Week 7: Multiple linear regression

ANTH 674: Research Design & Analysis in Anthropology
Professor Andrew Du
Andrew.Du2@colostate.edu
Office Hours: Thursdays, 9:00am–12:00pm
In person: GSB 312
Virtual: https://tinyurl.com/F22ANTH674

1

Statistics vignette

- How do we protect bomber planes from being shot down?

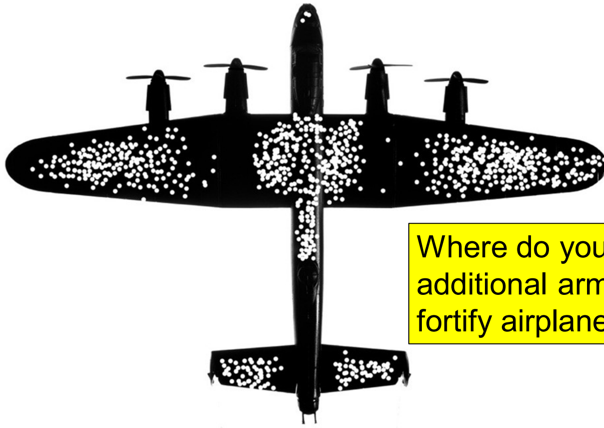
2

The setting is WW2…

3

4

Collected data on damage location



Where do you add additional armor to fortify airplane?

5

Abraham Wald

- Survivorship bias

We assume that

$$\lambda_1 q_i \le q_{i+1} \le \lambda_2 q_i ,$$

where $\lambda_1 < \lambda_2 < 1$ and such that the expression

$$\sum_{j=1}^{n} \frac{a_j}{\lambda_1^{\frac{j(j-1)}{2}}} < 1 - a_o$$

is satisfied.

"This story, like many World War II stories, starts with the Nazis hounding a Jew out of Europe and ends with the Nazis regretting it."

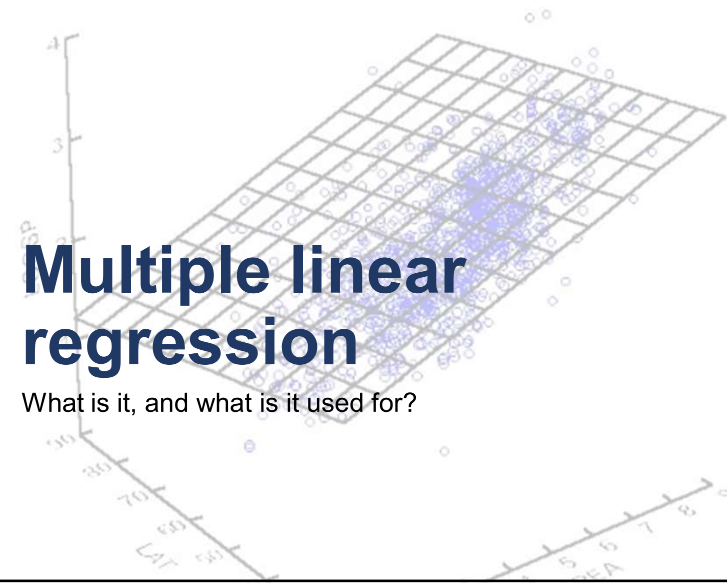
- Jordan Ellenberg (2014)

6

Lecture outline

- Multiple linear regression
  - What is it, and what is it used for?
  - How to interpret coefficients (w/ transformations)
- Interaction terms
- Assumptions & diagnostic plots
- The collinearity issue
- Variance partitioning

7



Multiple linear regression

What is it, and what is it used for?

8

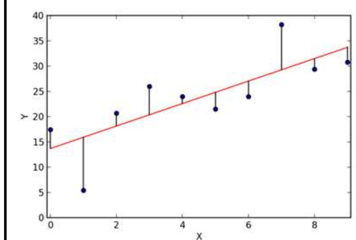
What is multiple regression?

- A general linear model where one continuous DV is a linear function of two or more IVs (which can be continuous or not)
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_n X_n + \varepsilon$
- E.g., `lm(qsec ~ hp + drat, data=mtcars)`
- Its importance will become clear when we go over how to interpret coefficients
- A lot of what you learned for simple linear regression will apply here!

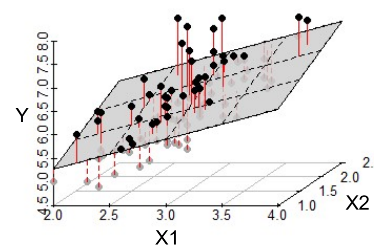
9

How are parameters estimated?

- Like simple linear regression, ordinary least squares (minimizes residuals)



w/ 2 IVs



Plane must pass through mean of DV & all IVs

10

How to interpret coefficients?

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_n X_n + \varepsilon$
- The intercept (β_0) is the value of DV (Y) when all IVs are set to zero
- β_1 ("coefficient" or "*partial* coefficient") is the change in DV (Y) as X_1 increases by 1, *holding all other IVs constant*
- Let's illustrate this with an example

11

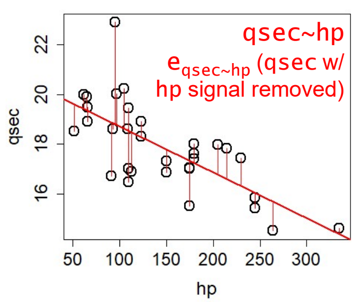
`qsec~hp+drat, data=mtcars`

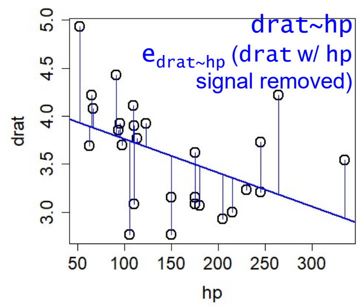
- `qsec = 24.45 – 0.02hp – 0.95drat`
- If hp = 2, `qsec = 24.45 – 0.04 – 0.95drat`
- `qsec = 24.41 – 0.95drat`
- Fixing other IVs at some value, basically shunts IVs to the intercept, leaving the slope to be interpreted as in simple regression

12

Another way of looking at it

- `qsec = 24.45 – 0.02hp – 0.95drat`
- $e_{qsec\sim hp}$ = 0 – 0.95$e_{drat\sim hp}$





13

Another way of looking at it

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- β_1 is the slope when the effects of X_2 have been removed (i.e., "partialed out") from Y and X_1
- Even the standard errors of the slope are the same!
- `drat` SE = 0.46 in both `qsec ~ hp + drat` <u>AND</u> $e_{qsec\sim hp}$ ~ $e_{drat\sim hp}$

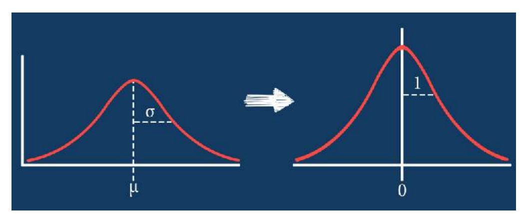
14

Questions?

15

Centering & scaling

- <u>Centering</u>: intercept interpreted as DV when all IVs are fixed at their mean values (= zero)
- <u>Scaling</u>: If IVs are in different units or differ by orders of magnitude, puts IVs on same scale, so they're comparable within the same model



16

Log-transformations

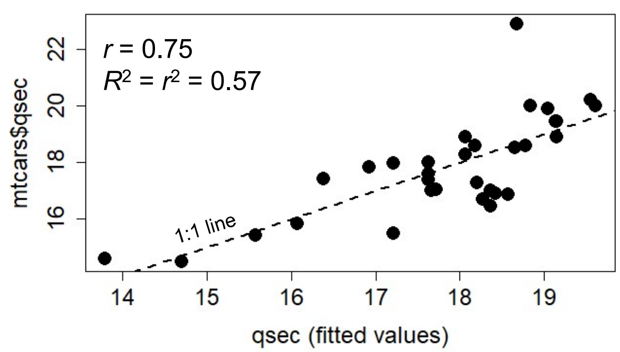
- Interpreted in *exactly* the same way as in simple linear regression, just with other IVs held constant

Coefficient of determination (R^2)

- R^2: proportion of variation in DV explained by IVs (goes from 0 to 1)
- Calculated in exactly the same way as in simple linear regression
- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$
- Also, R^2 = squared Pearson's correlation btw DV and fitted linear model values

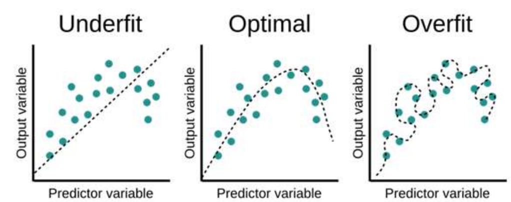
Coefficient of determination (R^2)

- `lm(qsec ~ hp + drat)`



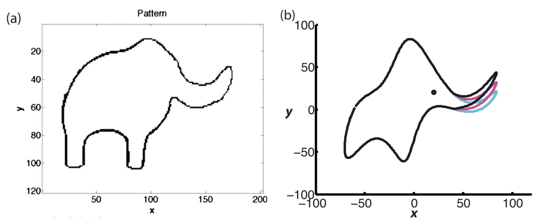
Overfitting

- Overfitting: when you fit the noise instead of the signal
- Happens when you have too few data points per parameter (general rule: need *at least* 20 data points per parameter)
- Parameters are like tuning dials, each one making a model more flexible



The folly of complex models

- Complex model: model w/ lots of parameters
- With enough parameters, model is flexible enough to fit **any** dataset
- “With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.”



John von Neumann

Mayer et al. 2010

21

22

Complexity and R^2

- As you add parameters, R^2 can only increase (can fit more DV variation/noise)
- So $R^2 = 0.9$ is not impressive if you have 100 parameters (unless you have 2,000 data points)!
- Adjusted R^2: adjusted for extra parameters (can’t be interpreted as proportion variation explained anymore)
- $R_a^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$

n = number of data points
p = number of IVs

23

Complexity and coefficients

- As you increase the number of model parameters the standard errors of your coefficients increase, all else being equal
- Less data points per parameter, so less information/power
- P-values go up
- Predictions become less precise
- No free lunch in statistics! Usually there is some trade-off (e.g., bias-variance trade-off)

24

What is multiple regression used for?

- Same as simple linear regression, but now extended to multiple IVs!

1. <u>Exploration</u>: just want to know what the coefficients are
2. <u>Hypothesis testing</u>: e.g., are coefficients significantly different from zero?
3. <u>Prediction</u>: what is DV when new IV values are input (need to cross-validate)?

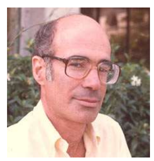
Hypothesis testing

- Interpretation same as simple linear regression
- IV is significant predictor of DV after accounting for other IVs

```
Call:
lm(formula = qsec ~ hp + drat, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8999 -0.6448 -0.0744  0.6948  4.2324 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.454022   1.949759  12.542 3.07e-13 ***
hp          -0.021777   0.003568  -6.103 1.20e-06 ***
drat        -0.948359   0.457564  -2.073   0.0472 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.217 on 29 degrees of freedom
Multiple R-squared:  0.5659,	Adjusted R-squared:  0.5359 
F-statistic:  18.9 on 2 and 29 DF,  p-value: 5.562e-06
```

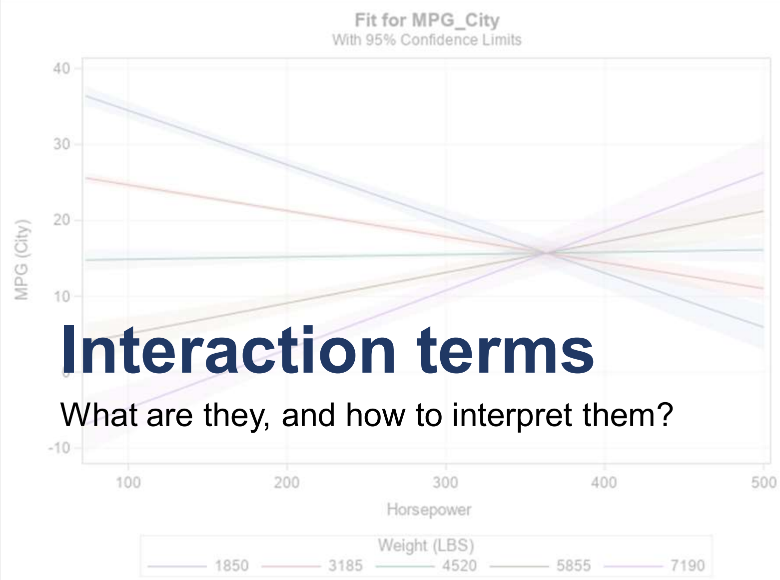
Model fits data significantly better than intercept-only model

Freedman's paradox



- Be careful when looking at many, many variables (5% will be significant w/ large slopes even if H_0 is true)
- Can go exploring (i.e., variable selection), but **<u>don't report P-values</u>**
- *A posteriori* hypotheses need to be confirmed w/ independent dataset!

Questions?

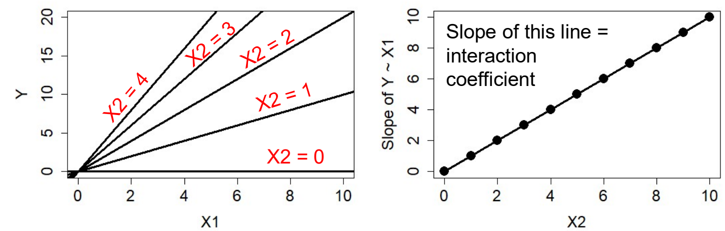


Interaction terms

What are they, and how to interpret them?

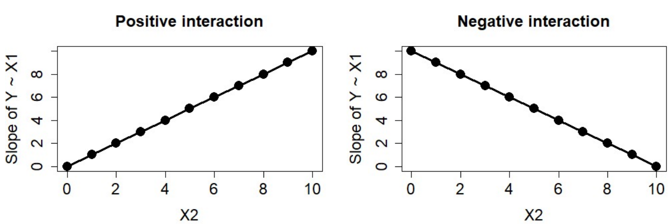
What are interaction terms?

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- If $Y \sim X_1$ slope changes as a function of X_2 (or $Y \sim X_2$ slope changes as a function of X_1), need an interaction term
- E.g., slope of marathon time ~ training affected by equipment used



How to interpret

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$
- If interaction term is positive, $Y \sim X_1$ slope increases w/ X_2 (& slope of $Y \sim X_2$ increases w/ X_1)
- If interaction is negative, $Y \sim X_1$ slope decreases w/ X_2 (& slope of $Y \sim X_2$ decreases w/ X_1)

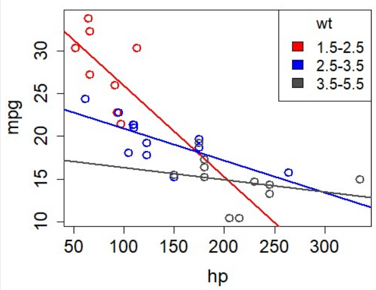


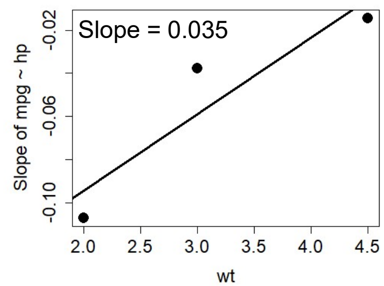
E.g., `mpg ~ hp * wt`

- `mpg=49.81-0.12hp-8.22wt+0.03(hp*wt)`
- All coefficients P < 0.001
- If `wt=1`,
  `mpg=49.81-0.12hp-8.22+0.03(hp)`
  `=41.59-0.09hp`
- If `wt=2`,
  `mpg=49.81-0.12hp-8.22*2+0.03(hp*2)`
  `=49.81-0.12hp-16.44+0.06hp`
  `=33.37-0.06hp`
- Interactions are symmetrical: all the above applies if you switch `hp` and `wt`

E.g., `mpg ~ hp * wt`

- `mpg=49.81−0.12hp−8.22wt+0.03(hp*wt)`
- All coefficients P < 0.001

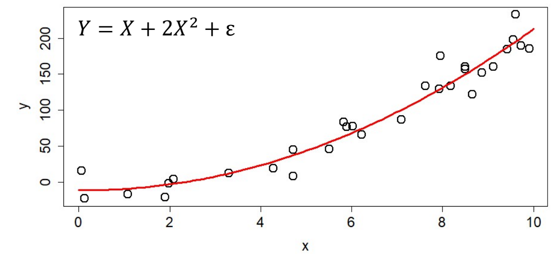




33

Variable can interact with itself

- $Y = \beta_0 + \beta_1 X + \beta_2 X^2$
- Slope of $Y \sim X$ increases as X increases (i.e., an accelerating curve)
- Known as **<u>quadratic regression</u>**



34

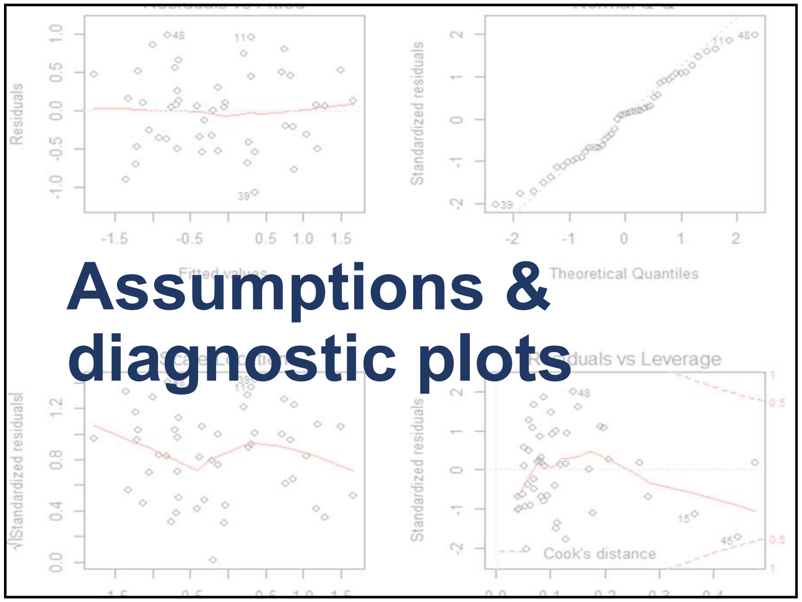
How do I know when to use interactions?

- Best justification is theory and expert knowledge
- Residual plots show some pattern (e.g., nonlinear relationship)

35

Questions?

36

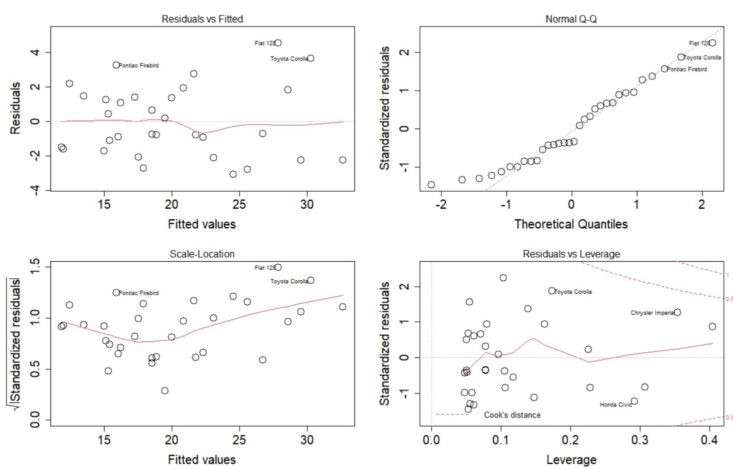


Assumptions & diagnostic plots

(Almost) exactly the same as simple linear regression!

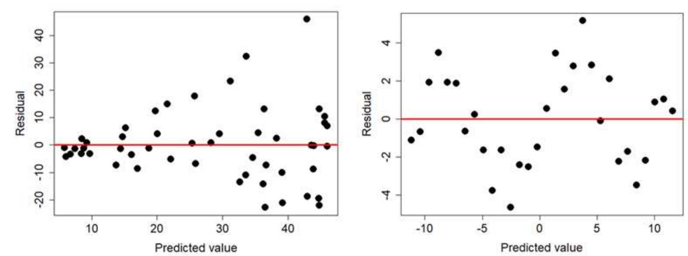
1. Relationship between DV and IVs is linear
2. IV measured without error
3. Error terms have mean = 0 and are normally distributed
4. Error terms drawn from population with the same variance (homoskedasticity)
5. Error terms are independent
6. No multicollinearity (explained shortly)

Use same diagnostic plots



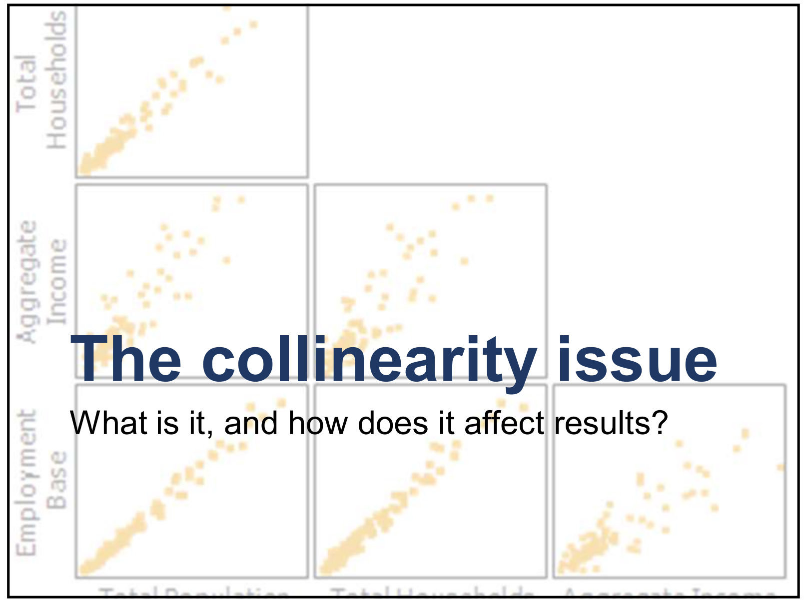
Violations affect results in the same way as simple regression

- E.g., coefficients estimates are unbiased and are not affected by heteroscedasticity or non-independent errors



Questions?

41



The collinearity issue

What is it, and how does it affect results?

42

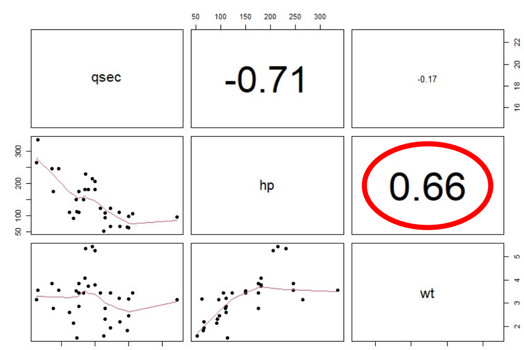
What is collinearity?

- Collinearity (AKA multicollinearity) is when two or more of your IVs are highly correlated
- This is **BAD**! Can screw up your coefficient estimates and P-values (i.e., the “bouncing betas” problem)

43

Let’s look at an example

- `qsec ~ hp + wt, data=mtcars`



44

Let's look at an example

- `qsec = 18.83 - 0.03hp + 0.94wt`
  P = 6.36e-08 P = 0.00137
- What do coefficients look like in simple linear regression?
- `qsec = 18.83 - 0.01hp`
  P = 5.77e-06
- `qsec = 18.83 - 0.32wt`
  P = 0.339

45

Effects of collinearity

- The more collinear your IVs, the more coefficients and P-values will change if an IV is added or dropped!
- Coefficient is change in DV w/ +1 in IV when all other IVs held constant, but not possible to hold other IVs constant w/ collinearity
- If IVs are completely independent, multiple regression coefficients will be identical to their simple regression counterparts

46

How to detect collinearity?

- Correlation coefficients between IVs
- Variation inflation factors (better when collinearity exists among >2 IVs)
- $VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is calculated from regressing X_j on all other IVs
- VIF ranges from 1 to infinity, and VIF > 10 is bad

47

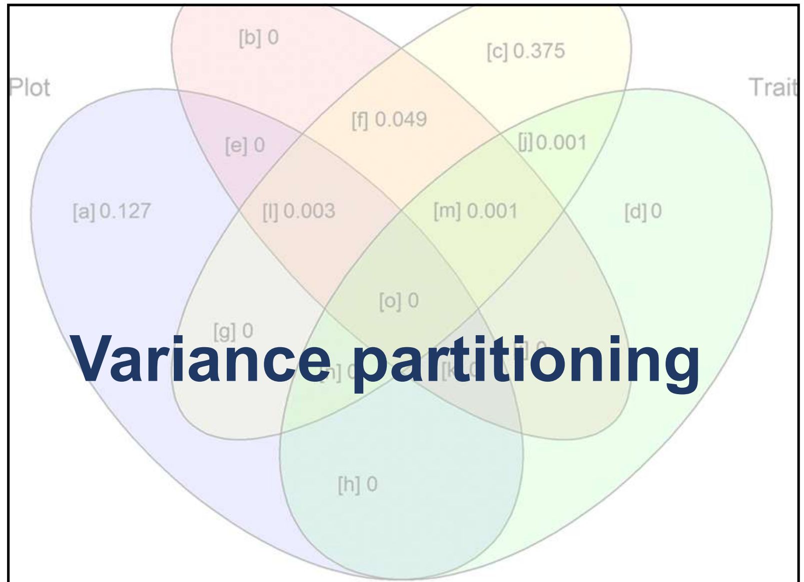
Solutions

- Drop all but one of the collinear IVs, as they offer redundant information
- Distill collinear IVs into a single IV, using principal components analysis
- Variance partitioning (if interested in variation explained, not estimated coefficients)
- Don't do anything
  - Your IVs of interest are unaffected by collinearity
  - Collinearity does not affect R^2 or predictions (if IVs are related in the same way in new dataset)

48

Questions?

49



Variance partitioning

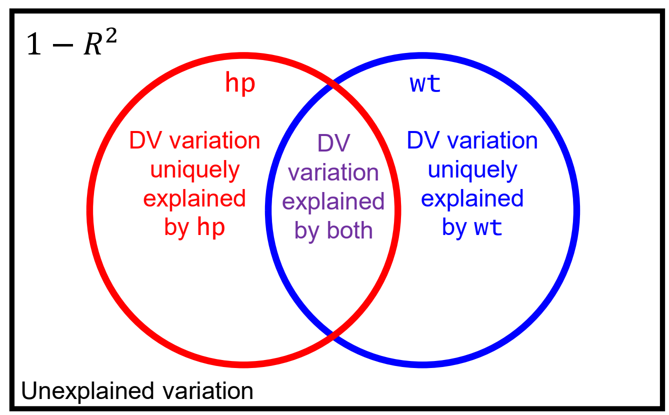
50

Variance partitioning

- Can see how much variation in DV is explained by each IV uniquely or jointly
- **<u>SUPER</u>** informative and criminally underused!
- Not affected by collinearity!

51

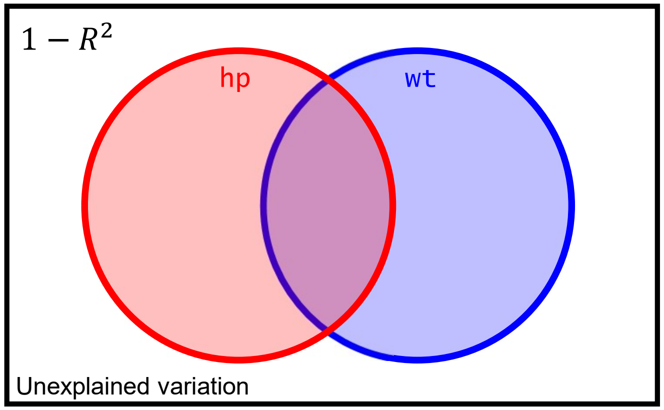
`mpg ~ hp + wt`



52

mpg ~ hp + wt

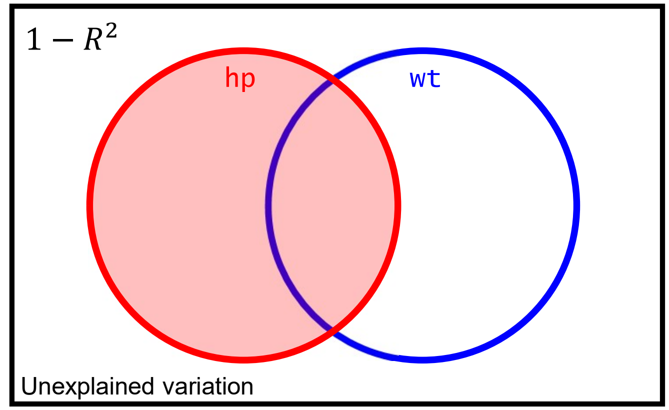
R^2 (both IVs)



53

mpg ~ hp + wt

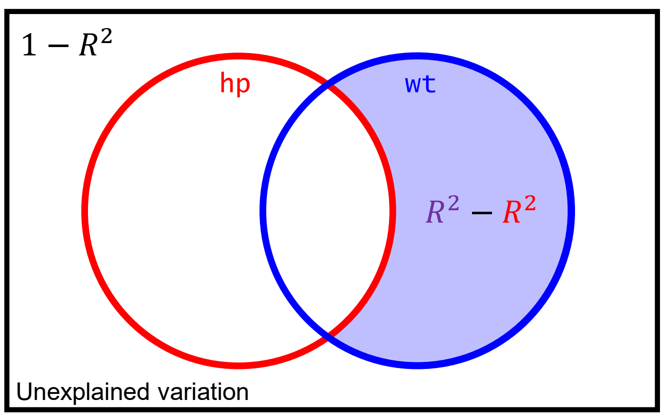
R^2 (both IVs)
R^2 (mpg~hp)



54

mpg ~ hp + wt

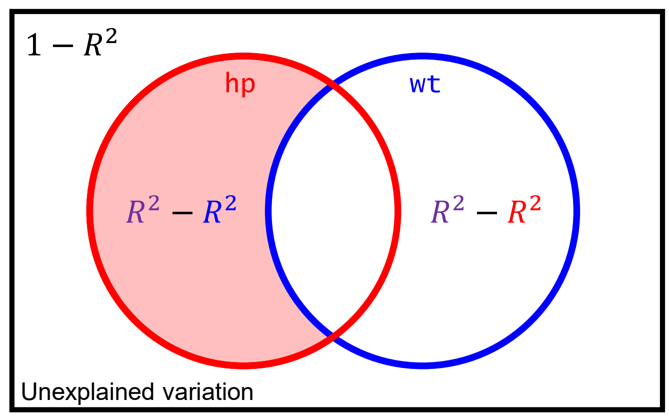
R^2 (both IVs)
R^2 (mpg~hp)



55

mpg ~ hp + wt

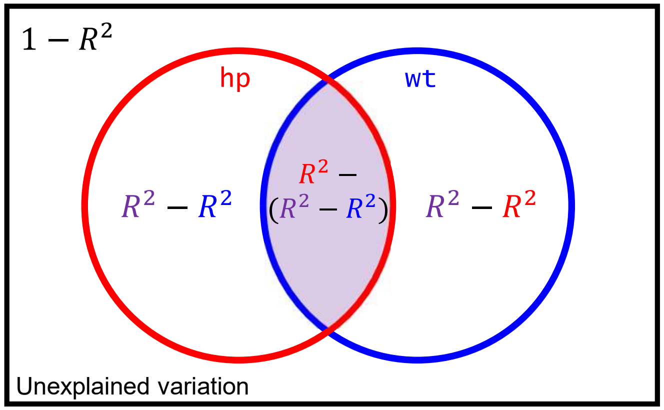
R^2 (both IVs)
R^2 (mpg~hp)
R^2 (mpg~wt)



56

`mpg ~ hp + wt`

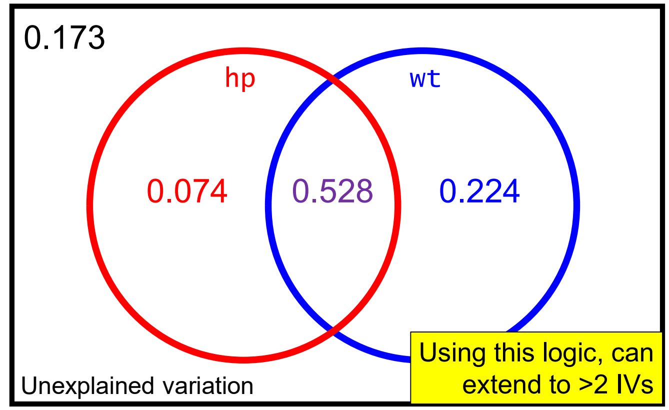
R^2 (both IVs)
R^2 (mpg~hp)
R^2 (mpg~wt)



57

`mpg ~ hp + wt`

R^2 (both IVs)
R^2 (mpg~hp)
R^2 (mpg~wt)



Using this logic, can extend to >2 IVs

58

Questions?

59

Summary

- Multiple regression used to estimate coefficient for an IV, while holding all other IVs constant
- Goals, assumptions, & interpretation of transformed variables are same as simple linear regression
- Be careful about including too many coefficients (e.g., Freedman's paradox, overfitting, imprecise coefficient estimates)
- Collinearity negatively affects coefficient estimates & P-values
- Variance partitioning is a useful tool in inference!

60