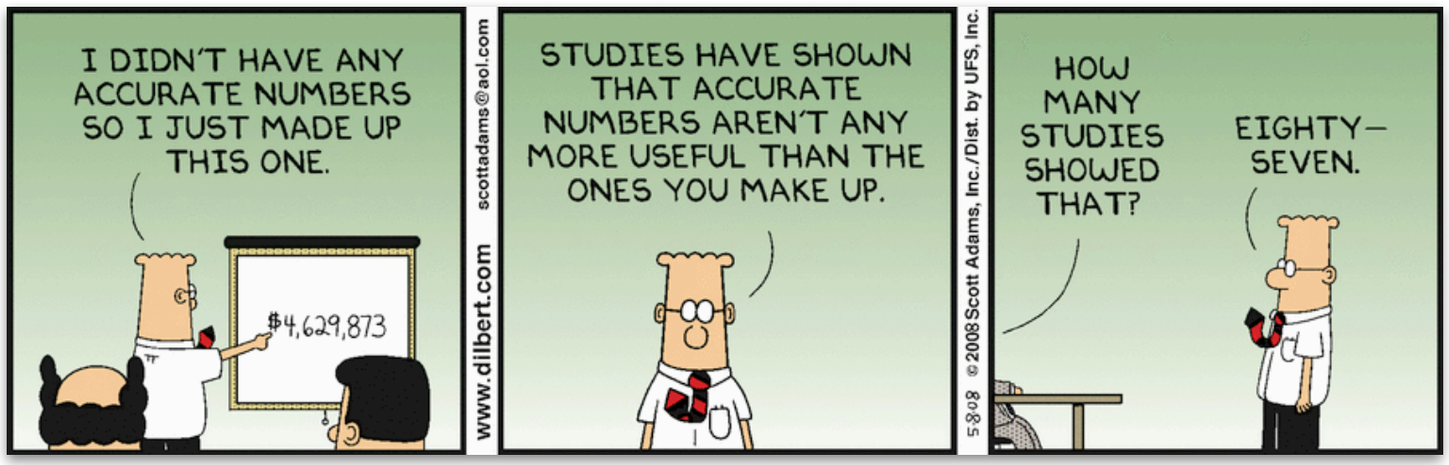
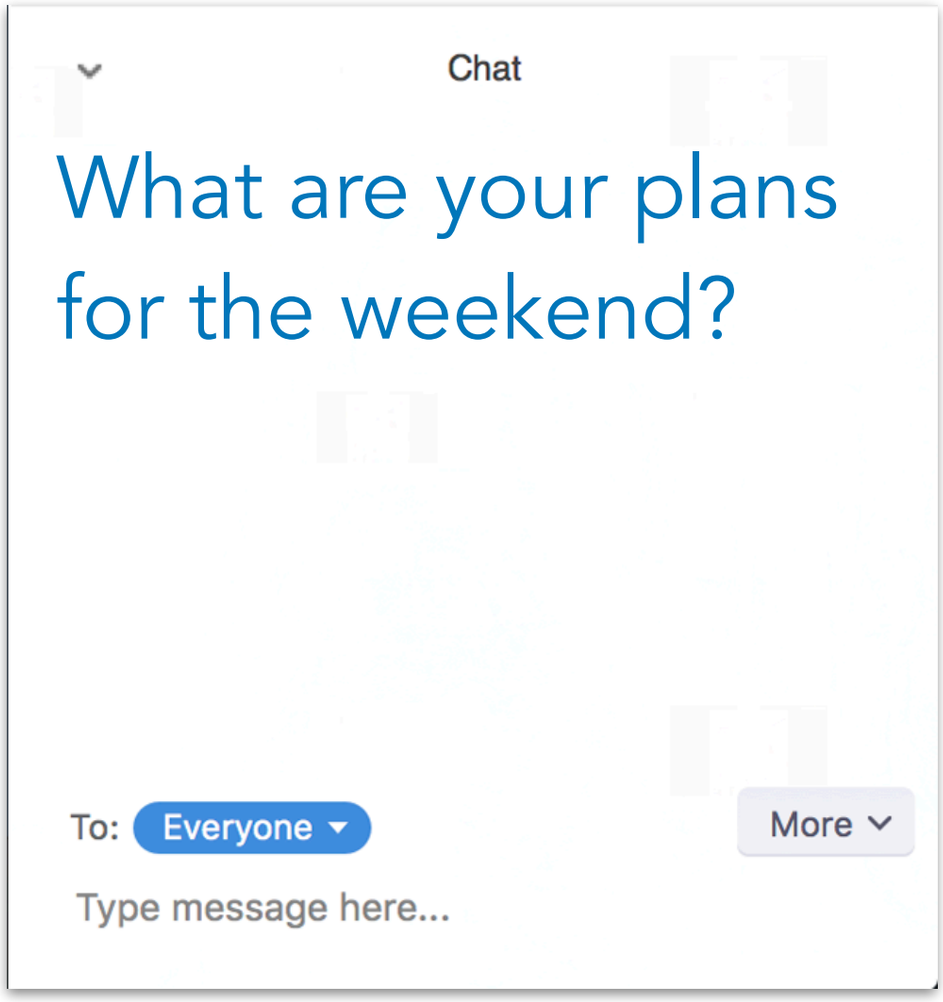
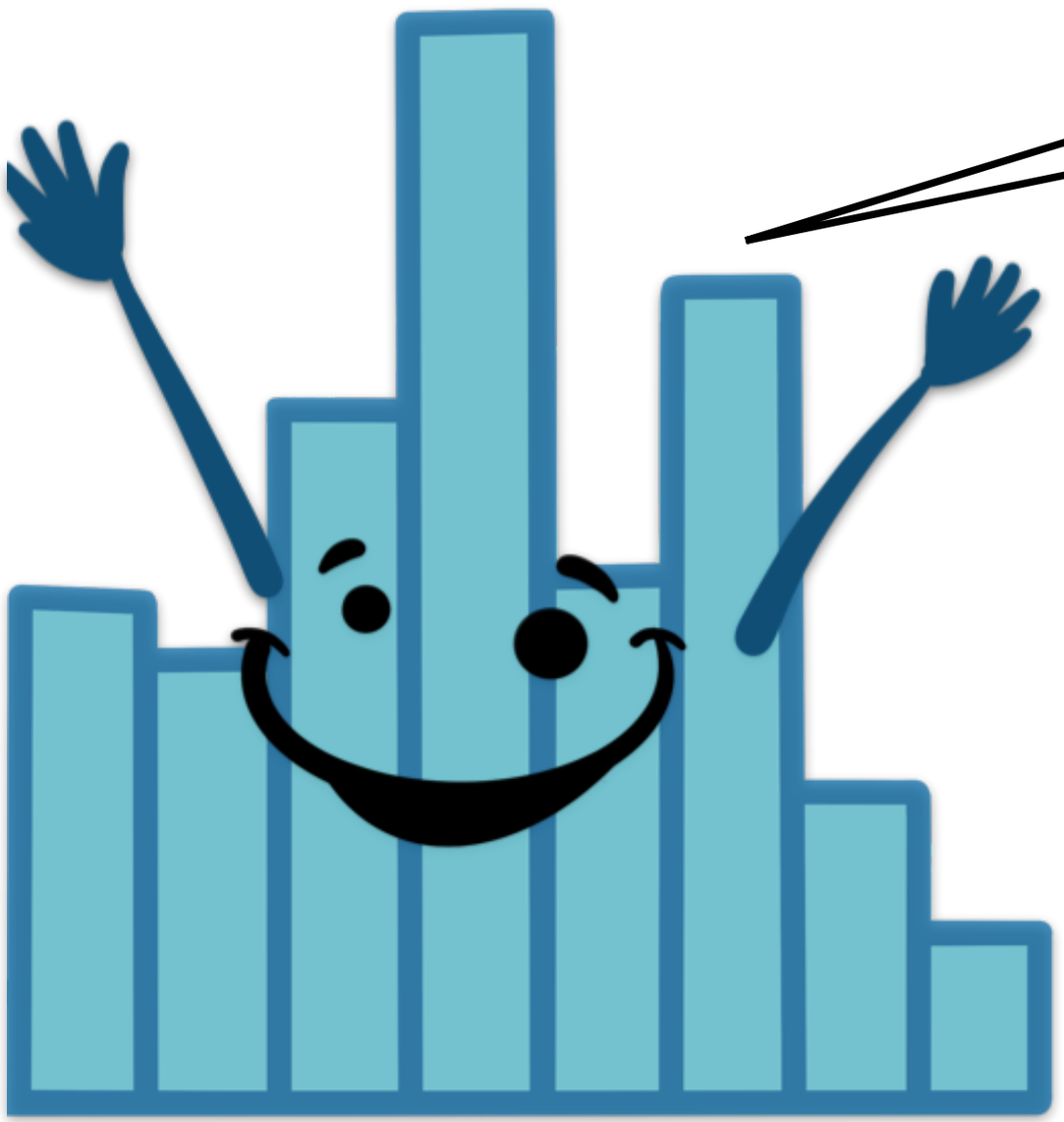
PSYCH 252: Statistical Methods for Behavioral and Social Sciences

Linear model 2





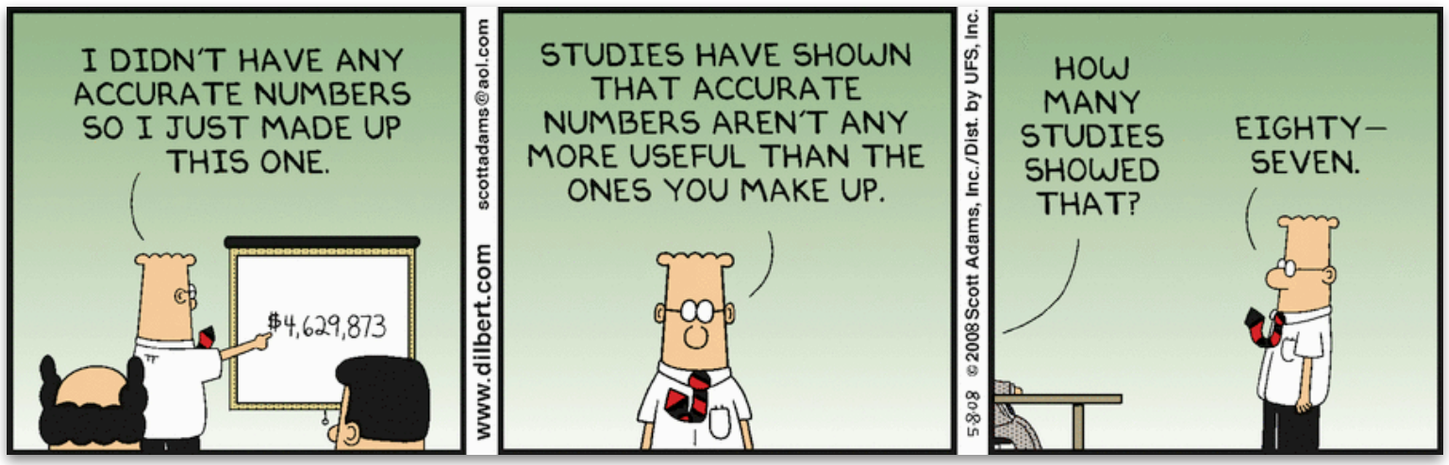
02/05/2021

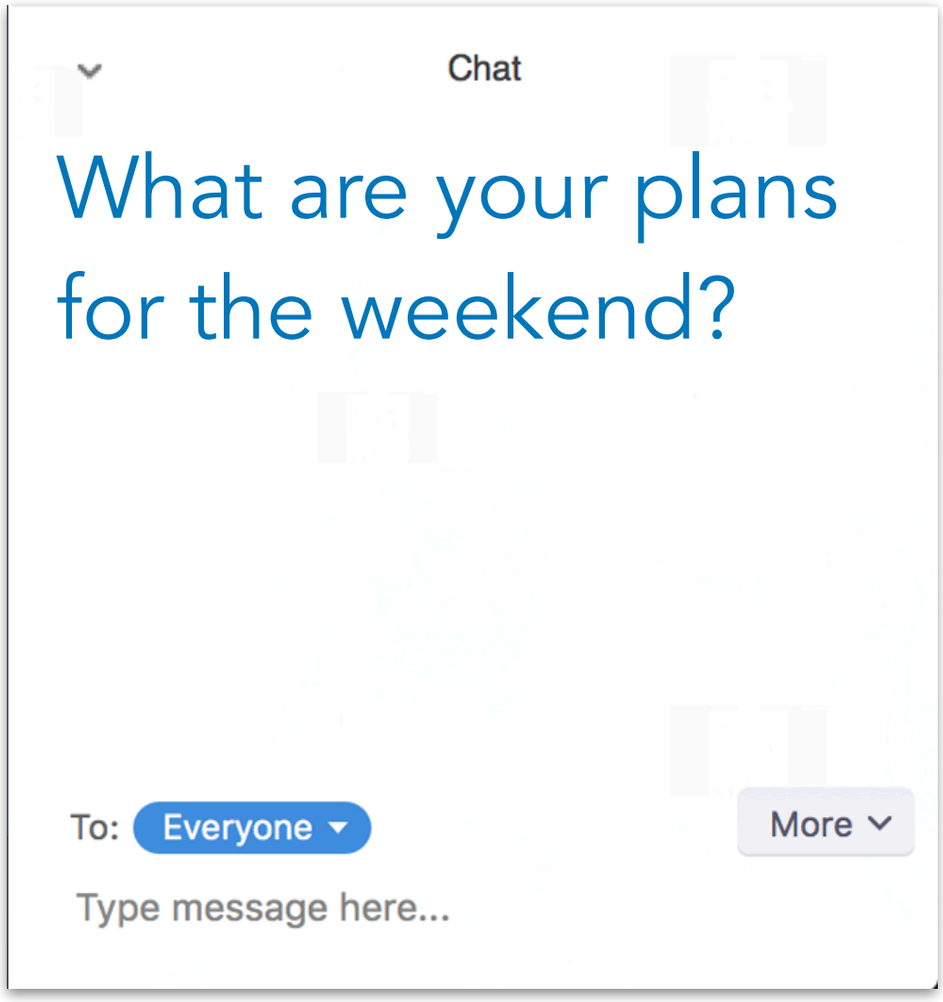


Remember to record the lecture!

PSYCH 252: Statistical Methods for Behavioral and Social Sciences

Linear model 2



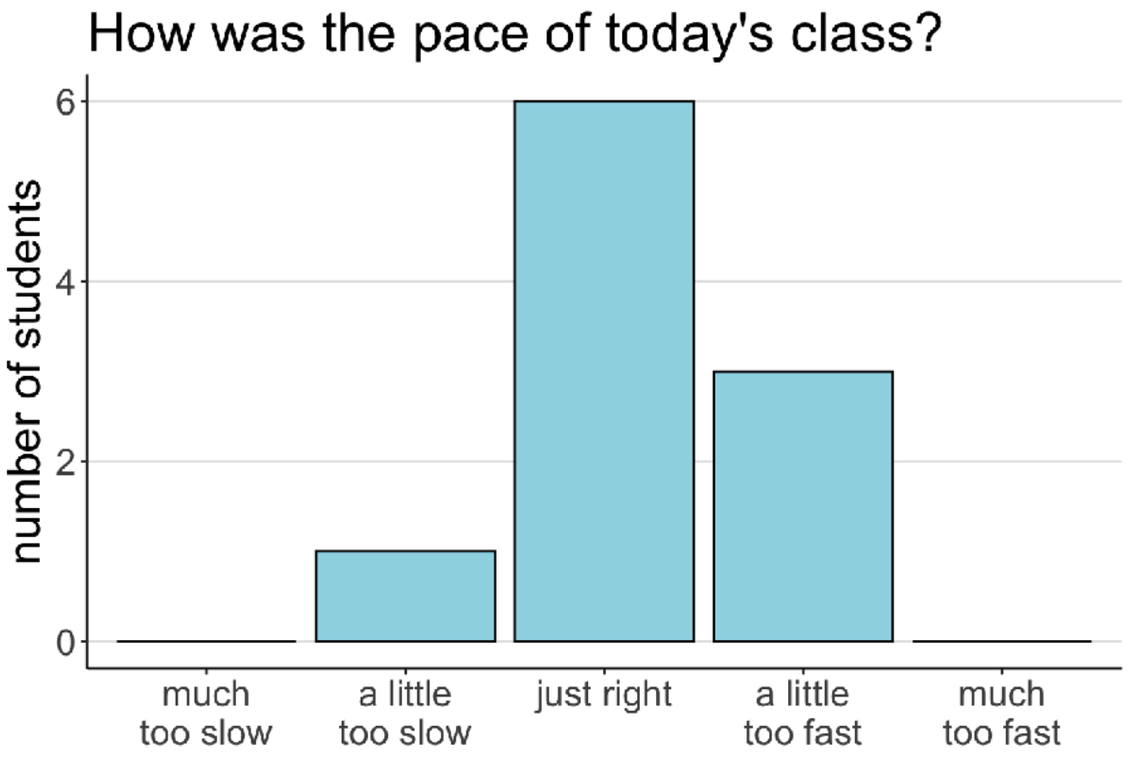


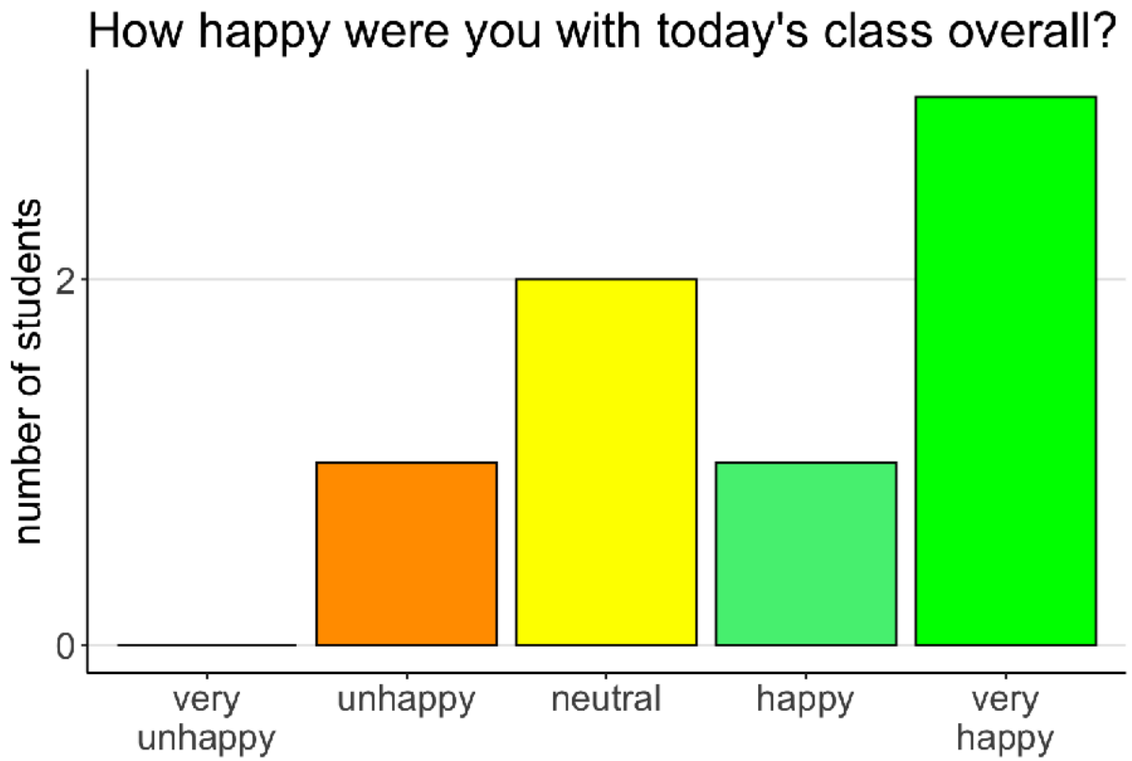
02/05/2021

Logistics

Feedback

Your feedback





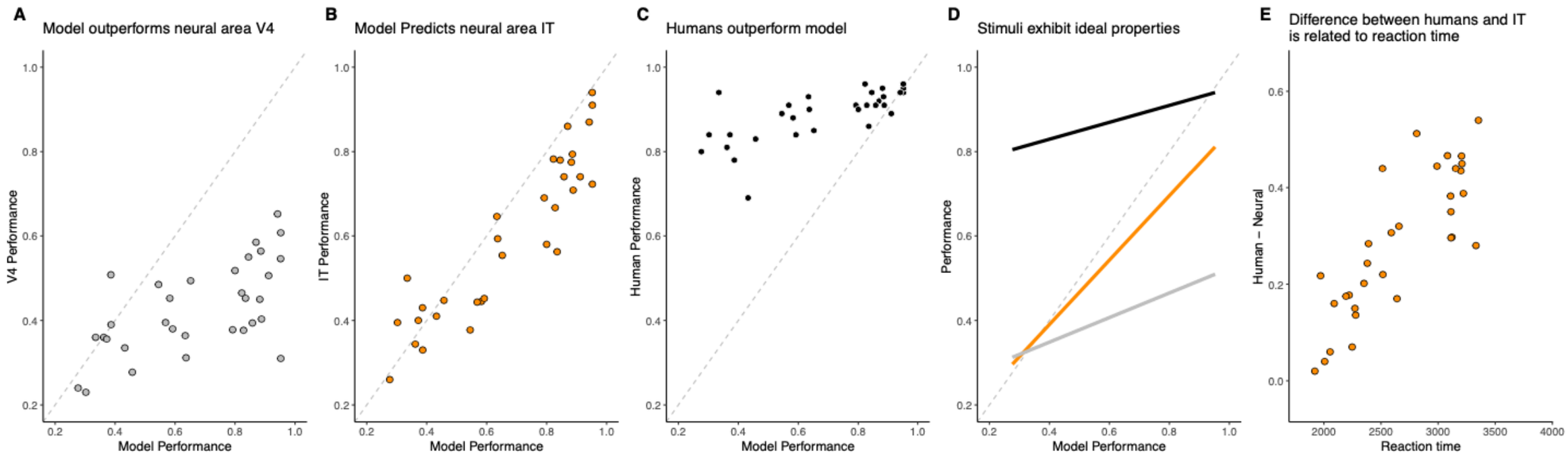
I found the breakout room exercise example that included suicide to be quite triggering, particularly given the "light-hearted" nature of the exercise and thus framing mental health as something that is not serious or something to "joke" about. Particularly given the tight coupling of students who struggle with mental health in university settings, I would recommend removing that example in future iterations of the class/exercise.

Thank you and I'm sorry!

Homework

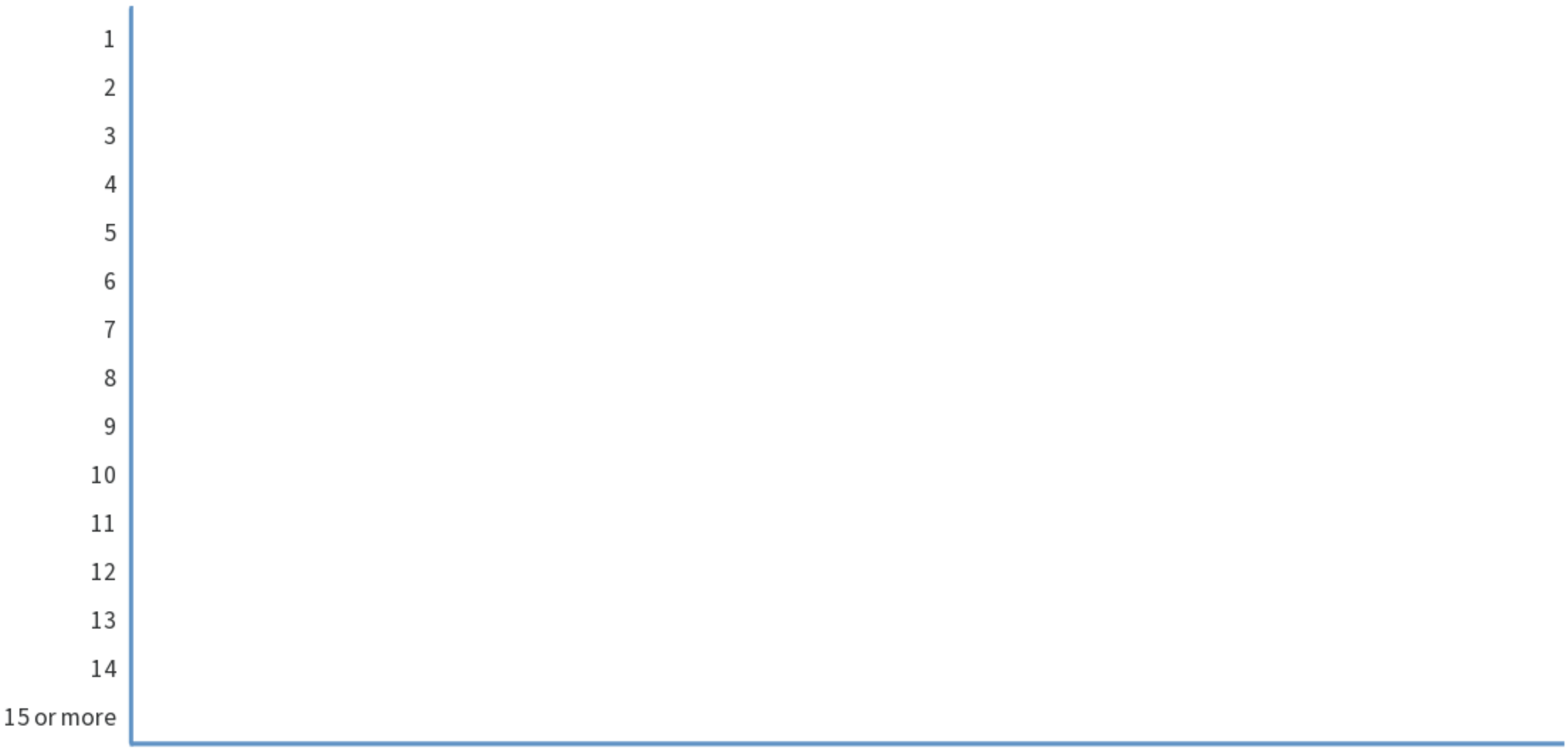
Homework 2

Grades are posted, and solutions are on Canvas.



Note: We've updated the colors here slightly from the original homework to make them easier to tell apart.

How many hours did it take you to complete Homework 3?



Homework 4

My name goes here

The names of the people I have worked with go here

2021-02-04 21:44:00

This homework is due by **Thursday, February 11th, 8:00pm**. Upload a pdf file to Canvas called `4_regression.pdf`

I. Simple linear regression and prediction

In this section and the next, we'll be revisiting the credit dataset.

```
# Load data
df.credit = read_csv("data/credit.csv") %>%
  clean_names()
```

1. Explore and visualize the data

Use the `ggpairs()` function from the `GGally` package to make scatterplots of all pairwise combinations of the numeric variables.

Tip: use `progress = FALSE` within the function to suppress unwanted progress bars of each plot

```
### YOUR CODE HERE ###

######################
```

That's perhaps a bit too congested to be useful. Recreate the plot, focusing on just ~5 of the variables that seem interesting to you.

```
### YOUR CODE HERE ###

######################
```

Does your plot tell you anything interesting about the data? Briefly describe one observation.

YOUR ANSWER HERE

II. Multiple linear regression and controls

In psychological research, people often run linear regressions in which the goal is to assess the relationship between two variables while "controlling" for other variables. These control variables could, for example, be age and gender. But how should we decide whether and which variables to control for? In this exercise, we will see what potential effects controlling for variables can have in different situations.

3

Now you are interested in whether age is a significant predictor of credit limit.

4. Interpreting model parameters

a) Build a simple linear regression model to predict `limit` from `age`. Is age a significant predictor of credit limit?

```
# Simple regression without control variables
### YOUR CODE HERE ###
fit.lm1 =
######################
```

YOUR ANSWER HERE

Then you realize that age is actually related to income, which is a strong predictor of credit limit, so you are interested in seeing whether age is related to credit limit controlling for income.

b) Build a multiple regression model to predict `limit` using both `age` and `income` as predictors. Is age still a significant predictor? What could be an explanation for the change, if there was any?

```
# Multiple regression with control variables
### YOUR CODE HERE ###
fit.lm2 =
fit.lm3 =
######################
```

YOUR ANSWER HERE

III. Interactions

We will be using the following dataset.

families.csv:

Data from a study of 68 companies, examining relationships between the quality of family-friendly programs at each company, the percentage of employees with families who use these programs, and employee satisfaction (all continuous variables).

Maybe find a homework buddy to team up with?!

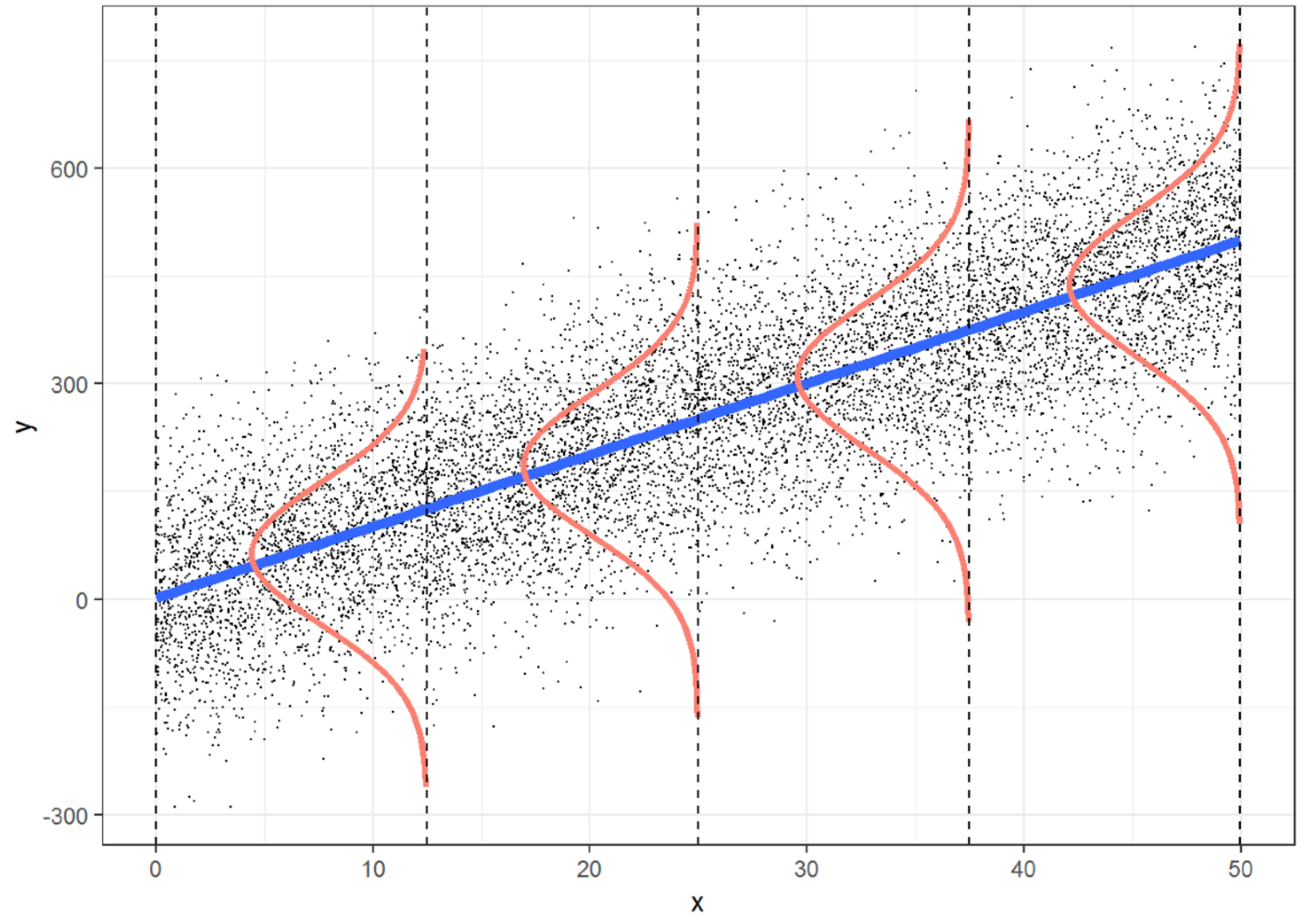
Plan for today

- Multiple regression
  - Model assumptions: no multi-collinearity
- Several continuous predictors
  - Hypothesis tests
  - Interpreting parameters
  - Reporting results
- One categorical predictor
- Both continuous and categorical predictors
- Interactions

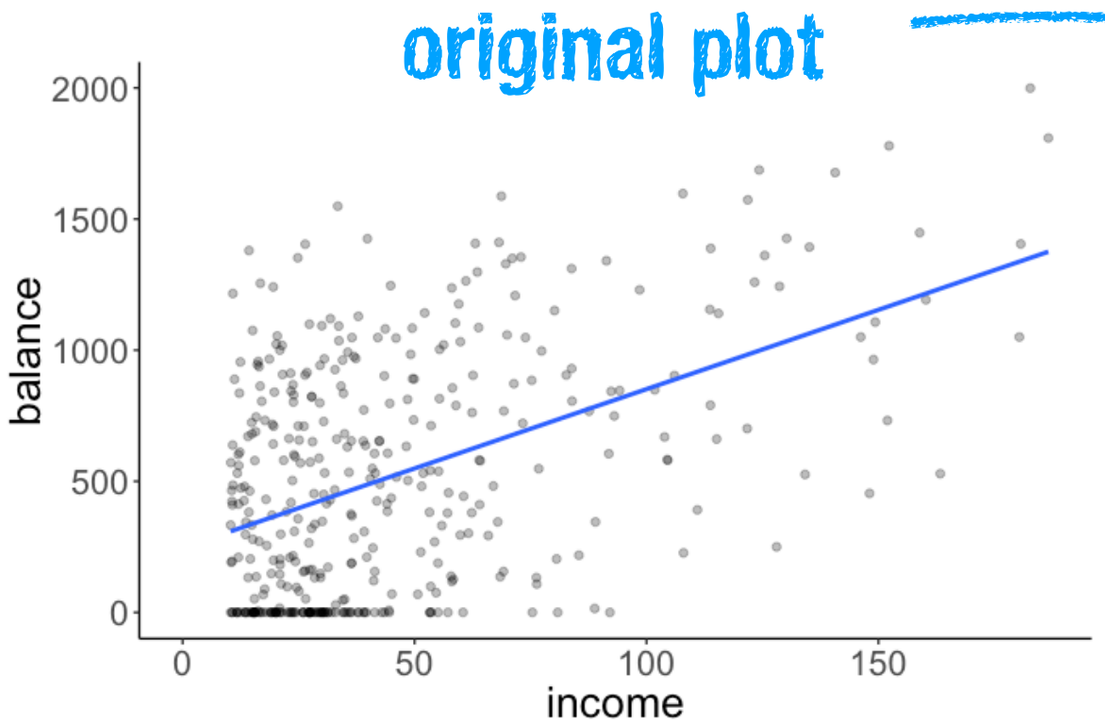
Multiple regression

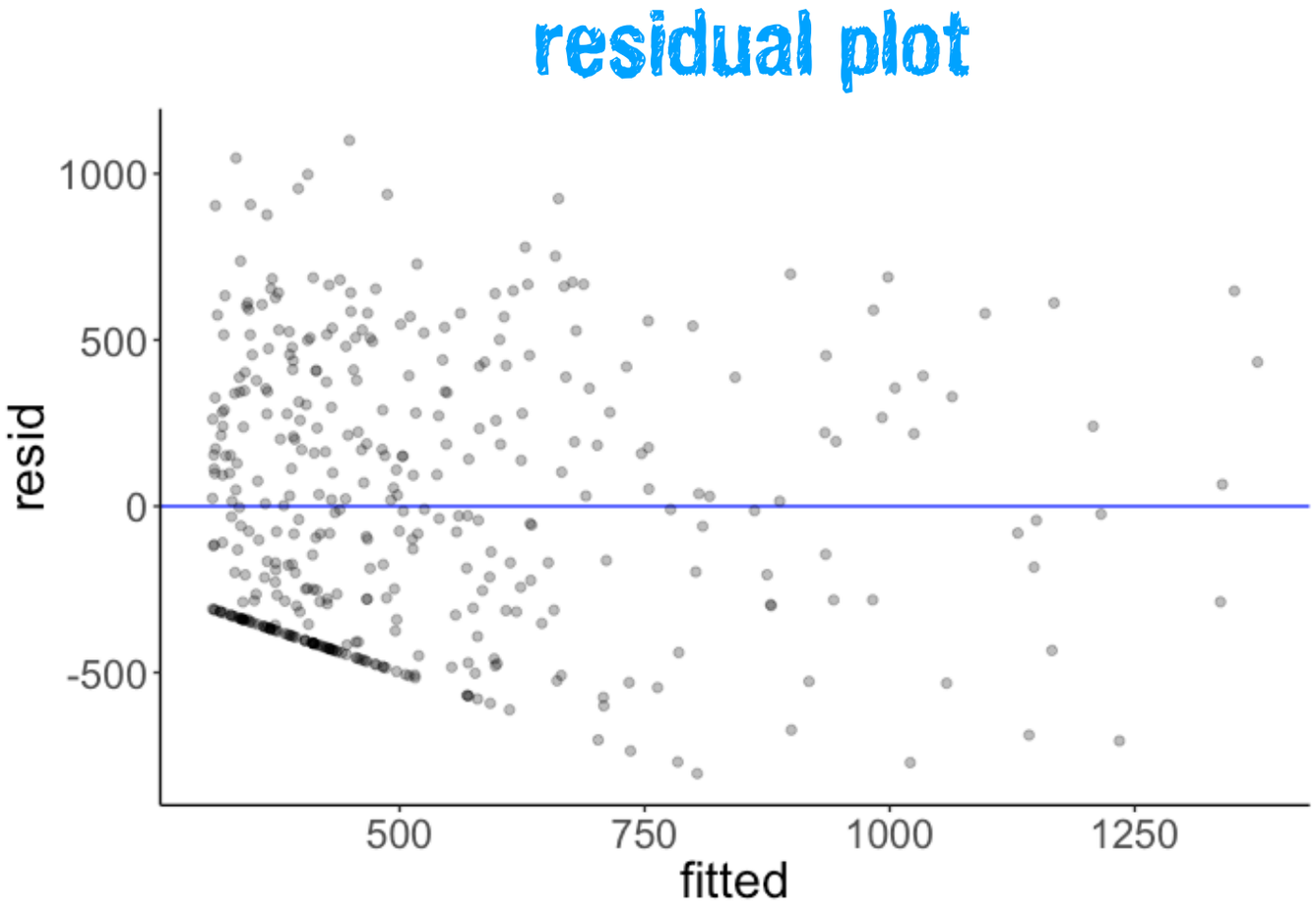
Model assumptions of simple regression

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

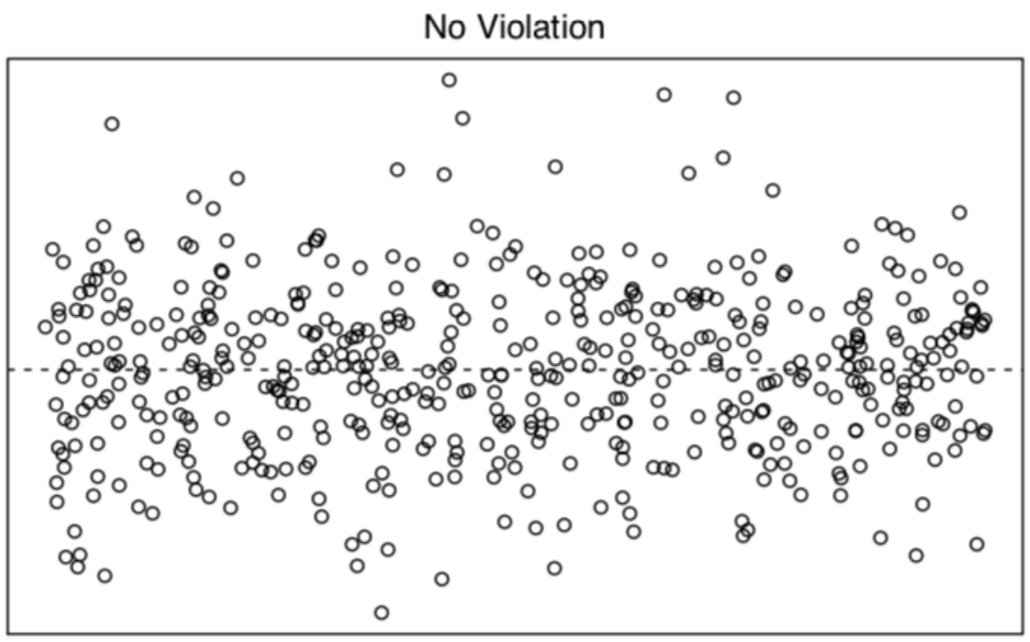


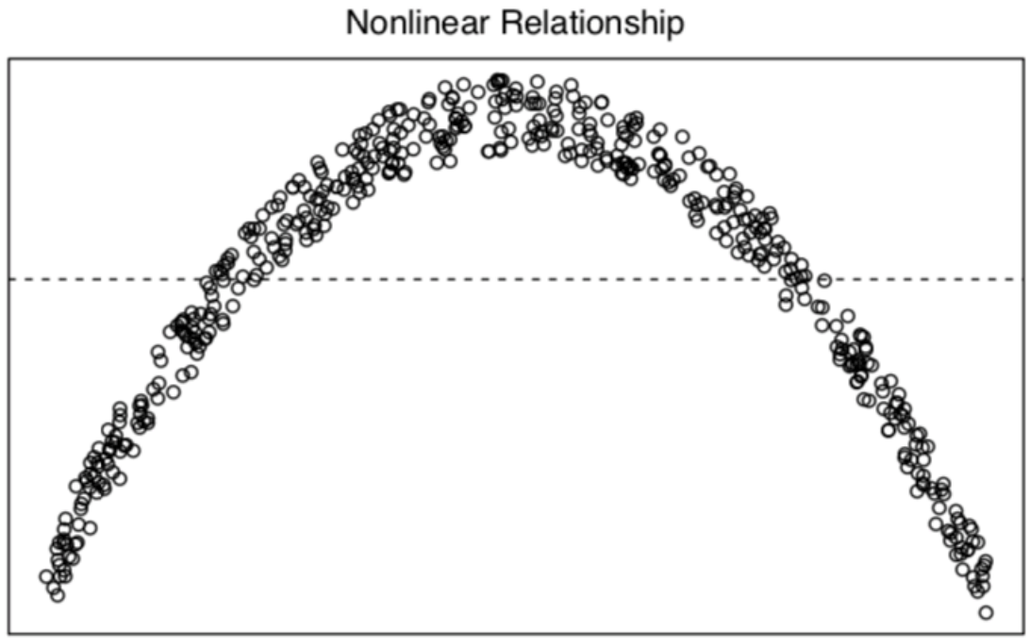
Model assumptions of simple regression

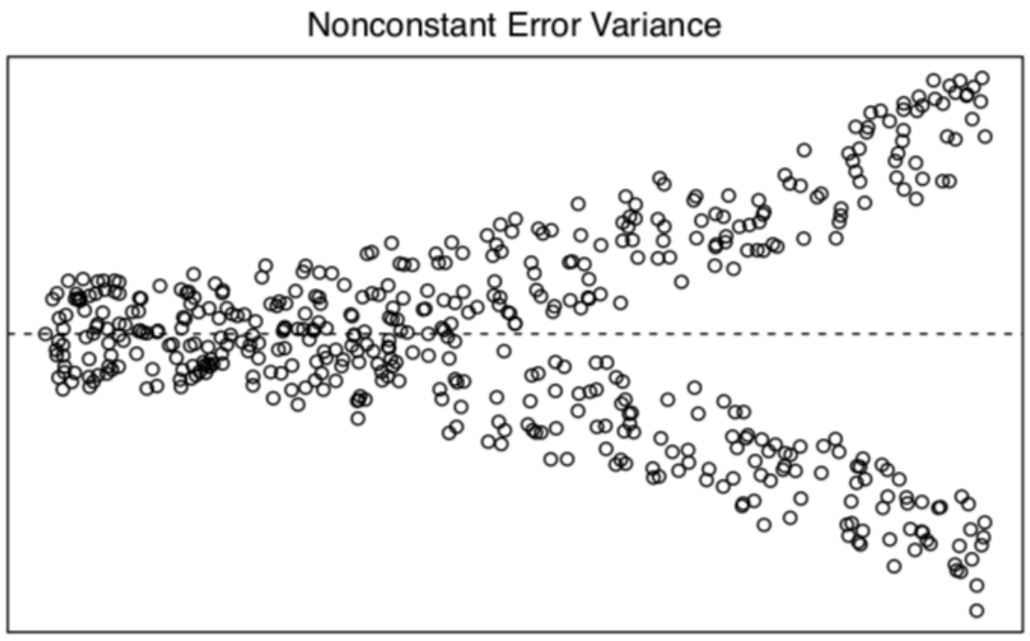


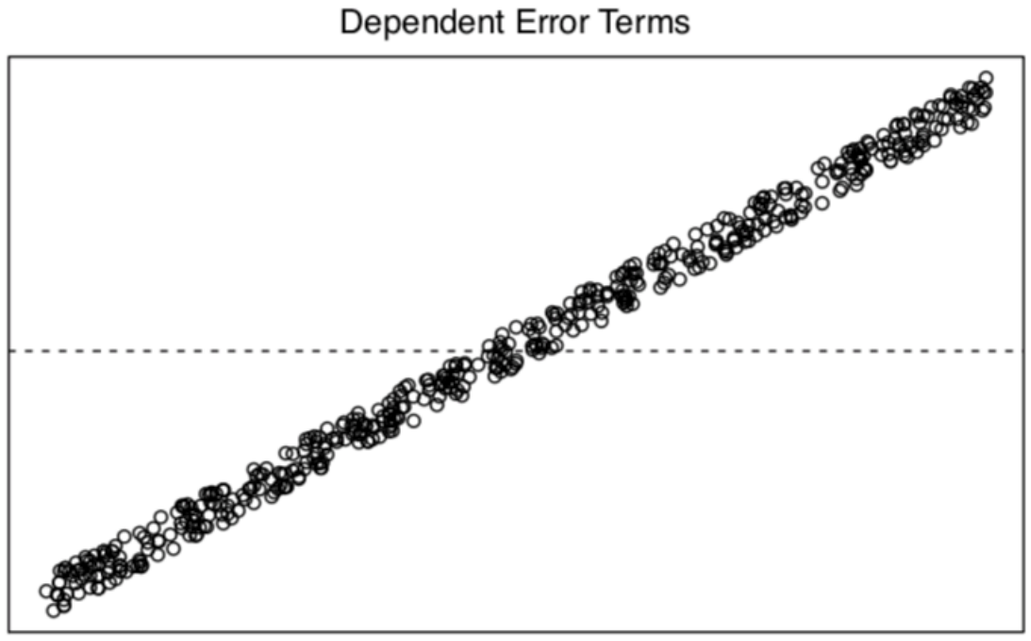


Model assumptions of simple regression



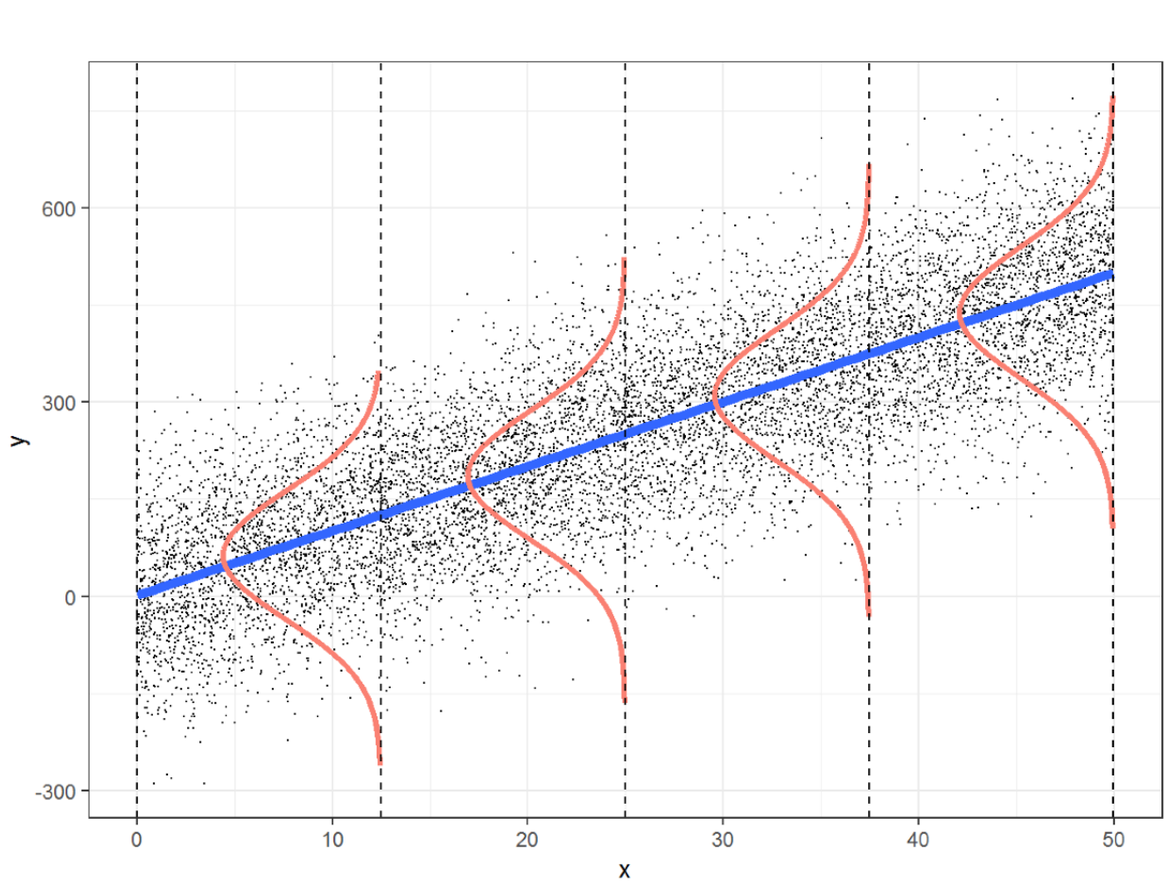






Assumptions of multiple regression

- 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

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

Simple regression

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

one predictor

Multiple regression

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

many predictors

Advertising data set

money spent on different media (x $1000)

sales (x1000)

- 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?

Visualizing correlations

Visualizing correlations

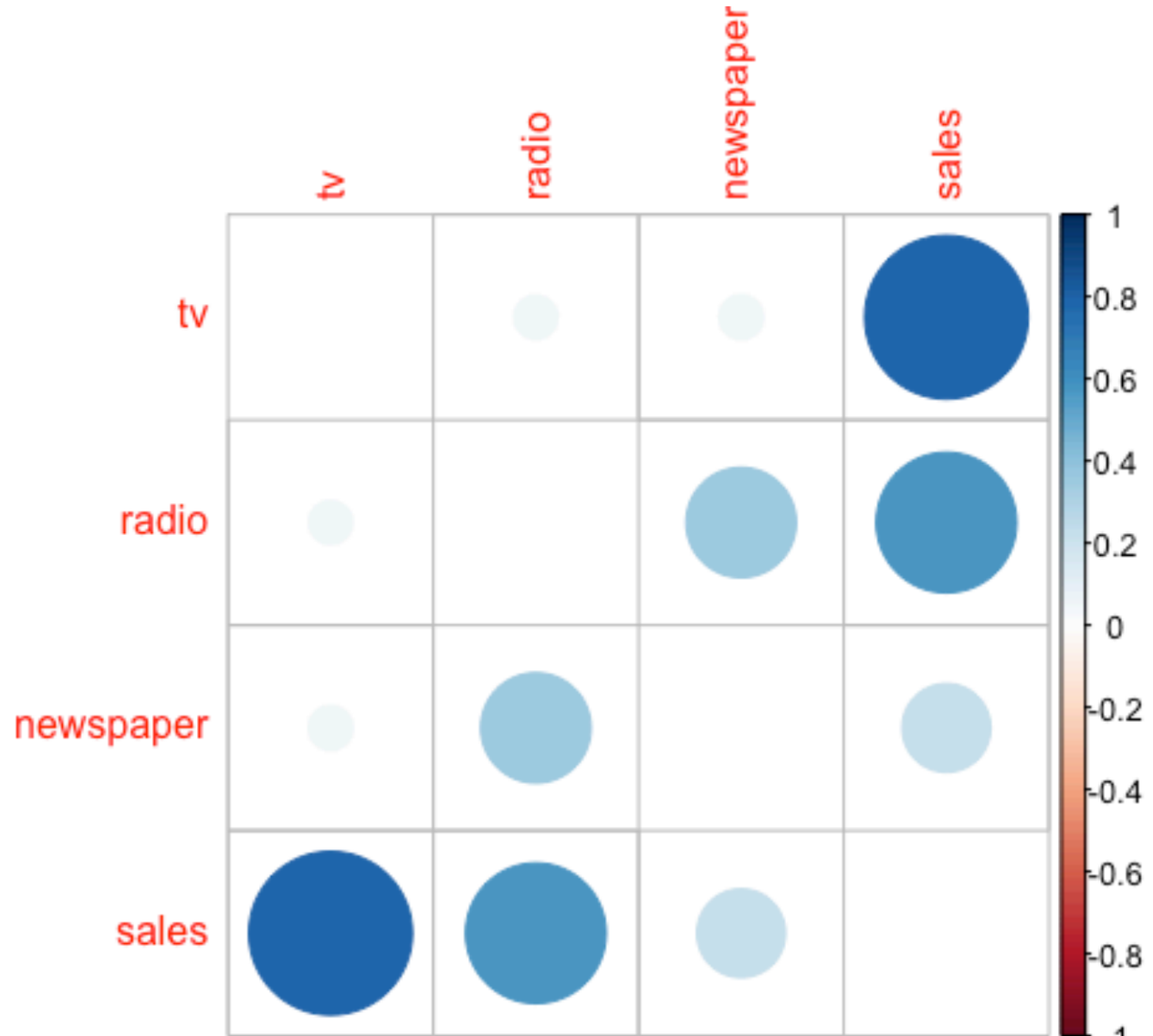
```
library("corrr")
```


```
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	

Visualizing correlations

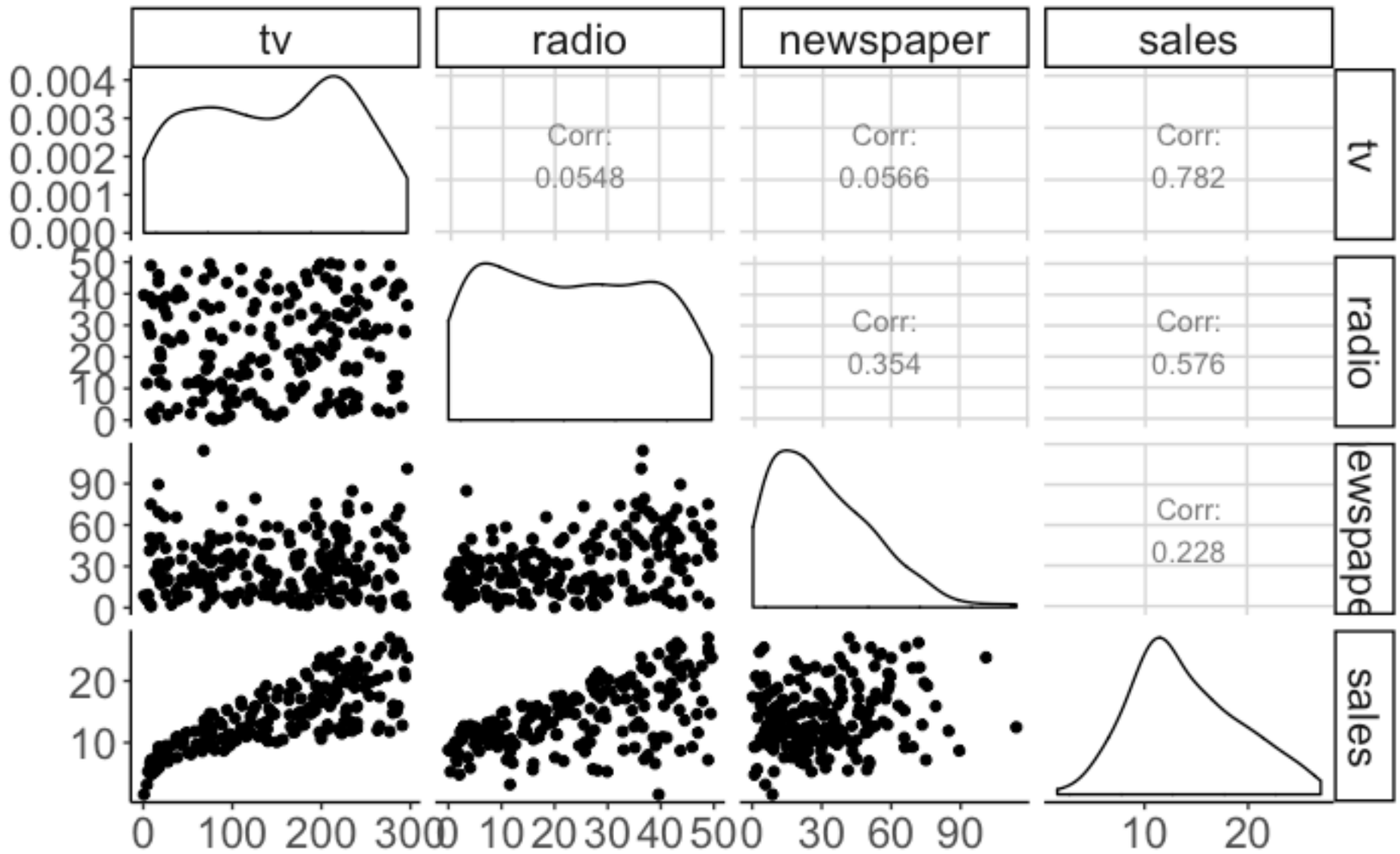
```
df.ads %>%
  select(-index) %>%
  correlate() %>%
  column_to_rownames() %>%
  as.matrix() %>%
  corrplot(diag = F)
```



Visualizing correlations

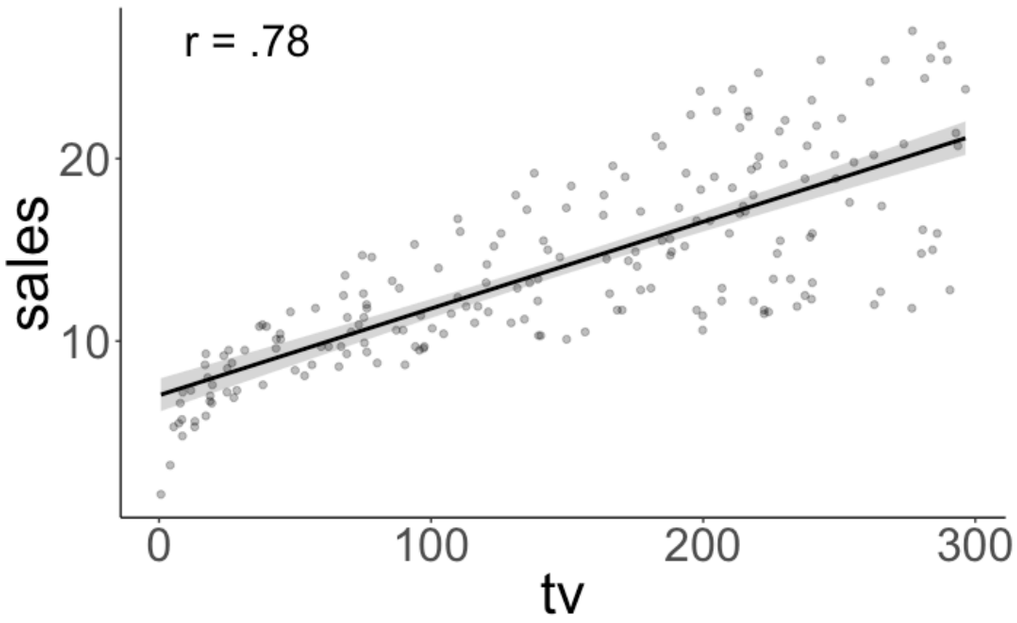
```
library("GGally")
```

```
1 df.ads %>%
2    select(-index) %>%
3    ggpairs()
```

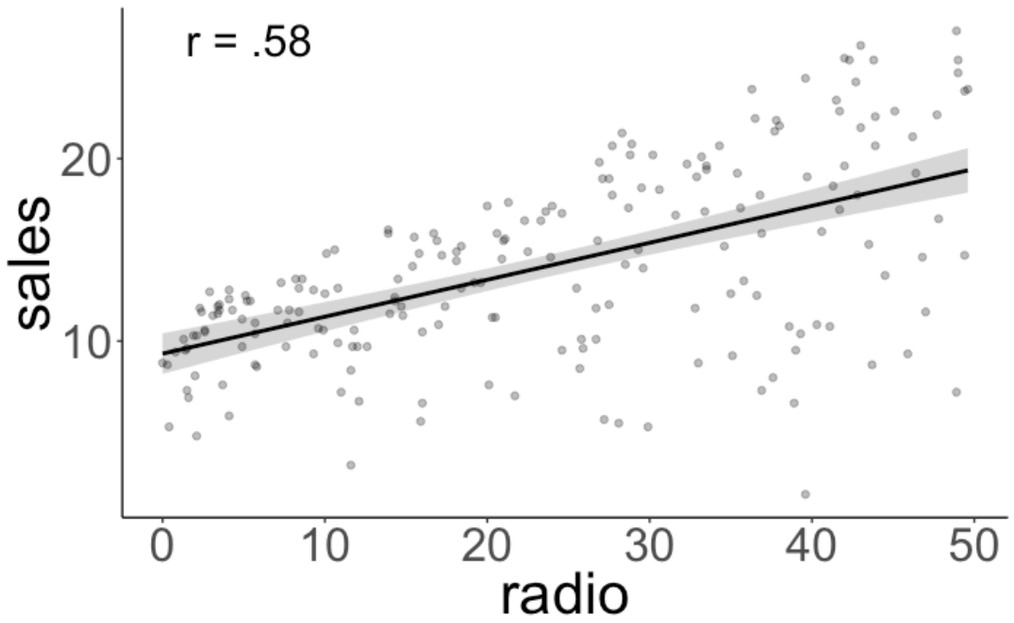


$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + b_2 \cdot \text{radio}_i + e_i$$

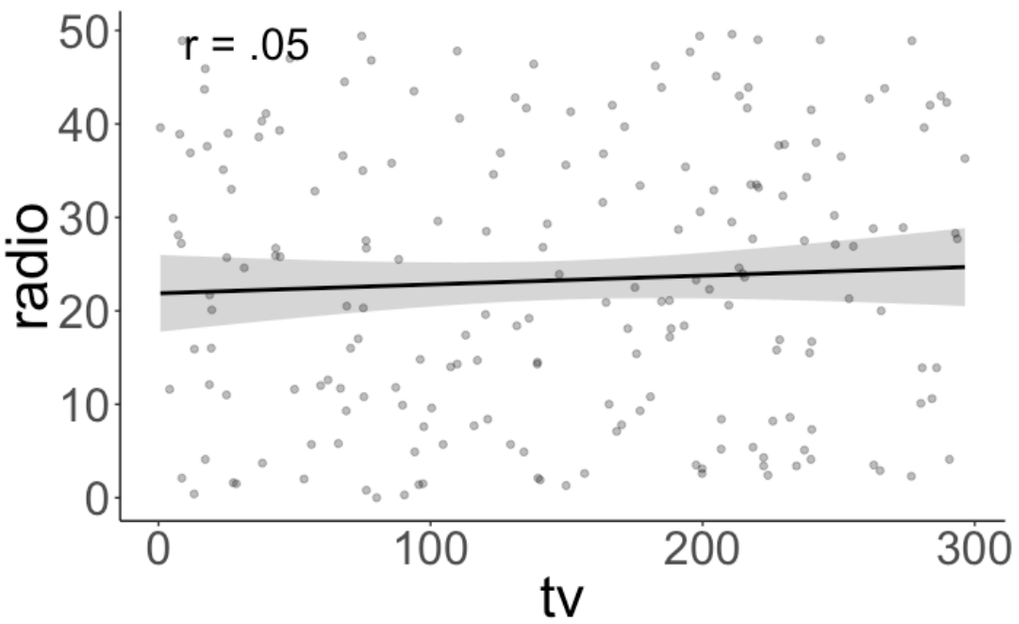
Relationship between TV ads and sales



Relationship between radio ads and sales



Relationship between TV ads and radio ads



predictors are not correlated, yay!

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.

Model C

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

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

Model A

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

```
# fit the models
fit_c = lm(sales ~ 1 + tv, data = df.ads)
fit_a = lm(sales ~ 1 + tv + radio, data = df.ads)

# do the F test
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.

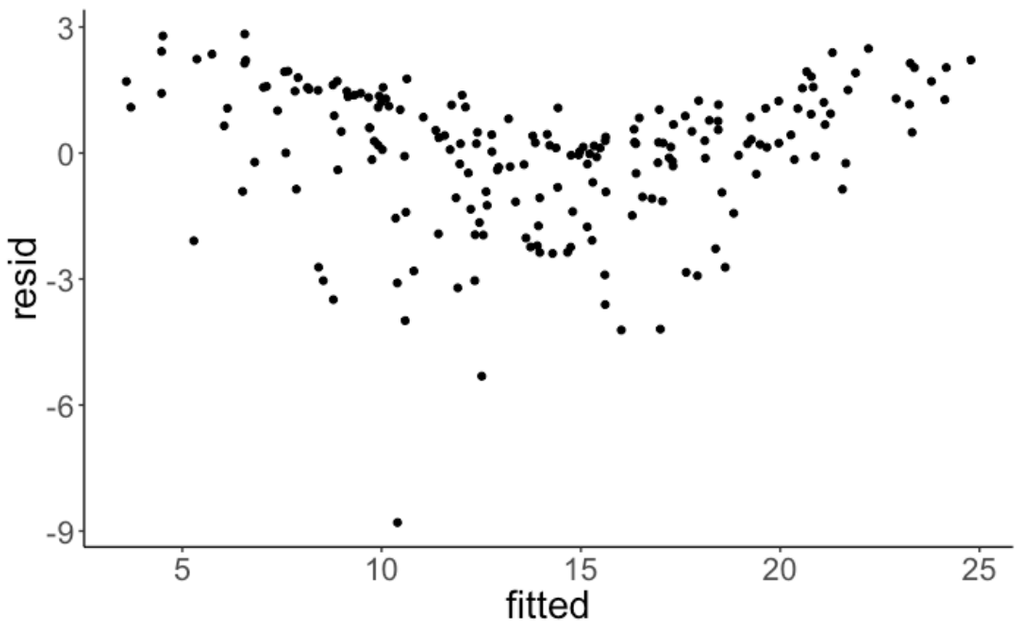
Augmented Model

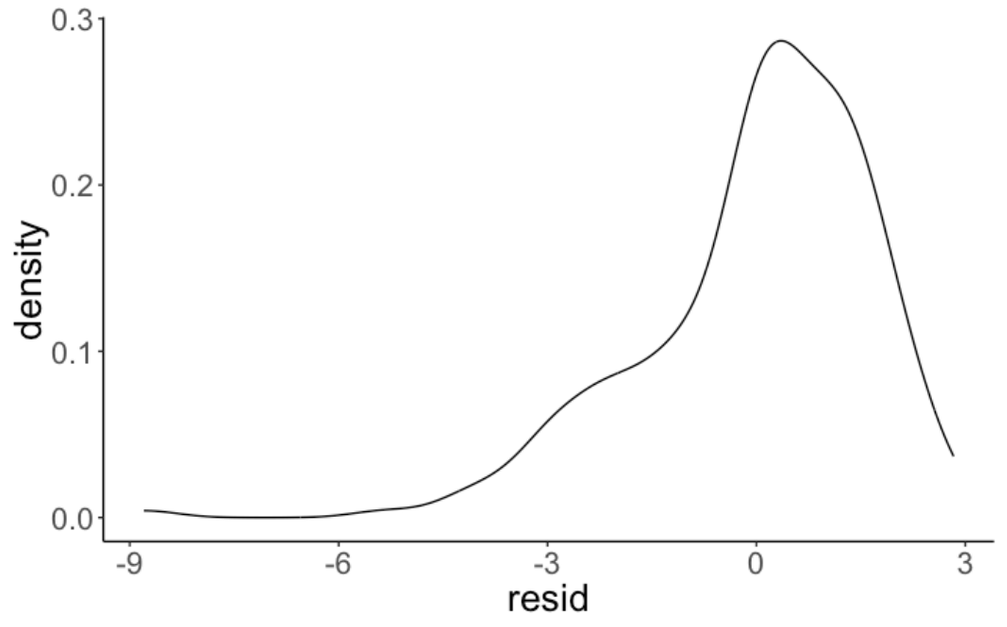
$$\text{sales}_i = b_0 + b_1 \cdot \text{tv}_i + b_2 \cdot \text{radio}_i + e_i$$

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

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

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

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

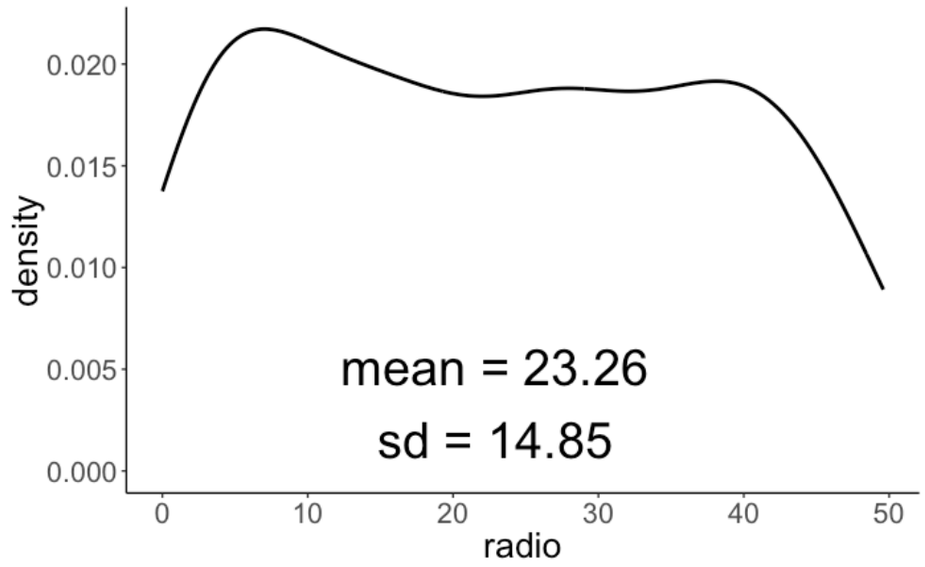
For a given amount of TV advertising, an additional \$1000 on radio advertising leads to an increase in sales by 190 units.

anything surprising?

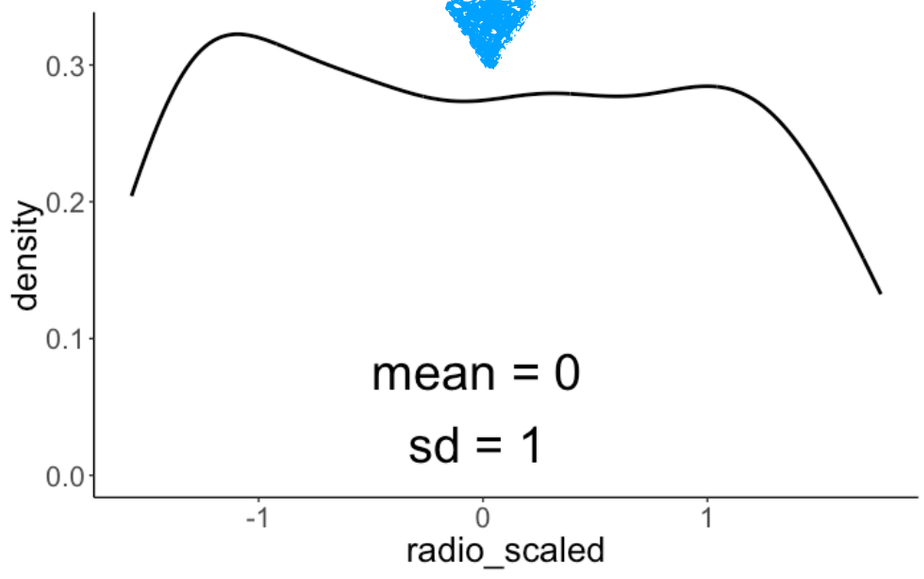
Standardizing predictors

$$z_i = \frac{Y_i - \overline{Y}}{s}$$

index	tv	radio	sales	tv_scaled	radio_scaled
1	230.1	37.8	22.1	0.97	0.98
2	44.5	39.3	10.4	-1.19	1.08
3	17.2	45.9	9.3	-1.51	1.52
4	151.5	41.3	18.5	0.05	1.21
5	180.8	10.8	12.9	0.39	-0.84
6	8.7	48.9	7.2	-1.61	1.73
7	57.5	32.8	11.8	-1.04	0.64
8	120.2	19.6	13.2	-0.31	-0.25
9	8.6	2.1	4.8	-1.61	-1.43
10	199.8	2.6	10.6	0.61	-1.39



$$z_i = \frac{Y_i - \overline{Y}}{s}$$



Interpreting the results

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

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	14.02	0.12	117.94	0	13.79	14.26
tv_scaled	3.93	0.12	32.91	0	3.69	4.16
radio_scaled	2.79	0.12	23.38	0	2.56	3.03

$$\widehat{\text{sales}}_i = 14.02 + 3.93 \cdot \text{tv\_z}_i + 2.79 \cdot \text{radio\_z}_i$$

For a given amount of TV advertising, increasing radio advertising by one standard deviation (23.26 x $1000) leads to an increase in sales by 2790 units.

Interpreting the results

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

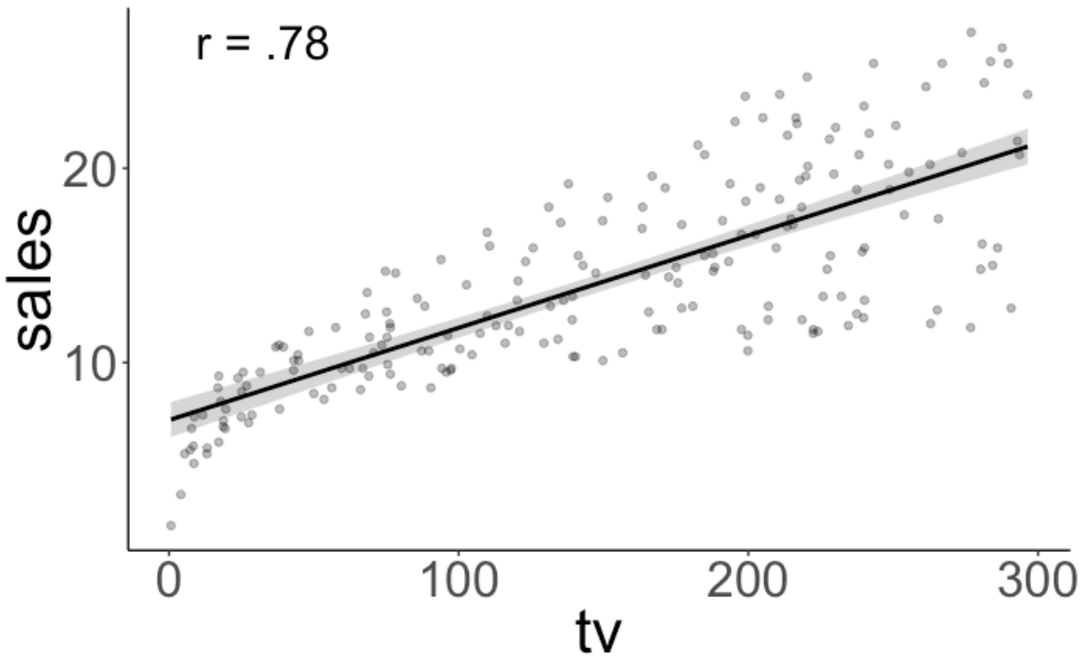
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	14.02	0.12	117.94	0	13.79	14.26
tv_scaled	3.93	0.12	32.91	0	3.69	4.16
radio_scaled	2.79	0.12	23.38	0	2.56	3.03

$$\widehat{\text{sales}}_i = 14.02 + 3.93 \cdot \text{tv\_z}_i + 2.79 \cdot \text{radio\_z}_i$$

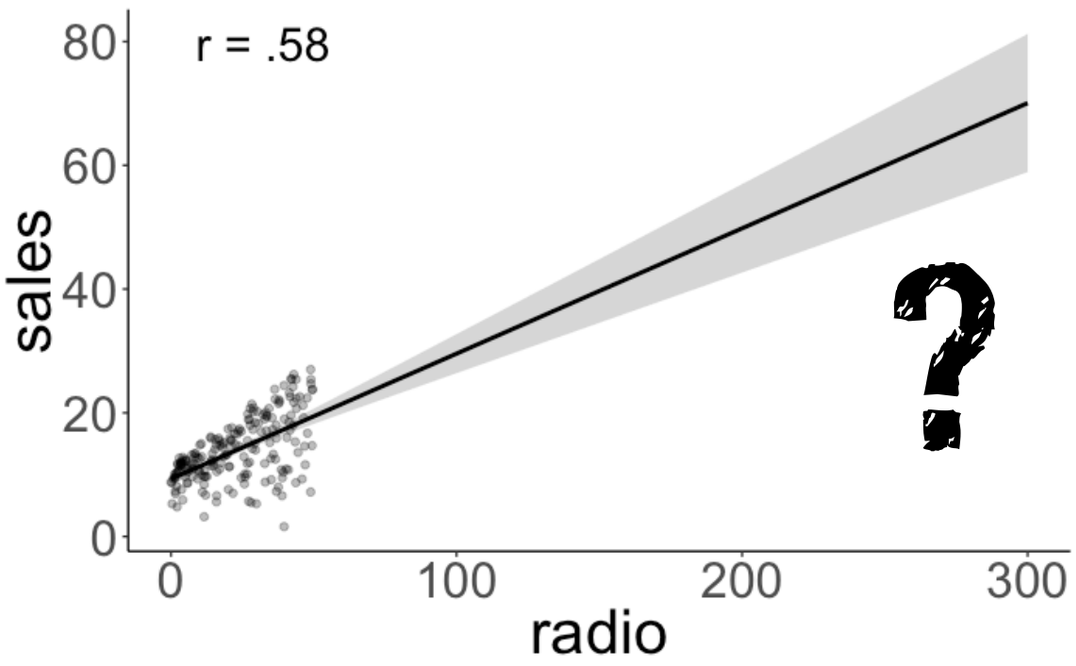
For a given amount of radio advertising, increasing TV advertising by one standard deviation (85.85 x $1000) leads to an increase in sales by approximately 3930 units.

On average, 14,020 units are being sold.

Question of extrapolation ...

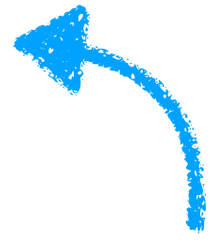


Radio ads give more bang for the buck but it's unclear whether this would extrapolate ...



Interpreting coefficients

- For multiple regression, the meaning of a coefficient is: How much is the outcome predicted to change for a unit increase in the predictor, holding all the other predictors fixed.
- Standardizing predictors can help with interpretation (particularly when comparing predictors with different units, ranges, ...).
- Standardizing coefficients means that you can compare the relative importance of each coefficient in a regression model.



well, sort of ...

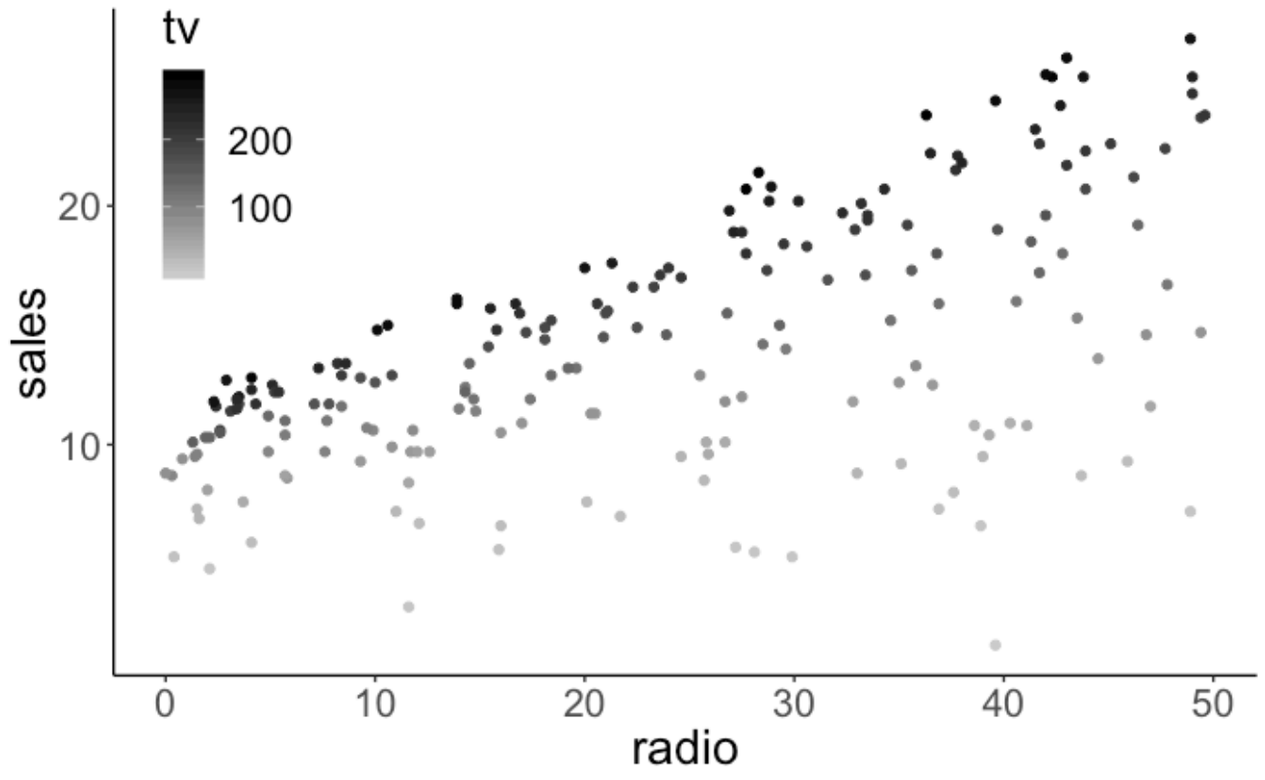
Interpreting coefficients

"The fundamental problem is that we cannot frame a question about relative importance in terms of a Model A versus Model C comparison because both models would necessarily have the same predictors. Hence, our machinery for statistical inference cannot address the question of relative importance.

Relative importance of predictor variables is a slippery concept. Any statement about relative importance must be accompanied by many caveats that it applies to this particular range of variables in this situation. As a result they are, in our opinion, seldom useful. **Thus, although it is tempting to compare the importance of predictors in multiple regression, it is almost always best not to do so.**"

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). Data analysis: A model comparison approach. Routledge.

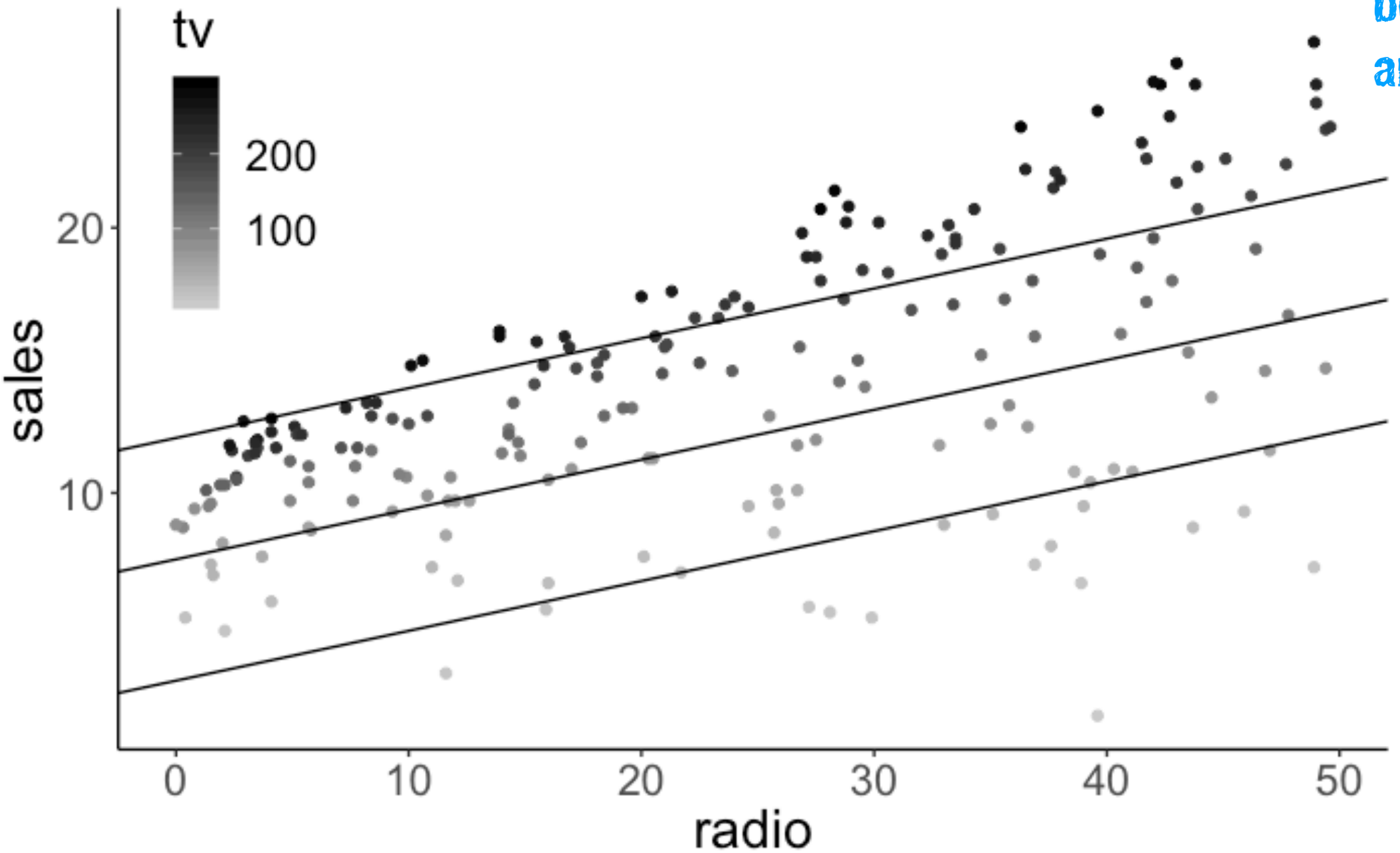
Reporting results



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 predicted to increase sales by 190 units [170, 200] (95% confidence intervals).

Visualizing the results



predicted relationship between radio and sales when:

tv = 200

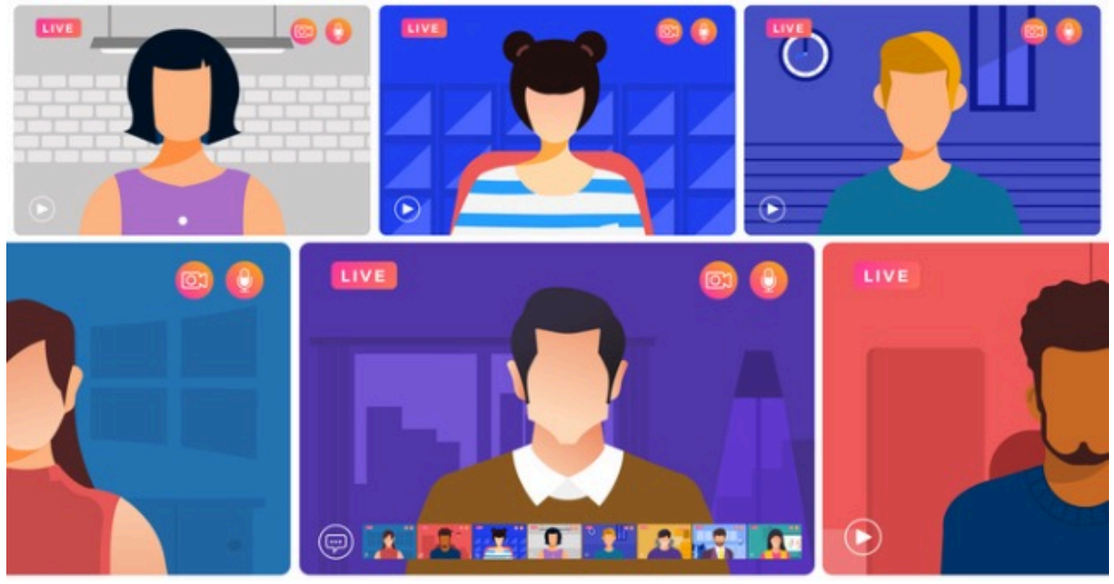
tv = 100

tv = 0

just imagine infinitely many more lines ...

Why can't I just run several simple regressions?

Breakout rooms



Task:

- What's the point of multiple regression?
- Why can't we just run several simple regressions instead?

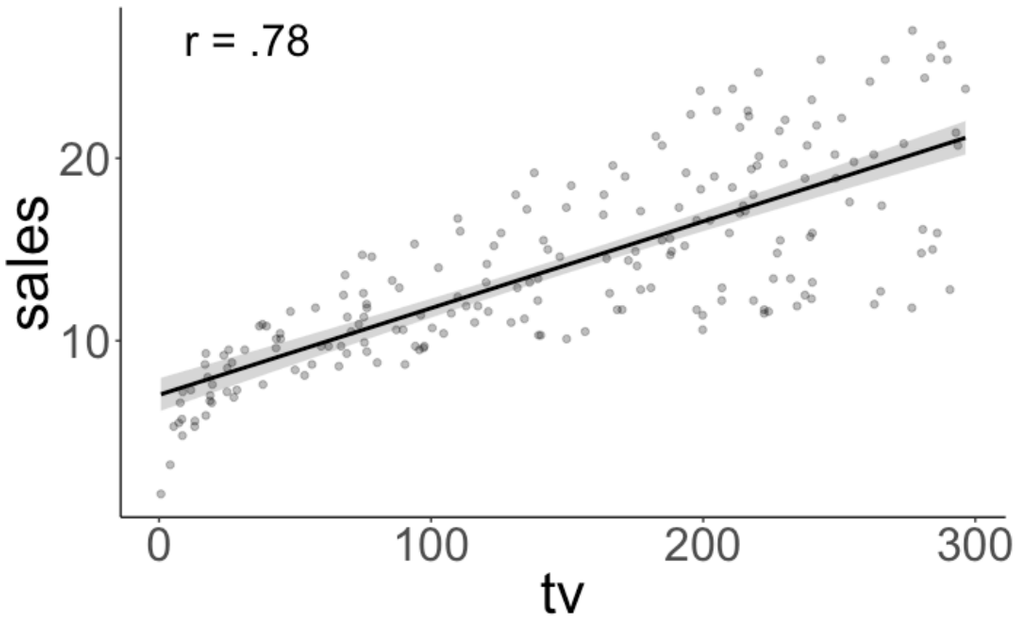
Discuss these questions with your breakout room group.

Size: ~3 people

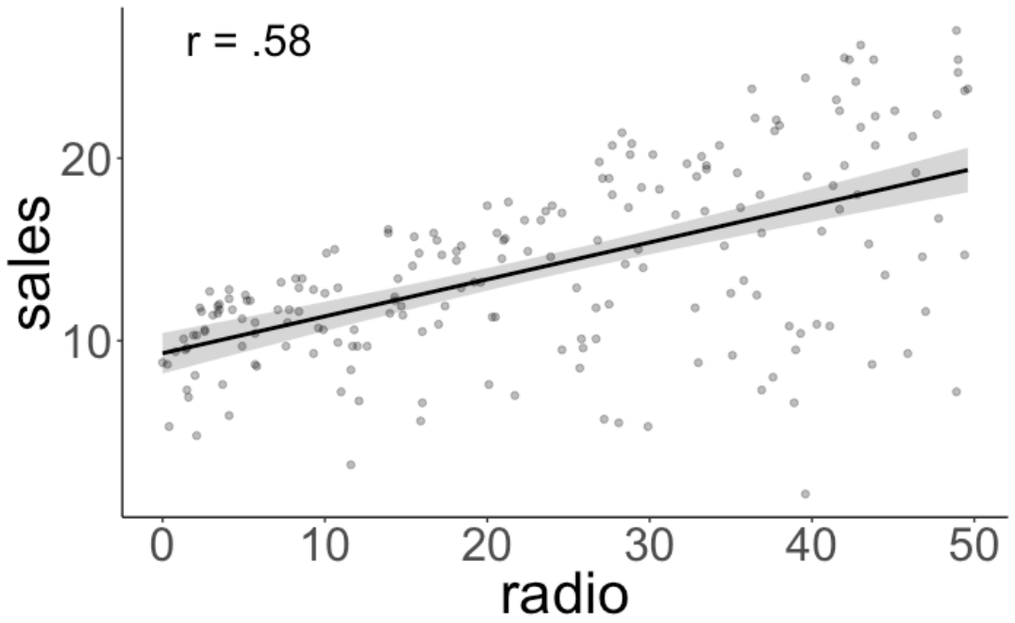
Time: 5 minutes

Report: I'll ask a few of you to share with the larger group.

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

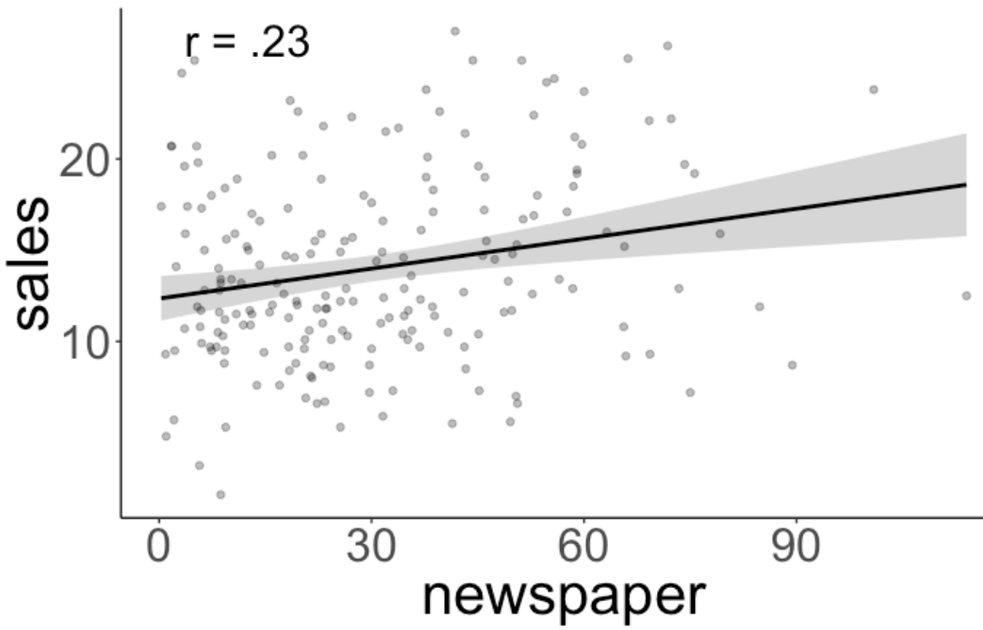
money spent on different media (x $1000)

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

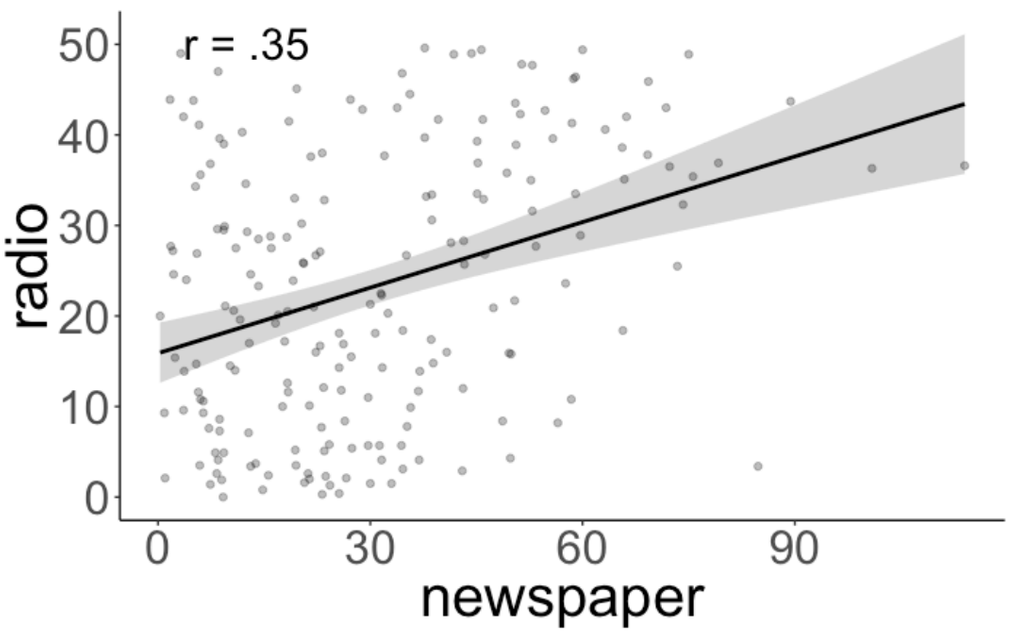
Are newspaper ads and sales related when controlling for radio ads and TV ads?

Relationship between newspaper ads and sales

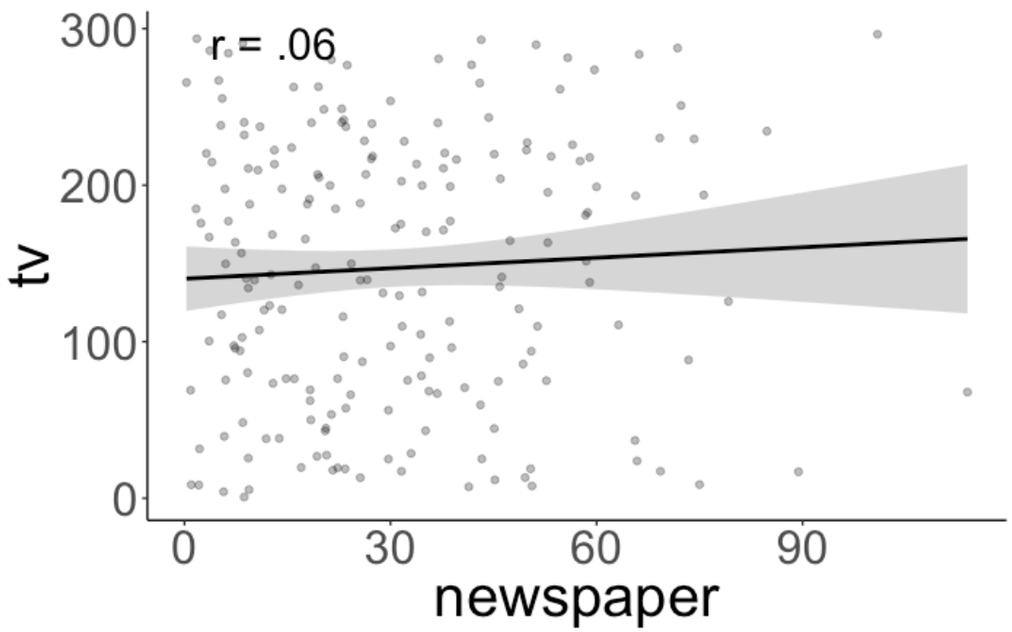


this is significant

Relationship between newspaper and radio ads



Relationship between newspaper and TV ads



```r
# fit the models
fit_c = lm(sales ~ 1 + tv + radio, data = df.ads)
fit_a = lm(sales ~ 1 + tv + radio + newspaper, data = df.ads)

# do the F test
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

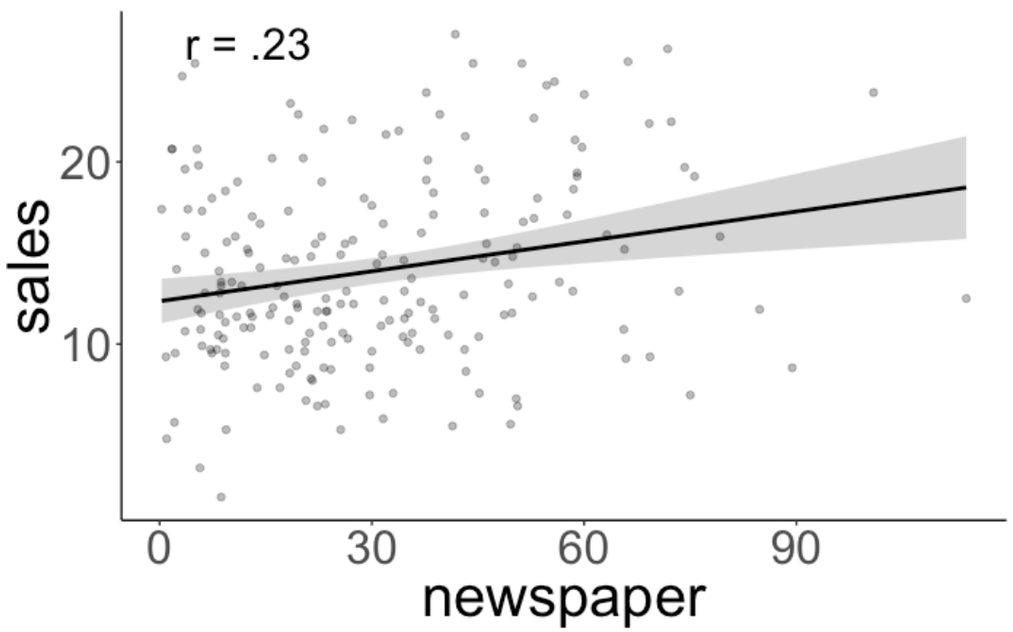
`sales ~ 1 + radio`

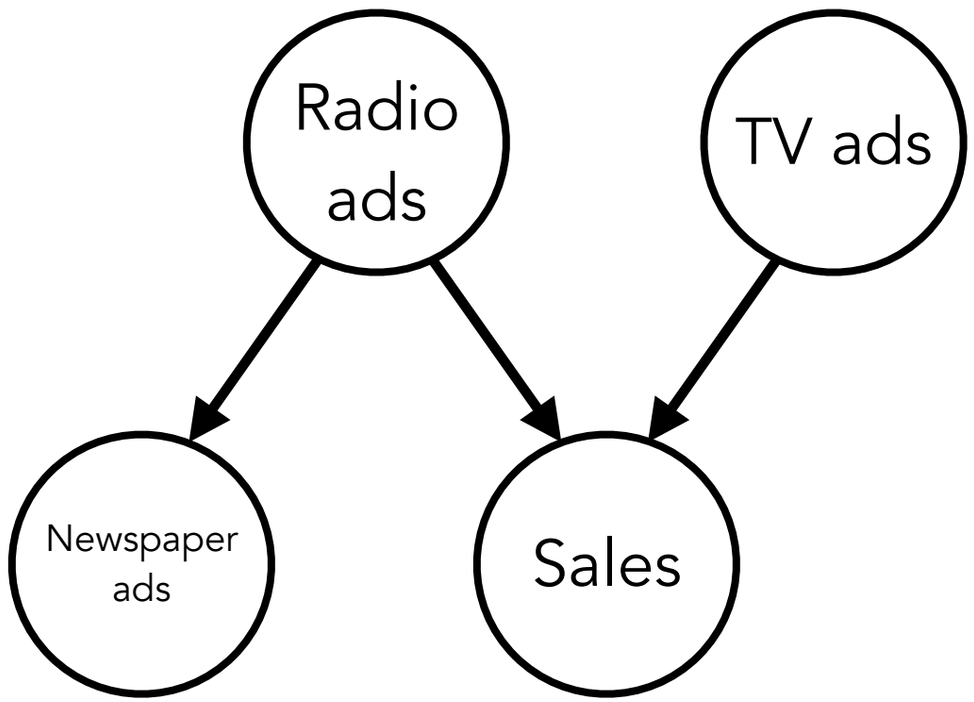
`sales ~ 1 + radio + newspaper`

it's not worth it

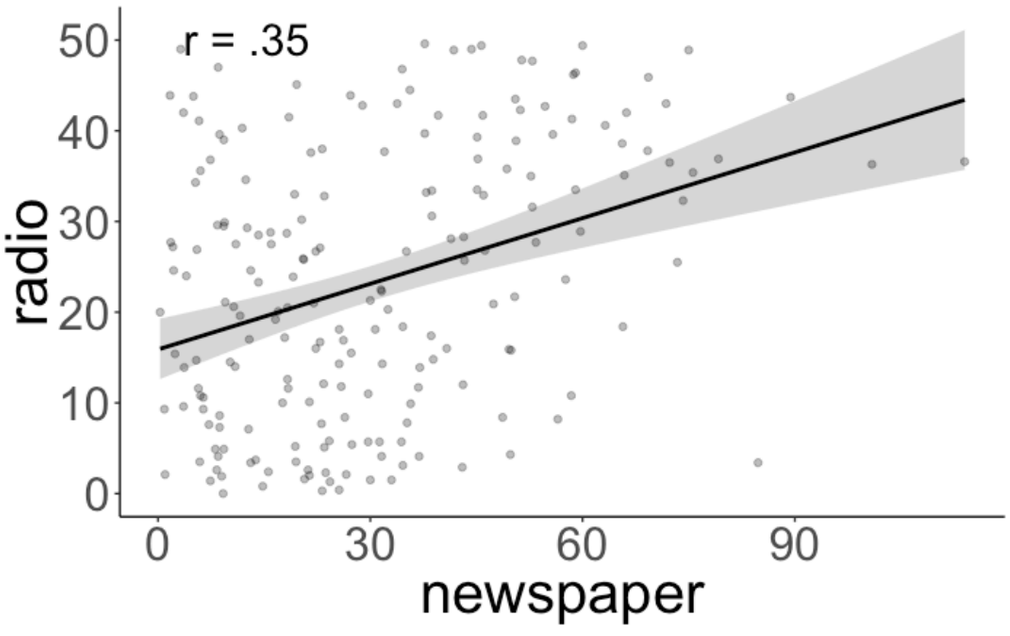
Are newspaper ads and sales related when controlling for radio ads and TV ads?

Relationship between newspaper ads and sales

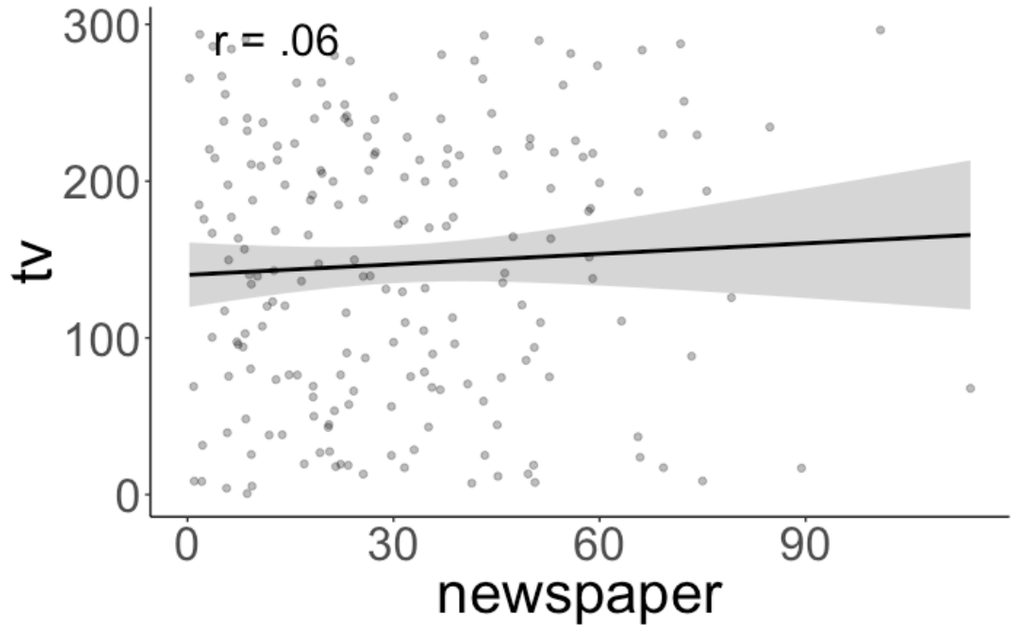




Relationship between newspaper and radio ads



Relationship between newspaper and TV ads



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

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

`nrow(df.credit) = 400`

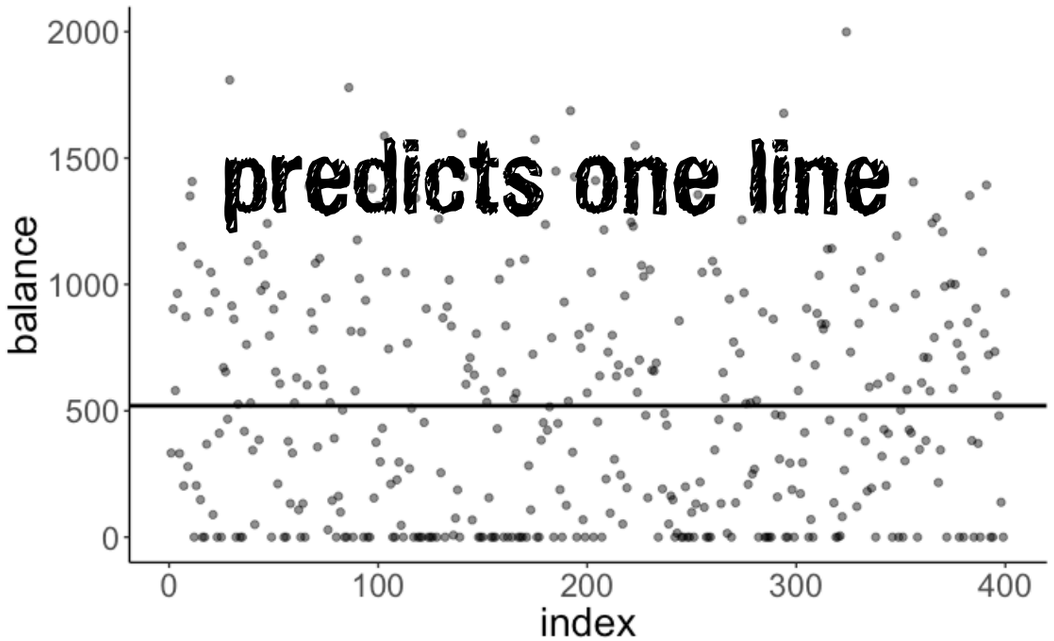
Do students have a different credit card balance from non-students?

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

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 520.02 + e_i$$

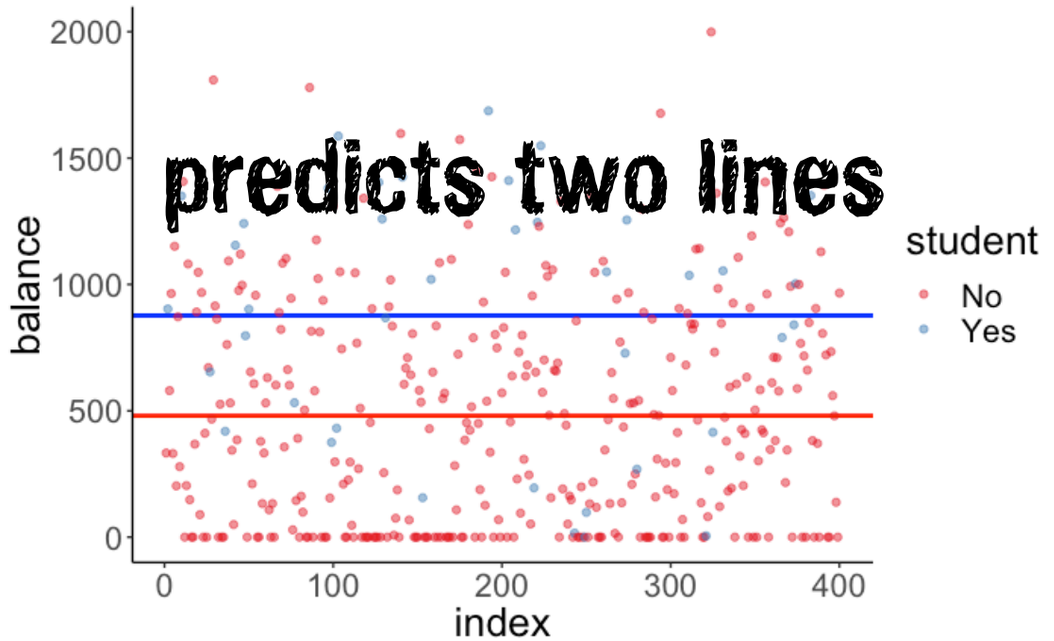
H_1: Students and non-students have different balances.

Model A

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

student

Model prediction



Fitted model

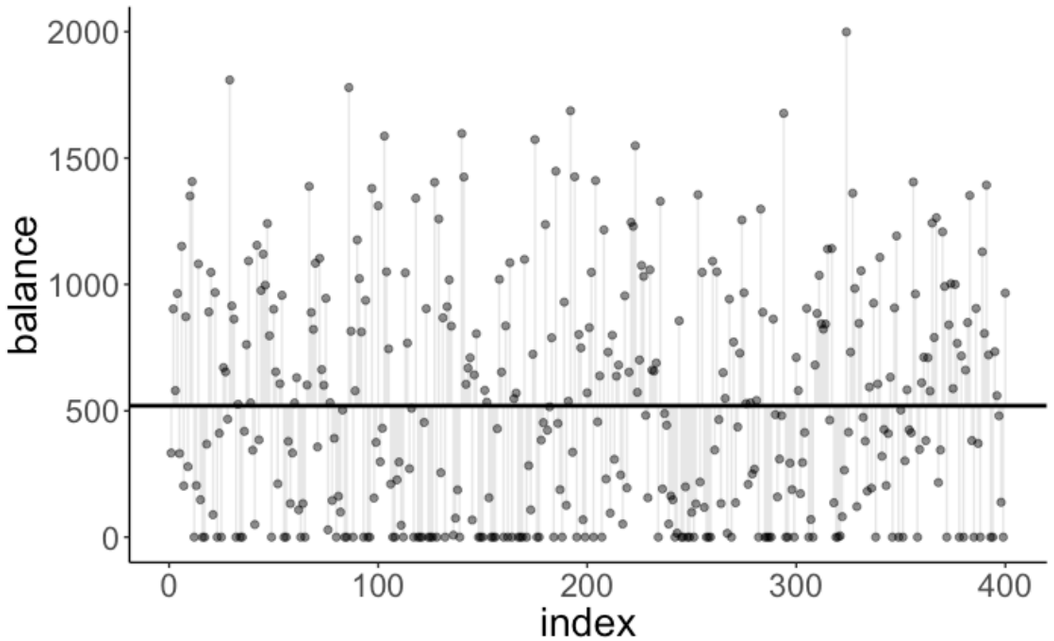
$$Y_i = 480.37 + 396.46 X_i + e_i$$

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

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 520.02 + e_i$$

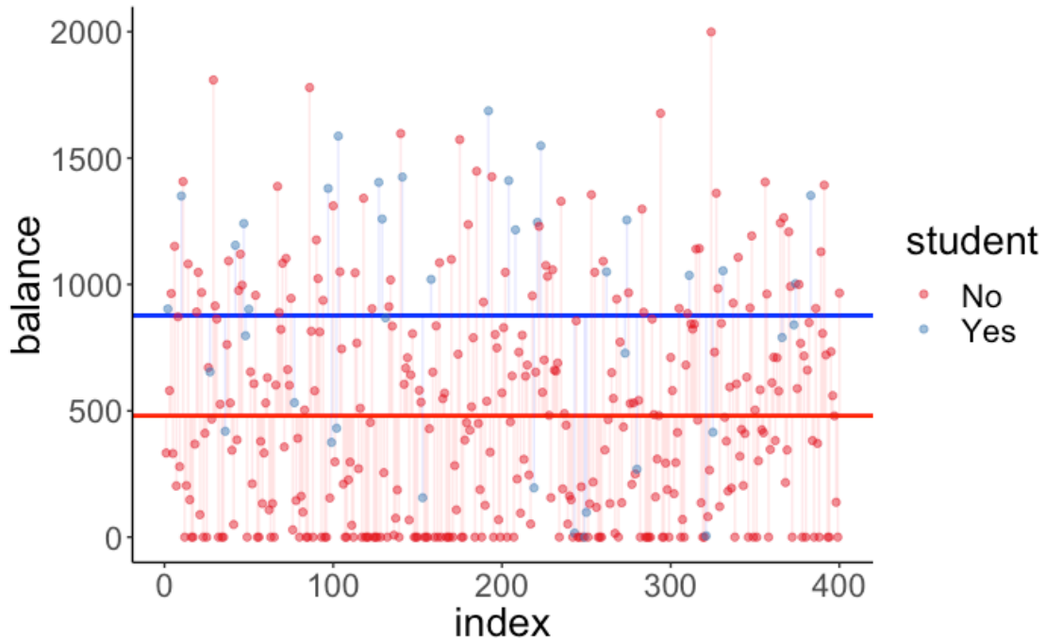
H_1: Students and non-students have different balances.

Model A

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

student

Model prediction



Fitted model

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

Worth it?

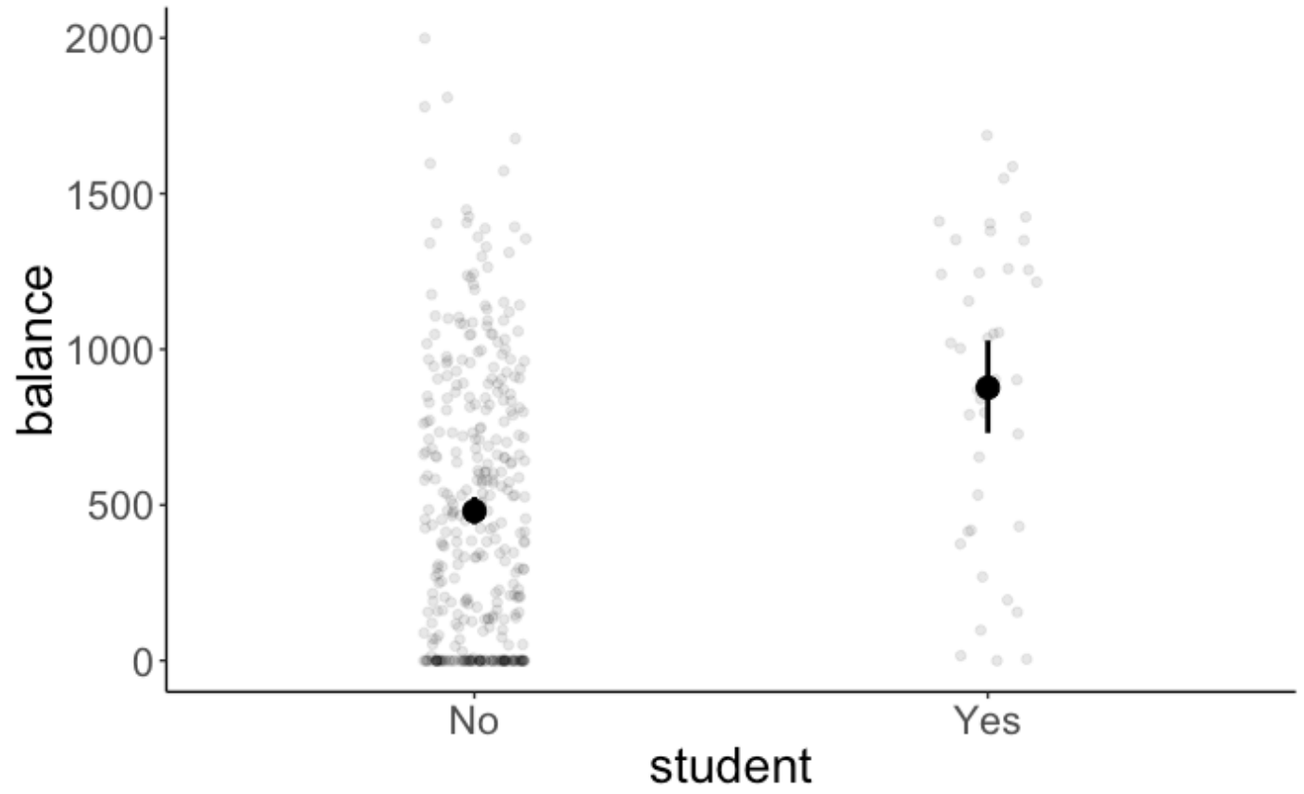
```r
# fit the models
fit_c = lm(balance ~ 1, data = df.credit)
fit_a = lm(balance ~ student, data = df.credit)

# run the F test
anova(fit_c, fit_a)
```

```
Analysis of Variance Table

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
```

Worth it!

Two sample t-test (with independent groups)

Reporting the results



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

Interpreting the model

```
1 fit_a = lm(balance ~ 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
No	0
No	0
No	0
No	0
No	0
No	0
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

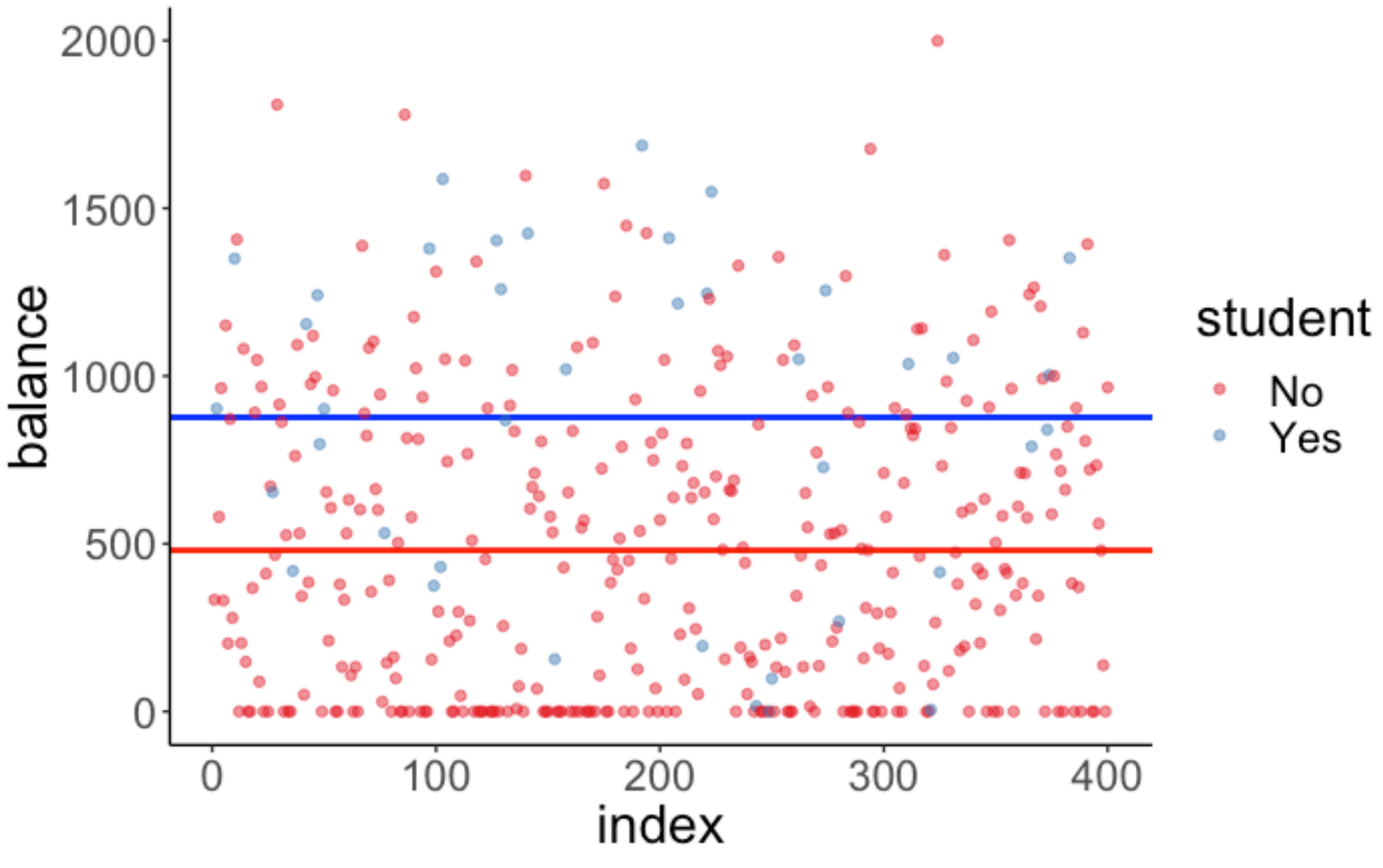
Dummy coding

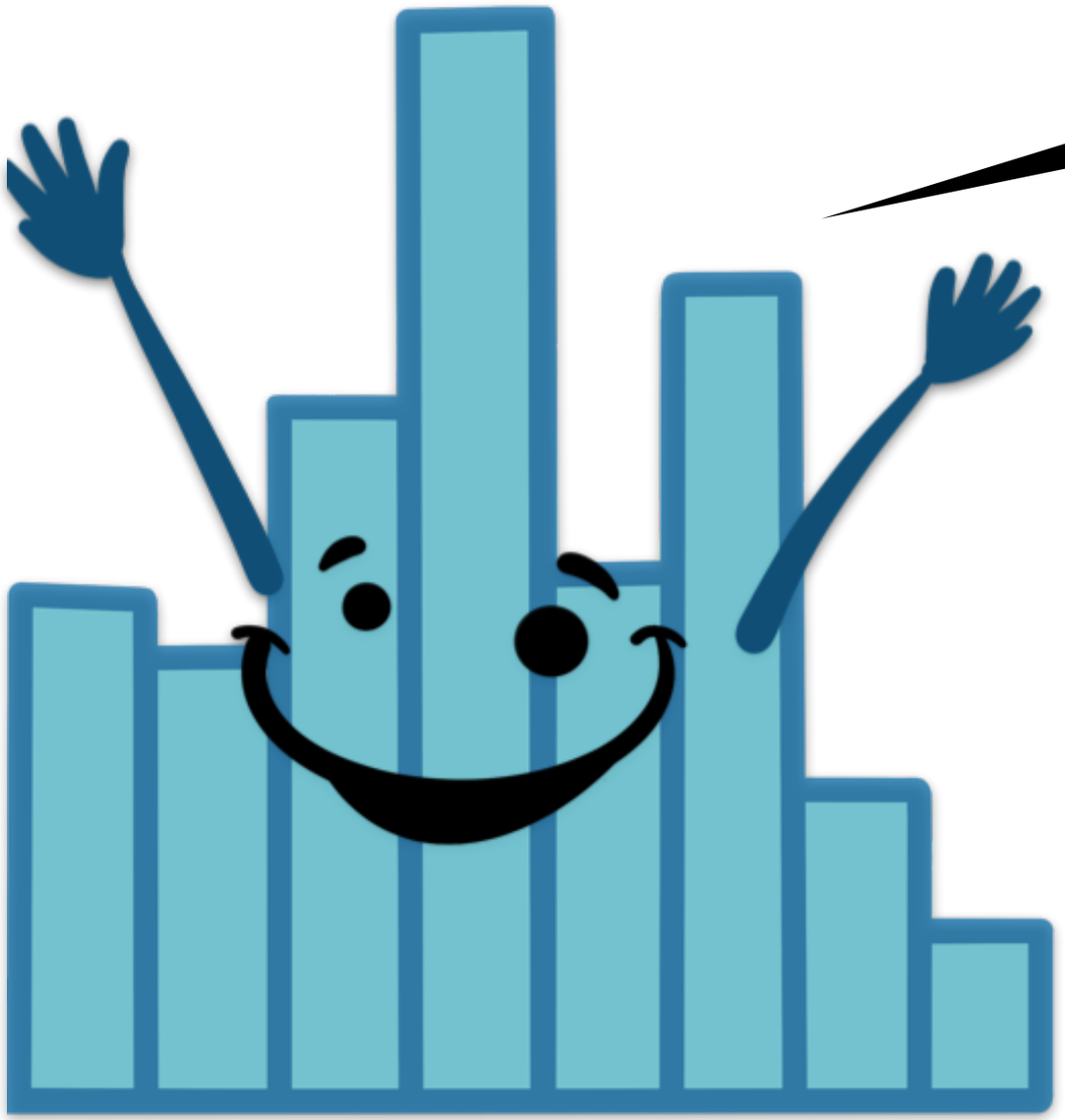
$$\hat{Y}_i = 480.37 + 396.46 \cdot \text{student\_dummy}_i$$

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
No	0
No	0
No	0
No	0
No	0
No	0
No	0
No	0
Yes	1





01:00

stretch break!

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

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

`nrow(df.credit) = 400`

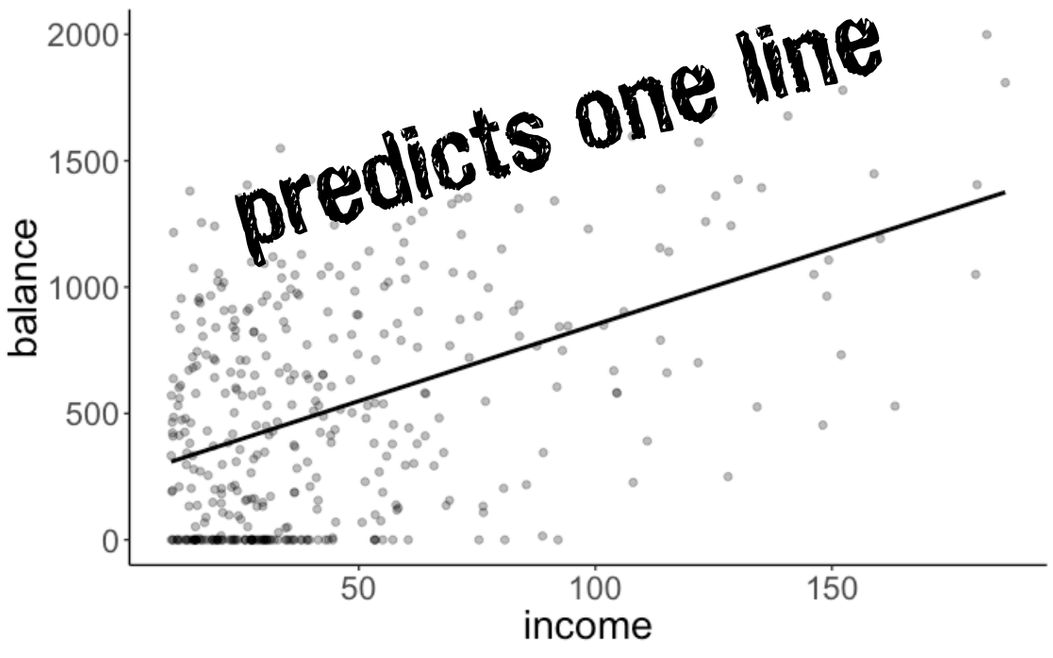
Do students have a different credit card balance from non-students, when controlling for income?

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

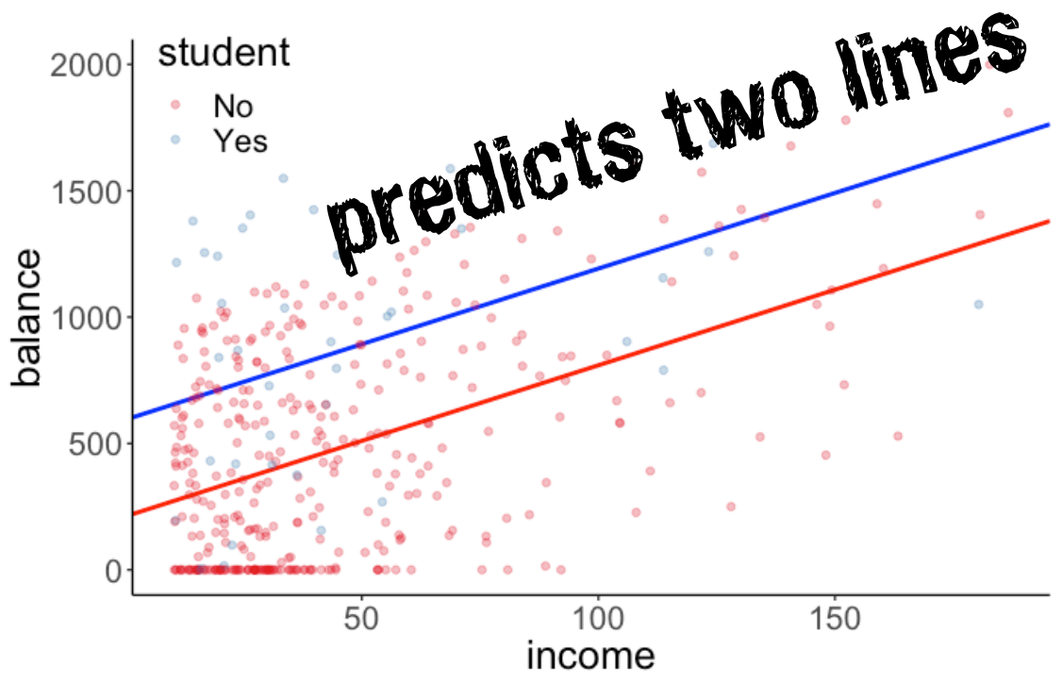
$$\widehat{\text{balance}}_i = 246.515 + 6.048 \cdot \text{income}_i$$

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

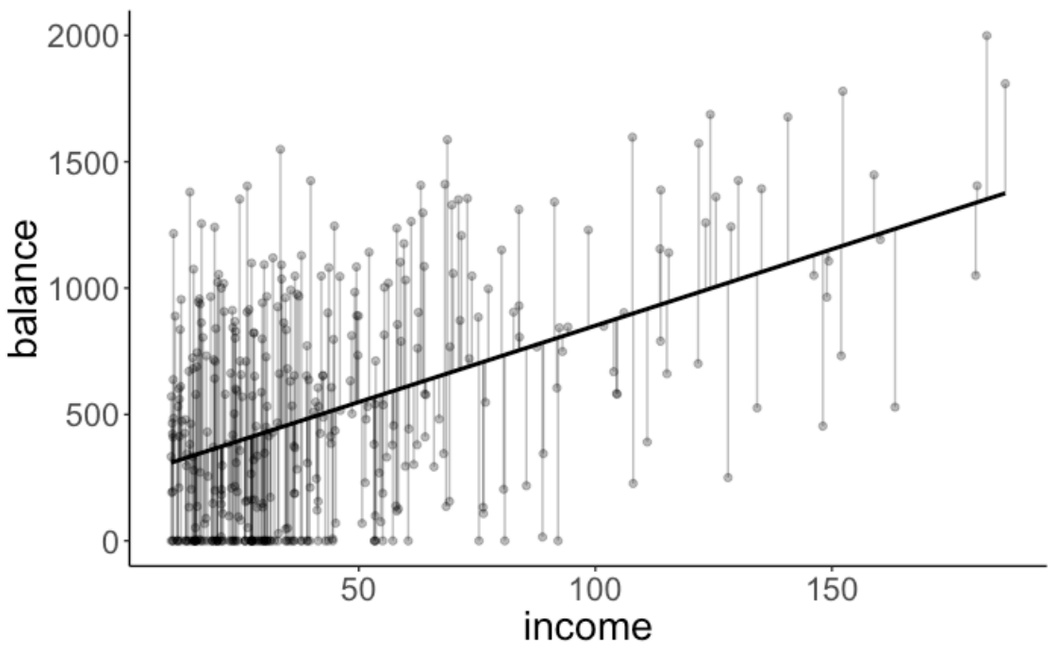
$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

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

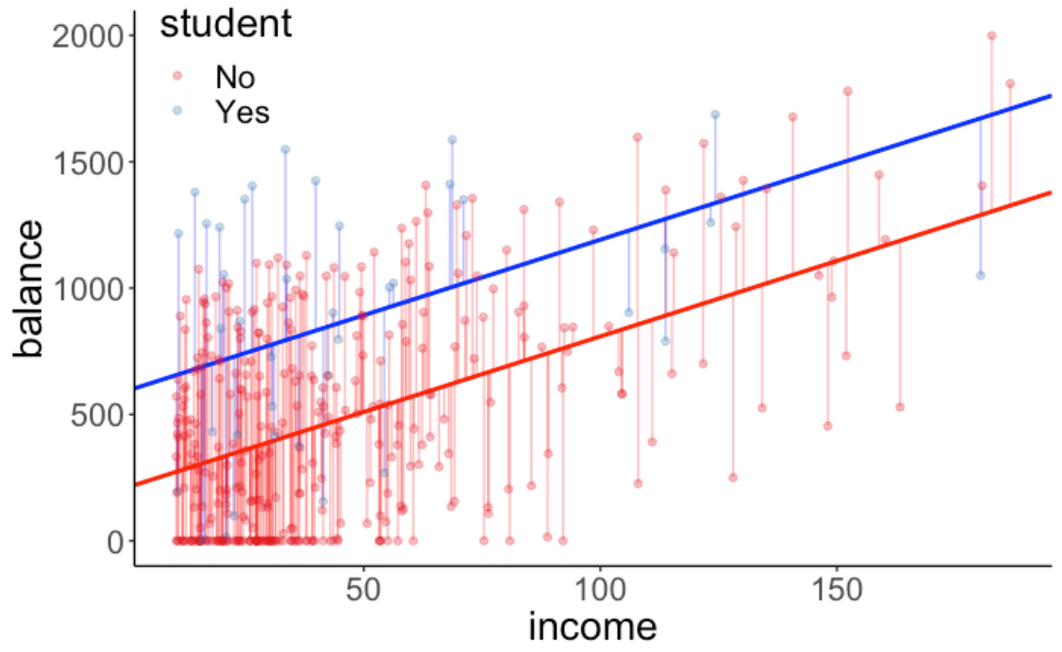
$$\widehat{\text{balance}}_i = 246.515 + 6.048 \cdot \text{income}_i$$

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

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

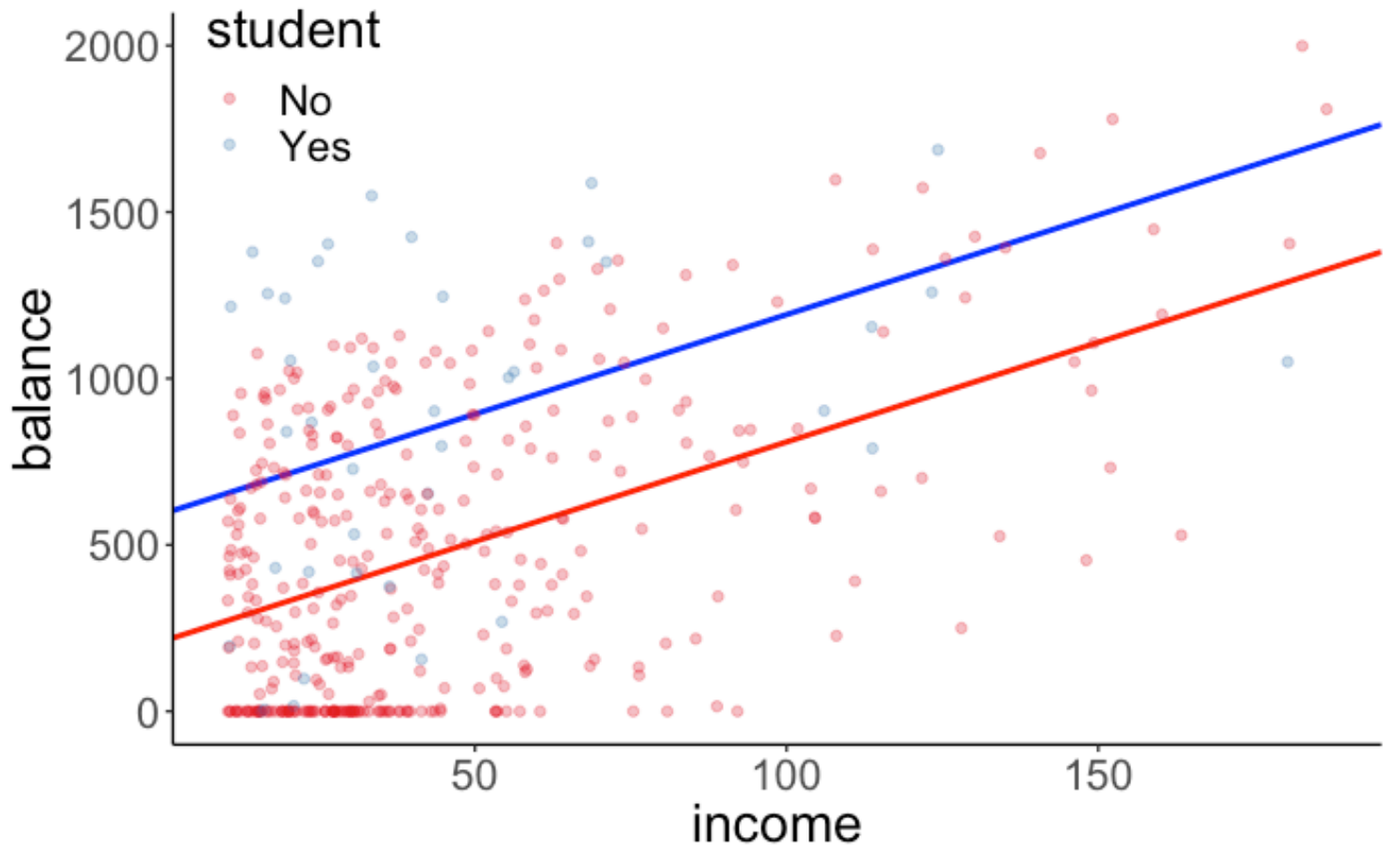
Worth it?

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

Worth it!

Interpretation

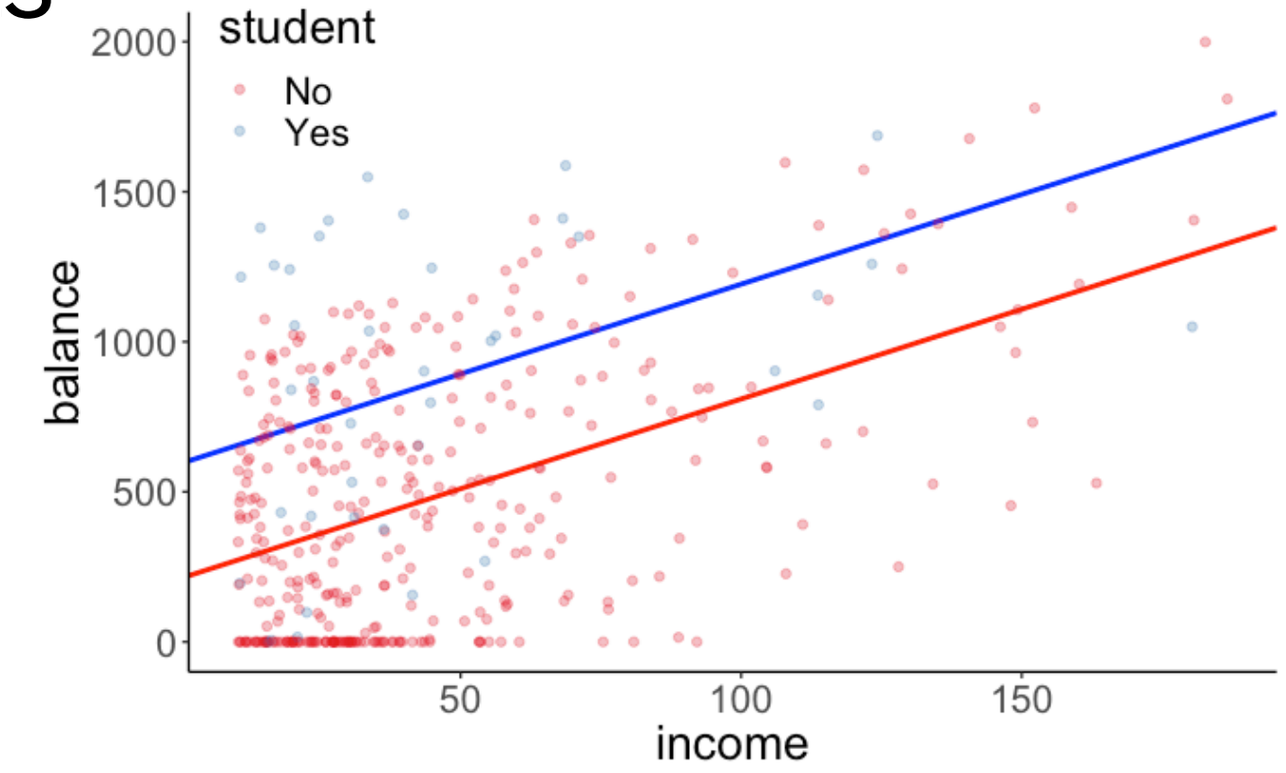


$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

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

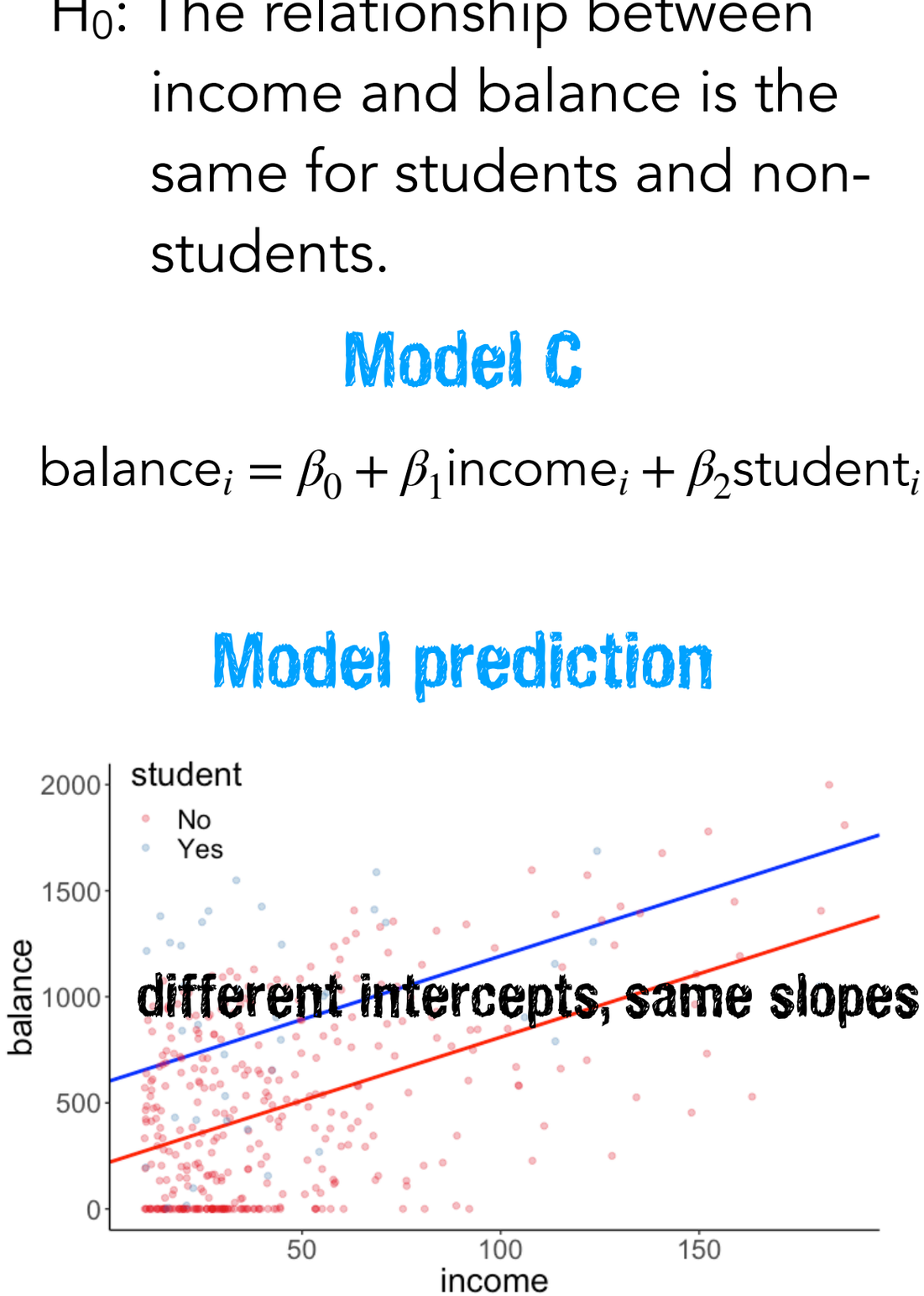
$$\widehat{\text{balance}}_i = b_0 + b_1\text{income}_i + b_2\text{student}_i + b_3(\text{income}_i \times \text{student}_i)$$

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

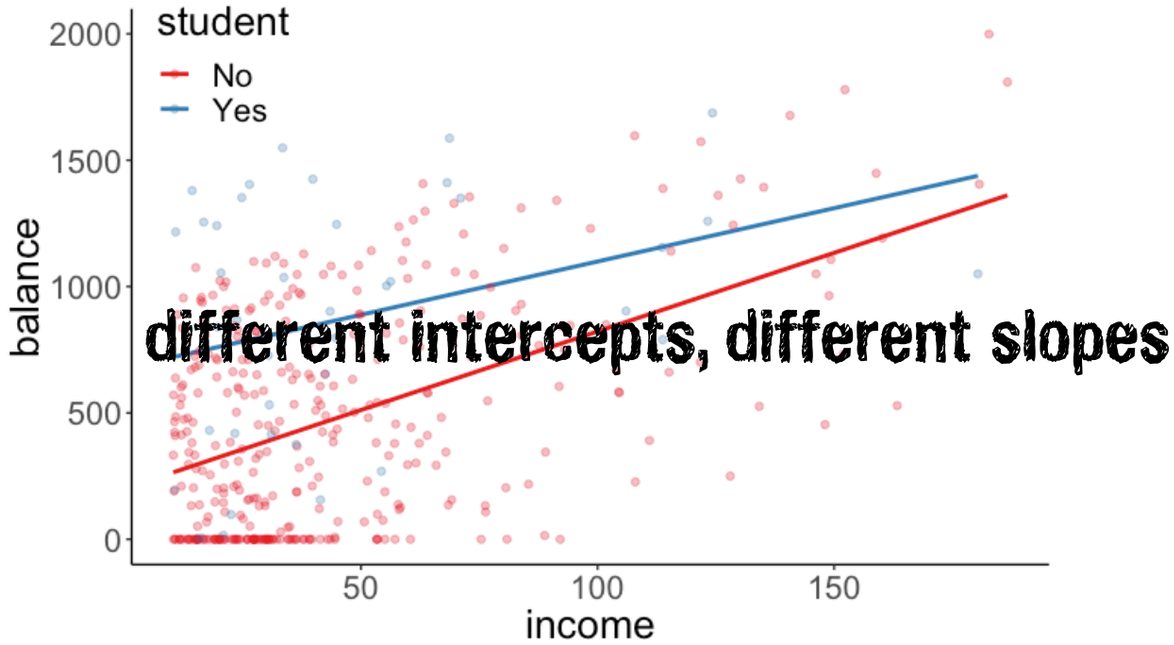
$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

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?

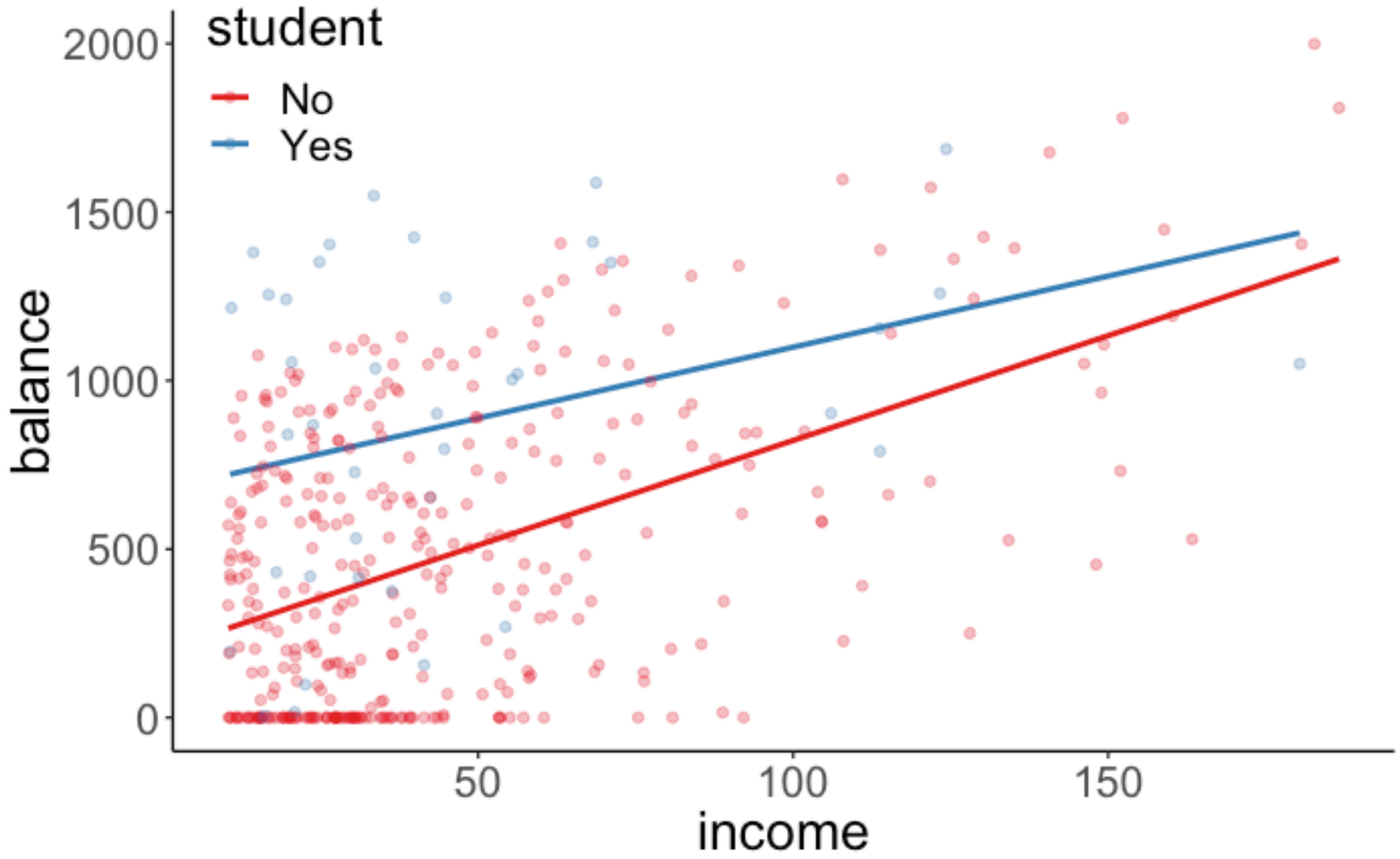
Is the relationship between level of income and balance different for students than it is for non-students?

```
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

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
```

not worth it!

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)
```

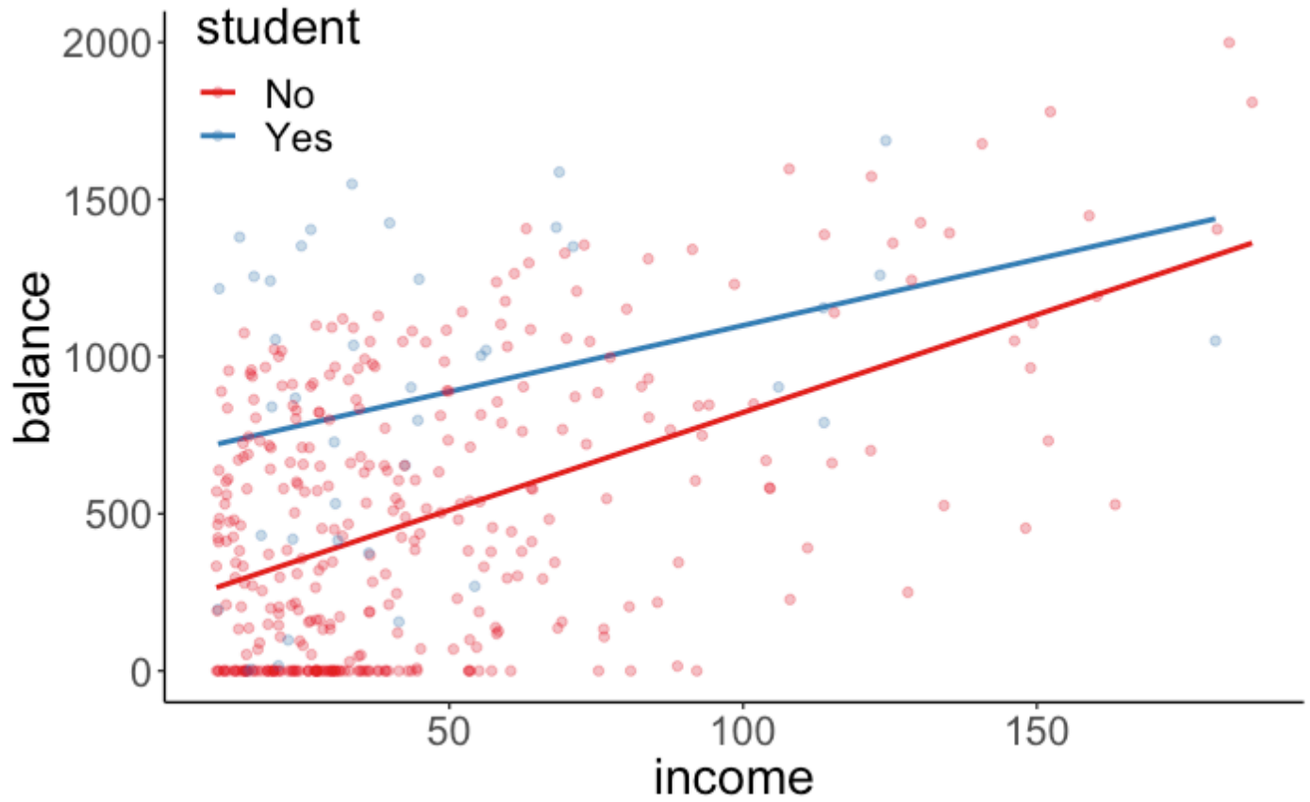
`fit1` and `fit2` are identical!

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

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

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

anova(fit_c, fit_a)
```

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

anova(fit_c, fit_a)
```

lm() output

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

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

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 ‘ ’ 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

```
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
```

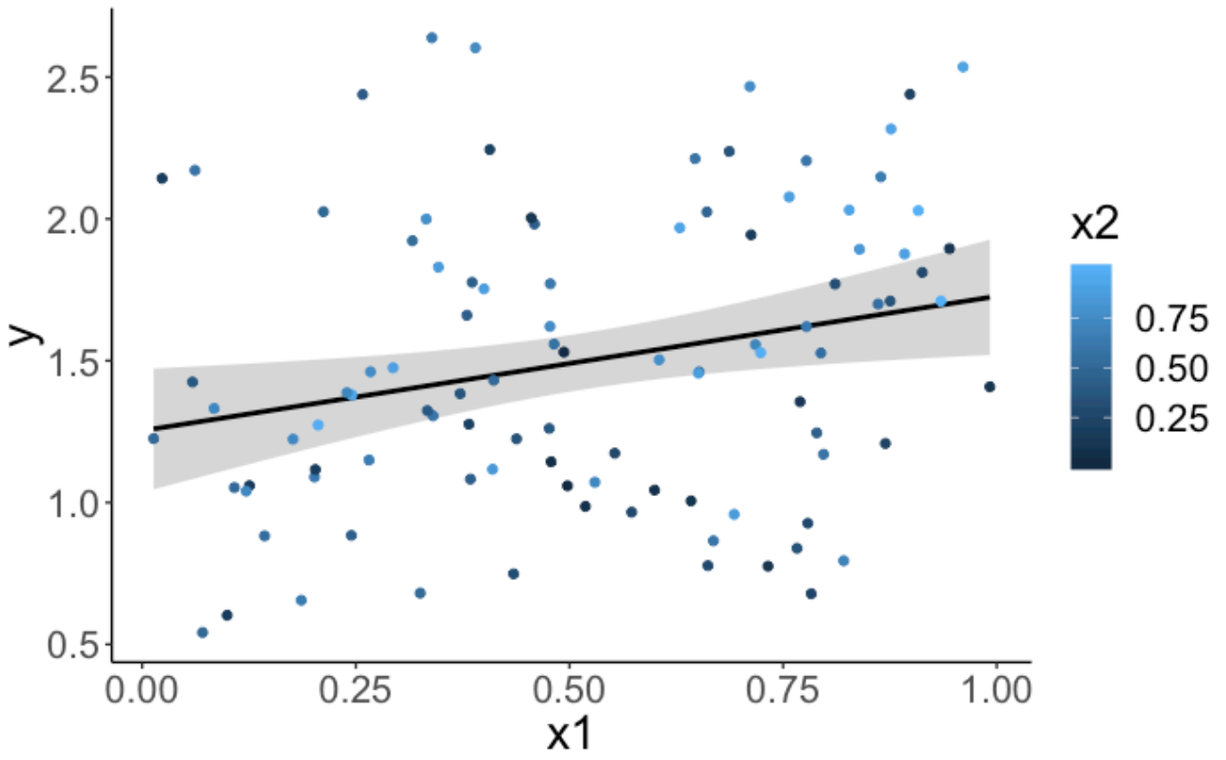
`anova()` gives me *F*s

but `lm()` gives me *t*s

```
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
```

F vs. *t* in `lm()` output

```r
# make example reproducible
set.seed(1)

# parameters
sample_size = 100
b0 = 1
b1 = 0.5
b2 = 0.5
sd = 0.5

# sample
df.data = tibble(
  participant = 1:sample_size,
  x1 = runif(sample_size, min = 0, max = 1),
  x2 = runif(sample_size, min = 0, max = 1),
  # simple additive model
  y = b0 + b1 * x1 + b2 * x2 + rnorm(sample_size, sd = sd)
)
```



$$Y_i = b_0 + b_1 \cdot x1_i + b_2 \cdot x2_i + e_i$$

F vs. *t* in `lm()` output

```
Call:
lm(formula = y ~ x1 + x2, data = df.data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9290 -0.3084 -0.0716  0.2676  1.1659

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9953     0.1395   7.133 1.77e-10 ***
x1            0.4654     0.1817   2.561  0.01198 *
x2            0.5072     0.1789   2.835  0.00558 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4838 on 97 degrees of freedom
Multiple R-squared:  0.1327,  Adjusted R-squared:  0.1149
F-statistic: 7.424 on 2 and 97 DF,  p-value: 0.001
```

Is x1 a significant predictor, controlling for x2?

F vs. *t* in `lm()` output

without x1

with x1

```
1 # fit models
2 model_compact = lm(formula = y ~ 1 + x2,
3                    data = df.data)
4
5 model_augmented = lm(formula = y ~ 1 + x1 + x2,
6                      data = df.data)
7
8 # compare models using the F-test
9 anova(model_compact, model_augmented)
```

```
Analysis of Variance Table

Model 1: y ~ 1 + x2
Model 2: y ~ 1 + x1 + x2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     98 24.235
2     97 22.700  1    1.5347 6.558 0.01198 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F vs. *t* in `lm()` output

```
Call:
lm(formula = y ~ x1 + x2, data = df.data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9290 -0.3084 -0.0716  0.2676  1.1659

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9953     0.1395   7.133 1.77e-10 ***
x1            0.4654     0.1817   2.561  0.01198 *
x2            0.5072     0.1789   2.835  0.00558 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4838 on 97 degrees of freedom
Multiple R-squared:  0.1327,  Adjusted R-squared:  0.1149
F-statistic: 7.424 on 2 and 97 DF,  p-value: 0.001
```

deterministic mapping between t and F

$$t^2 = F$$

$$2.561^2 = 6.558$$

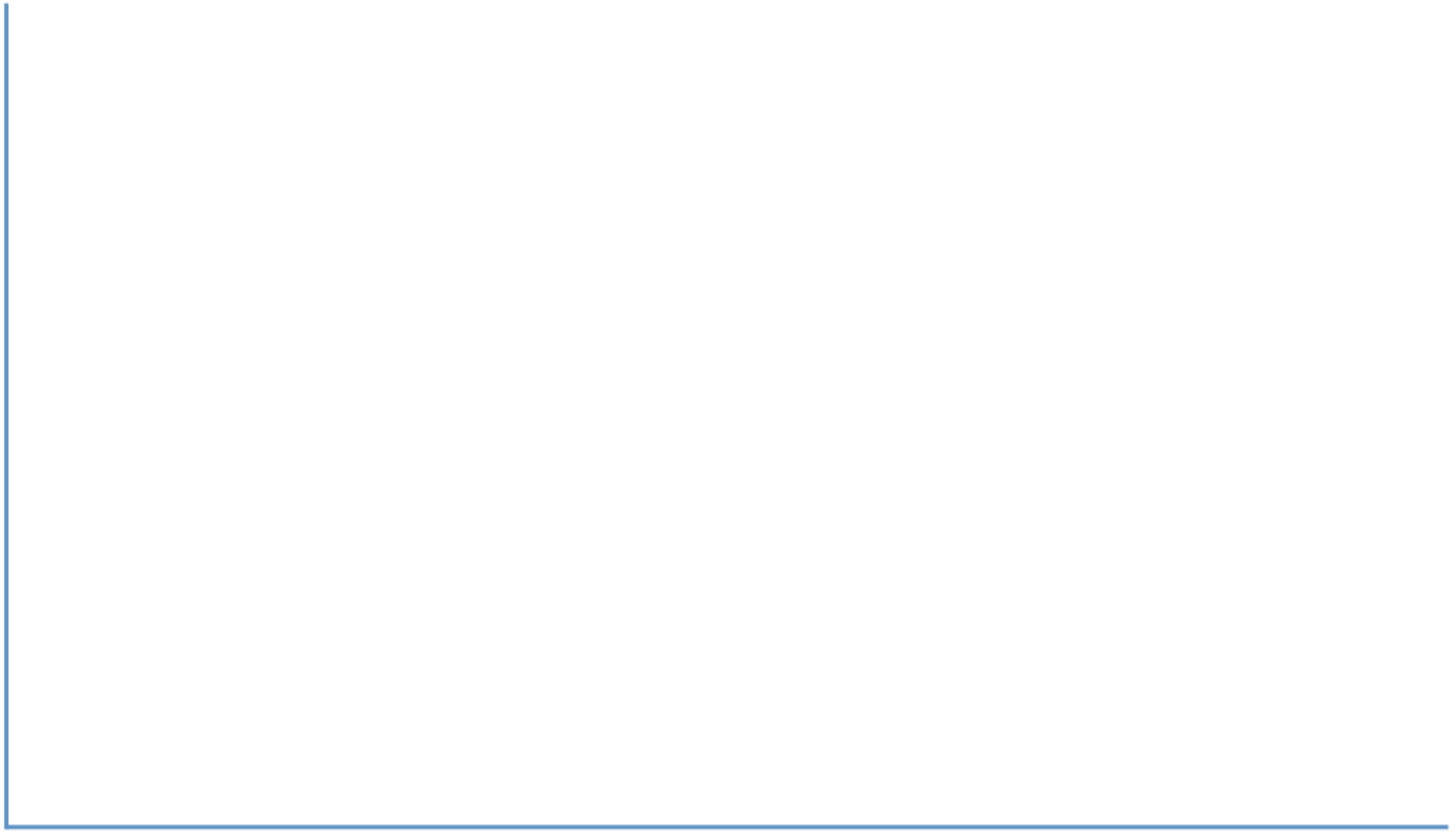
```
Analysis of Variance Table

Model 1: y ~ 1 + x2
Model 2: y ~ 1 + x1 + x2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     98 24.235
2     97 22.700  1    1.5347 6.558 0.01198 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary

- Multiple regression
  - Model assumptions: no multi-collinearity
- Several continuous predictors
  - Hypothesis tests
  - Interpreting parameters
  - Reporting results
- One categorical predictor
- Both continuous and categorical predictors
- Interactions

Feedback

How was the pace of today's class?



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!