Introduction to multiple linear regression

Nicholas Reich, UMass-Amherst Biostatistics

Outline

Multiple regression

- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

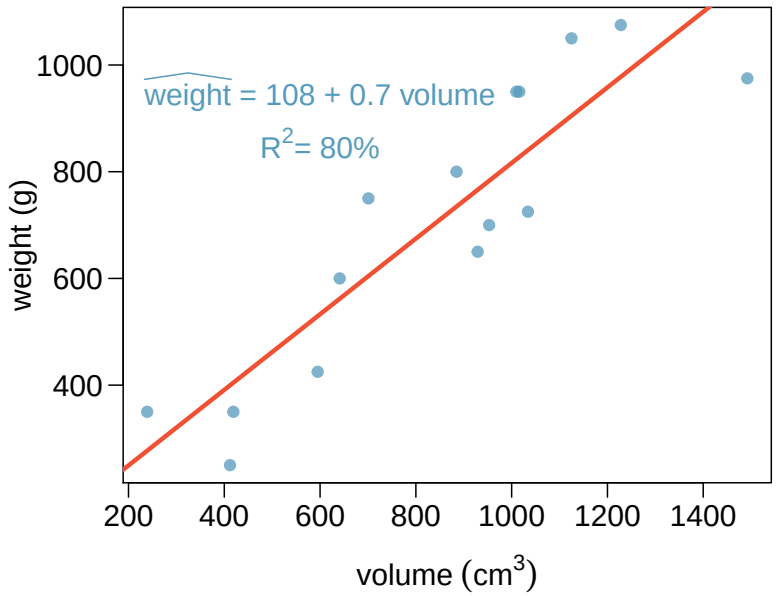
Weights of books

	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) “Data Analysis and Graphics Using R”

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

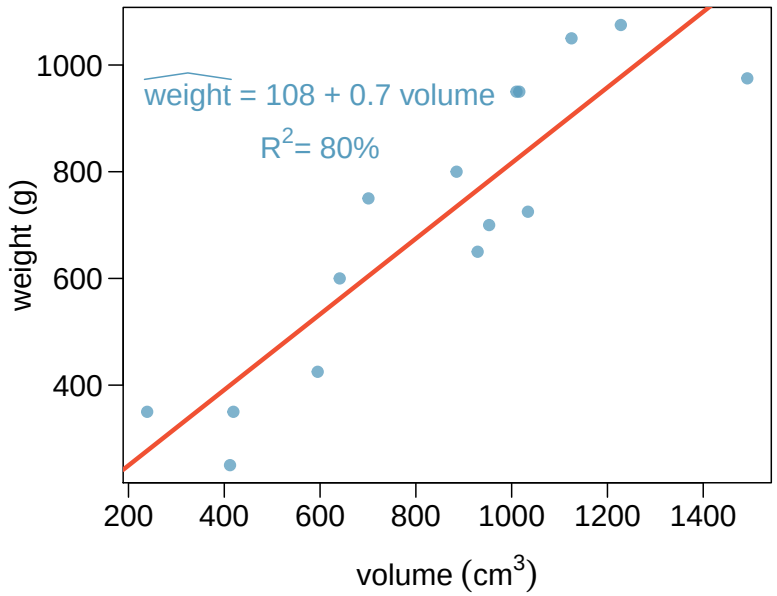
(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

(b) *Books that are 10 cm^3 over average are expected to weigh 7 g over average.*

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

Modeling weights of books using volume

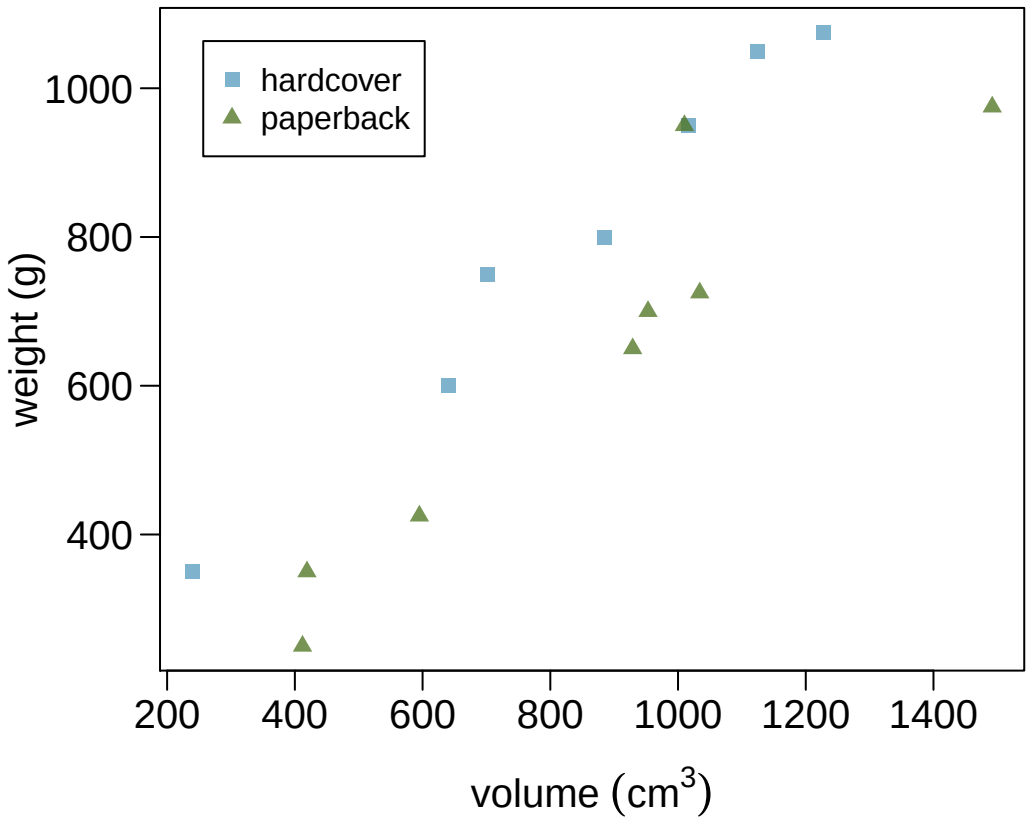
somewhat abbreviated output...

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06


Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared: 0.8026, Adjusted R-squared: 0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

Weights of hardcover and paperback books

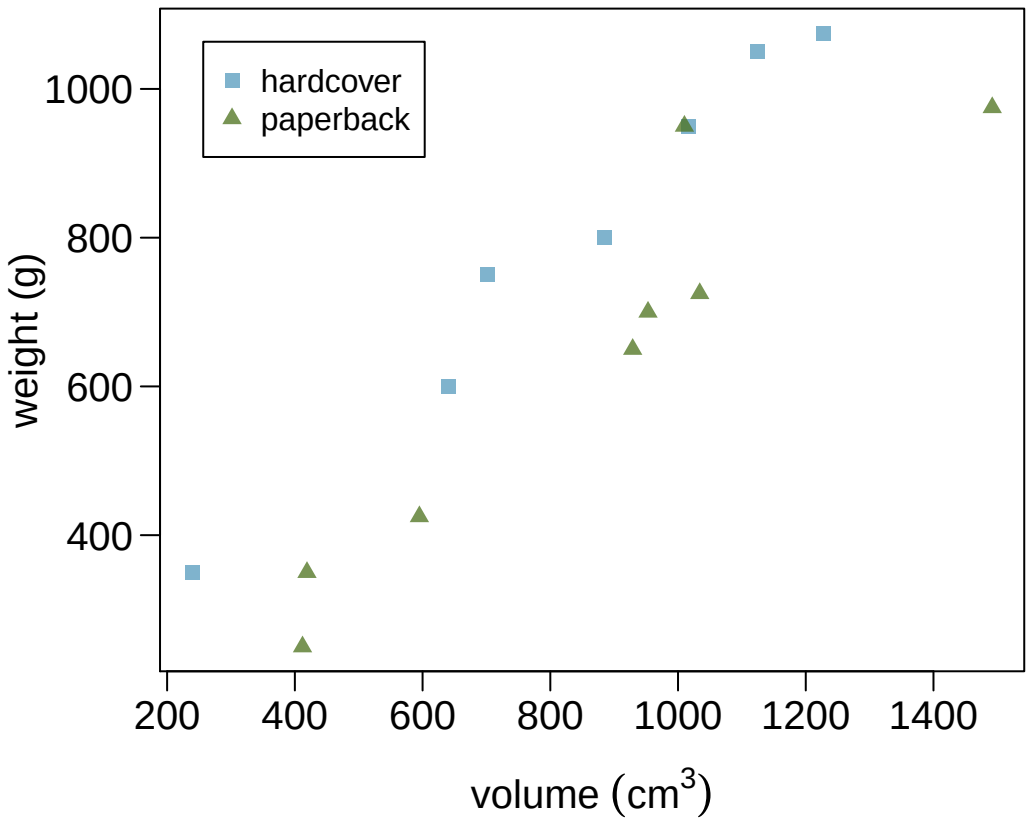
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Weights of hardcover and paperback books

Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?

Paperbacks generally weigh less than hardcover books after controlling for the book's volume.

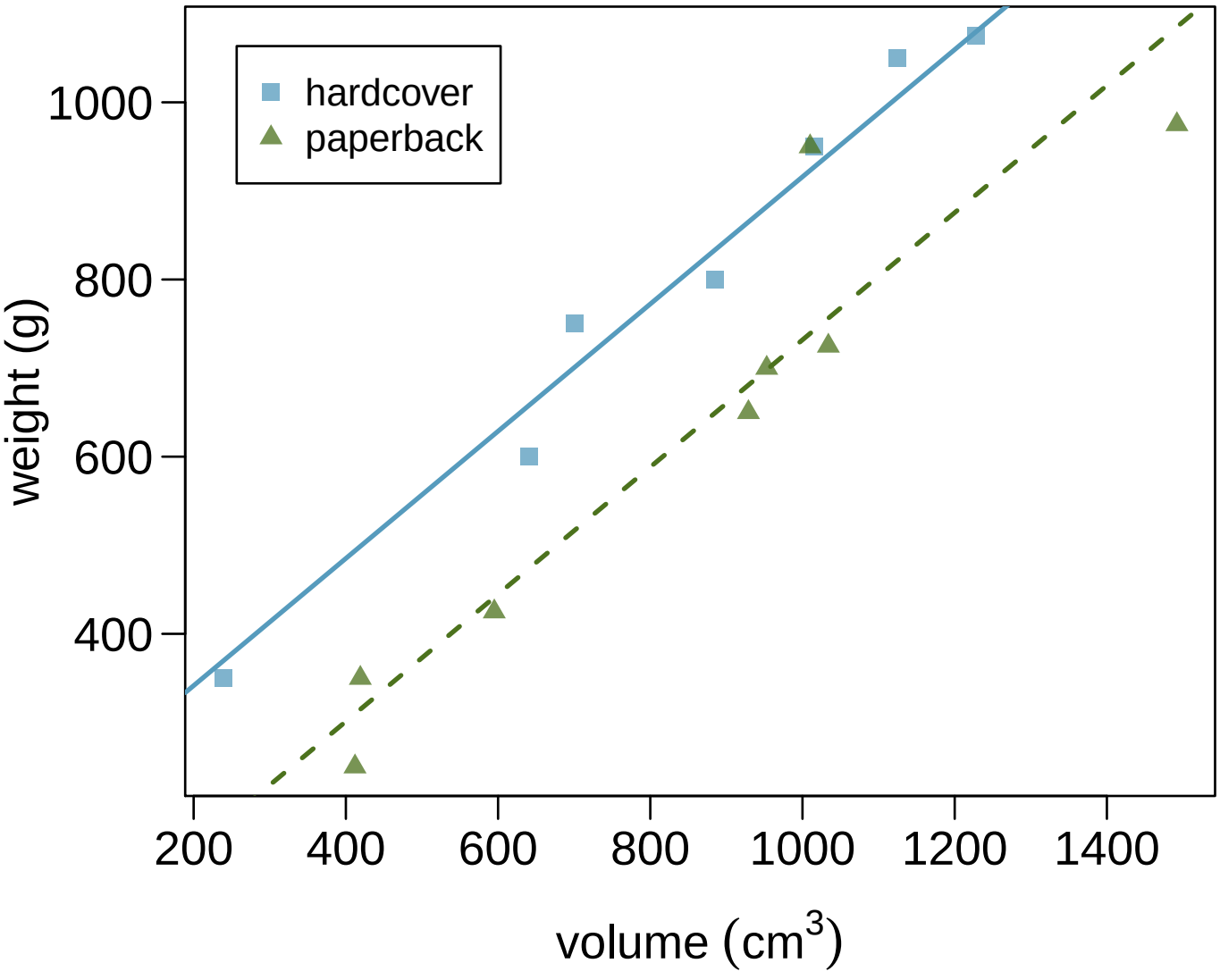


Modeling weights of books using volume and cover type

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***


Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275, Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Visualising the linear model



Determining the reference level

Based on the regression output below, which level of `cover` is the reference level? Note that `pb`: paperback.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) paperback

(b) hardcover

Determining the reference level

Based on the regression output below, which level of `cover` is the reference level? Note that `pb`: paperback.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) paperback

(b) *hardcover*

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) response: weight, explanatory: volume, cover type

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) *response: weight, explanatory: volume, cover type*

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

1. For *hardcover* books: plug in *0* for `cover`

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 0$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. For *paperback* books: plug in *1* for `cover`

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 1$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

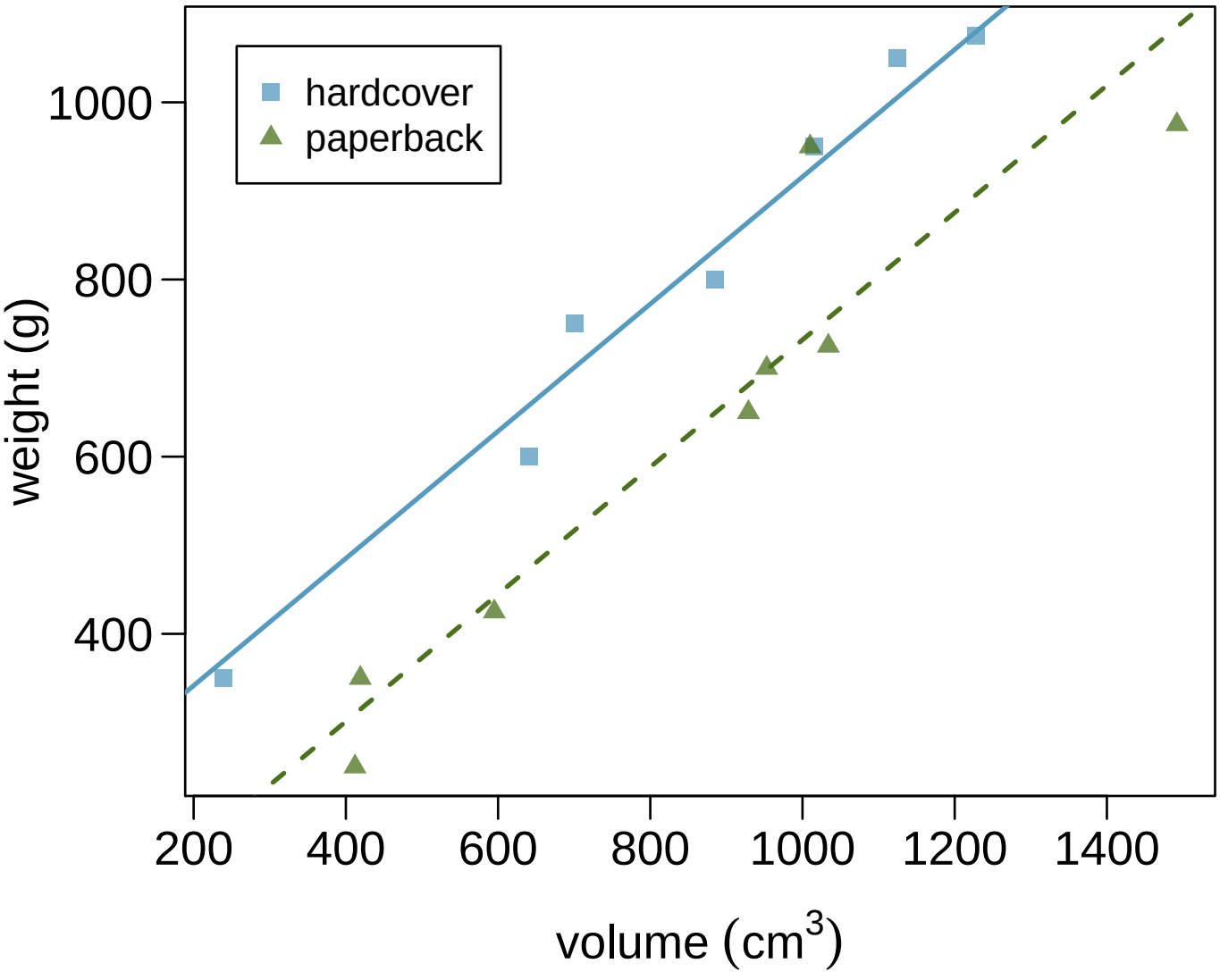
1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. For *paperback* books: plug in *1* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 1 \\ &= 13.91 + 0.72\ volume \end{aligned}$$

Visualising the linear model



Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.
  - Obviously, the intercept does not make sense in context. It only serves to adjust the height of the line.

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) 197.96 + 0.72 * 600 - 184.05 * 1

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) *197.96 + 0.72 * 600 - 184.05 * 1* $= 445.91$ *grams*

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Another example: Modeling kid's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

Interpreting the slope

What is the correct interpretation of the slope for mom's IQ?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

, kids with mothers whose IQs are one point higher tend to score on average 0.56 points higher.

Interpreting the slope

What is the correct interpretation of the slope for mom's IQ?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else held constant, kids with mothers whose IQs are one point higher tend to score on average 0.56 points higher.

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Kids whose moms haven't gone to HS, did not work during the first three years of the kid's life, have an IQ of 0 and are 0 yrs old are expected on average to score 19.59. Obviously, the intercept does not make any sense in context.

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else being equal, kids whose moms worked during the first three year's of the kid's life

(a) are estimated to score 2.54 points lower

(b) are estimated to score 2.54 points higher

than those whose moms did not work.

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

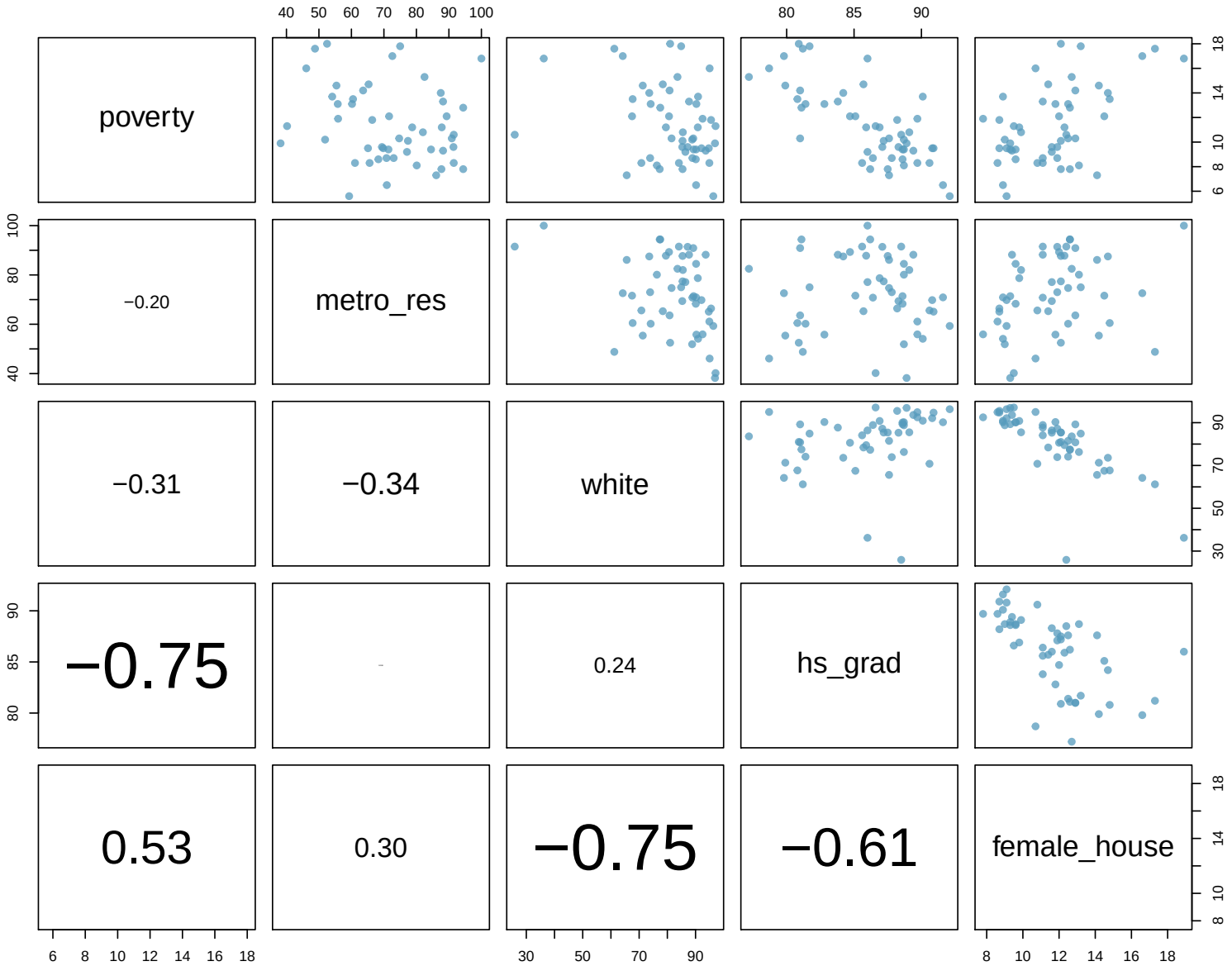
All else being equal, kids whose moms worked during the first three year's of the kid's life

(a) are estimated to score 2.54 points lower

(b) *are estimated to score 2.54 points higher*

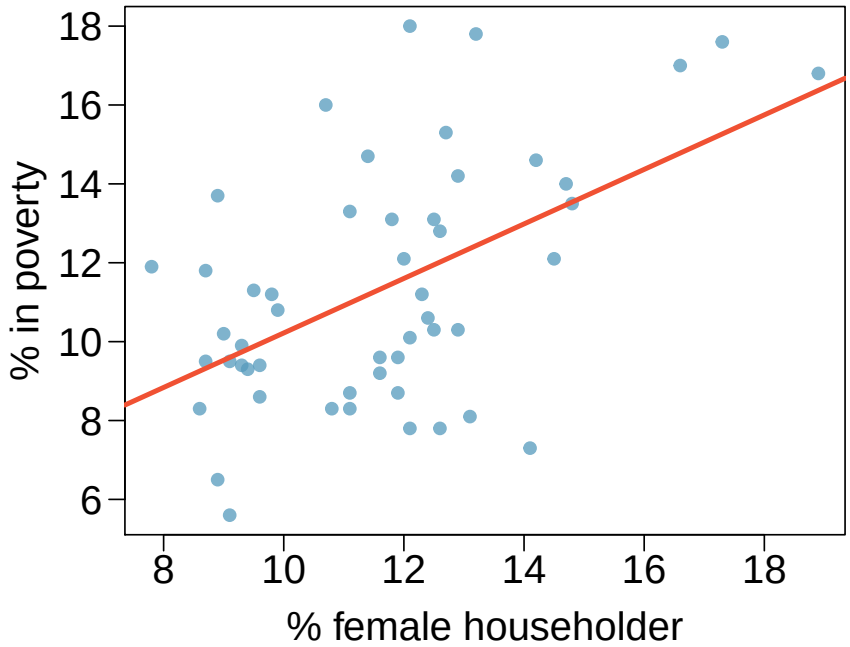
than those whose moms did not work.

Revisit: Modeling poverty



Predicting poverty using % female householder

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00



$$R = 0.53$$

$$R^2 = 0.53^2 = 0.28$$

Another look at R^2

R^2 can be calculated in three ways:

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Using *ANOVA* we can calculate the explained variability and total variability in y.

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares of y: $SS_{Total} = \sum (y - \bar{y})^2 = 480.25$ → *total variability*

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares of y: $SS_{Total} = \sum (y - \bar{y})^2 = 480.25$ → *total variability*

Sum of squares of residuals: $SS_{Error} = \sum e_i^2 = 347.68$ → *unexplained variability*

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares of y: $SS_{Total} = \sum (y - \bar{y})^2 = 480.25$ → *total variability*

Sum of squares of residuals: $SS_{Error} = \sum e_i^2 = 347.68$ → *unexplained variability*

Sum of squares of x: $SS_{Model} = SS_{Total} - SS_{Error}$ → *explained variability*

$= 480.25 - 347.68 = 132.57$

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\text{Sum of squares of } y\text{: } SS_{Total} = \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability}$$

$$\text{Sum of squares of residuals: } SS_{Error} = \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability}$$

$$\text{Sum of squares of } x\text{: } SS_{Model} = SS_{Total} - SS_{Error} \rightarrow \textit{explained variability}$$

$$= 480.25 - 347.68 = 132.57$$

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28 \checkmark$$

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

- *For single-predictor linear regression, having three ways to calculate the same value may seem like overkill.*
- *However, in multiple linear regression, we can't calculate R^2 as the square of the correlation between x and y because we have multiple xs.*
- *And next we'll learn another measure of explained variability, adjusted R^2, that requires the use of the third approach, ratio of explained and unexplained variability.*

Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

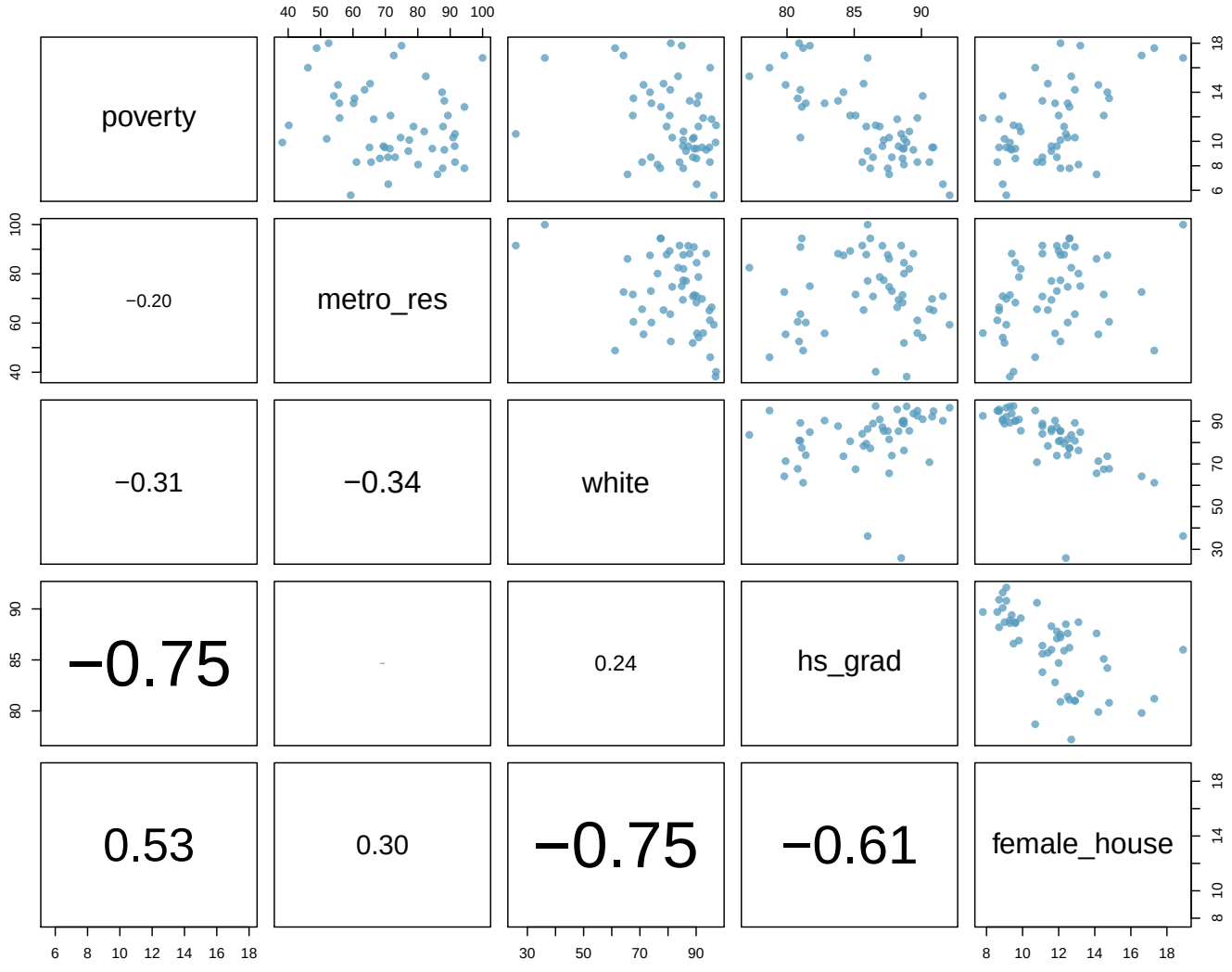
Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 8.21}{480.25} = 0.29$$

Does adding the variable `white` to the model add valuable information that wasn't provided by `female_house`?



Collinearity between explanatory variables

poverty vs. % female head of household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00

poverty vs. % female head of household and % female hh

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

Collinearity between explanatory variables

poverty vs. % female head of household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	*0.69*	0.16	4.32	0.00

poverty vs. % female head of household and % female hh

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	*0.89*	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When any variable is added to the model R^2 increases.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right)
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right)
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74 \\
&= 0.26
\end{aligned}
$$

Outline

Beauty in the classroom

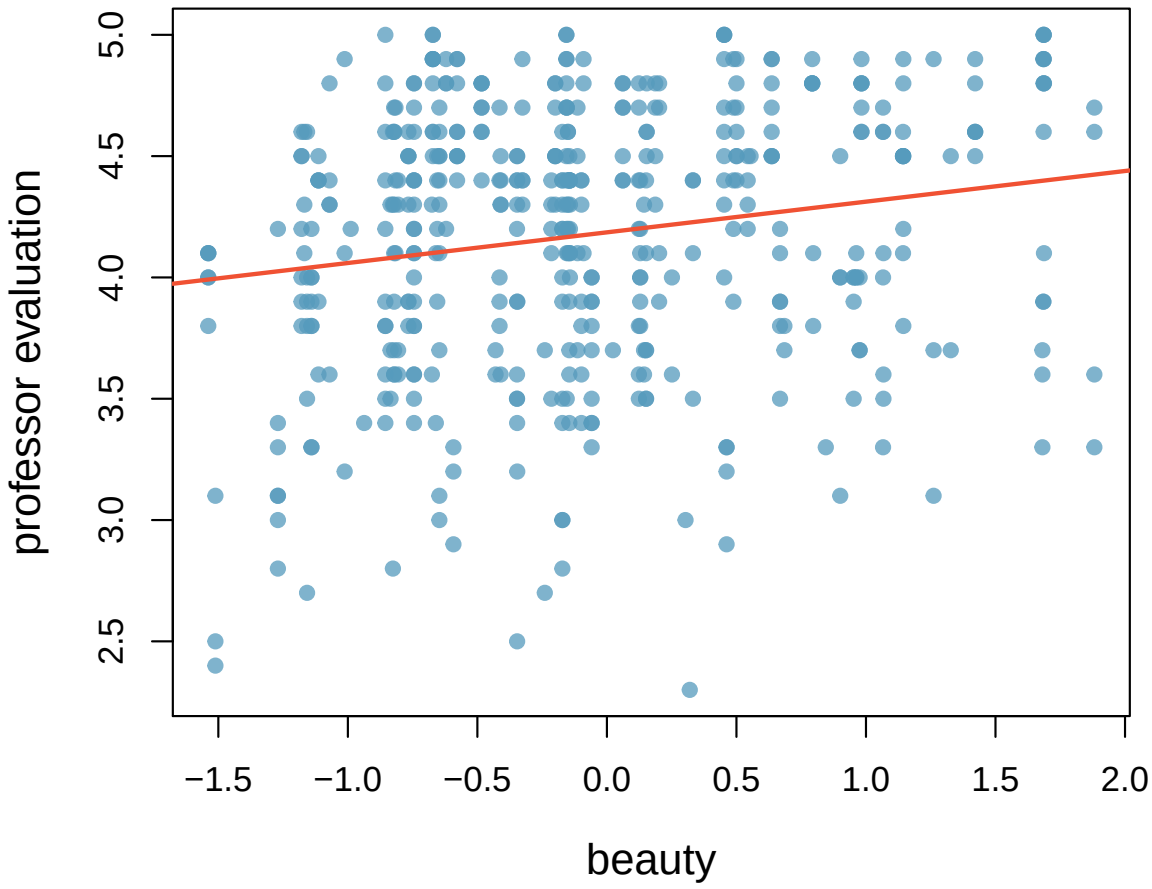
- Data: Student evaluations of instructors' beauty and teaching quality for 463 courses at the University of Texas.
- Evaluations conducted at the end of semester, and the beauty judgements were made later, by six students who had not attended the classes and were not aware of the course evaluations (2 upper level females, 2 upper level males, one lower level female, one lower level male).

Hamermesh & Parker. (2004)"Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity"

Economics Education Review.

Professor rating vs. beauty

Professor evaluation score (higher score means better) vs. beauty score (a score of 0 means average, negative score means below average, and a positive score above average):



Which of the below is correct based on the model output?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00
$R^2 = 0.0336$				

(a) Model predicts 3.36% of professor ratings correctly.

(b) Beauty is not a significant predictor of professor evaluation.

(c) An increase of 1 point in a professor's beauty score is associated with a 0.13 point increase in their evaluation score.

(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(e) The correlation coefficient could be $\sqrt{0.0336} = 0.18$ or -0.18, we can't tell which is correct.

Which of the below is correct based on the model output?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00
$R^2 = 0.0336$				

(a) Model predicts 3.36% of professor ratings correctly.

(b) Beauty is not a significant predictor of professor evaluation.

(c) *An increase of 1 point in a professor's beauty score is associated with a 0.13 point increase in their evaluation score.*

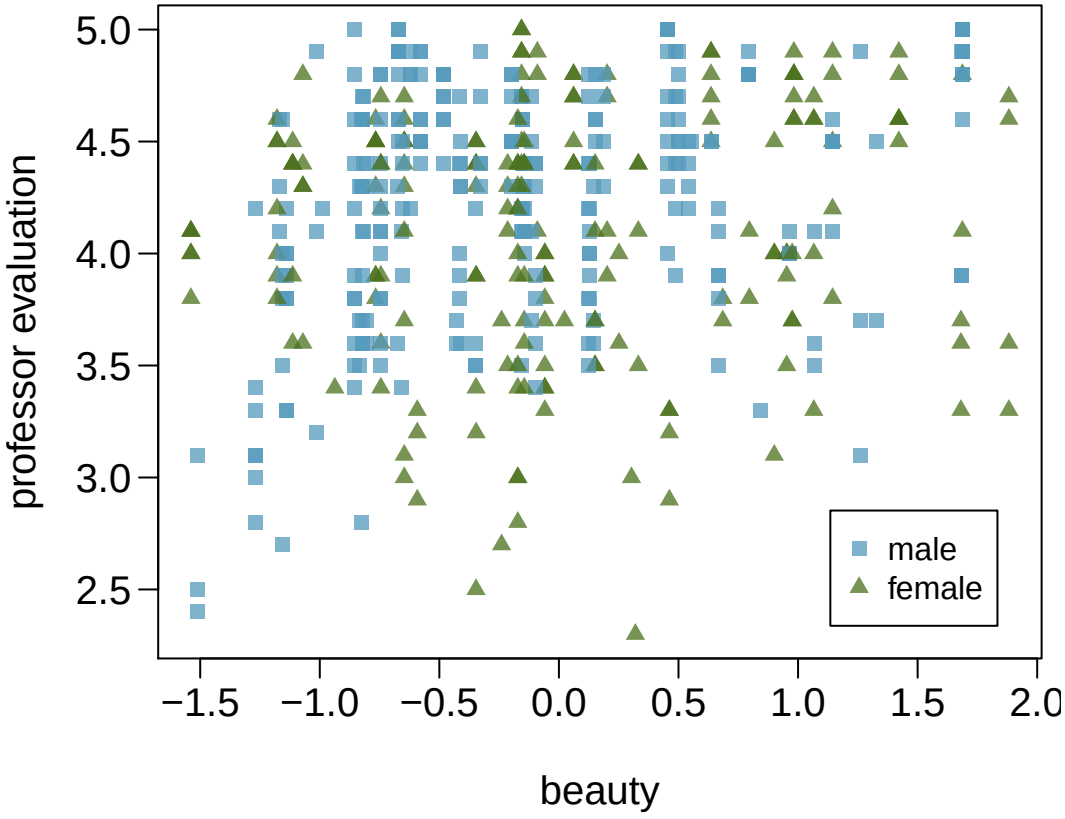
(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(e) The correlation coefficient could be $\sqrt{0.0336} = 0.18$ or -0.18, we can't tell which is correct.

Exploratory analysis

Any interesting features?

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

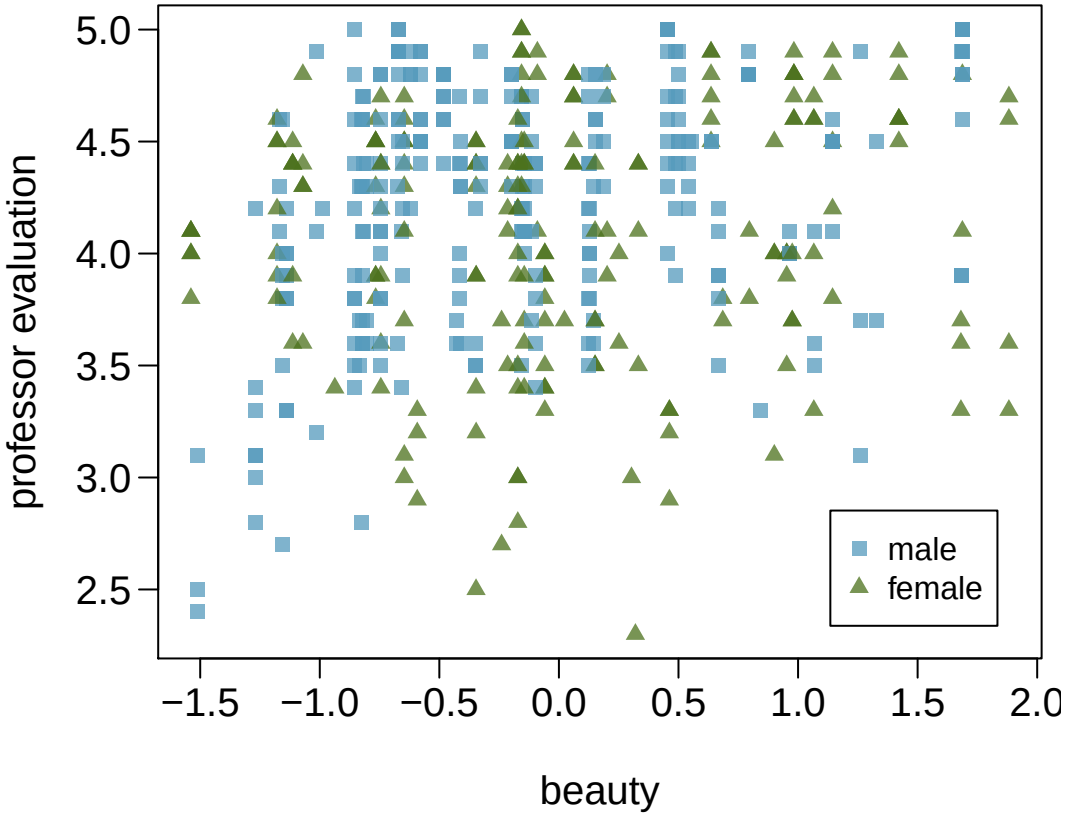


Exploratory analysis

Any interesting features?

Few females with very low beauty scores.

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?



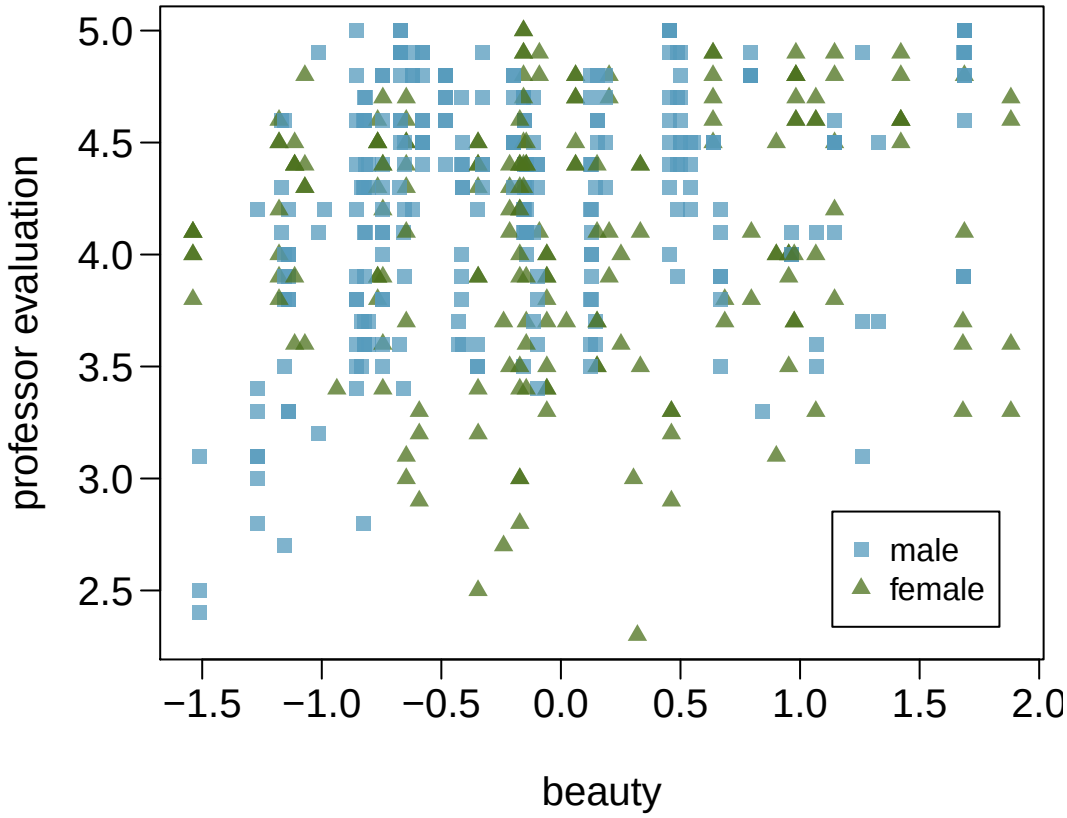
Exploratory analysis

Any interesting features?

Few females with very low beauty scores.

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

Difficult to tell from this plot only.



Professor rating vs. beauty + gender

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00
R^2_{adj} = 0.057				

(a) higher

(b) lower

(c) about the same

Professor rating vs. beauty + gender

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00
R^2_{adj} = 0.057				

(a) *higher → Beauty held constant, male professors are rated 0.17 points higher on average than female professors.*

(b) lower

(c) about the same

Full model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [1]	0.1511	0.0749	2.02	0.04
lower.yes [2]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [3]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [4]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

[1] `formal`: picture wearing tie&jacket/blouse, levels: `yes`, `no`

[2] `lower`: lower division course, levels: `yes`, `no`

[3] `students`: number of students

[4] `tenure`: tenure status, levels: `non-tenure track`, `tenure track`, `tenured`

Hypotheses

Just as the interpretation of the slope parameters take into account all other variables in the model, the hypotheses for testing for significance of a predictor also takes into account all other variables.

$H_0 : B_i = 0$ when other explanatory variables are included in the model.

$H_A : B_i \neq 0$ when other explanatory variables are included in the model.

Assessing significance: numerical variables

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) If we hold all other variables in the model constant, there is strong evidence that professor's age is associated with their rating.

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: numerical variables

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) *If we hold all other variables in the model constant, there is strong evidence that professor's age is associated with their rating.*

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) *All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.*

Assessing significance

Which predictors do not seem to meaningfully contribute to the model, i.e. may not be significant predictors of professor's rating score?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes	0.1511	0.0749	2.02	0.04
lower.yes	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students	-0.0004	0.0004	-1.03	0.30
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

Why are you building a model in the first place?

Model selection: realities

All models are wrong. Some are more useful than others.

- George Box

- If we are asking which is the “true” model, we will have a bad time
- In practice, issues with sample size, collinearity, and available predictors are real problems
- It is often possible to differentiate between better models and less-good models, though
- The key decisions in model selection almost always involve balancing model complexity with the potential for overfitting.

Basic idea for model selection

A very general algorithm

- Specify a "class" of models
- Define a criterion to quantify the fit of each model in the class
- Select the model that optimizes the criterion you're using
- Subject the selected model to model checking/diagnostics, possibly adjust interpretations as needed.

Again, we're focusing on $f(x)$ in the model specification. Once you've selected a model, you should subject it to regression diagnostics – which might change or augment the class of models you specify or alter your criterion.

Classes of models

Some examples of classes of models

- Linear models including all subsets of $x_1, ..., x_p$
- Linear models including all subsets of $x_1, ..., x_p$ and their first order interactions
- ...

Popular criteria

- Adjusted R^2
- Residual mean square error
- Akaike Information Criterion (AIC)
- Bayes Information Criterion (BIC)
- Cross-validated error (similar to Prediction RSS, aka PRESS)
- F- or t-tests (via stepwise selection)
- Likelihood ratio tests (F-tests)

Sidebar: Confusing notation about p

p can mean different things

- p can be the number of covariates you have in your model (not including your column of 1s and the intercept
- p can be the number of betas you estimate, including β_0.

In these slides, p is the former: the number of covariates.

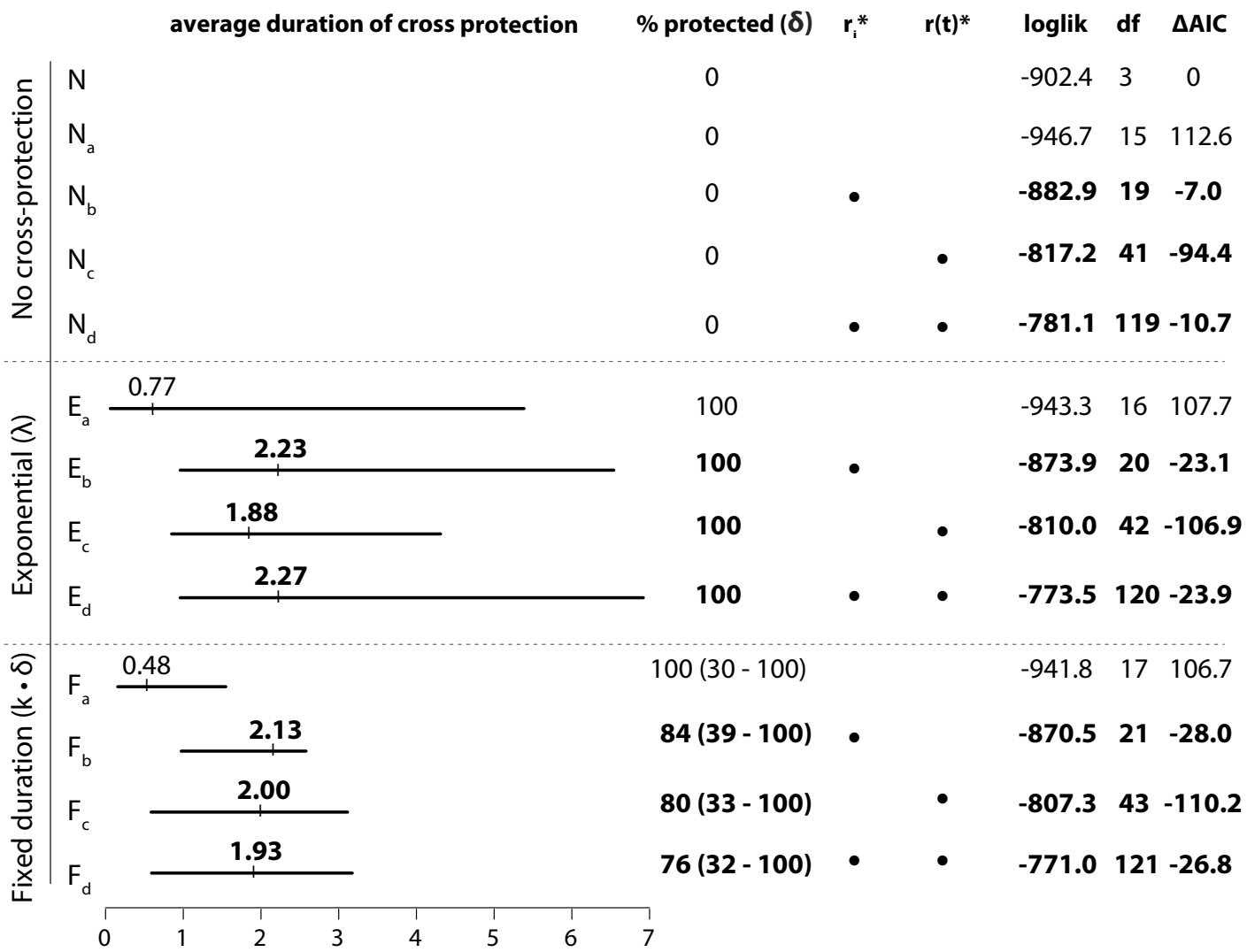
AIC

AIC ("Akaike Information Criterion") measures goodness-of-fit through RSS (equivalently, log likelihood) and penalizes model size:

$$AIC = n \log(RSS/n) + 2(p + 1)$$

- Small AIC's are better, but scores are not directly interpretable
- Penalty on model size tries to induce *parsimony*

Example of AIC in practice



r_i = serotype-specific transmission parameters included
r(t) = seasonal transmission parameters included
loglik = log likelihood for the given model
df = degrees of freedom of the model
ΔAIC = change in Akaike Information Criterion over null model

Reich et al. (2013) Journal of the Royal Society Interface

BIC

BIC ("Bayes Information Criterion") similarly measures goodness-of-fit through RSS (equivalently, log likelihood) and penalizes model size:

$$BIC = n\log(RSS/n) + (p+1)\log(n)$$

- Small BIC's are better, but scores are not directly interpretable
- AIC and BIC measure goodness-of-fit through RSS, but use different penalties for model size. They won't always give the same answer

Bonus link! Bolker on AIC vs. BIC

Example of BIC in practice

Step	Number of Predictors in Model	Breslow's Thickness	DCCD	Ulceration	Age	Nodal Status[a]	Localization	Gender	BIC
1	7	<0.0001	0.0068	0.0009	0.0051	0.0371	0.1380	0.8052	1,657.8
2	6	<0.0001	0.0069	0.0008	0.0050	0.0340	0.1035	—	1,650.9
3	5	<0.0001	0.0011	0.0008	0.0054	0.0475	—	—	1,646.6
4	4	<0.0001	<0.0001	0.0005	0.0127	—	—	—	1,643.6
5	3	<0.0001	<0.0001	0.0002	—	—	—	—	1,642.9
6	2	<0.0001	<0.0001	—	—	—	—	—	1,649.8
7	1	<0.0001	—	—	—	—	—	—	1,679.1

p-Values are for testing whether a hazard ratio equals 1; low BIC identifies best model.
[a]As determined by routine histopathology.
doi:10.1371/journal.pmed.1001604.t004

Vasantha and Venkatesan (2014) PLoS ONE

Example of model selection in practice

TABLE 2. Results of unrestricted longitudinal latent class analysis in the Medical Research Council 1946 National Survey of Health and Development (pooled sexes, *n* = 3,272)

	Three classes (LLCA*-3)	Four classes (LLCA-4)	Five classes (LLCA-5)
Sequential model comparisons (*T* + 1 classes vs. *T* classes)	3 vs. 2	4 vs. 3	5 vs. 4
Log-likelihood value for model with *T* + 1 classes	–3,243.605	–3,211.173	–3,201.380
Log-likelihood value for model with *T* classes	–3,344.440	–3,243.605	–3,211.173
–2 difference in log-likelihood	201.669	64.863	19.587
Difference in no. of parameters (*T* + 1 classes vs. *T* classes)	7	8	8
Lo-Mendell-Rubin adjusted LRT* value	198.171	63.877	19.289
Lo-Mendell-Rubin adjusted LRT *p* value	<0.0001	<0.0001	0.0322
Bootstrap LRT *p* value	<0.01	<0.01	>0.50
Chi-square goodness-of-fit tests			
Degrees of freedom	43	36	29
LRT χ^2	123.588	58.725	39.138
p value	<0.0001	0.0098	0.0990
Bootstrap *p* value†	<0.01	0.02	0.11
Pearson χ^2	132.431	49.416	35.966
p value	<0.0001	0.0674	0.1746
Bootstrap *p* value†	<0.01	0.10	0.40
Information criterion‡			
Akaike's Information Criterion	6,527.210	6,476.347	*6,470.760*
Bayesian Information Criterion	6,649.073	*6,640.862*	6,677.927
Sample-size-adjusted Bayesian Information Criterion	6,585.524	*6,555.071*	6,569.894
Entropy	0.856	0.913	0.897
Condition number§	$0.120E^{-03}$	$0.783E^{-03}$	$0.379E^{-03}$

* LLCA, longitudinal latent class analysis; LRT, likelihood ratio test.
† Bootstrap *p* values were based on 200 resamples.
‡ Minimum values are shown in italic type.
§ Condition number = ratio of the largest eigenvalue to the smallest eigenvalue for the Fisher information

Sequential variable selection methods

PROCEED WITH CAUTION: Stepwise selection methods are dangerous if you want accurate inferences

- General idea: add/remove variables sequentially.
- There are many potential models – usually exhausting the model space is difficult or infeasible
- Stepwise methods don't consider all possibilities
- One paper* showed that stepwise analyses produced models that...
  - represented noise 20-75% of the time
  - contained <50% of actual predictors
  - correlation btw predictors $\longrightarrow$ including more predictors
  - number of predictors correlated with number of noise predictors included

* Derksen and Keselman (1992) British J Math Stat Psych

A more modern approach to variable selection

Penalized regression (a.k.a. “shrinkage”, “regularization”)

- adds an explicit penalty to the least squares criterion
- keeps regression coefficients from being too large, or can shrink coefficients to zero
- Keywords for methods: LASSO, Ridge Regression
- More in Biostat Methods 3 (fall semester)!

Whole branches of modern statistics are devoted to figuring out what to do when $p \geq n$.

Outline

Modeling conditions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

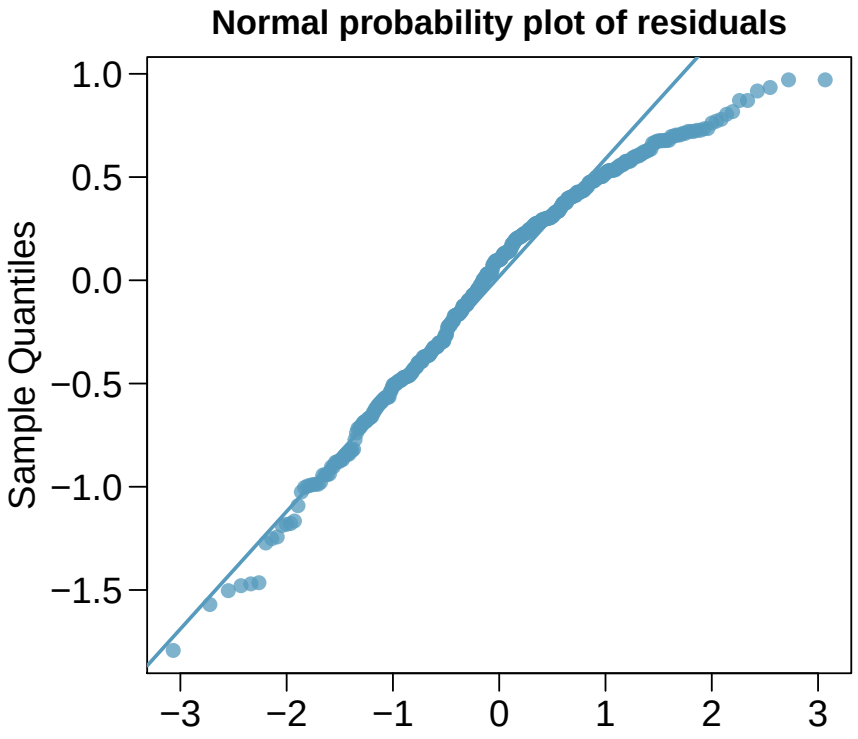
The model depends on the following conditions

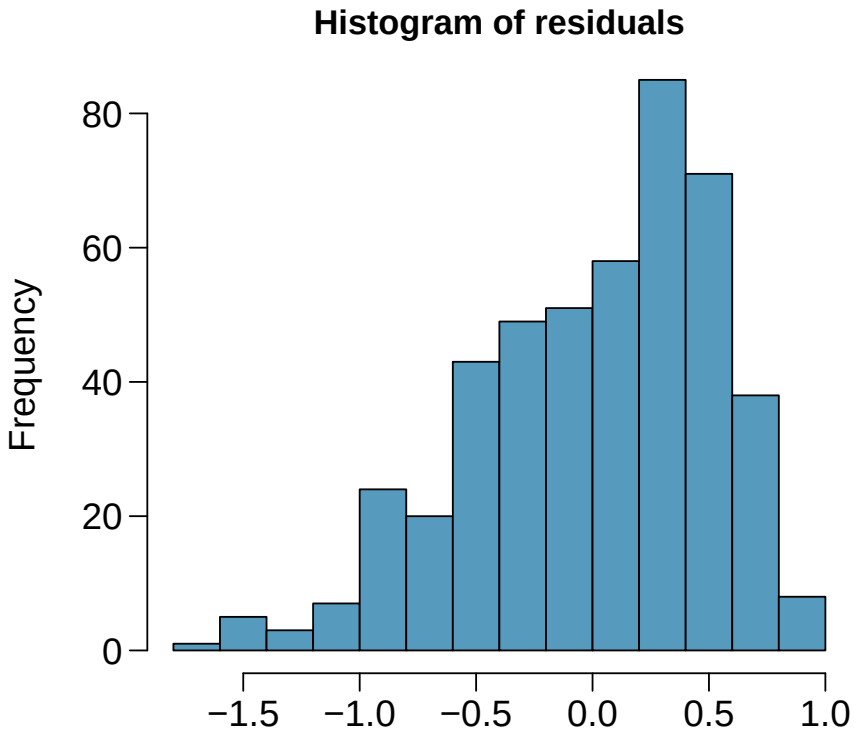
1. residuals are nearly normal (primary concern relates to residuals that are outliers)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

normal probability plot and/or histogram of residuals:

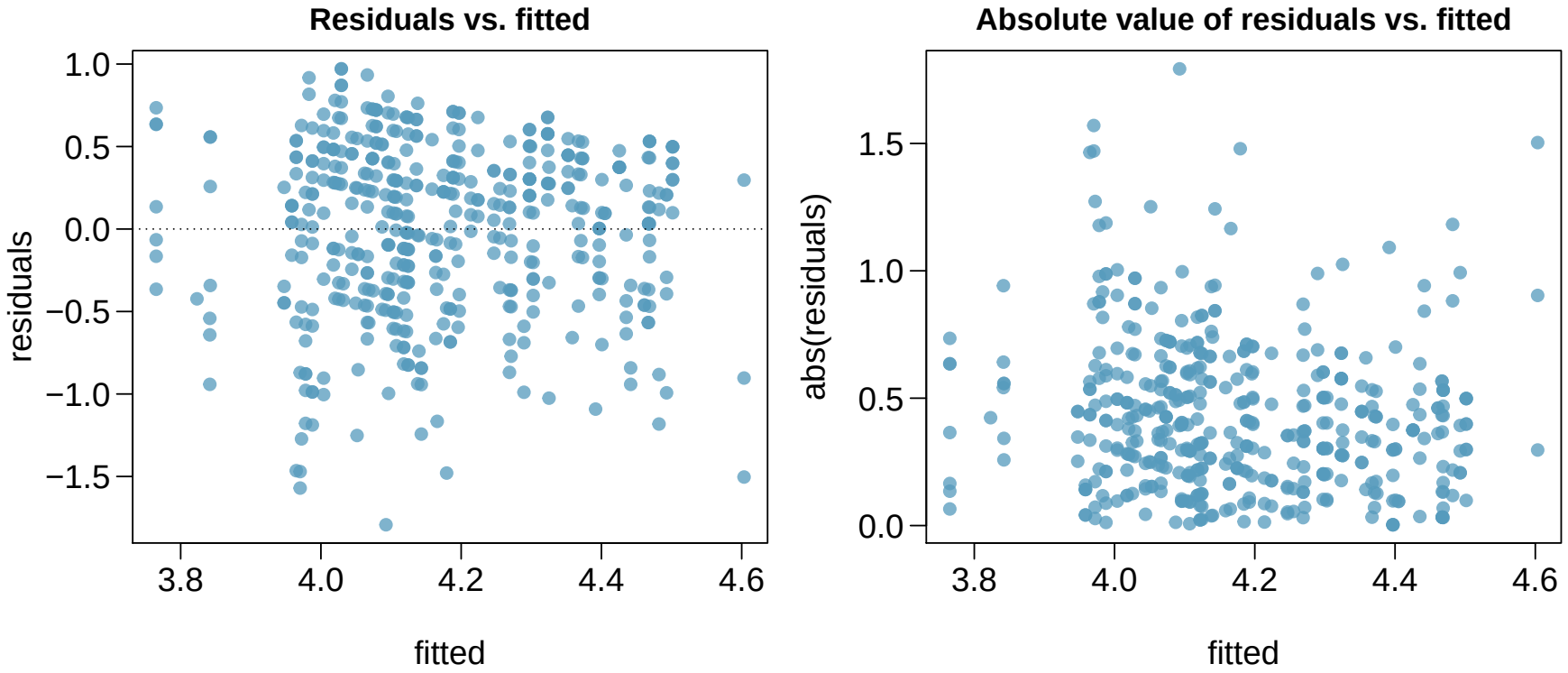




Does this condition appear to be satisfied?

(2) constant variability in residuals

scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted):



Does this condition appear to be satisfied?

Checking constant variance - recap

- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

Checking constant variance - recap

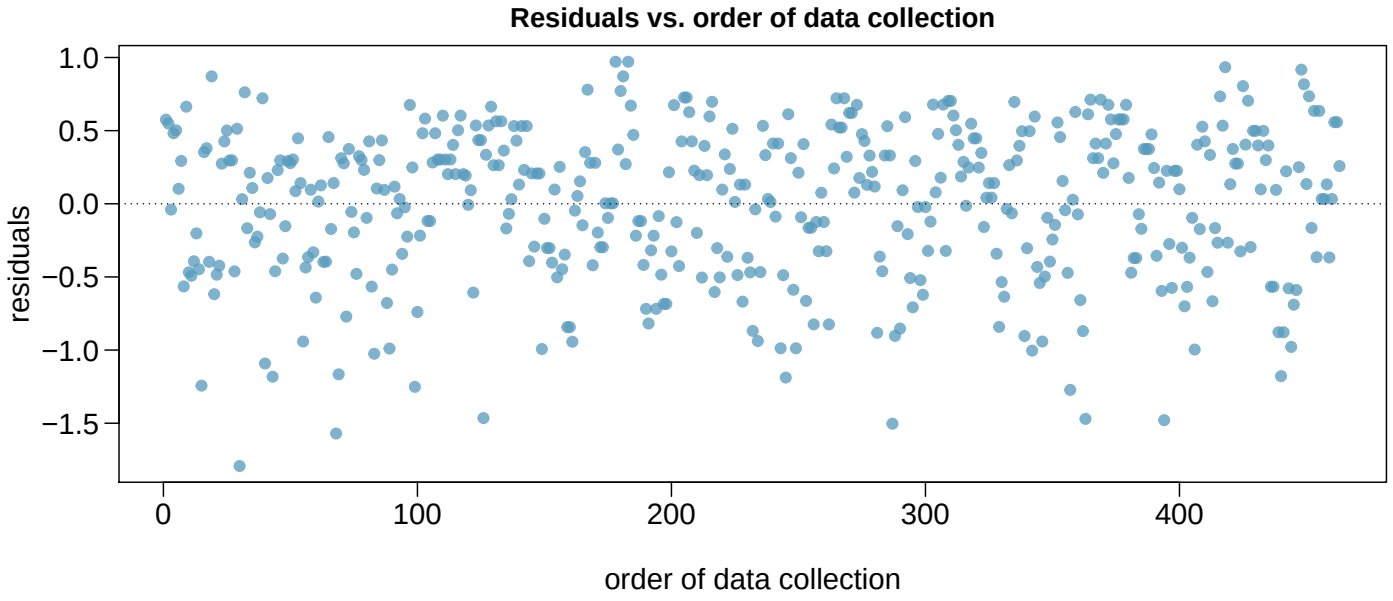
- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

(3) independent residuals

scatterplot of residuals vs. order of data collection:



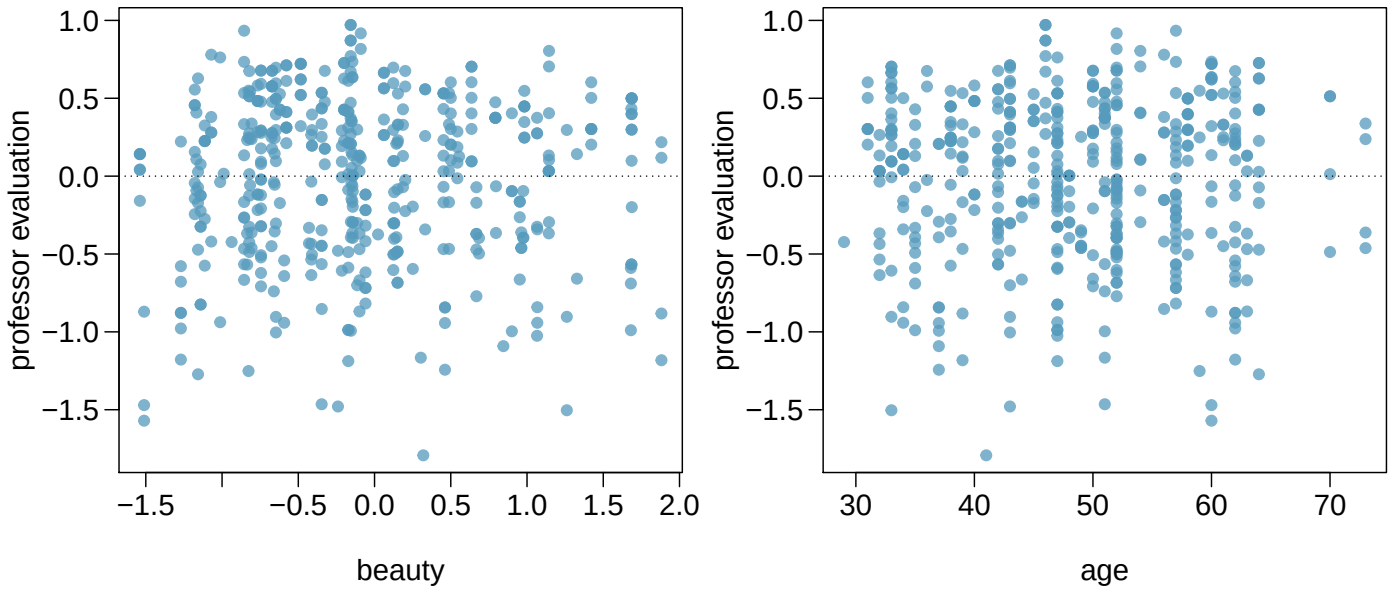
Does this condition appear to be satisfied?

More on the condition of independent residuals

- Checking for independent residuals allows us to indirectly check for independent observations.
- If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.
- This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

scatterplot of residuals vs. each (numerical) explanatory variable:



Does this condition appear to be satisfied?

Note: We use residuals instead of the predictors on the y-axis so that we can still check for linearity without worrying about other possible violations like collinearity between the predictors.