

Before we get started...

Study groups are okay...

But do NOT directly copy others' answers!

Plagiarism won't be tolerated.

Project presentation

- Live presentation sessions in-class: June 17-18 (Thur & Fri)
- Each presentation should be $\leq 10 \text{ minutes}$
- If you can't make it to the live sessions, OR you prefer not to present live, OR you want a back-up presentation...
- You can upload a presentation video (e.g., screen recording with voice over) to the shared Duke Box folder
- Name format:
FirstnameLastname_FirstnameLastname.xxx ("xxx" is the proper file extension name)
- e.g., *FanBu_RickPresman.mp4*
- You may also upload your presentation slides to the folder (with the same name format)

Chapter 9: Multiple and logistic regression

STA 101, Summer I 2021, Duke University

Derived from OpenIntro slides, developed by Mine Çetinkaya-Rundel of OpenIntro.
Edited under the CC BY-SA license.

Introduction to multiple regression

Multiple regression

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

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level*: east
- *Intercept*: The estimated average poverty percentage in eastern states is 11.17%

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level*: east
- *Intercept*: The estimated average poverty percentage in eastern states is 11.17%
 - This is the value we get if we plug in *0* for the explanatory variable

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level*: east
- *Intercept*: The estimated average poverty percentage in eastern states is 11.17%
 - This is the value we get if we plug in 0 for the explanatory variable
- *Slope*: The estimated average poverty percentage in western states is 0.38% higher than eastern states.

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level*: east
- *Intercept*: The estimated average poverty percentage in eastern states is 11.17%
 - This is the value we get if we plug in 0 for the explanatory variable
- *Slope*: The estimated average poverty percentage in western states is 0.38% higher than eastern states.
 - Then, the estimated average poverty percentage in western states is $11.17 + 0.38 = 11.55\%$.

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level*: east
- *Intercept*: The estimated average poverty percentage in eastern states is 11.17%
 - This is the value we get if we plug in *0* for the explanatory variable
- *Slope*: The estimated average poverty percentage in western states is 0.38% higher than eastern states.
 - Then, the estimated average poverty percentage in western states is $11.17 + 0.38 = 11.55\%$.
 - This is the value we get if we plug in *1* for the explanatory variable

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

- (a) northeast
- (b) midwest
- (c) west
- (d) south
- (e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

- (a) *northeast*
- (b) midwest
- (c) west
- (d) south
- (e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

- (a) northeast
- (b) midwest
- (c) west
- (d) south
- (e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

- (a) *northeast*
- (b) midwest
- (c) west
- (d) south
- (e) cannot tell

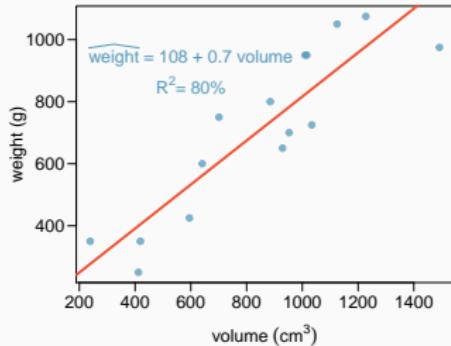
Weights of books

	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



Weights of books (cont.)

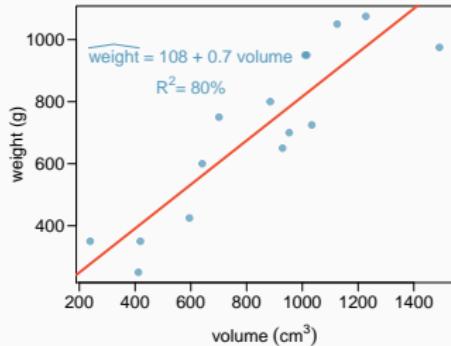
The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



- (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

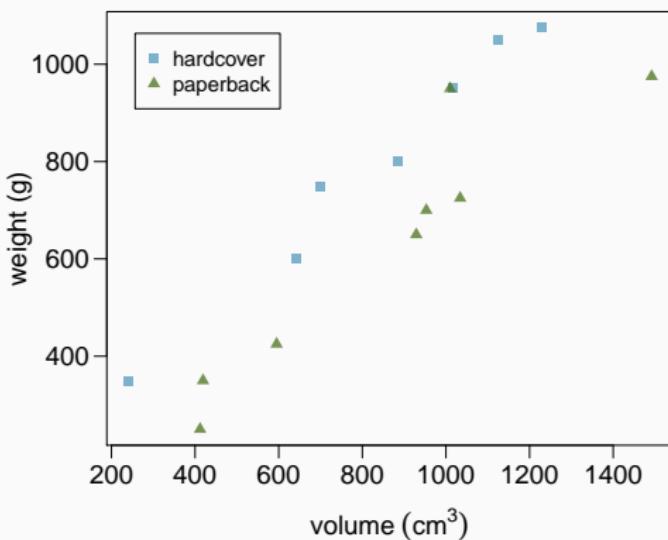
Residual standard error: 123.9 on 13 degrees of freedom

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

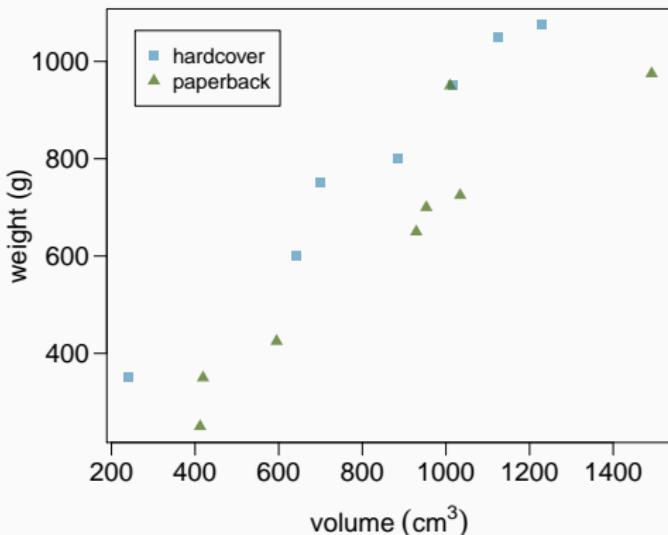
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

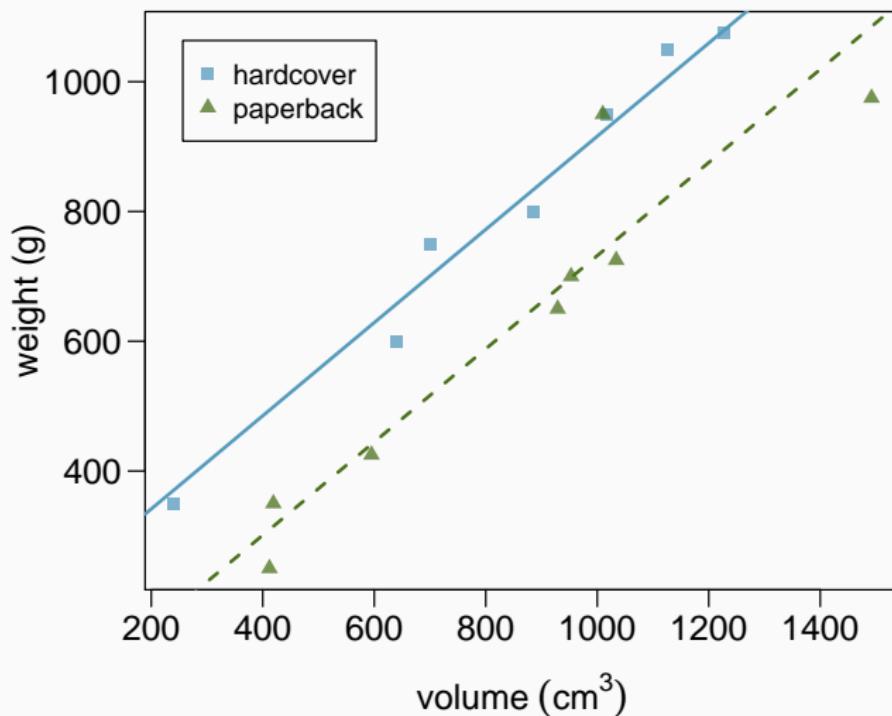
1. For *hardcover* books: plug in *0* for cover

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

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

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

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 cm³?

	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 cm³?

	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) $197.96 + 0.72 * 600 - 184.05 * 1 = 445.91 \text{ grams}$
- (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 kids' 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

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 years 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?

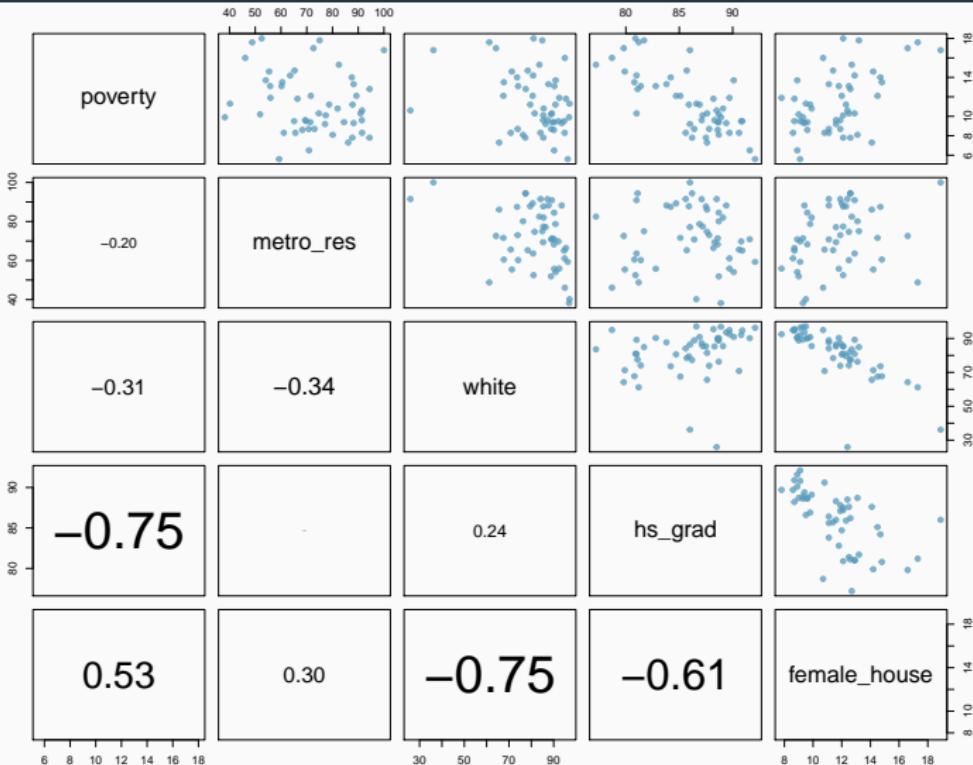
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

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

- (a) are estimated to score 2.54 points lower
- (b) *are estimated to score 2.54 points higher*

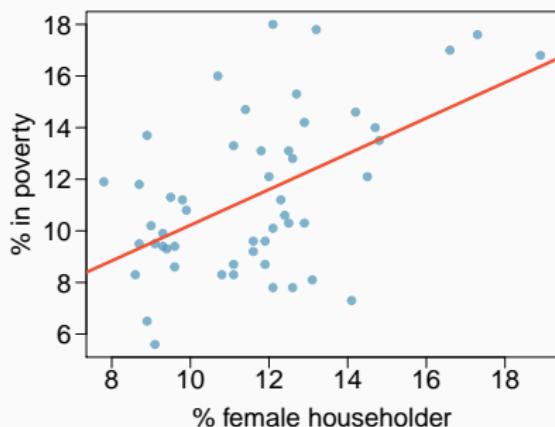
than those whose moms did not work.

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

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

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

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			

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

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

$$\begin{aligned}\text{Sum of squares of } x: SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \text{explained variability} \\ &= 480.25 - 347.68 = 132.57\end{aligned}$$

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			

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

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

$$\begin{aligned}\text{Sum of squares of } x: SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \text{explained variability} \\ &= 480.25 - 347.68 = 132.57\end{aligned}$$

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

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 x s.
- 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

<i>Linear model:</i>	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

<i>ANOVA:</i>	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			

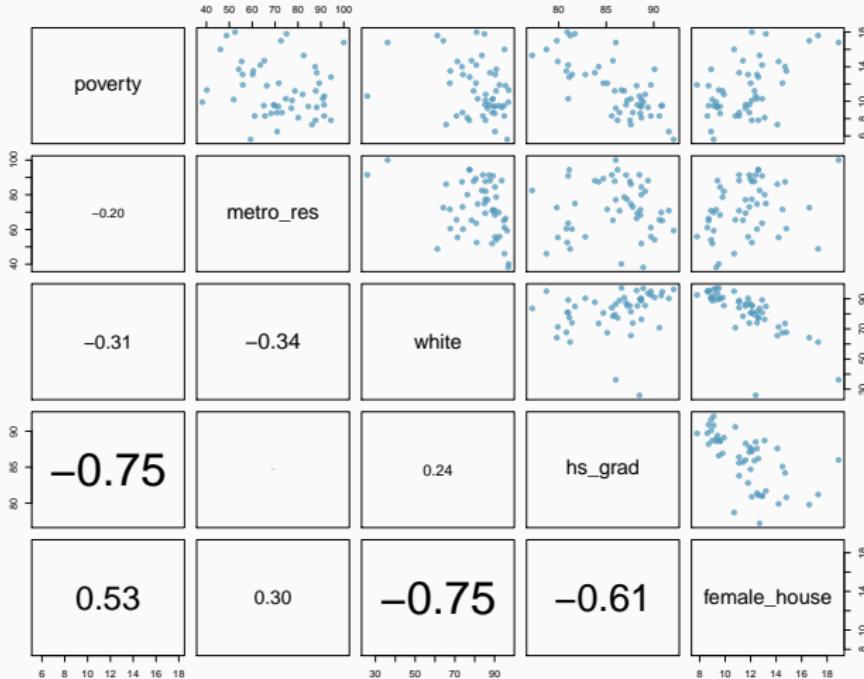
Predicting poverty using % female hh + % white

<i>Linear model:</i>	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

<i>ANOVA:</i>	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

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

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_{adj}^2 will always be smaller than R^2 .
- R_{adj}^2 applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R_{adj}^2 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			

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

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			

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

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			

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

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			

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

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			

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

Model selection

Beauty in the classroom

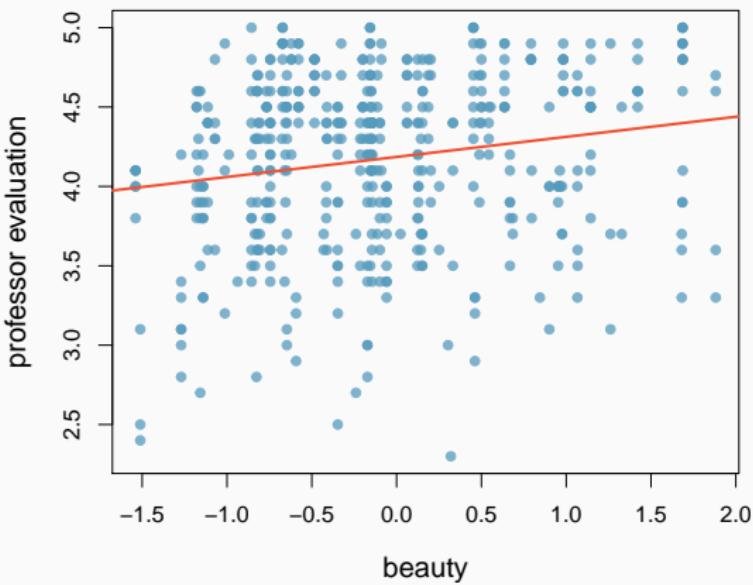
- Data: Student evaluations of instructors' beauty and teaching quality for 463 courses at the University of Texas.
- Evaluations conducted at the end of semester, and the beauty judgements were made later, by six students who had not attended the classes and were not aware of the course evaluations (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) Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.
- (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

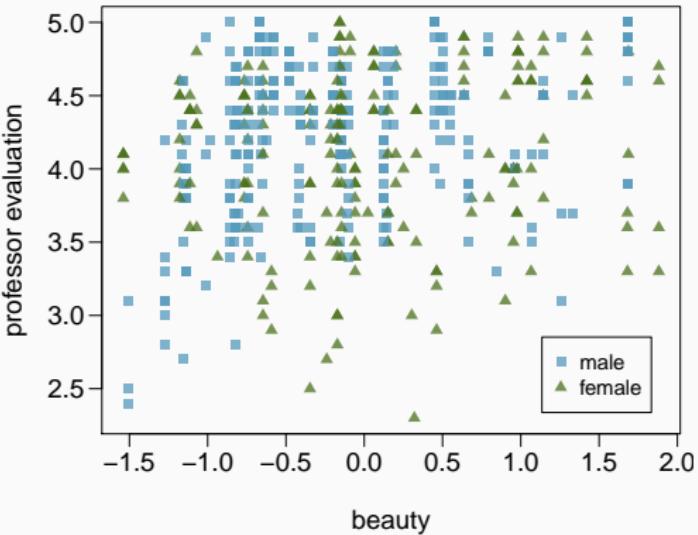
$$R^2 = 0.0336$$

- (a) Model predicts 3.36% of professor ratings correctly.
- (b) Beauty is not a significant predictor of professor evaluation.
- (c) *Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.*
- (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 evalu-
ated higher, lower, or about
the same as female profes-
sors?

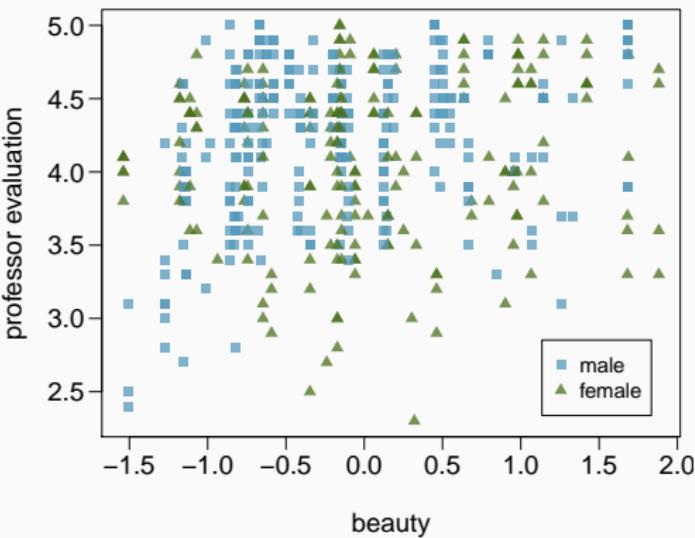


Exploratory analysis

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?



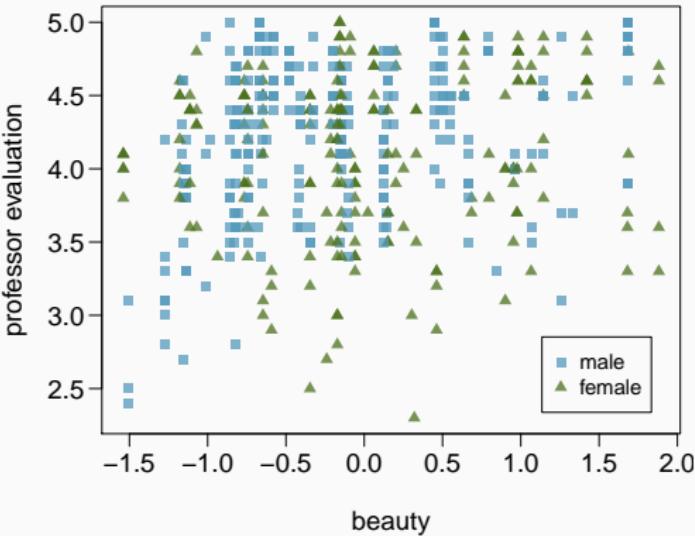
Exploratory analysis

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

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

²lower: lower division course, levels: yes, no

³students: number of students

⁴tenure: 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.

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

$H_A : \beta_i \neq 0$ 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
...				

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
...				

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

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. What can be said about the variable tenure?

	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

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. What can be said about the variable tenure?

	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

1. Reference level is “non tenure track”.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. What can be said about the variable tenure?

	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

1. Reference level is “non tenure track”.
2. All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. What can be said about the variable tenure?

	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

1. Reference level is “non tenure track”.
2. All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.
3. All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track 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

Model selection strategies

Based on what we've learned so far, what are some ways you can think of that can be used to determine which variables to keep in the model and which to leave out?

Backward-elimination

1. Start with the full model
2. Drop one variable at a time and record R^2_{adj} of each smaller model
3. Pick the model with the highest increase in R^2_{adj}
4. Repeat until none of the models yield an increase in R^2_{adj}

Backward-elimination

Full

| beauty + gender + age + formal + lower + native + minority + students + tenure

0.0839

Backward-elimination

Full	beauty + gender + age + formal + lower + native + minority + students + tenure	0.0839
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	0.0842
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733

Backward-elimination

Full	beauty + gender + age + formal + lower + native + minority + students + tenure	0.0839
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	0.0842
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	0.0846
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725

Backward-elimination

Full	beauty + gender + age + formal + lower + native + minority + students + tenure	0.0839
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	0.0842
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	0.0846
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	0.0855
	beauty + gender + age + formal + native + students	0.0713

Backward-elimination

Full	beauty + gender + age + formal + lower + native + minority + students + tenure	0.0839
Step 1	gender + age + formal + lower + native + minority + students + tenure beauty + age + formal + lower + native + minority + students + tenure beauty + gender + formal + lower + native + minority + students + tenure beauty + gender + age + lower + native + minority + students + tenure beauty + gender + age + formal + native + minority + students + tenure beauty + gender + age + formal + lower + minority + students + tenure beauty + gender + age + formal + lower + native + students + tenure beauty + gender + age + formal + lower + native + minority + tenure beauty + gender + age + formal + lower + native + minority + students	0.0642 0.0557 0.0706 0.0777 0.0837 0.0788 0.0842 0.0838 0.0733
Step 2	gender + age + formal + lower + native + students + tenure beauty + age + formal + lower + native + students + tenure beauty + gender + formal + lower + native + students + tenure beauty + gender + age + lower + native + students + tenure beauty + gender + age + formal + native + students + tenure beauty + gender + age + formal + lower + native + tenure beauty + gender + age + formal + lower + native + students	0.0647 0.0543 0.0708 0.0776 0.0846 0.0844 0.0725
Step 3	gender + age + formal + native + students + tenure beauty + age + formal + native + students + tenure beauty + gender + formal + native + students + tenure beauty + gender + age + native + students + tenure beauty + gender + age + formal + students + tenure beauty + gender + age + formal + native + tenure beauty + gender + age + formal + native + students	0.0653 0.0534 0.0707 0.0786 0.0756 0.0855 0.0713
Step 4	gender + age + formal + native + tenure beauty + age + formal + native + tenure beauty + gender + formal + native + tenure beauty + gender + age + native + tenure beauty + gender + age + formal + tenure beauty + gender + age + formal + native	0.0667 0.0553 0.0723 0.0806 0.0773 0.0713

step function in R

The step function in R does a similar backward elimination process, however it uses a different metric called AIC (Akaike Information Criterion) instead of adjusted R^2 to do the model selection.

Call:

```
lm(formula = profevaluation ~ beauty + gender + age + formal +  
    native + tenure, data = d)
```

Coefficients:

(Intercept)	beauty	gendermale
4.628435	0.105546	0.208079
age	formalyes	native non english
-0.008844	0.132422	-0.243003
tenure	tenure track	tenure tenured
-0.206784	-0.175967	

step function in R

The step function in R does a similar backward elimination process, however it uses a different metric called AIC (Akaike Information Criterion) instead of adjusted R^2 to do the model selection.

Call:

```
lm(formula = profevaluation ~ beauty + gender + age + formal +  
    native + tenure, data = d)
```

Coefficients:

(Intercept)	beauty	gendermale
4.628435	0.105546	0.208079
age	formalyes	native non english
-0.008844	0.132422	-0.243003
tenure	tenure track	tenure tenured
-0.206784	-0.175967	

Best model: beauty + gender + age + formal + native + tenure

Forward selection

1. Start with regressions of response vs. each explanatory variable
2. Pick the model with the highest R^2_{adj}
3. Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
4. Repeat until the addition of any of the remaining variables does not result in a higher R^2_{adj}

- Backward elimination with the p-value approach:
 1. Start with the full model
 2. Drop the variable with the highest p-value and refit a smaller model
 3. Repeat until all variables left in the model are significant
- Forward selection with the p-value approach:
 1. Start with regressions of response vs. each explanatory variable
 2. Pick the variable with the lowest significant p-value
 3. Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
 4. Repeat until any of the remaining variables does not have a significant p-value

Backward-elimination: p – value approach

Step	Variables included & p-value									
	beauty	gender male	age	formal yes	lower yes	native nonenglish	minority yes	students	tenure tenure track	tenure tenured
Full	0.00	0.00	0.01	0.04	0.29	0.06	0.35	0.30	0.02	0.02

Backward-elimination: p – value approach

Step	Variables included & p-value									
	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
	male			yes	yes	nonenglish	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	0.35	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
	male			yes	yes	nonenglish			tenure track	tenured
	0.00	0.00	0.01	0.04	0.38	0.03		0.34	0.02	0.01

Backward-elimination: p – value approach

Step	Variables included & p-value									
	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
	male			yes	yes	nonenglish	yes		track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	0.35	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
	male			yes	yes	nonenglish			track	tenured
	0.00	0.00	0.01	0.04	0.38	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
	male			yes		nonenglish			track	tenured
	0.00	0.00	0.01	0.05		0.02		0.44	0.01	0.01

Backward-elimination: p – value approach

Step	Variables included & p-value									
	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
	male	male	yes	yes	yes	nonenglish	yes	tenure track	tenured	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	0.35	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native	students	tenure	tenure	tenure
	male	male	yes	yes	yes	nonenglish	tenure track	tenured	tenured	tenured
	0.00	0.00	0.01	0.04	0.38	0.03	0.34	0.02	0.01	0.01
Step 2	beauty	gender	age	formal	native		students	tenure	tenure	tenure
	male	male	yes	yes	nonenglish	0.02	0.44	tenure track	tenured	tenured
	0.00	0.00	0.01	0.05			0.01	0.01	0.01	0.01
Step 3	beauty	gender	age	formal	native		tenure	tenure	tenure	tenure
	male	male	yes	yes	nonenglish	0.02	tenure track	tenured	tenured	tenured
	0.00	0.00	0.01	0.06			0.01	0.01	0.01	0.01

Backward-elimination: p – value approach

Step	Variables included & p-value									
	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
	male	male	yes	yes	yes	nonenglish	yes	tenure track	tenured	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	0.35	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native	students	tenure	tenure	tenure
	male	male	yes	yes	yes	nonenglish	tenure track	tenured	tenured	tenured
	0.00	0.00	0.01	0.04	0.38	0.03	0.34	0.02	0.01	0.01
Step 2	beauty	gender	age	formal		native	students	tenure	tenure	tenure
	male	male	yes		nonenglish	tenure track	tenured	tenured	tenured	tenured
	0.00	0.00	0.01	0.05	0.02	0.44	0.01	0.01	0.01	0.01
Step 3	beauty	gender	age	formal		native	tenure	tenure	tenure	tenure
	male	male	yes		nonenglish	tenure track	tenured	tenured	tenured	tenured
	0.00	0.00	0.01	0.06	0.02	0.01	0.01	0.01	0.01	0.01
Step 4	beauty	gender	age			native	tenure	tenure	tenure	tenure
	male	male			nonenglish	tenure track	tenured	tenured	tenured	tenured
	0.00	0.00	0.01		0.06	0.01	0.01	0.01	0.01	0.01

Backward-elimination: p – value approach

Step	Variables included & p-value									
	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
Full	beauty	gender male	age 0.01	formal yes	lower yes	native nonenglish	minority yes	students 0.30	tenure tenure track	tenure tenured
	0.00	0.00	0.01	0.04	0.29	0.06	0.35		0.02	0.02
Step 1	beauty	gender male	age 0.01	formal yes	lower yes	native nonenglish		students 0.34	tenure tenure track	tenure tenured
	0.00	0.00	0.01	0.04	0.38	0.03		0.34	0.02	0.01
Step 2	beauty	gender male	age 0.01	formal yes		native nonenglish		students 0.44	tenure tenure track	tenure tenured
	0.00	0.00	0.01	0.05		0.02		0.44	0.01	0.01
Step 3	beauty	gender male	age 0.01	formal yes		native nonenglish			tenure tenure track	tenure tenured
	0.00	0.00	0.01	0.06		0.02			0.01	0.01
Step 4	beauty	gender male	age 0.01			native nonenglish			tenure tenure track	tenure tenured
	0.00	0.00	0.01			0.06			0.01	0.01
Step 5	beauty	gender male	age 0.01						tenure tenure track	tenure tenured
	0.00	0.00	0.01						0.01	0.01

Backward-elimination: p – value approach

Step	Variables included & p-value									
	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
	male	male	yes	yes	yes	nonenglish	yes	tenure track	tenured	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	0.35	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native	students	tenure	tenure	tenure
	male	male	yes	yes	yes	nonenglish	tenure track	tenured	tenured	tenured
	0.00	0.00	0.01	0.04	0.38	0.03	0.34	0.02	0.01	0.01
Step 2	beauty	gender	age	formal		native	students	tenure	tenure	tenure
	male	male	yes		nonenglish	tenure track	tenured	tenured	tenured	tenured
	0.00	0.00	0.01	0.05	0.02	0.44	0.01	0.01	0.01	0.01
Step 3	beauty	gender	age	formal		native	tenure	tenure	tenure	tenure
	male	male	yes		nonenglish	tenure track	tenured	tenured	tenured	tenured
	0.00	0.00	0.01	0.06	0.02	0.01	0.01	0.01	0.01	0.01
Step 4	beauty	gender	age			native	tenure	tenure	tenure	tenure
	male	male			nonenglish	tenure track	tenured	tenured	tenured	tenured
	0.00	0.00	0.01		0.06	0.01	0.01	0.01	0.01	0.01
Step 5	beauty	gender	age				tenure	tenure	tenure	tenure
	male	male				tenure track	tenured	tenured	tenured	tenured
	0.00	0.00	0.01			0.01	0.01	0.01	0.01	0.01

Best model: beauty + gender + age + tenure

Adjusted R^2 vs. p-value approaches

- The two approaches are similar, but they sometimes lead to different models, with the adjusted R^2 approach tending to include *more* predictors in the final model.

Adjusted R^2 vs. p-value approaches

- The two approaches are similar, but they sometimes lead to different models, with the adjusted R^2 approach tending to include *more* predictors in the final model.
- When the sole goal is to improve prediction accuracy, use R^2 . Common in machine learning applications.
- P-value approach is preferred if
 - we care about which variables are statistically significant predictors
 - OR interest is in producing a simpler model at the potential cost of a little prediction accuracy

Adjusted R^2 vs. p-value approaches

- The two approaches are similar, but they sometimes lead to different models, with the adjusted R^2 approach tending to include *more* predictors in the final model.
- When the sole goal is to improve prediction accuracy, use R^2 . Common in machine learning applications.
- P-value approach is preferred if
 - we care about which variables are statistically significant predictors
 - OR interest is in producing a simpler model at the potential cost of a little prediction accuracy
- Still have to verify the model conditions (coming up) regardless of the approach of model selection

Checking model conditions using graphs

Modeling conditions

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

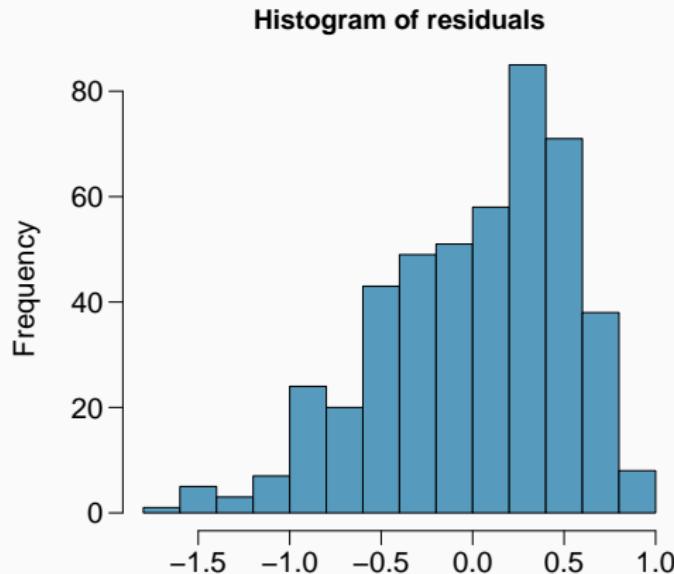
The model depends on the following conditions

1. residuals are nearly normal (less important for larger data sets)
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, where the graphs are also called diagnostic plots.

(1) nearly normal residuals

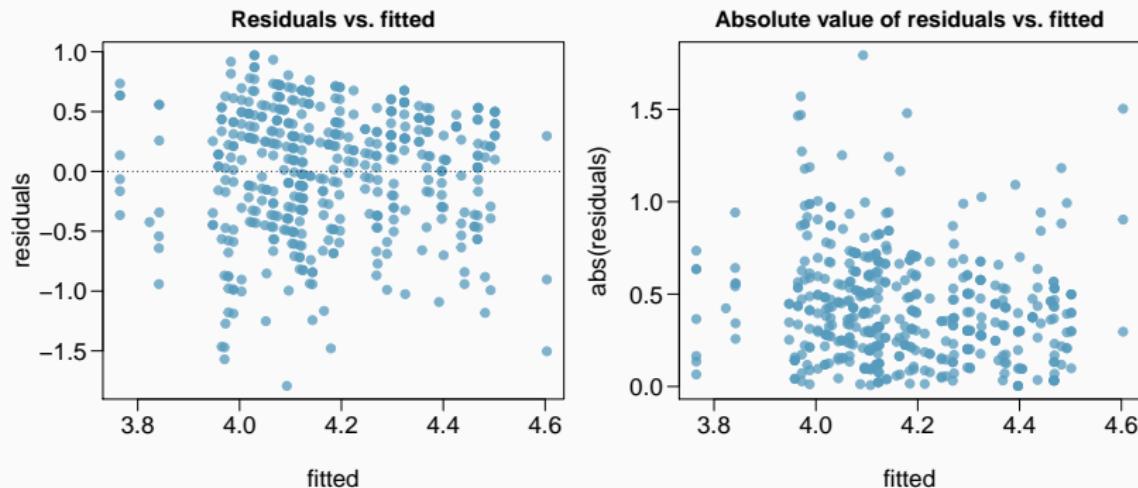
normal probability plot and/or histogram of residuals:



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

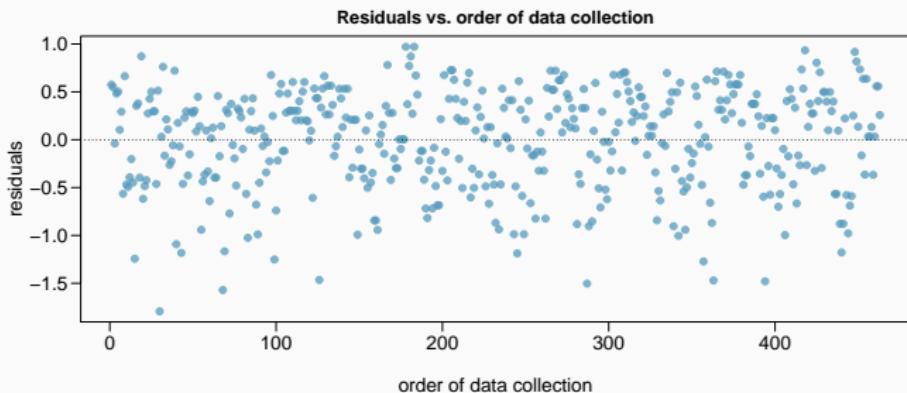
- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

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:



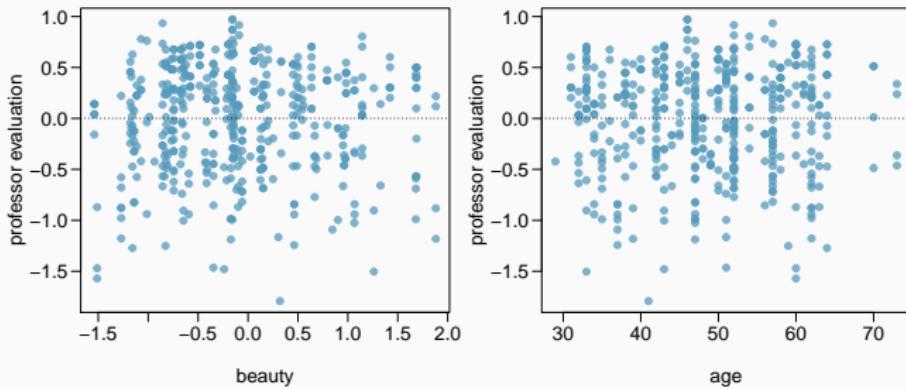
Does this condition appear to be satisfied?

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.

Several options for improving a model

- Transforming variables
- Seeking out additional variables to fill model gaps
- Using more advanced methods that would account for challenges around inconsistent variability or nonlinear relationships between predictors and the outcome

Transformations

If the concern with the model is non-linear relationships between the explanatory variable(s) and the response variable, transforming the response variable can be helpful.

- Log transformation ($\log y$)
- Square root transformation (\sqrt{y})
- Inverse transformation ($1/y$)
- Truncation (cap the max value possible)

It is also possible to apply transformations to the explanatory variable(s), however such transformations tend to make the model coefficients even harder to interpret.

Models can be wrong, but useful

All models are wrong, but some are useful. - George Box

- No model is perfect, but even imperfect models can be useful, as long as we are clear and report the model's shortcomings.
- If conditions are grossly violated, we should not report the model results, but instead consider a new model, even if it means learning more statistical methods or hiring someone who can help.

Logistic regression

Regression so far ...

At this point we have covered:

- Simple linear regression
 - Relationship between numerical response and a numerical or categorical predictor

Regression so far ...

At this point we have covered:

- Simple linear regression
 - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
 - Relationship between numerical response and multiple numerical and/or categorical predictors

Regression so far ...

At this point we have covered:

- Simple linear regression
 - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
 - Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven't seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Regression so far ...

At this point we have covered:

- Simple linear regression
 - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
 - Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven't seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.).

We now discuss binary responses (categorical response with 2 levels).

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

Odds

For some event E ,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = \frac{x}{x+y}, \quad P(E^c) = \frac{y}{x+y}$$

Example - Donner Party

- Donner and Reed families travelled from Springfield, Illinois to Sacramento, California by covered wagon.
- Got delayed twice at difficult crossings, and then got stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October.
- By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd ed)

Example - Donner Party - Data

	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
:	:	:	:
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs. Gender:

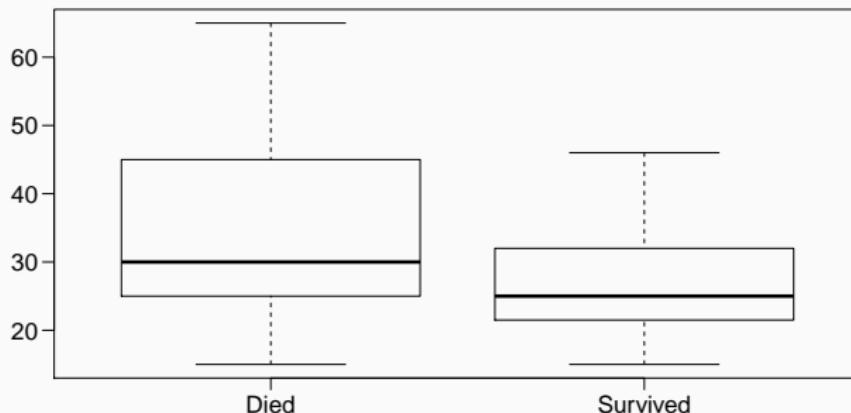
	Male	Female
Died	20	5
Survived	10	10

Example - Donner Party - EDA

Status vs. Gender:

	Male	Female
Died	20	5
Survived	10	10

Status vs. Age:



Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a Bernoulli distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model
 - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
3. A link function that relates the linear model to the parameter of the outcome distribution
 - $g(p) = \eta$ or $p = g^{-1}(\eta)$

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a Bernoulli distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a Bernoulli distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p . There are a variety of options but the most commonly used is the logit function.

Logit function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞ .

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Bern}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Example - Donner Party - Model

In R we fit a GLM in the same was as a linear model except using `glm` instead of `lm` and we must also specify the type of GLM to fit using the `family` argument.

```
summary(glm(Status ~ Age, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.81852   0.99937   1.820   0.0688 .
## Age        -0.06647   0.03222  -2.063   0.0391 *
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 56.291 on 43 degrees of freedom
## AIC: 60.291
##
## Number of Fisher Scoring iterations: 4
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times 0 \\ \frac{p}{1-p} &= \exp(1.8185) = 6.16 \end{aligned}$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

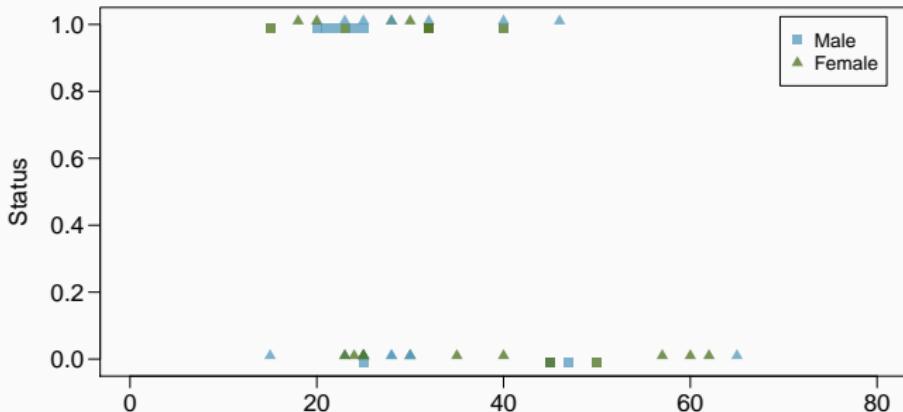
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

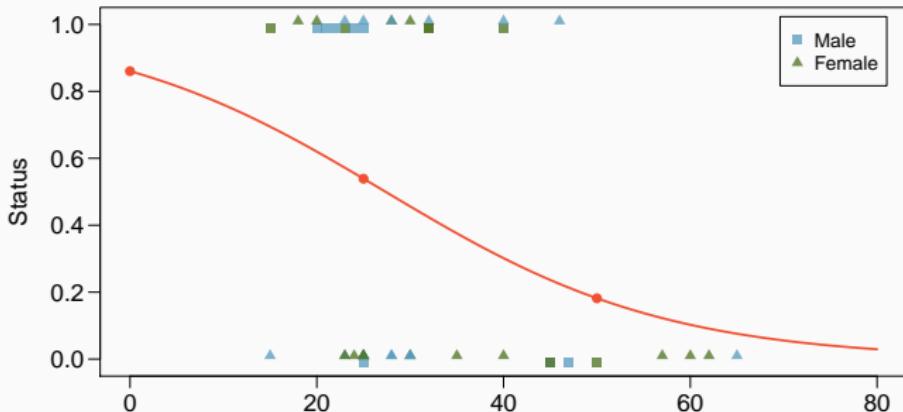
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Simple interpretation is only possible in terms of log odds and log odds ratios for intercept and slope terms.

Intercept: The **log odds** of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.

Slope: For a unit increase in age (being 1 year older) how much will the **log odds** change (or how much will the **log odds ratio** be), not particularly intuitive. More often than not we care only about sign and relative magnitude.

Example - Donner Party - Interpretation - Slope

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.8185 - 0.0665(x + 1) \\ &= 1.8185 - 0.0665x - 0.0665\end{aligned}$$

$$\log\left(\frac{p_2}{1-p_2}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = -0.0665$$

$$\log\left(\frac{p_1}{1-p_1} \middle| \frac{p_2}{1-p_2}\right) = -0.0665$$

$$\frac{p_1}{1-p_1} \middle| \frac{p_2}{1-p_2} = \exp(-0.0665) = 0.94$$

Example - Donner Party - Age and Gender

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age        -0.07820   0.03728  -2.097   0.0359 *
## SexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## 
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Gender slope: When the other predictors are held constant this is the log odds ratio between the given level (Female) and the reference level (Male).

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

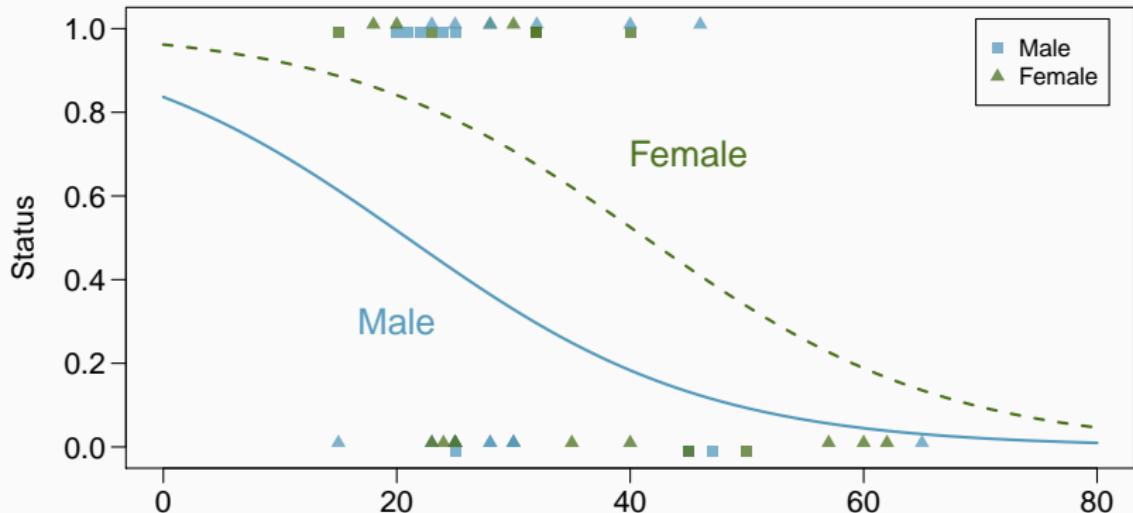
Male model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{O} \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

Female model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{O} \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

Example - Donner Party - Gender Models (cont.)



Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age        -0.07820   0.03728  -2.097   0.0359 *
## SexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## 
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age        -0.07820   0.03728  -2.097   0.0359 *
## SexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## 
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Note: The model output does not include any F-statistic, as a general rule there are not single model hypothesis tests for GLM models.

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We are however still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z test.

Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta}_{age} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\begin{aligned} p\text{-value} &= P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10) \\ &= 2 \times 0.0178 = 0.0359 \end{aligned}$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds (or “log odds ratio”) per unit increase in the predictor.

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds (or “log odds ratio”) per unit increase in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds (or “log odds ratio”) per unit increase in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

$$\exp(CI) = (\exp(-0.1513), \exp(-0.0051)) = (0.8596, 0.9949)$$

Interpretation: odds but not probabilities!

The most common mistake made when interpreting logistic regression is to treat (log) odds ratios as ratios of probabilities.

Interpretation: odds but not probabilities!

The most common mistake made when interpreting logistic regression is to treat (log) odds ratios as ratios of probabilities.

Females are *not* $\exp(1.6) = 4.95$ times more likely than males to survive in the Donnor Party.

Interpretation: odds but not probabilities!

The most common mistake made when interpreting logistic regression is to treat (log) odds ratios as ratios of probabilities.

Females are *not* $\exp(1.6) = 4.95$ times more likely than males to survive in the Donnor Party.

But rather, the survival odds ratio between females and males is $\exp(1.6) = 4.95$.

Interpretation: odds but not probabilities!

The most common mistake made when interpreting logistic regression is to treat (log) odds ratios as ratios of probabilities.

Females are *not* $\exp(1.6) = 4.95$ times more likely than males to survive in the Donnor Party.

But rather, the survival odds ratio between females and males is $\exp(1.6) = 4.95$.

Relative risk v.s. odds ratio:

$$RR = \frac{P(\text{survive|female})}{P(\text{survive|male})}$$

$$OR = \frac{P(\text{survive|female})/[1 - P(\text{survive|female})]}{P(\text{survive|male})/[1 - P(\text{survive|male})]}$$