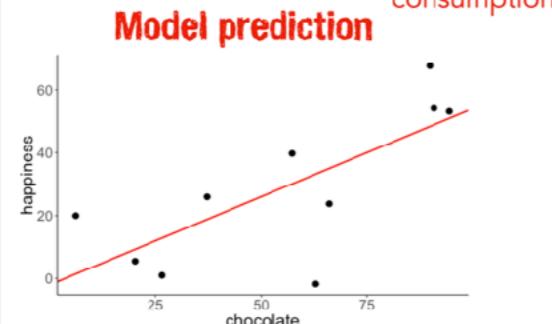
$$Y_i = 28.88 + e_i$$

H_1 : Chocolate consumption and happiness are related.

Model A

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

Model prediction



Fitted model

$$Y_i = -2.04 + 0.56X_i + e_i$$

Decide whether it's **worth it**

- To compute the F statistic, we need:

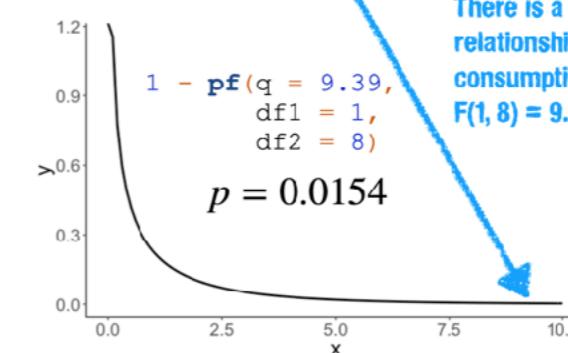
- PRE = 0.54
- PC = 1
- PA = 2
- $n = 10$

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

$$= \frac{0.54/(2 - 1)}{(1 - 0.54)/(10 - 2)}$$

$$= 9.39$$

There is a significant relationship between chocolate consumption and happiness
 $F(1, 8) = 9.39, p = .0154$.



Quick recap: Regression (in R)

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

$$\text{balance}_i = b_0 + b_1 \cdot \text{income}_i + e_i$$

```
fit = lm(formula = balance ~ 1 + income, data = df.credit)
```

outcome
intercept
predictor
data
(doesn't need to be specified explicitly)

broom: turn messy model outputs into tidy TIBBLES!



@allisonhorst

anova (fit_c, fit_a)

```
Analysis of Variance Table

Model 1: balance ~ 1
Model 2: balance ~ 1 + income
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     399 84339912
2     398 66208745  1  18131167 108.99 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\text{PRE} = 1 - \frac{66208745}{84339912} \approx 0.215$$

The augmented model reduces the error by 21.5%.

```
lm(balance ~ 1 + income, data = df.credit) %>%
  summary()
```

R²

```
Residual standard error: 407.9 on 398 degrees of freedom
Multiple R-squared:  0.215, Adjusted R-squared:  0.213
F-statistic: 109 on 1 and 398 DF,  p-value: < 2.2e-16
```

The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)

```
fit_c = lm(formula = balance ~ 1,
           data = df.credit)
```

2. Fit model parameters to the data

```
fit_a = lm(formula = balance ~ 1 + income,
           data = df.credit)
```

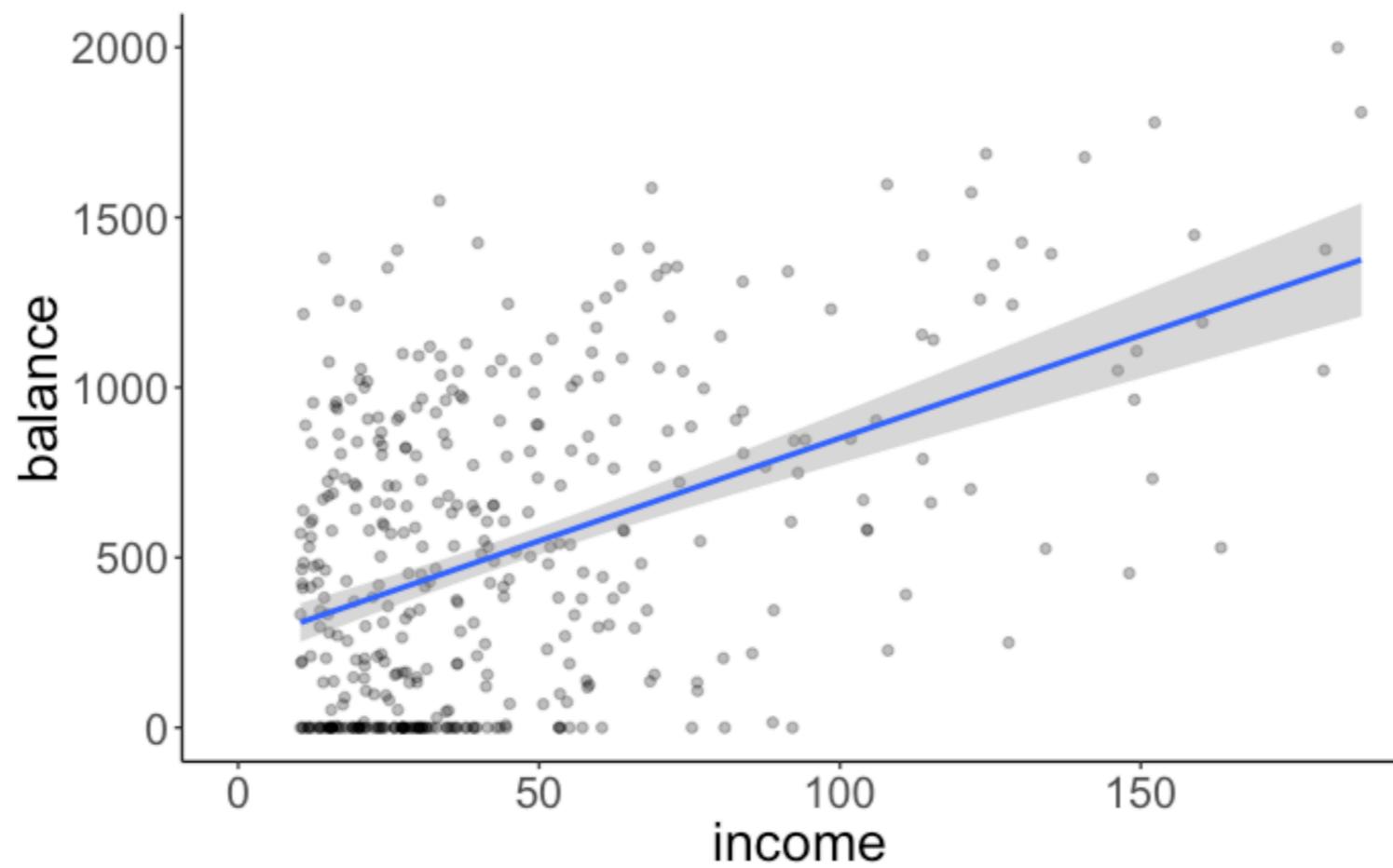
3. Calculate the proportional reduction of error (PRE) in our sample

anova (fit_c, fit_a)

4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

Quick recap: Reporting results

plot



statistical

test There is a significant relationship between a person's income and the average balance on their credit cards

$$F(1, 389) = 108.99, p < .001, r = .463.$$

effect size
measure

With each additional \$1000 of income, the average balance is predicted to increase by \$6.05 [4.91, 7.19] (95% CI).

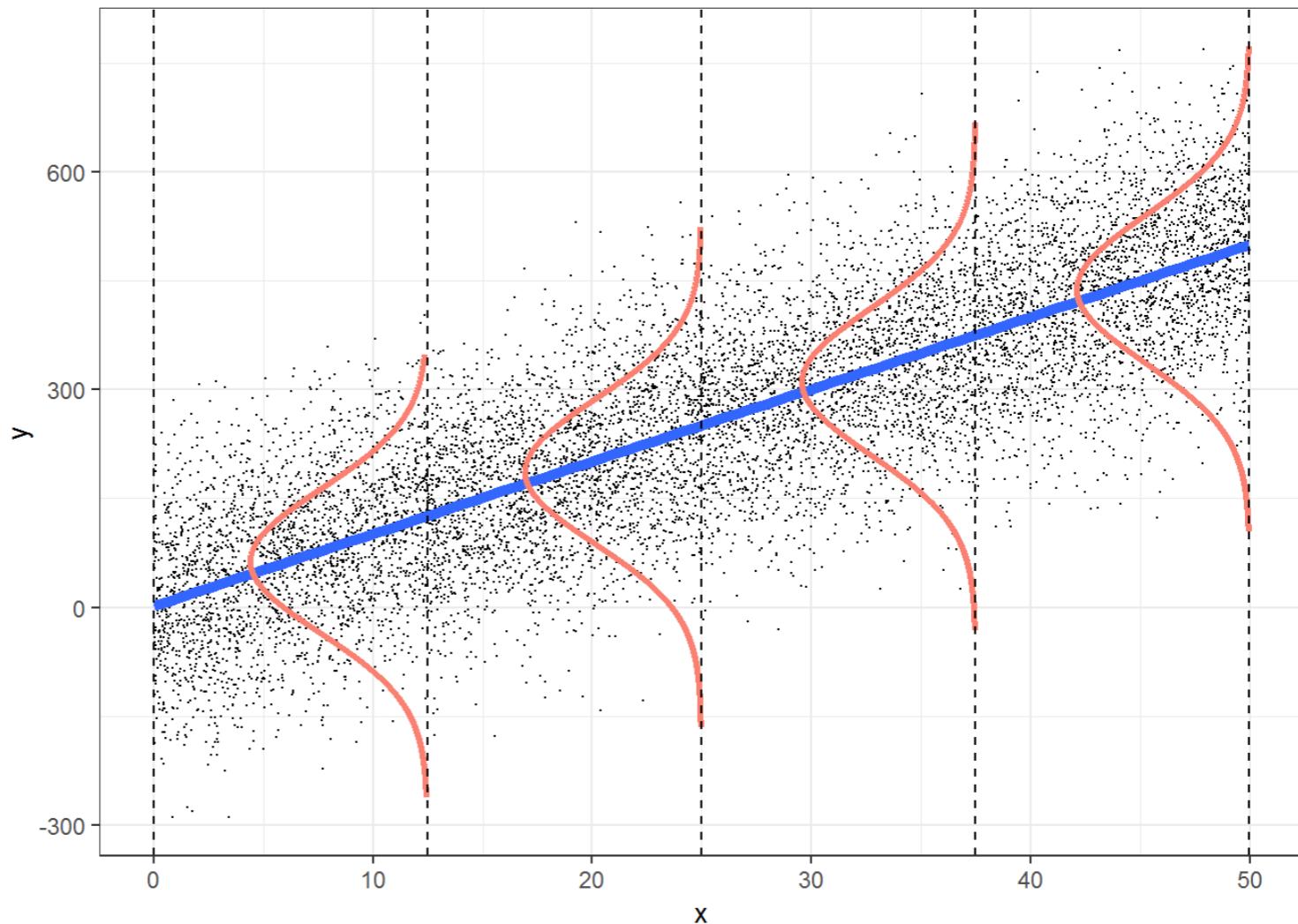
parameter estimate with confidence interval

Multiple regression

Model assumptions of **simple** regression

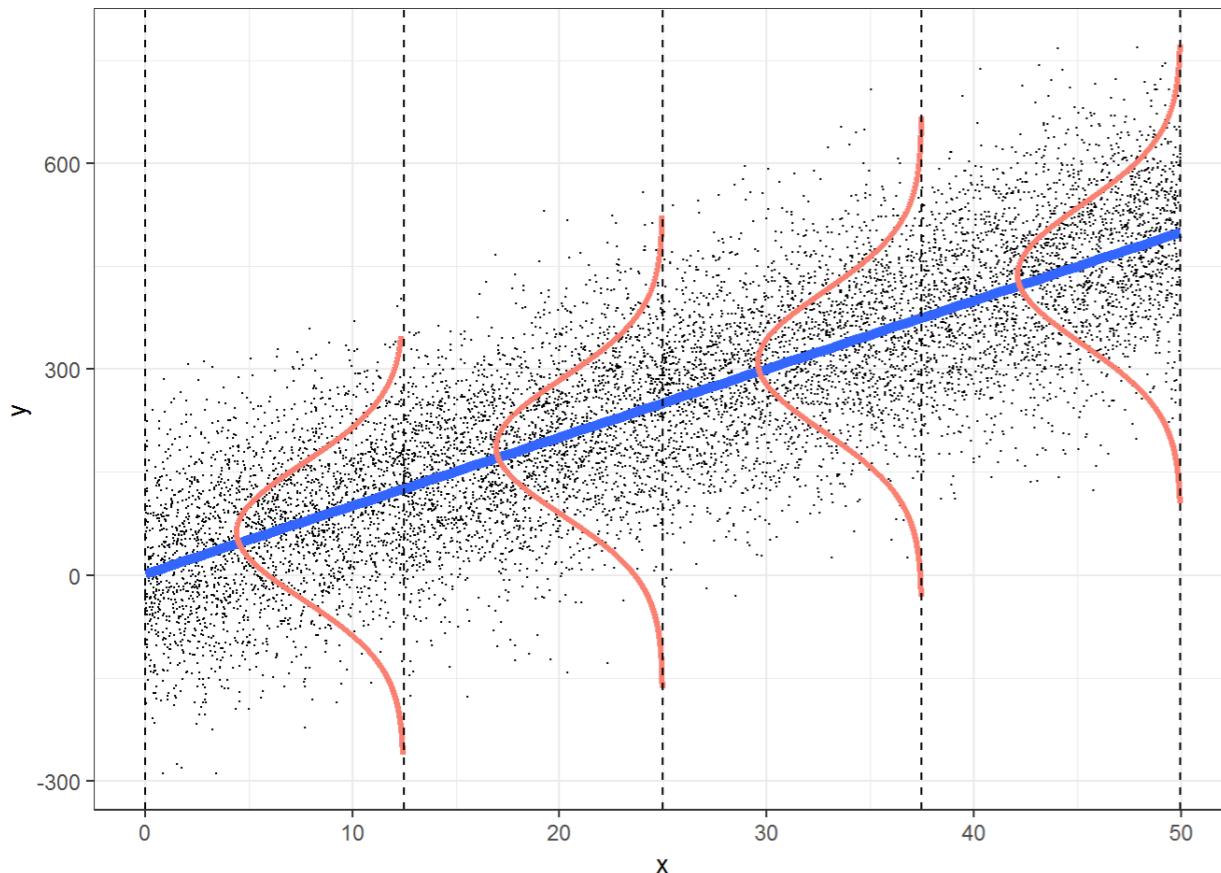
- independent observations
- Y is continuous
- errors are normally distributed
- errors have constant variance
- error terms are uncorrelated

the dependent variable doesn't need to be normally distributed!



Assumptions of multiple regression

- independent observations
- Y is continuous
- errors are normally distributed
- errors have constant variance
- error terms are uncorrelated
- **no multicollinearity**



predictors in the model should not be highly correlated with each other



Linear model

Data = Model + Error

Simple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

one predictor

Multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \epsilon_i$$

many predictors

Advertising data set

money spent on
different media
(x \$1000)

sales
(x 1000)

- Combine several predictors to explain an outcome variable of interest

index	tv	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75.0	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1.0	4.8
10	199.8	2.6	21.2	10.6

Model C

Simple regression

$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + e_i$$

Model A

Multiple regression

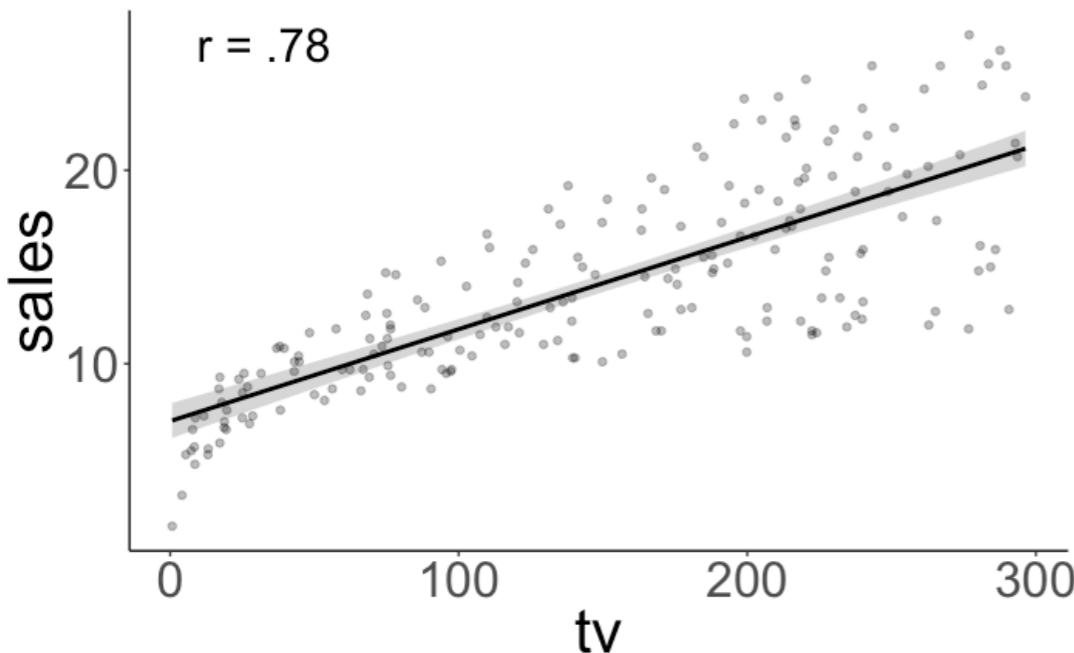
$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + b_2 \cdot \text{radio}_i + e_i$$

Can we predict sales better when we consider radio ads in addition to TV ads?

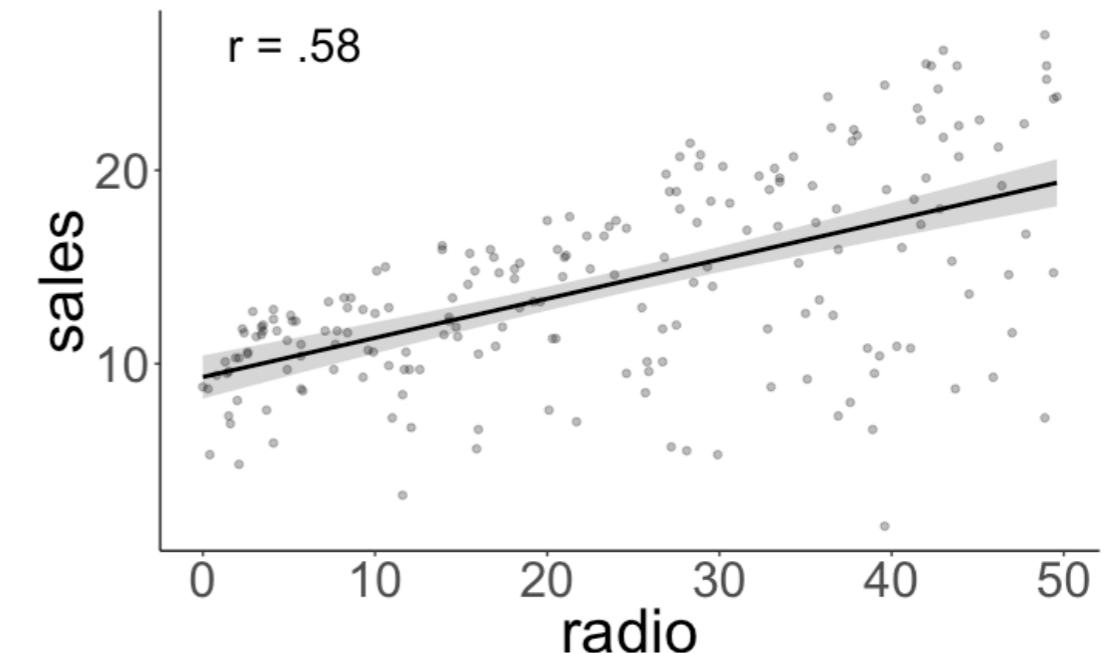
"Controlling" for TV ads, do radio ads explain any of the additional variance in sales?

$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + b_2 \cdot \text{radio}_i + e_i$$

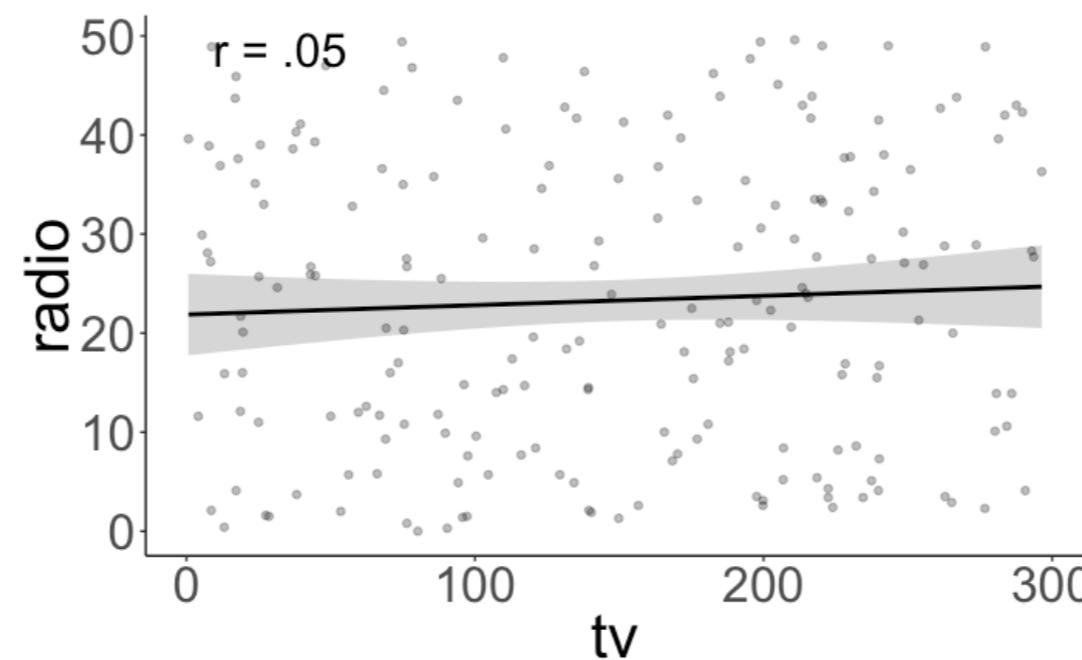
Relationship between TV ads and sales



Relationship between radio ads and sales



Relationship between TV ads and radio ads



predictors are not correlated, yay!

Visualizing correlations

```
library("corr")
```



```
1 df.credit %>%
2   select_if(is.numeric) %>%
3   correlate() %>%
4   rearrange() %>%
5   shave() %>%
6   fashion()
```

rowname	index	newspaper	radio	sales	tv
index					
newspaper	-0.15				
radio	-0.11	0.35			
sales	-0.05	0.23	0.58		
tv	0.02	0.06	0.05	0.78	

Can we predict sales better when we consider radio in addition to TV ads?

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

H_0 : Radio ads and sales are not related once we control for TV ads.

H_1 : Radio ads and sales are related even when we control for TV ads.

Model C

$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + e_i$$

Model A

$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + b_2 \cdot \text{radio}_i + e_i$$

```
1 # fit the models
2 fit_c = lm(sales ~ 1 + tv, data = df.ads)
3 fit_a = lm(sales ~ 1 + tv + radio, data = df.ads)
4
5 # do the F test
6 anova(fit_c, fit_a)
```

we reject the H_0

Analysis of Variance Table

Model 1: sales ~ 1 + tv

Model 2: sales ~ 1 + tv + radio

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	198	2102.53			
2	197	556.91	1	1545.6	546.74 < 2.2e-16 ***

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

Evaluating the model: Model fit

fit_a %>%
glance()

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.897	0.896	1.681	859.618	0	3	-386.197	780.394	793.587	556.914	197

r.squared	The percent of variance explained by the model
adj.r.squared	r.squared adjusted based on the degrees of freedom
sigma	The square root of the estimated residual variance
statistic	F-statistic
p.value	p-value from the F test, describing whether the full regression is significant
df	Degrees of freedom used by the coefficients
logLik	the data's log-likelihood under the model
AIC	the Akaike Information Criterion
BIC	the Bayesian Information Criterion
deviance	deviance
df.residual	residual degrees of freedom

Evaluating the model: Model fit

```
fit_a %>%  
  glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.897	0.896	1.681	859.618	0	3	-386.197	780.394	793.587	556.914	197

Compact Model

$$\text{sales}_i = b_0 + e_i$$

The augmented model reduces 89.7% of the error compared to a compact model that just predicts the mean.

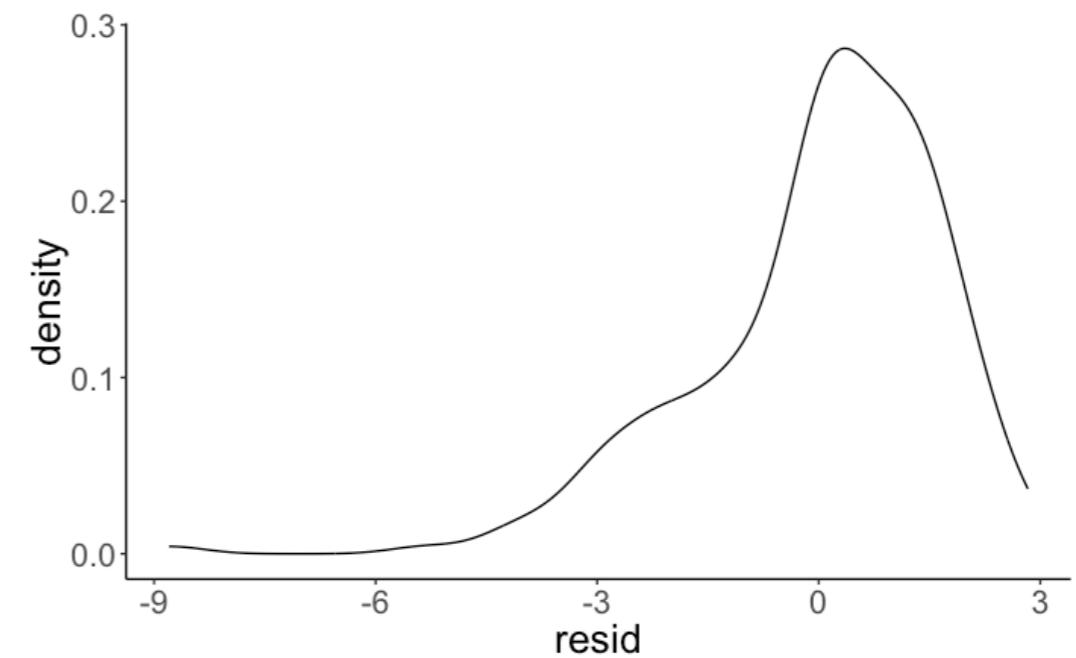
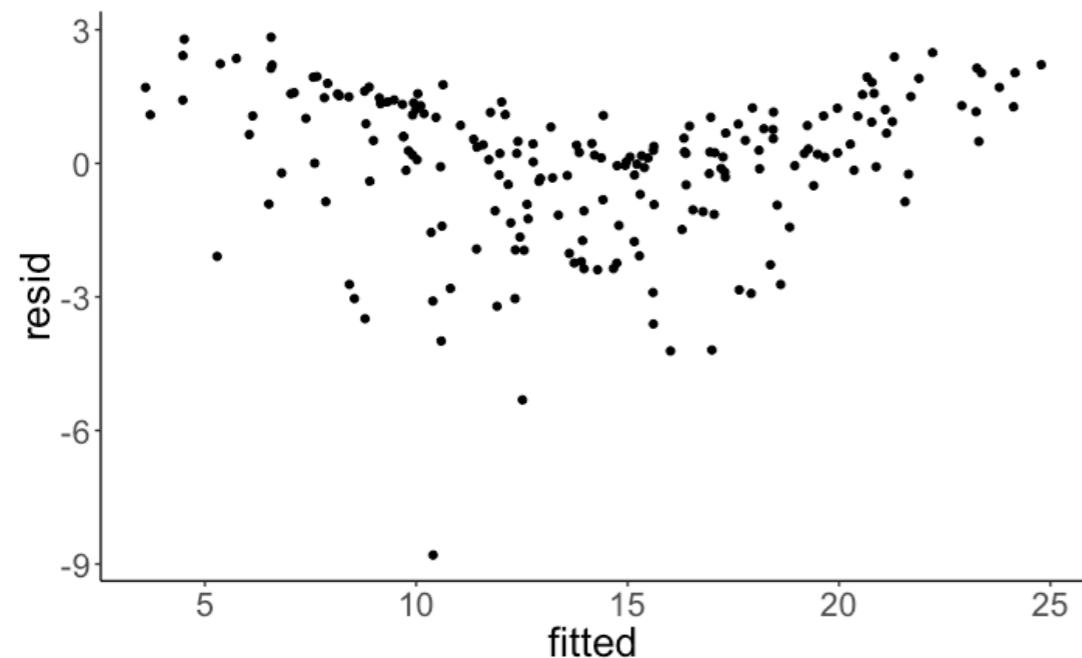
Augmented Model

$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + b_2 \cdot \text{radio}_i + e_i$$

$$\text{PRE} = 1 - \frac{\text{SSE}(A)}{\text{SSE}(C)} = R^2$$

Evaluating the model: Residual plots

resid = sales - fitted



OKish overall

Interpreting the results

```
fit_a %>%
  tidy(conf.int = T)
```

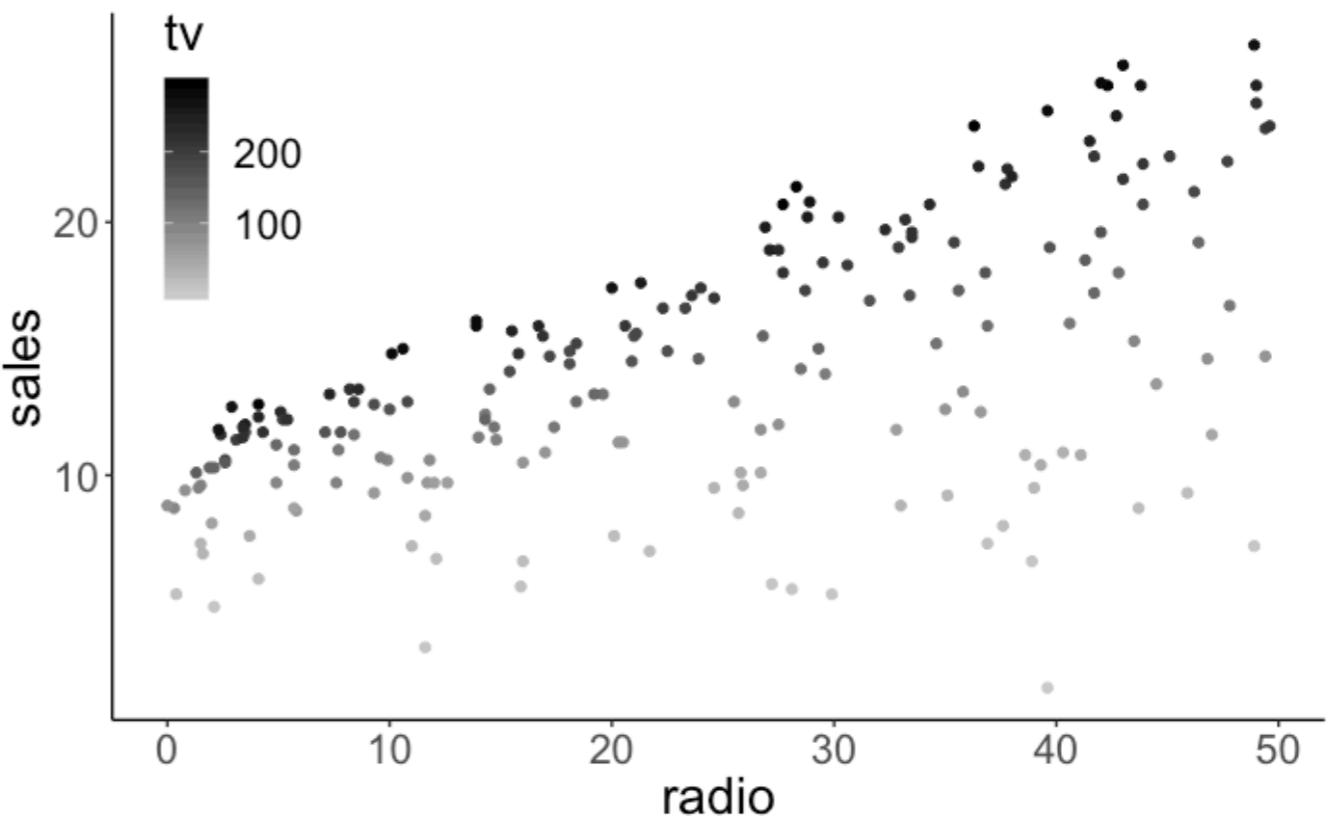
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.92	0.29	9.92	0	2.34	3.50
tv	0.05	0.00	32.91	0	0.04	0.05
radio	0.19	0.01	23.38	0	0.17	0.20

$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + b_2 \cdot \text{radio}_i + e_i$$

$$\widehat{\text{sales}}_i = 2.92 + 0.05 \cdot \text{tv}_i + 0.19 \cdot \text{radio}_i$$

For a given amount of TV advertising, an additional \$1000 on radio advertising is associated with an increase in sales by 190 units.

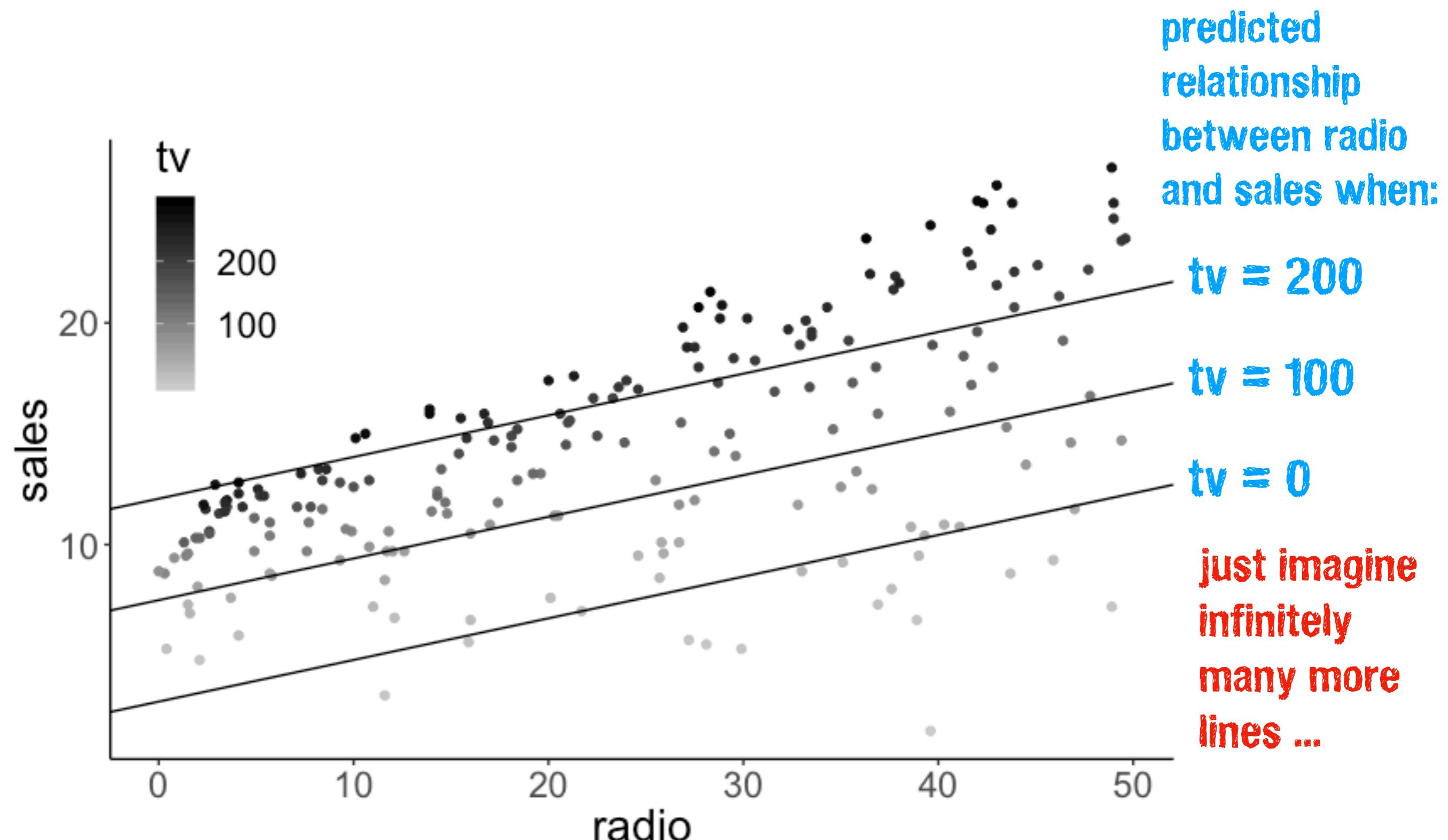
Reporting results



There is a significant relationship between sales and radio ads, controlling for TV ads $F(1, 197) = 546.74, p < .001$.

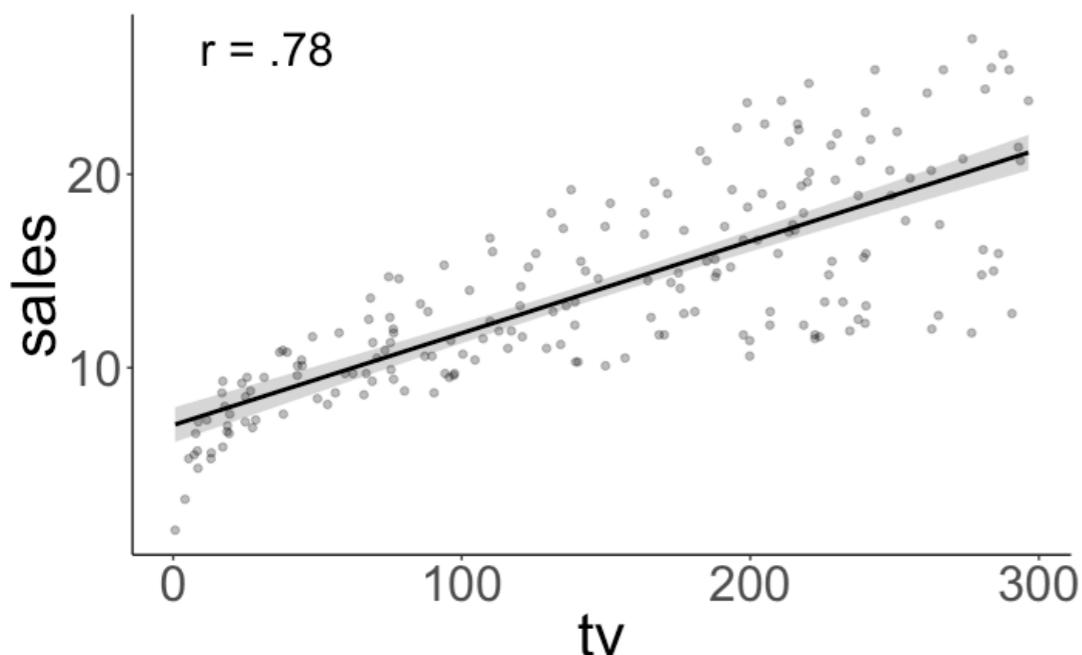
Holding TV ads fixed, an increase in \$1000 on radio ads is associated with an increase in sales by 190 units [170, 200] (95% confidence intervals).

Visualizing the results

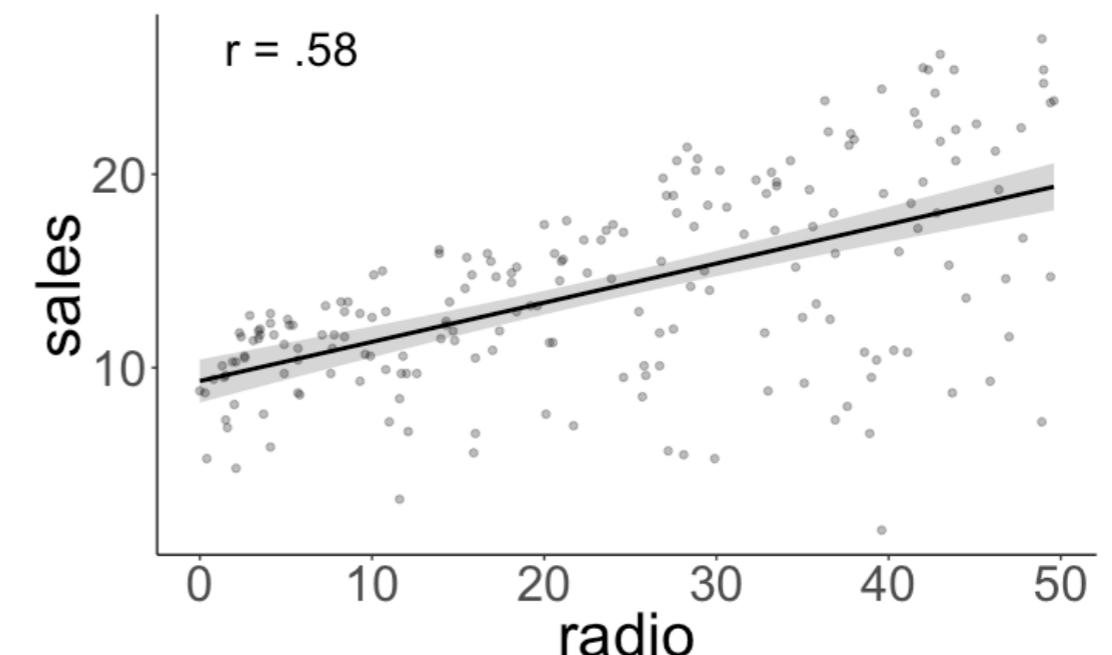


**Why can't I just run several
simple regressions?**

Relationship between TV ads and sales



Relationship between radio ads and sales



We found that both TV ads and radio ads were related to sales.

But did we need to run a multiple regression? Could we not just have looked at correlations?

Advertising data set

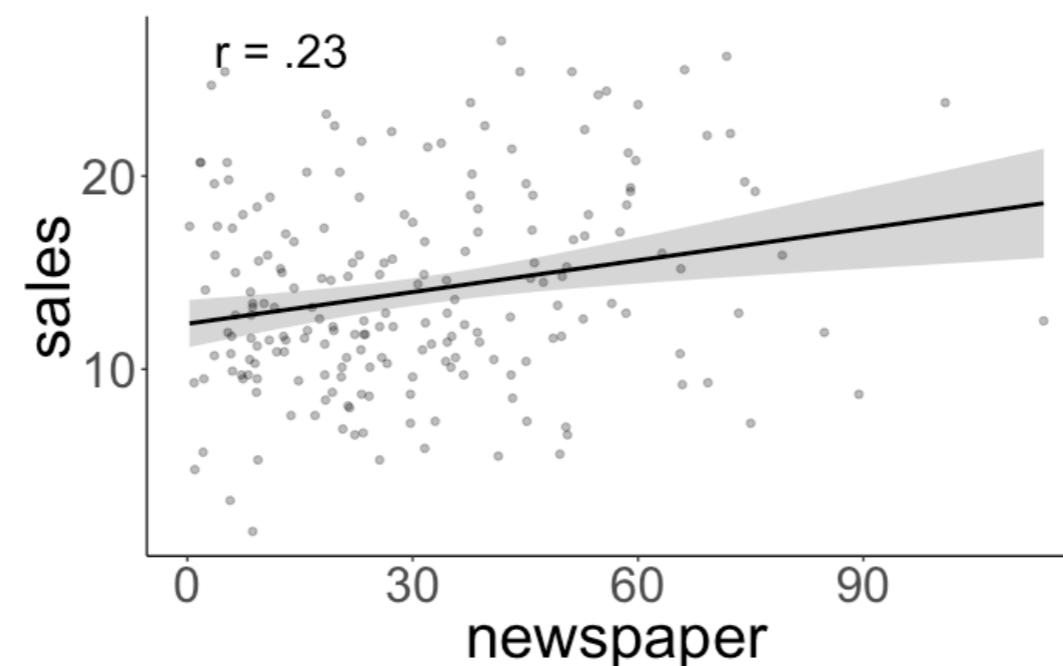
money spent on
different media
(x \$1000)

sales
(x1000)

index	tv	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75.0	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1.0	4.8
10	199.8	2.6	21.2	10.6

Are newspaper ads and sales related when controlling for radio ads and TV ads?

Relationship between newspaper ads and sales



this
is
significant

```

1 # fit the models
2 fit_c = lm(sales ~ 1 + tv + radio, data = df.ads)
3 fit_a = lm(sales ~ 1 + tv + radio + newspaper, data = df.ads)
4
5 # do the F test
6 anova(fit_c, fit_a)

```

Analysis of Variance Table

Model 1: sales ~ 1 + tv + radio

Model 2: sales ~ 1 + tv + radio + newspaper

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	197	556.91				
2	196	556.83	1	0.088717	0.0312	0.8599

it's not worth it

sales ~ 1 + tv

sales ~ 1 + tv + newspaper

it's worth it

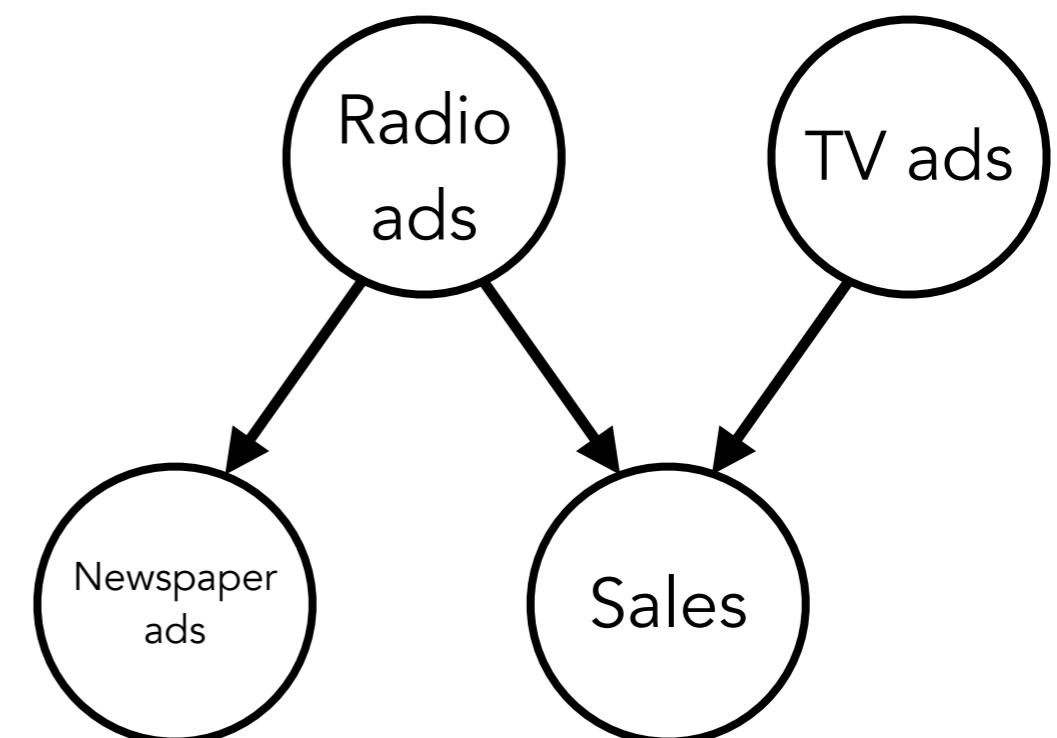
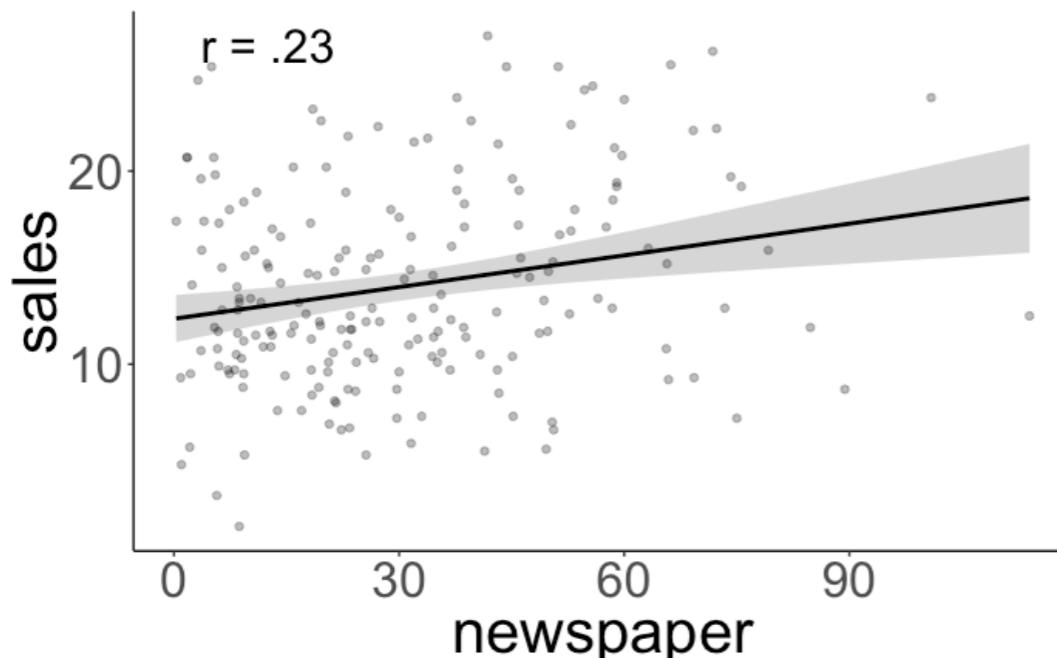
sales ~ 1 + radio

sales ~ 1 + radio + newspaper

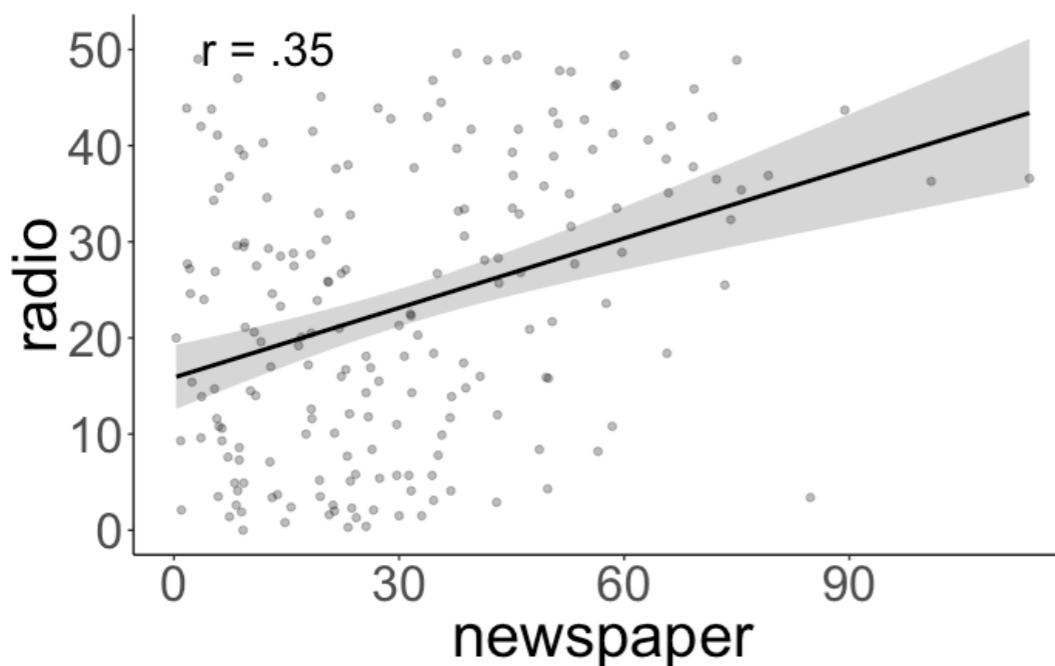
it's not worth it

Are newspaper ads and sales related when controlling for radio ads and TV ads?

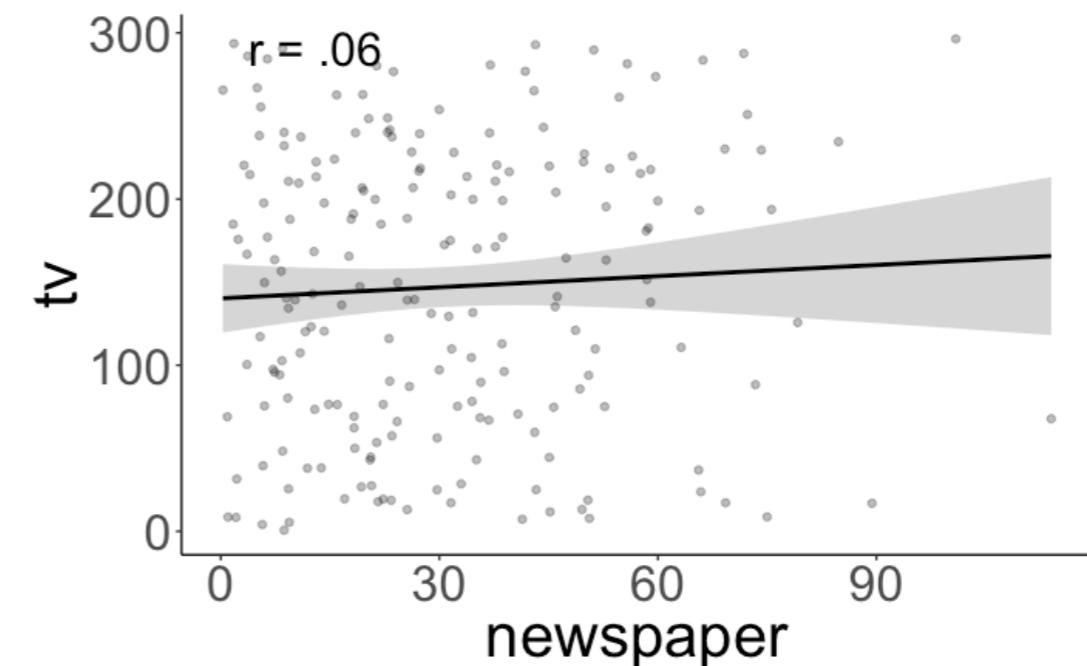
Relationship between newspaper ads and sales



Relationship between newspaper and radio ads

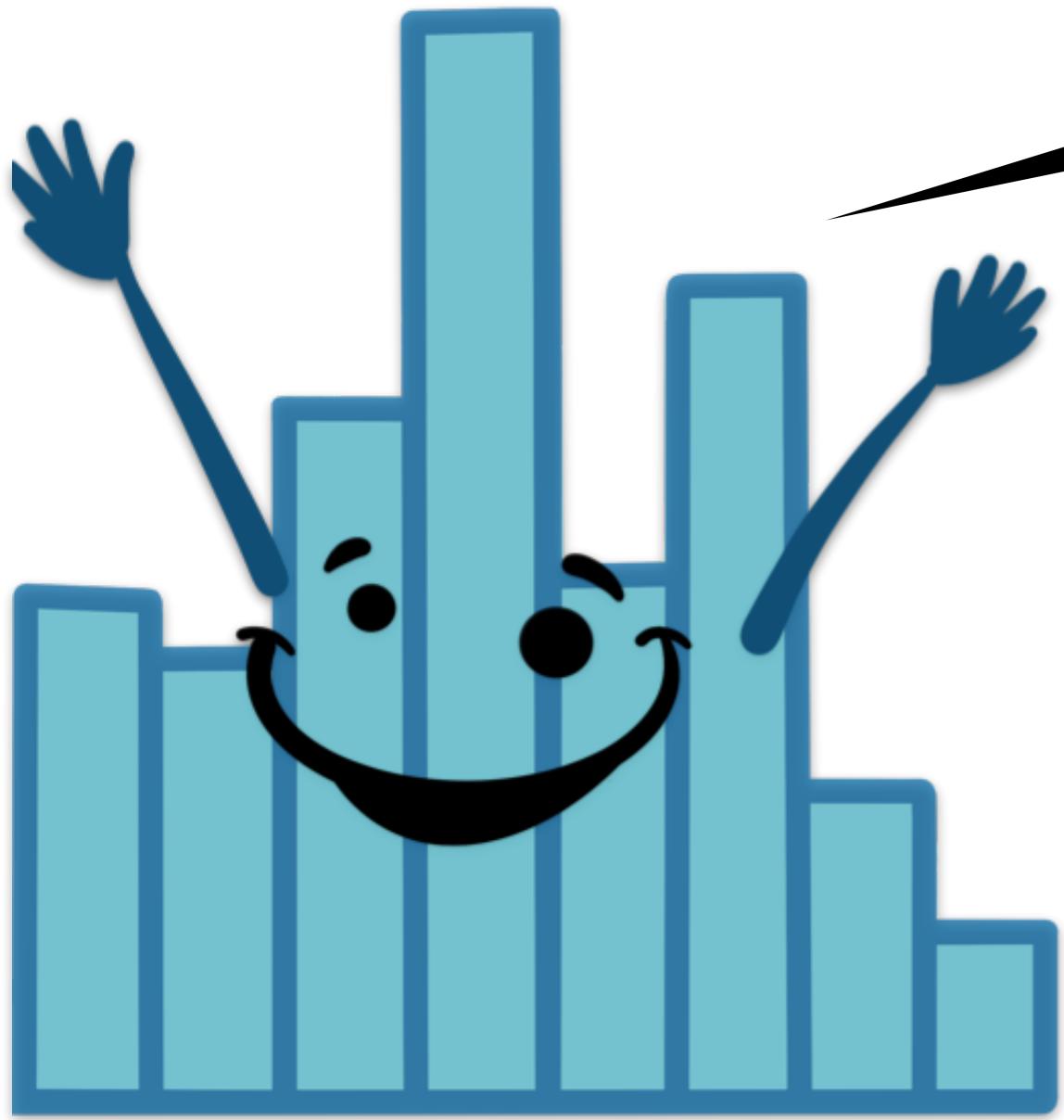


Relationship between newspaper and TV ads



02:00

stretch break!



Categorical predictors

Credit data set

df.credit

index	income	limit	rating	cards	age	education	gender	student	married	ethnicity	balance
1	14.89	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.03	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.59	7075	514	4	71	11	Male	No	No	Asian	580
4	148.92	9504	681	3	36	11	Female	No	No	Asian	964
5	55.88	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	80.18	8047	569	4	77	10	Male	No	No	Caucasian	1151
7	21.00	3388	259	2	37	12	Female	No	No	African American	203
8	71.41	7114	512	2	87	9	Male	No	No	Asian	872
9	15.12	3300	266	5	66	13	Female	No	No	Caucasian	279
10	71.06	6819	491	3	41	19	Female	Yes	Yes	African American	1350

nrow(df.credit) = 400

Do students have a different credit card balance from non-students?

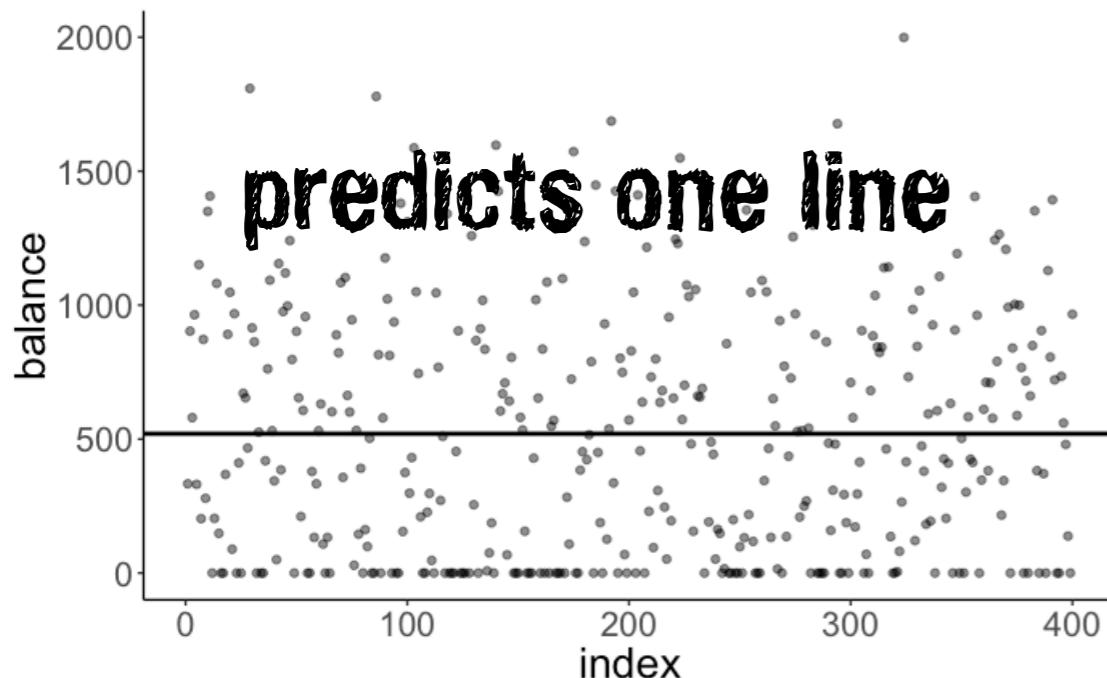
variable	description
income	in thousand dollars
limit	credit limit
rating	credit rating
cards	number of credit cards
age	in years
education	years of education
gender	male or female
student	student or not
married	married or not
ethnicity	African American, Asian, Caucasian
balance	average credit card debt in dollars

H_0 : Students and non-students have the same balance.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 520.02 + e_i$$

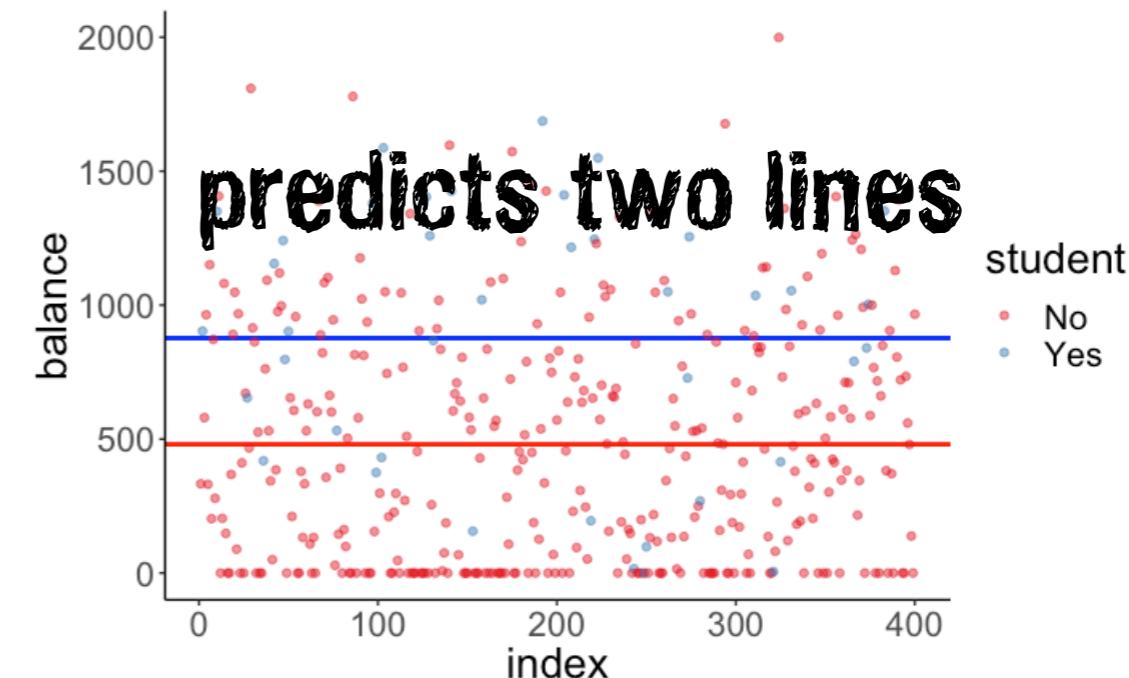
H_1 : Students and non-students have different balances.

Model A

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

 student

Model prediction



Fitted model

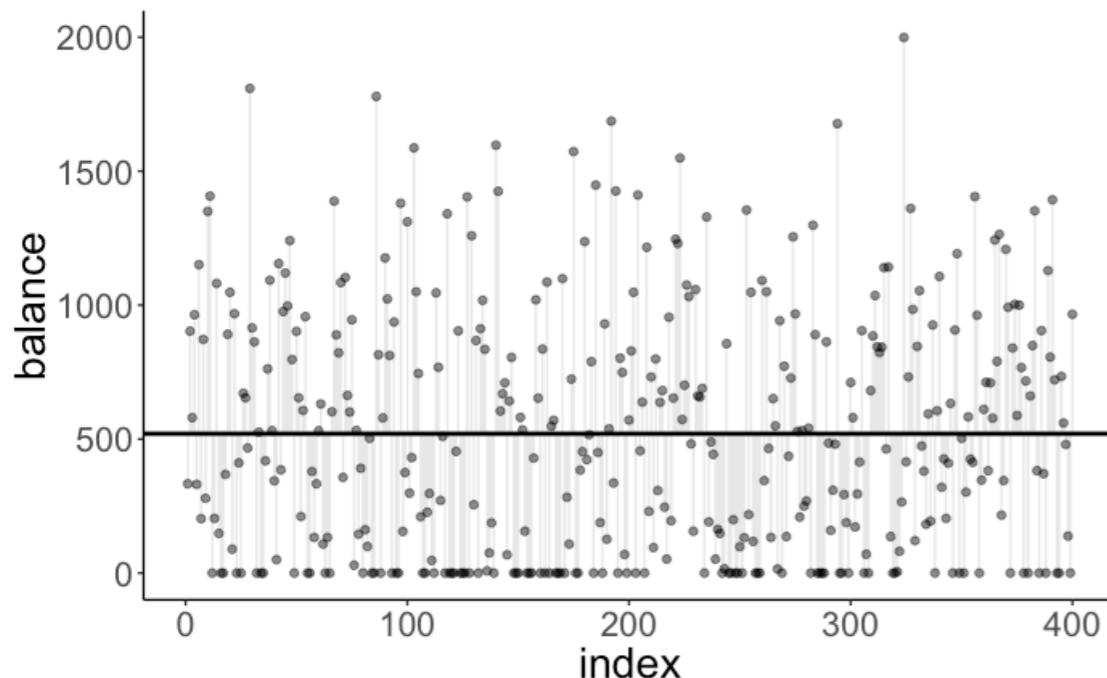
$$Y_i = 480.37 + 396.46X_i + e_i$$

H_0 : Students and non-students have the same balance.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 520.02 + e_i$$

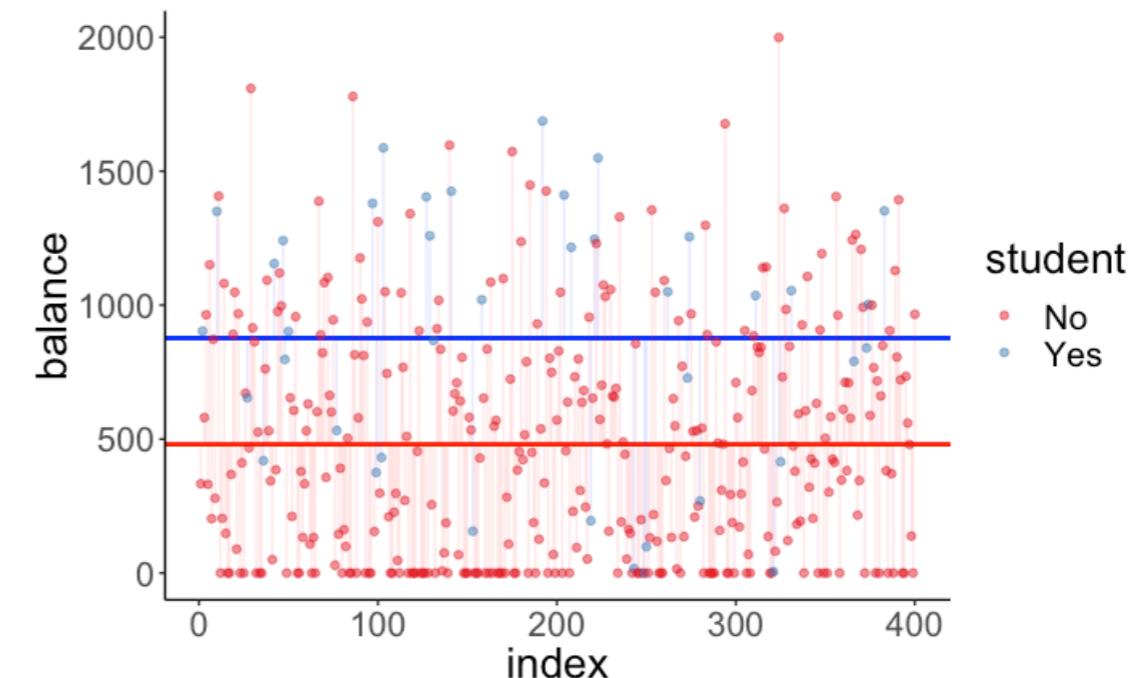
H_1 : Students and non-students have different balances.

Model A

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

 student

Model prediction



Fitted model

$$Y_i = 480.37 + 396.46X_i + e_i$$

Worth it?

```
1 # fit the models  
2 fit_c = lm(balance ~ 1, data = df.credit)  
3 fit_a = lm(balance ~ 1 + student, data = df.credit)  
4  
5 # run the F test  
6 anova(fit_c, fit_a)
```

Analysis of Variance Table

Worth it!

Model 1: balance ~ 1

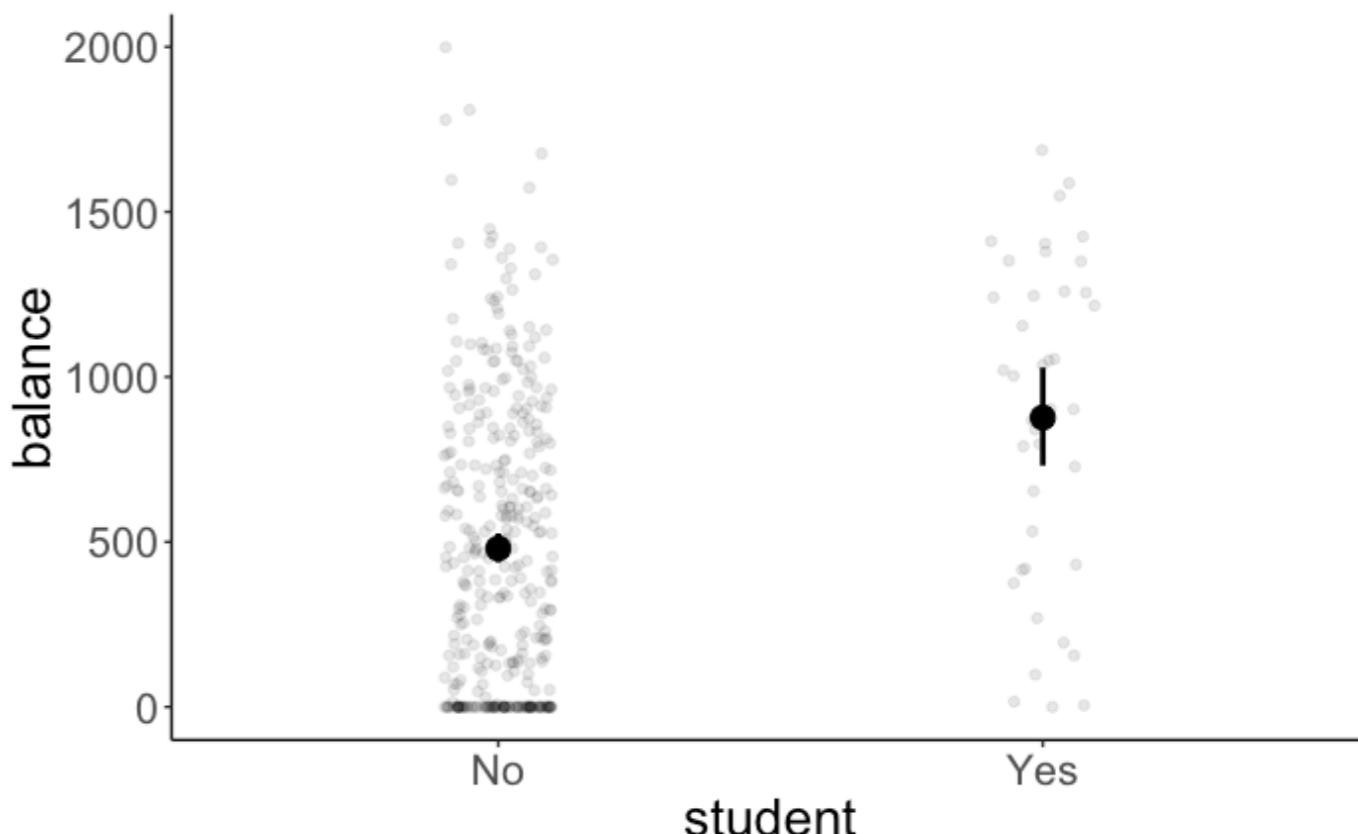
Model 2: balance ~ student

	Res.Df	RSS	Df	Sum of Sq	F	Pr (>F)	
1	399	84339912					
2	398	78681540	1	5658372	28.622	1.488e-07	***

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

Two sample t-test (with independent groups)

Reporting the results



Students have a significantly higher average credit card balance ($\text{Mean} = 876.83, SD = 490.00$) than non-students ($\text{Mean} = 480.37, SD = 439.41$), $F(1, 398) = 28.622, p < .001$.

Interpreting the model

```
1 fit_a = lm(balance ~ 1 + student, data = df.credit)
2 fit_a %>%
3   summary()
```

Call:

```
lm(formula = balance ~ student, data = df.credit)
```

Residuals:

Min	1Q	Median	3Q	Max
-876.82	-458.82	-40.87	341.88	1518.63

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	480.37	23.43	20.50	< 2e-16 ***
studentYes	396.46	74.10	5.35	1.49e-07 ***

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

Residual standard error: 444.6 on 398 degrees of freedom

Multiple R-squared: 0.06709, Adjusted R-squared: 0.06475

F-statistic: 28.62 on 1 and 398 DF, p-value: 1.488e-07

Dummy coding



Dummy coding

$$\hat{Y}_i = 480.37 + 396.46 \cdot \text{student_dummy}_i$$

if student = "No" $\hat{Y}_i = 480.37$

if student = "Yes" $\hat{Y}_i = 480.37 + 396.46 = 876.83$

student	student_dummy
No	0
Yes	1
No	0
Yes	1

- Reference category is coded as 0, the other category is coded as 1
- When thrown into an `lm()`, R automatically turns character columns into factors, and determines the reference category alphabetically

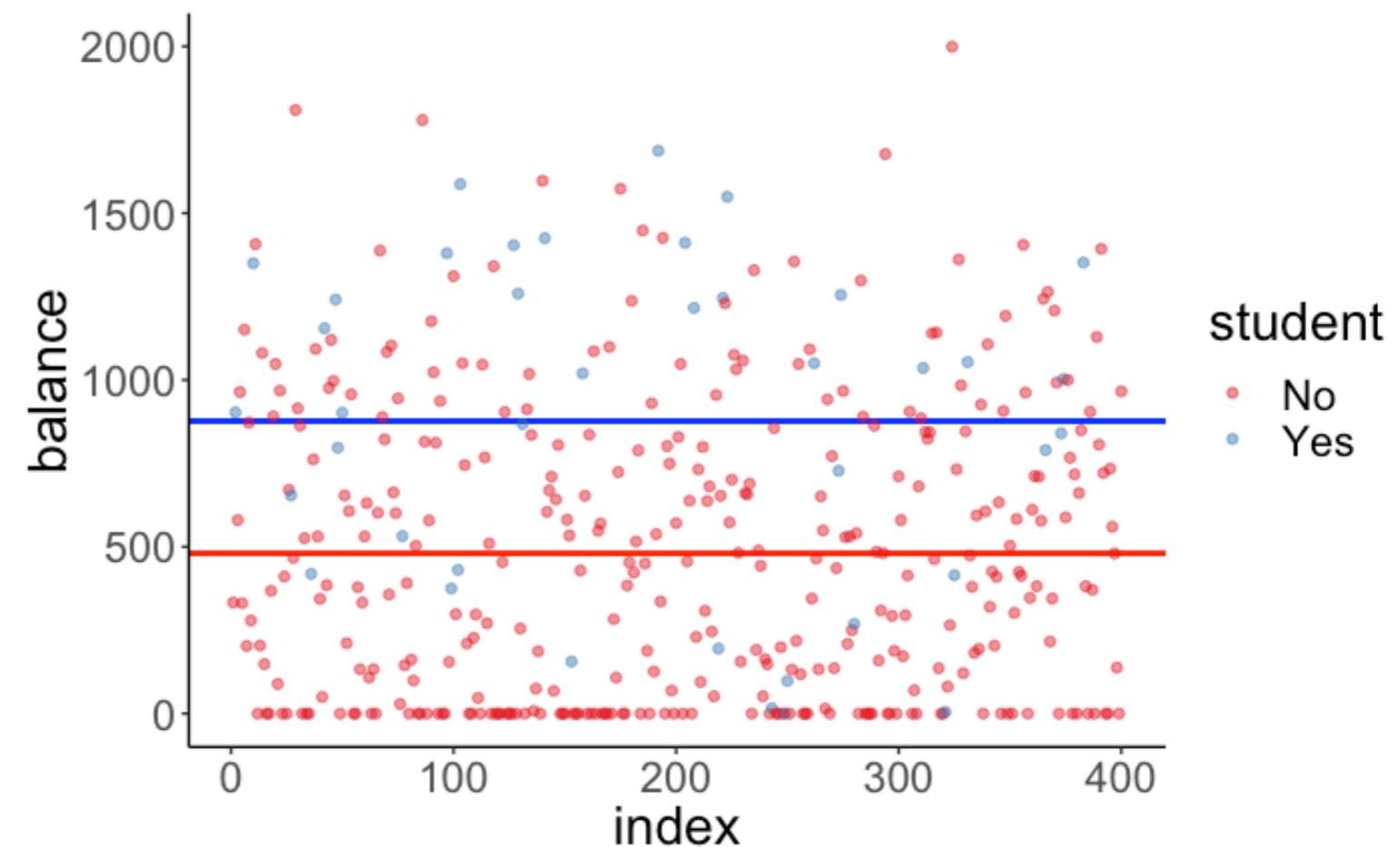
Dummy coding

$$\hat{Y}_i = 480.37 + 396.46 \cdot \text{student_dummy}_i$$

if student = "No" $\hat{Y}_i = 480.37$

if student = "Yes" $\hat{Y}_i = 480.37 + 396.46 = 876.83$

student	student_dummy
No	0
Yes	1
No	0
Yes	1



Categorical and continuous predictor

Credit data set

df.credit

index	income	limit	rating	cards	age	education	gender	student	married	ethnicity	balance
1	14.89	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.03	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.59	7075	514	4	71	11	Male	No	No	Asian	580
4	148.92	9504	681	3	36	11	Female	No	No	Asian	964
5	55.88	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	80.18	8047	569	4	77	10	Male	No	No	Caucasian	1151
7	21.00	3388	259	2	37	12	Female	No	No	African American	203
8	71.41	7114	512	2	87	9	Male	No	No	Asian	872
9	15.12	3300	266	5	66	13	Female	No	No	Caucasian	279
10	71.06	6819	491	3	41	19	Female	Yes	Yes	African American	1350

nrow(df.credit) = 400

Do students have a different credit card balance from non-students, when controlling for income?

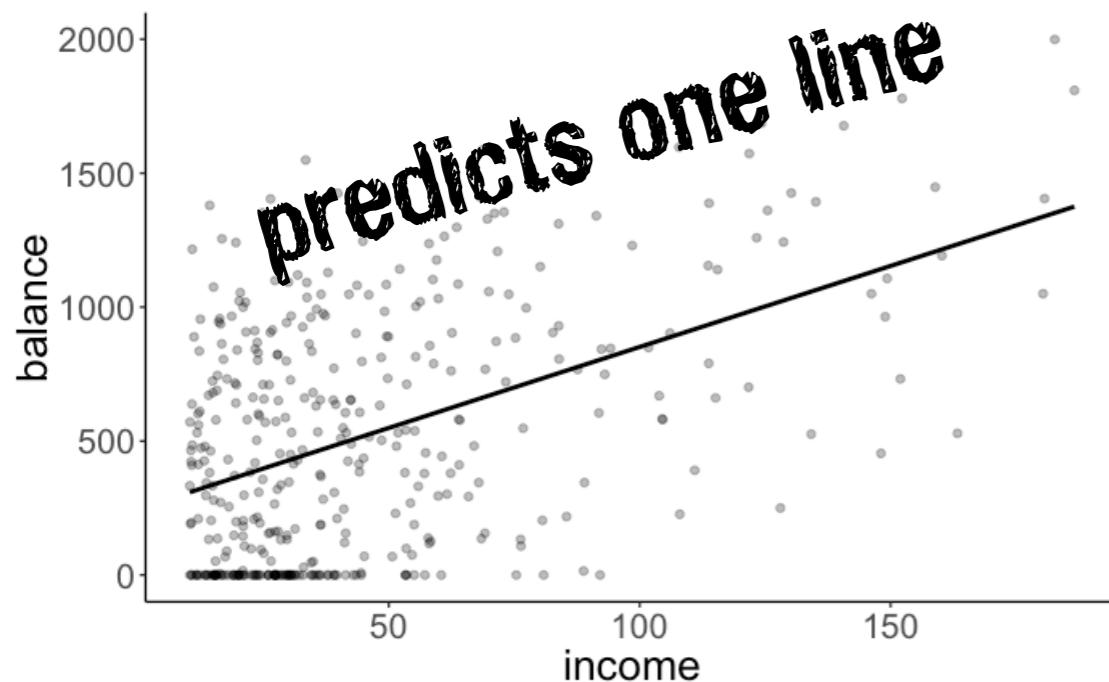
variable	description
income	in thousand dollars
limit	credit limit
rating	credit rating
cards	number of credit cards
age	in years
education	years of education
gender	male or female
student	student or not
married	married or not
ethnicity	African American, Asian, Caucasian
balance	average credit card debt in dollars

H_0 : Students and non-students have the same balance, when controlling for income.

Model C

$$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \epsilon_i$$

Model prediction



Fitted model

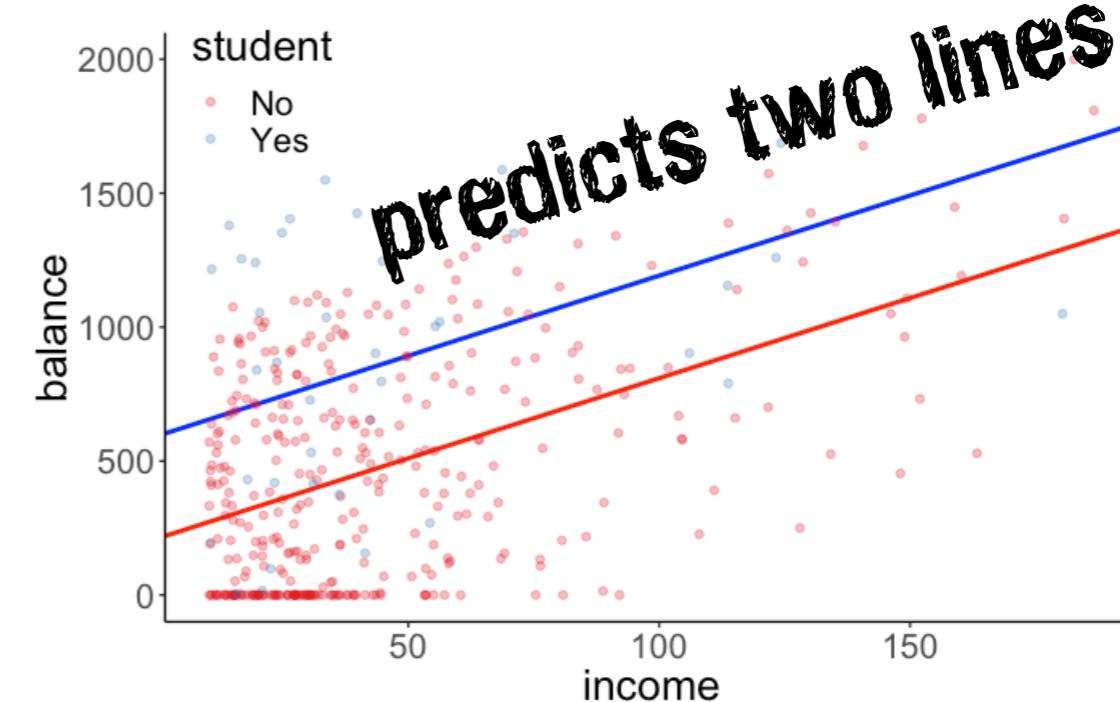
$$\widehat{\text{balance}}_i = 246.515 + 6.048 \cdot \text{income}_i$$

H_1 : Students and non-students have different balances, when controlling for income.

Model A

$$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{student}_i + \epsilon_i$$

Model prediction



Fitted model

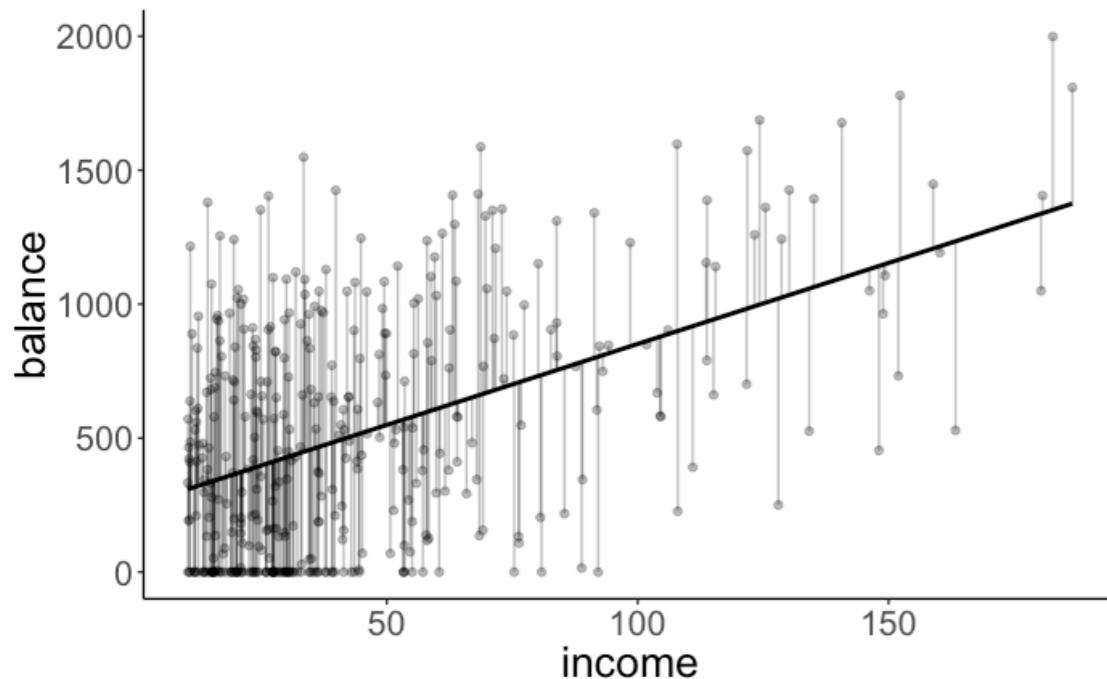
$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

H_0 : Students and non-students have the same balance, when controlling for income.

Model C

$$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \epsilon_i$$

Model prediction



Fitted model

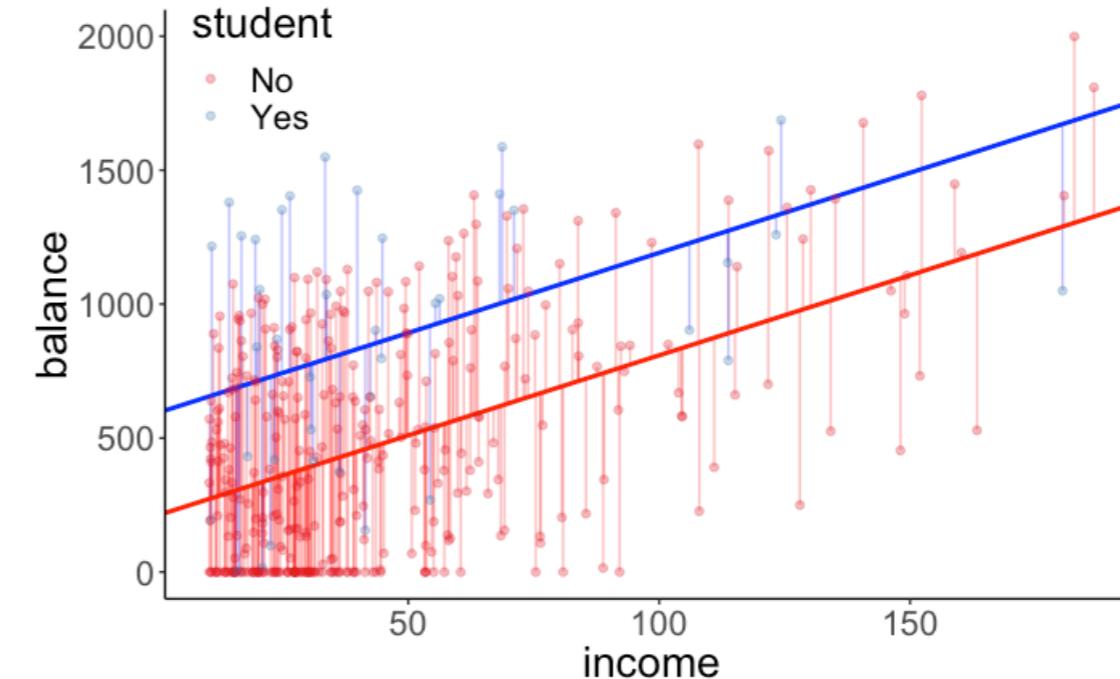
$$\widehat{\text{balance}}_i = 246.515 + 6.048 \cdot \text{income}_i$$

H_1 : Students and non-students have different balances, when controlling for income.

Model A

$$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{student}_i + \epsilon_i$$

Model prediction



Fitted model

$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

Worth it?

```
1 # fit the models
2 fit_c = lm(balance ~ 1 + income, df.credit)
3 fit_a = lm(balance ~ 1 + income + student, df.credit)
4
5 # run the F test
6 anova(fit_c, fit_a)
```

Analysis of Variance Table

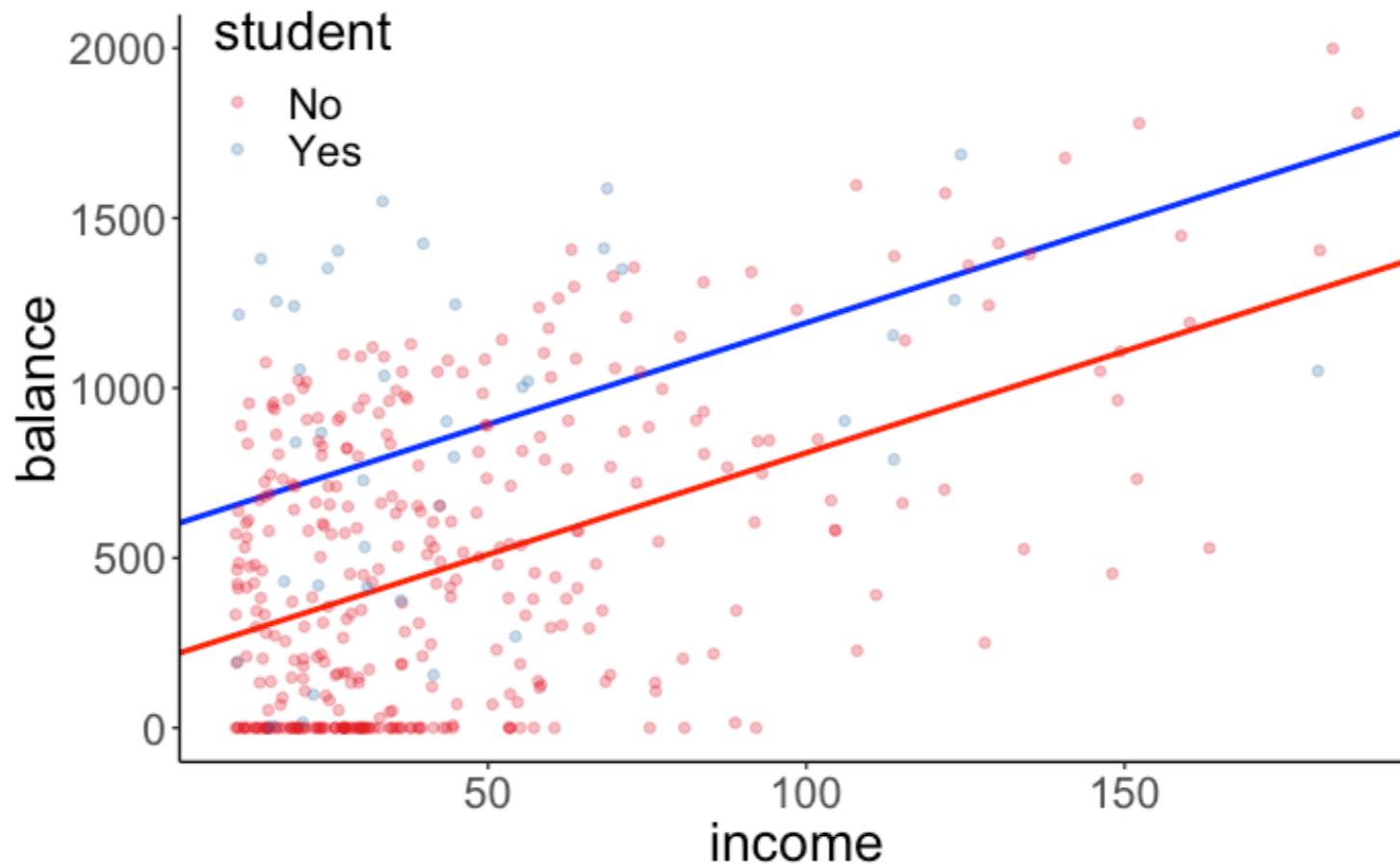
Model 1: balance ~ 1 + income

Model 2: balance ~ 1 + income + student

Res.Df	RSS	Df	Sum of Sq	F	Pr (>F)	
1	398	66208745				
2	397	60939054	1	5269691	34.331	9.776e-09 ***
<hr/>						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Worth it!

Interpretation

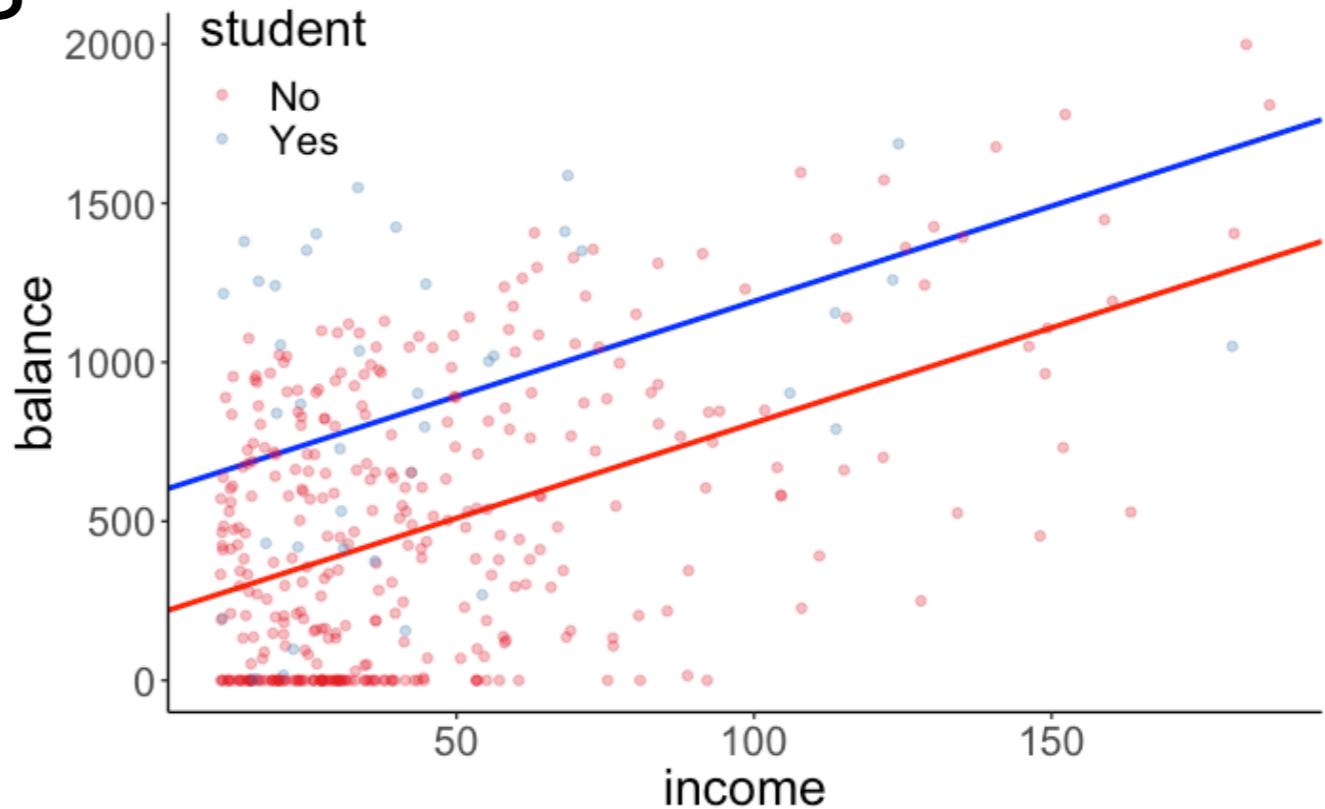


$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

if student = "No" $\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i$

if student = "Yes"
$$\begin{aligned}\widehat{\text{balance}}_i &= 211.14 + 5.98 \cdot \text{income}_i + 382.67 \\ &= 211.14 + 382.67 + 5.98 \cdot \text{income}_i \\ &= 593.81 + 5.98 \cdot \text{income}_i\end{aligned}$$

Reporting the results



Controlling for income, students have a significantly higher average credit card balance (Mean = 876.83, SD = 490.00) than non-students (Mean = 480.37, SD = 439.41), $F(1, 397) = 34.331$, $p < .001$.

Interactions

Is the relationship between level of income and balance different for students than it is for non-students?

Compact Model

$$\widehat{\text{balance}}_i = b_0 + b_1 \text{income}_i + b_2 \text{student}_i$$

Augmented Model

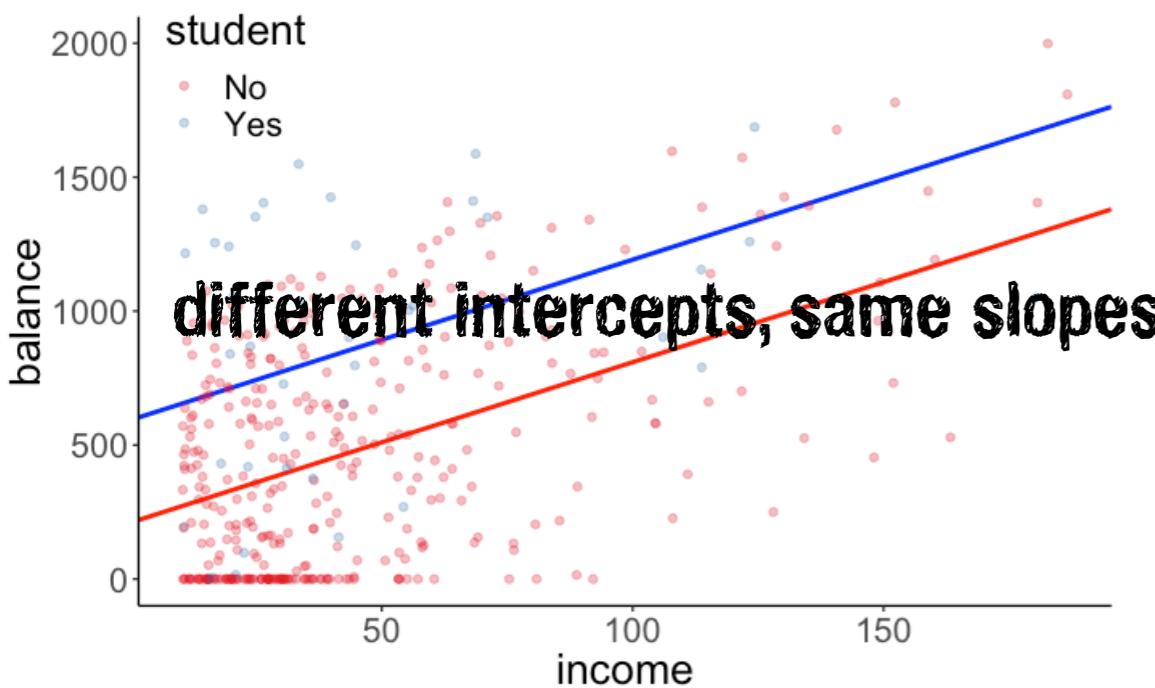
$$\widehat{\text{balance}}_i = b_0 + b_1 \text{income}_i + b_2 \text{student}_i + b_3 (\text{income}_i \times \text{student}_i)$$

H_0 : The relationship between income and balance is the same for students and non-students.

Model C

$$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{student}_i + \epsilon_i$$

Model prediction



Fitted model

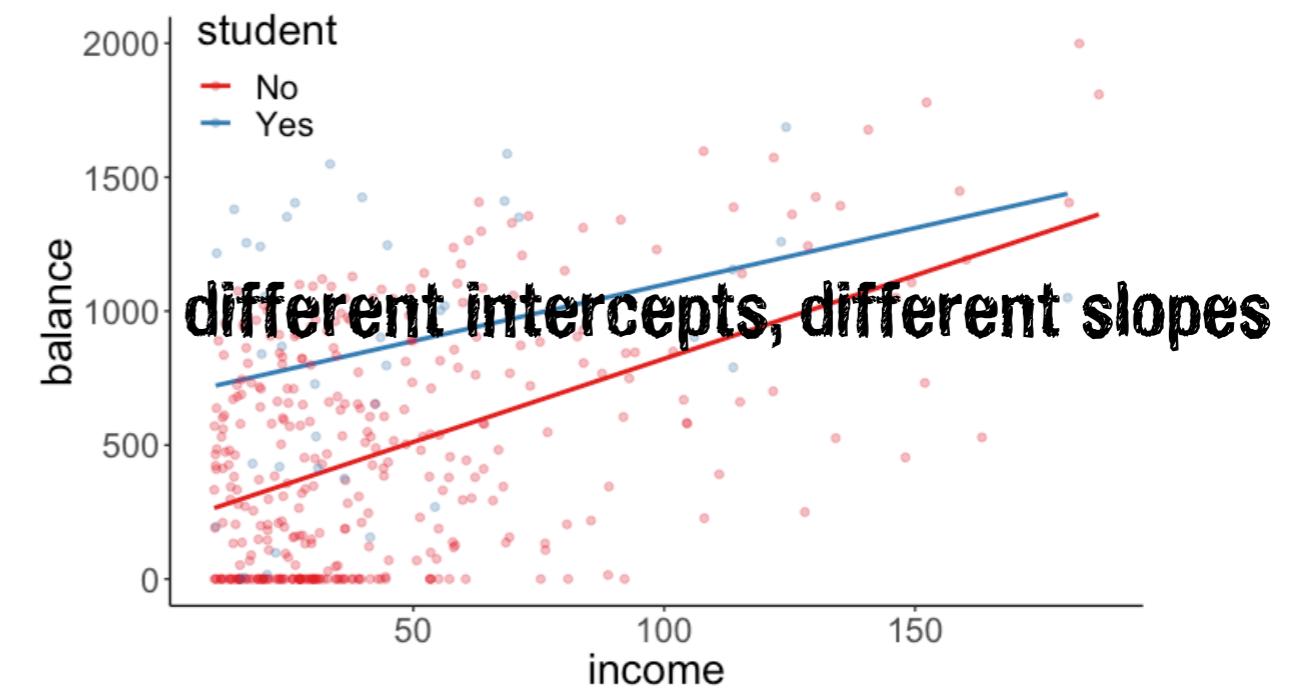
$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

H_1 : The relationship between income and balance differs between students and non-students.

Model A

$$\widehat{\text{balance}}_i = b_0 + b_1 \text{income}_i + b_2 \text{student}_i + b_3 (\text{income}_i \times \text{student}_i)$$

Model prediction



Fitted model

$$\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot \text{student}_i - 2.00 \cdot (\text{income}_i \times \text{student}_i)$$

Worth it?

Is the relationship between level of income and balance different for students than it is for non-students?

```
1 # fit models
2 fit_c = lm(formula = balance ~ income + student, data = df.credit)
3 fit_a = lm(formula = balance ~ income * student, data = df.credit)
4
5 # F-test
6 anova(fit_c, fit_a)
```

Analysis of Variance Table

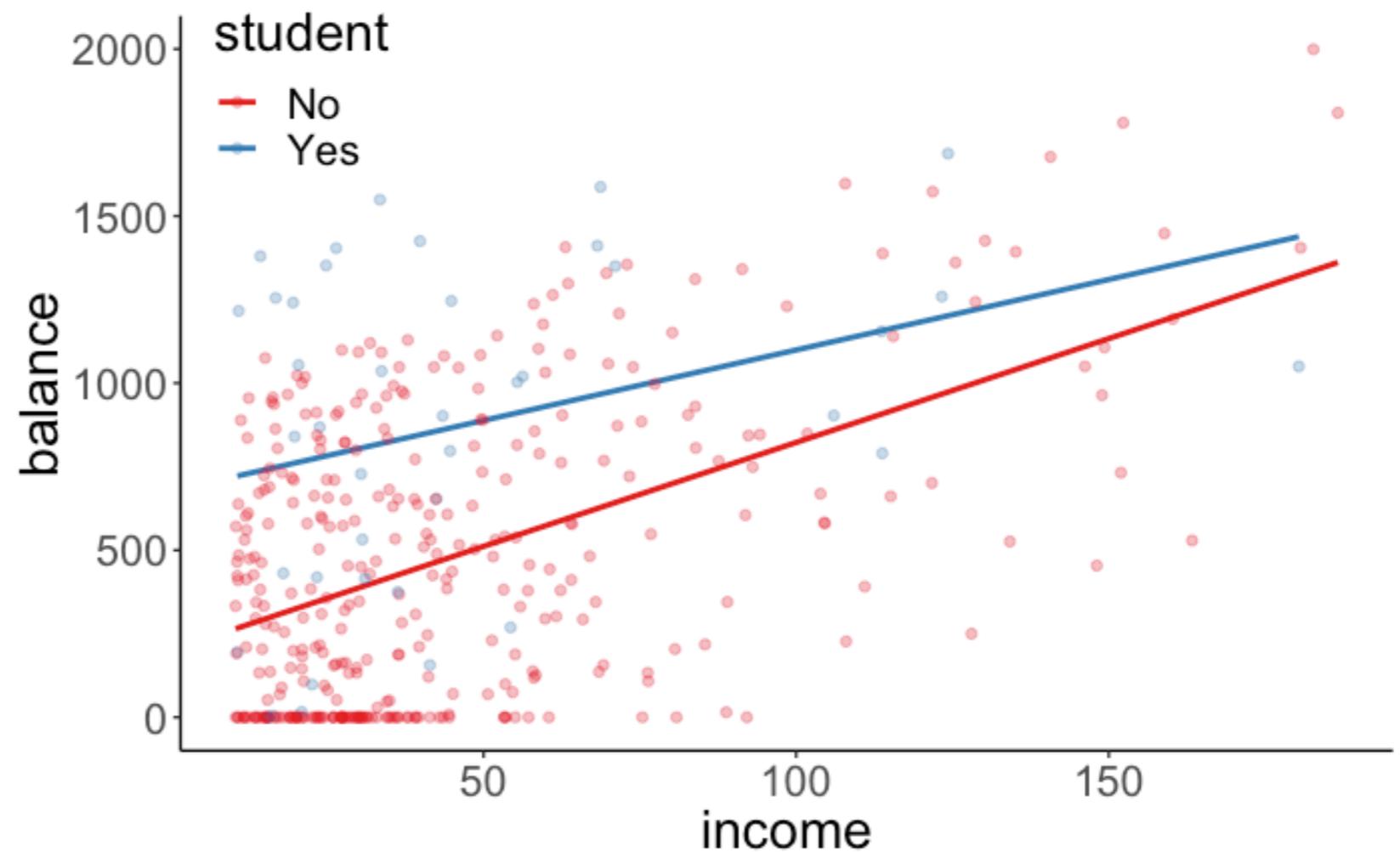
not worth it!

Model 1: balance ~ income + student

Model 2: balance ~ income * student

	Res.Df	RSS	Df	Sum of Sq	F	Pr (>F)
1	397	60939054				
2	396	60734545	1	204509	1.3334	0.2489

Interpretation



$$\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot \text{student}_i - 2.00 \cdot (\text{income}_i \times \text{student}_i)$$

if student = "No" $\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i$

if student = "Yes"

$$\begin{aligned}\widehat{\text{balance}}_i &= 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot 1 - 2.00 \cdot (\text{income}_i \times 1) \\ &= 677.3 + 6.22 \cdot \text{income}_i - 2.00 \cdot \text{income}_i \\ &= 677.3 + 4.22 \cdot \text{income}_i\end{aligned}$$

Interpretation

```
fit1 = lm(formula = balance ~ income + student + income:student, data = df.credit)
```

Explicitly encode the interaction

```
1 df.credit %>%
2   mutate(student_dummy = ifelse(student == "No", 0, 1)) %>%
3   mutate(income_student = income * student_dummy) %>%
4   select(balance, income, student, student_dummy, income_student)
```

balance	income	student	student_dummy	income_student
333	14.89	No	0	0.00
903	106.03	Yes	1	106.03
580	104.59	No	0	0.00
964	148.92	No	0	0.00
331	55.88	No	0	0.00
1151	80.18	No	0	0.00
203	21.00	No	0	0.00
872	71.41	No	0	0.00
279	15.12	No	0	0.00
1350	71.06	Yes	1	71.06

```
fit2 = lm(formula = balance ~ income + student + income_student, data = df.credit)
```

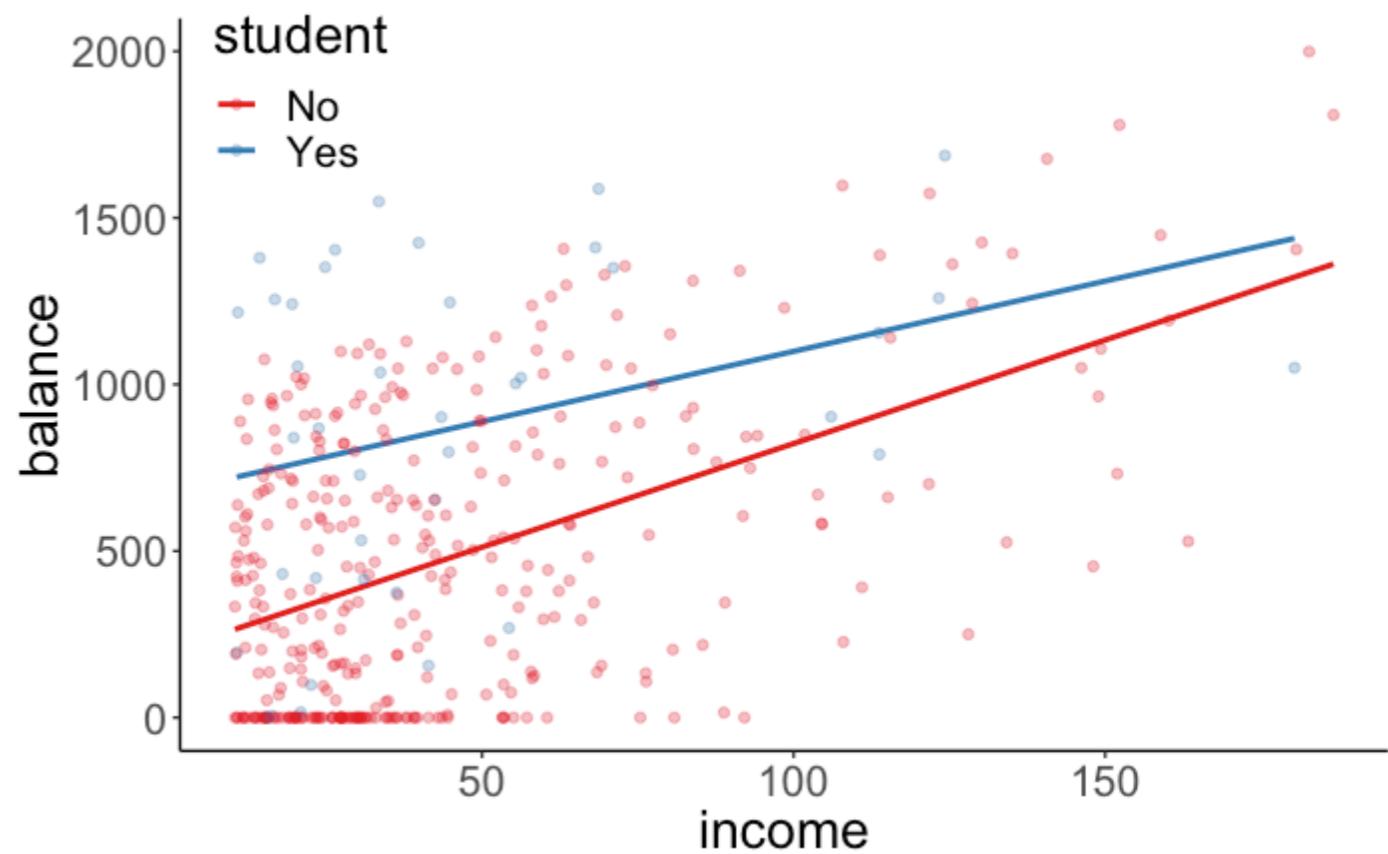
fit1 and fit2 are identical!

How to report results of interaction

There is no significant difference in the relationship between income and balance for students versus non-students, $F(1, 396) = 1.33, p = 0.25$.

For *students*, an increase in \$1000 income is associated with an increase in \$4.21 of average credit card balance.

For *non-students*, an increase in \$1000 income is associated with an increase in \$6.22 of average credit card balance.



lm () output

lm() output

```
1 lm(formula = balance ~ income + student + income:student, data = df.credit) %>%
2   summary()
```

```
Call:
lm(formula = balance ~ income + student + income:student,
data = df.credit)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-773.39	-325.70	-41.13	321.65	814.04

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	200.6232	33.6984	5.953	5.79e-09 ***
income	6.2182	0.5921	10.502	< 2e-16 ***
studentYes	476.6758	104.3512	4.568	6.59e-06 ***
income:studentYes	-1.9992	1.7313	-1.155	0.249

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
Residual standard error: 391.6 on 396 degrees of freedom
```

```
Multiple R-squared: 0.2799, Adjusted R-squared: 0.2744
```

```
F-statistic: 51.3 on 3 and 396 DF, p-value: < 2.2e-16
```



```
1 fit_c = lm(formula = balance ~ student + income:student, data = df.credit)
2 fit_a = lm(formula = balance ~ income + student + income:student, data = df.credit)
3
4 anova(fit_c, fit_a)
```

```
1 fit_c = lm(formula = balance ~ income + student, data = df.credit)
2 fit_a = lm(formula = balance ~ income + student + income:student, data = df.credit)
3
4 anova(fit_c, fit_a)
```

lm() output

```
1 lm(formula = balance ~ income + student + income:student, data = df.credit) %>%
2   summary()
```

```
Call:
lm(formula = balance ~ income + student + income:student,
data = df.credit)
```

Residuals:

Min	1Q	Median	3Q	Max
-773.39	-325.70	-41.13	321.65	814.04

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	200.6232	33.6984	5.953	5.79e-09 ***
income	6.2182	0.5921	10.502	< 2e-16 ***
studentYes	476.6758	104.3512	4.568	6.59e-06 ***
income:studentYes	-1.9992	1.7313	-1.155	0.249

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

Residual standard error: 391.6 on 396 degrees of freedom

Multiple R-squared: 0.2799, Adjusted R-squared: 0.2744

F-statistic: 51.3 on 3 and 396 DF, p-value: < 2.2e-16

```
1 fit_c = lm(formula = balance ~ 1, data = df.credit)
2 fit_a = lm(formula = balance ~ income + student + income:student, data = df.credit)
3
4 anova(fit_c, fit_a)
```

Analysis of Variance Table

```
Model 1: balance ~ 1
Model 2: balance ~ 1 + income
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     399 84339912
2     398 66208745  1  18131167 108.99 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

**deterministic mapping
between t and F**

$$t^2 = F$$

$$10.44^2 = 108.99$$

anova () gives me F s ?
but lm () gives me ts ?

```
Call:
lm(formula = balance ~ 1 + income, data = df.credit)

Residuals:
    Min      1Q  Median      3Q      Max 
-803.64 -348.99 -54.42  331.75 1100.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 246.5148    33.1993   7.425 6.9e-13 ***
income       6.0484     0.5794 10.440 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 407.9 on 398 degrees of freedom
Multiple R-squared:  0.215,    Adjusted R-squared:  0.213 
F-statistic: 109 on 1 and 398 DF,  p-value: < 2.2e-16
```

lm() output

```
1 lm(formula = balance ~ income + student + income:student, data = df.credit) %>%
2   summary()
```

```
Call:
lm(formula = balance ~ income + student + income:student, data =
df.credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-773.39 -325.70 -41.13  321.65  814.04 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 200.6232   33.6984   5.953 5.79e-09 ***
income        6.2182    0.5921  10.502 < 2e-16 ***
studentYes   476.6758  104.3512   4.568 6.59e-06 ***
income:studentYes -1.9992    1.7313  -1.155    0.249  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared:  0.2799,    Adjusted R-squared:  0.2744 
F-statistic: 51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

- runs many hypothesis tests at the same time
- increases the danger of making a type-I error (incorrectly rejecting the H_0)
- will not give us p-values for mixed effects models ...

The model comparison approach

- allows to formulate hypotheses as specific comparisons between candidate models
- is more flexible: we could test a model with 2 predictors vs. one with 4 predictors
- gives us insight into the underlying statistical procedure

lm() output

very important

```
Call:  
lm(formula = balance ~ income + student + income:student,  
data = df.credit)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-773.39 -325.70 -41.13  321.65  814.04  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 200.6232   33.6984   5.953 5.79e-09 ***  
income       6.2182    0.5921  10.502 < 2e-16 ***  
studentYes  476.6758  104.3512   4.568 6.59e-06 ***  
income:studentYes -1.9992    1.7313  -1.155   0.249  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '  
1  
  
Residual standard error: 391.6 on 396 degrees of freedom  
Multiple R-squared:  0.2799, Adjusted R-squared:  0.2744  
F-statistic: 51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

what does this mean?

not the overall effect
of income!

instead the predicted
effect of income for
non-students

we'll talk more about the difference between simple/conditional
effects and main effects next time!

Summary

- Quick recap
- Multiple regression
 - two continuous predictors
 - one categorical predictor
 - one continuous and one categorical predictor
- Interactions
- `lm()` output

Feedback

How was the pace of today's class?

much a little just a little much
too too right too too
slow slow

How happy were you with today's class overall?



What did you like about today's class? What could be improved next time?

Thank you!