Introduction to multiple regression

Multiple regression

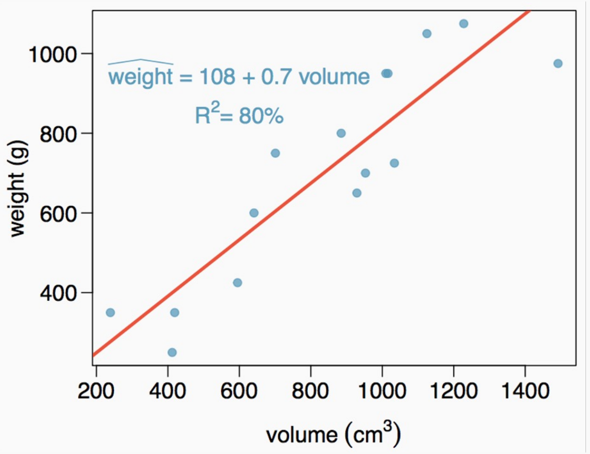
- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

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

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output.



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

Modeling weights of books using volume

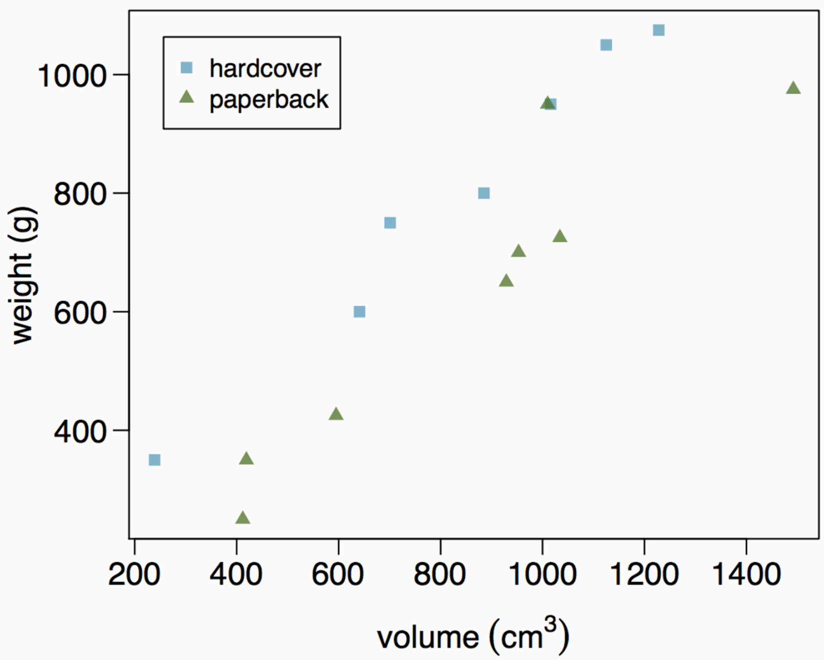
somewhat abbreviated output...

```
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

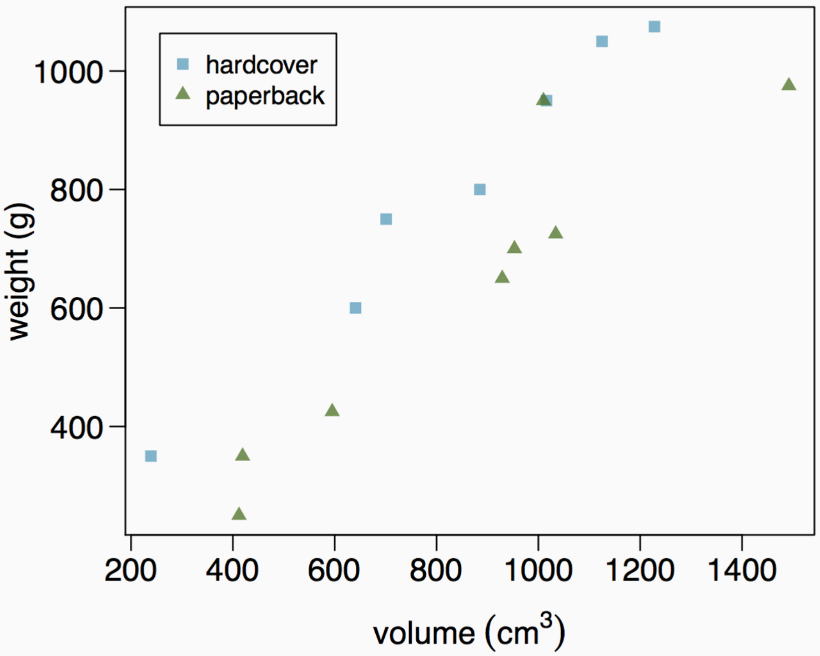
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



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.



Qualitative predictors

How can we include hardcover/paperback in our regression?

When a predictor takes two (categorical) values, we create a dummy variable which takes the values 0/1:

$$cover{:}pb = \begin{cases} 0, if\ cover\ is\ hardcover \\ 1, if\ cover\ is\ paperback \end{cases}$$

*Which value we select for 0 is arbitrary, and is called the **reference** value*

Qualitative predictors

How can we include hardcover/paperback in our regression?

When a predictor takes two (categorical) values, we create a dummy variable which takes the values 0/1:

$$cover{:}pb = \begin{cases} 0, if\ cover\ is\ hardcover \\ 1, if\ cover\ is\ paperback \end{cases}$$

How can we interpret β_1 *in this case*

Assume $y = \beta_0 + \beta_1 x + \epsilon_i,$ *where* $y\ is\ weight, x\ is\ cover{:}pb$

Then

$$\mathrm{y_i} = \begin{cases} \beta_0 + \epsilon_i, & if\ the\ cover\ is\ hardcover \\ \beta_0 + \beta_1 + \epsilon_i, & if\ the\ cover\ is\ paperback \end{cases}$$

Paperback books are on average β_1 *units heavier than hardcover books*

Do we expect β_1 *to be positive or negative?*

Qualitative predictors

How can we include hardcover/paperback in our regression?

When a predictor takes two (categorical) values, we create a dummy variable which takes the values 0/1:

$$cover{:}pb = \begin{cases} 0, if\ cover\ is\ hardcover \\ 1, if\ cover\ is\ paperback \end{cases}$$

How can we interpret β_1 *in this case*

Assume $y = \beta_0 + \beta_1 x + \epsilon_i,$ *where* $y\ is\ weight, x\ is\ cover{:}pb$

Then

$$\mathrm{y_i} = \begin{cases} \beta_0 + \epsilon_i, & if\ the\ cover\ is\ hardcover \\ \beta_0 + \beta_1 + \epsilon_i, & if\ the\ cover\ is\ paperback \end{cases}$$

Paperback books are on average β_1 *units heavier than hardcover books*

Do we expect β_1 *to be positive or negative?*

Qualitative predictors

When a predictor takes three (categorical) values (e.g., religion: Muslim/Christian/Atheist) we create two dummy variables

$$rel{:}muslim = \begin{cases} 0, if\ subject\ is\ NOT\ a\ Muslim \\ 1, if\ subject\ is\ a\ Muslim \end{cases}$$

$$rel{:}christian = \begin{cases} 0, if\ subject\ is\ NOT\ a\ Christian \\ 1, if\ subject\ is\ a\ Christian \end{cases}$$

Why not three?

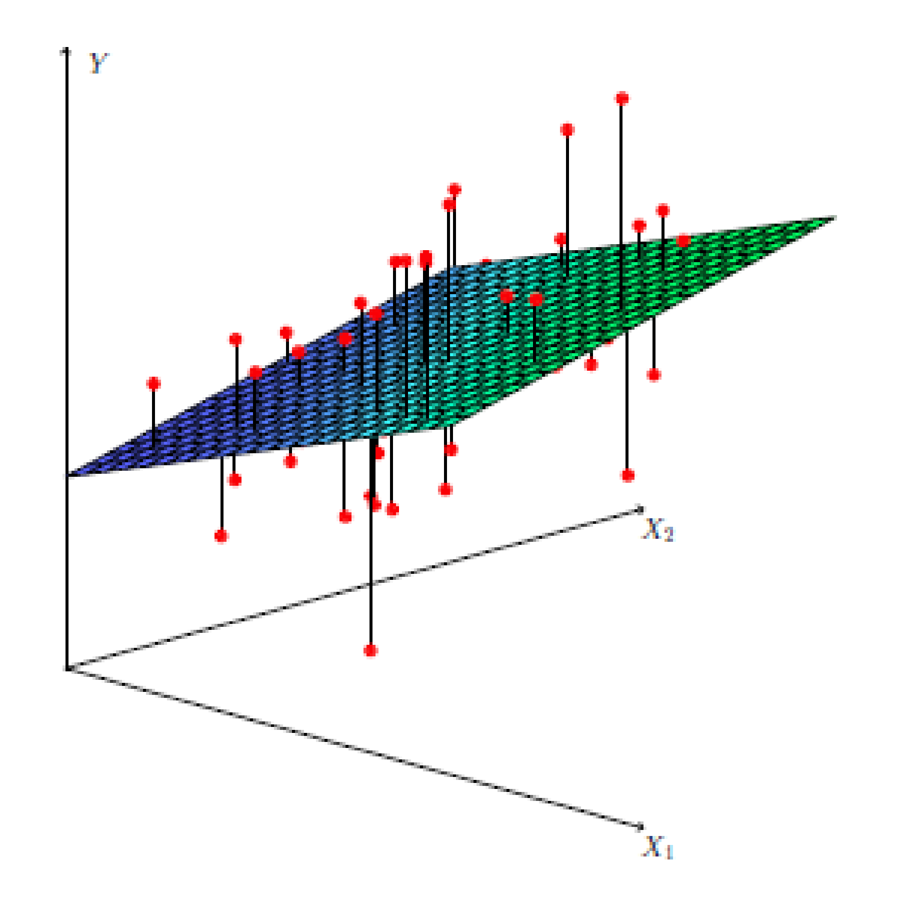
Modeling weights of books using volume and cover type

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



Still a least squares solution.

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 variance
3. residuals are independent
4. each variable is linearly related to the outcome

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

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

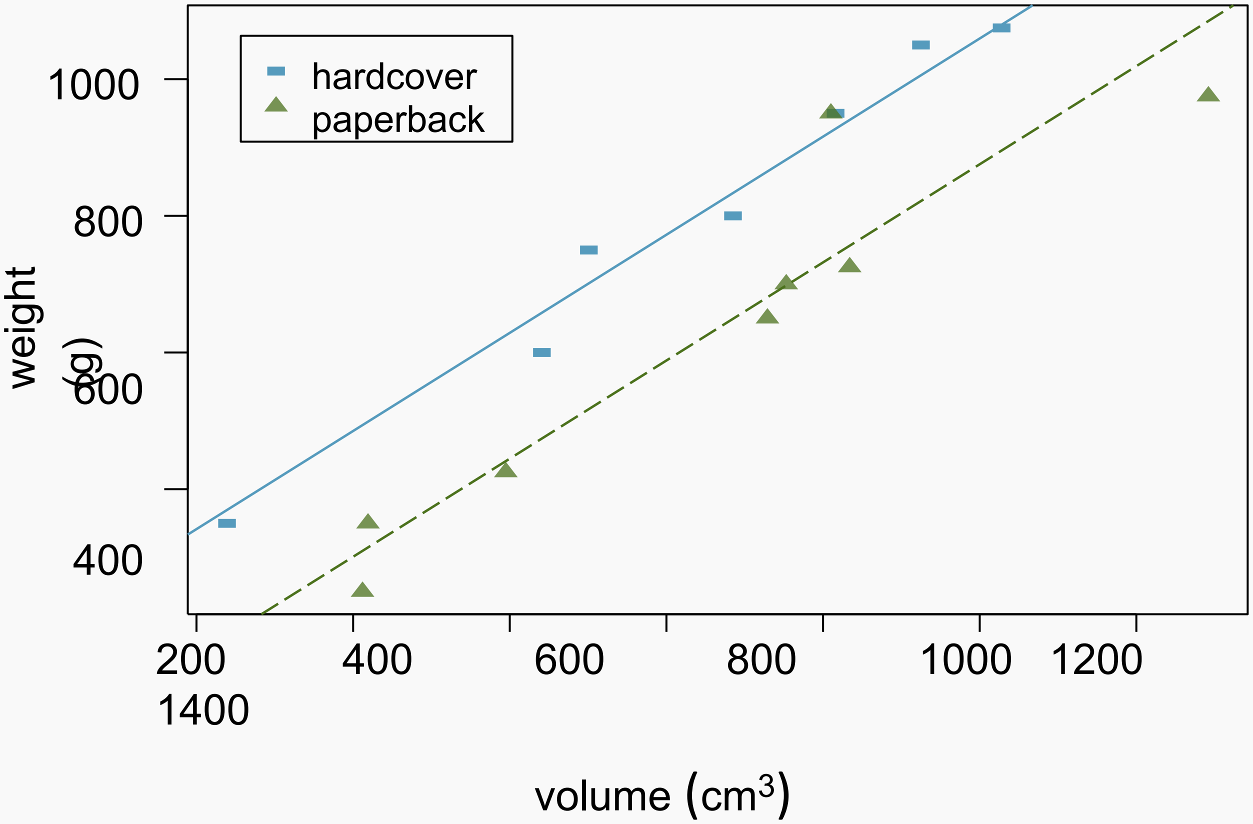
1. For hardcover books: plug in **0** for cover

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

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



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

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*

(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

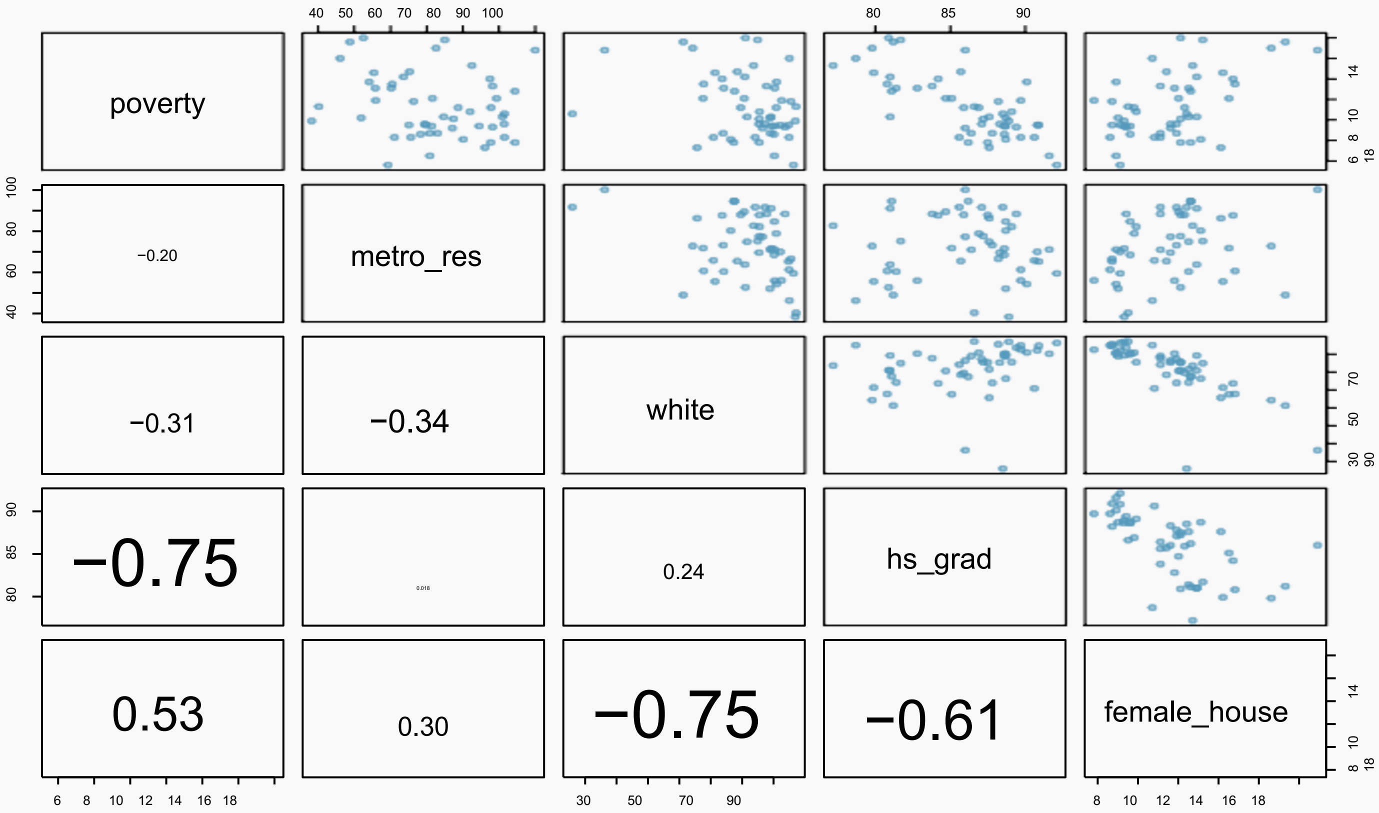
Another Example: Predicting Poverty

Response variable: Percentage of residents living in poverty

Explanatory variables:

- Percentage of residents living in a metropolitan area
- Percentage of residents that are white
- Percentage of residents that are high-school graduates
- Percentage of residents that live in a single-parent, female-led household

Revisit: Modeling poverty



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

$$R^2 = 1 - \frac{SS_{REG}}{SS_{TOT}}$$

Sum of squares

Sum of squares of y: $SS_{TOT} = \sum_i (y - \overline{y})^2 = 480.25$

Sum of squares of residuals : $SS_{REG} = \sum_i (y - \hat{y})^2 = 347.68$

$R^2 = 1 - \frac{SS_{REG}}{SS_{TOT}}$ = 0.29

R squared

- *For single-predictor linear regression, this is the square of the calculation coefficient.*
- *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.*

Adjusted R^2

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

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

But it may be the case that the variable is not really informative
(at all or in the context of the other variables)

This is called overfitting and can lead to falsely large R^2

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.

R^2 vs. adjusted R^2

Adjusted R^2

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

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

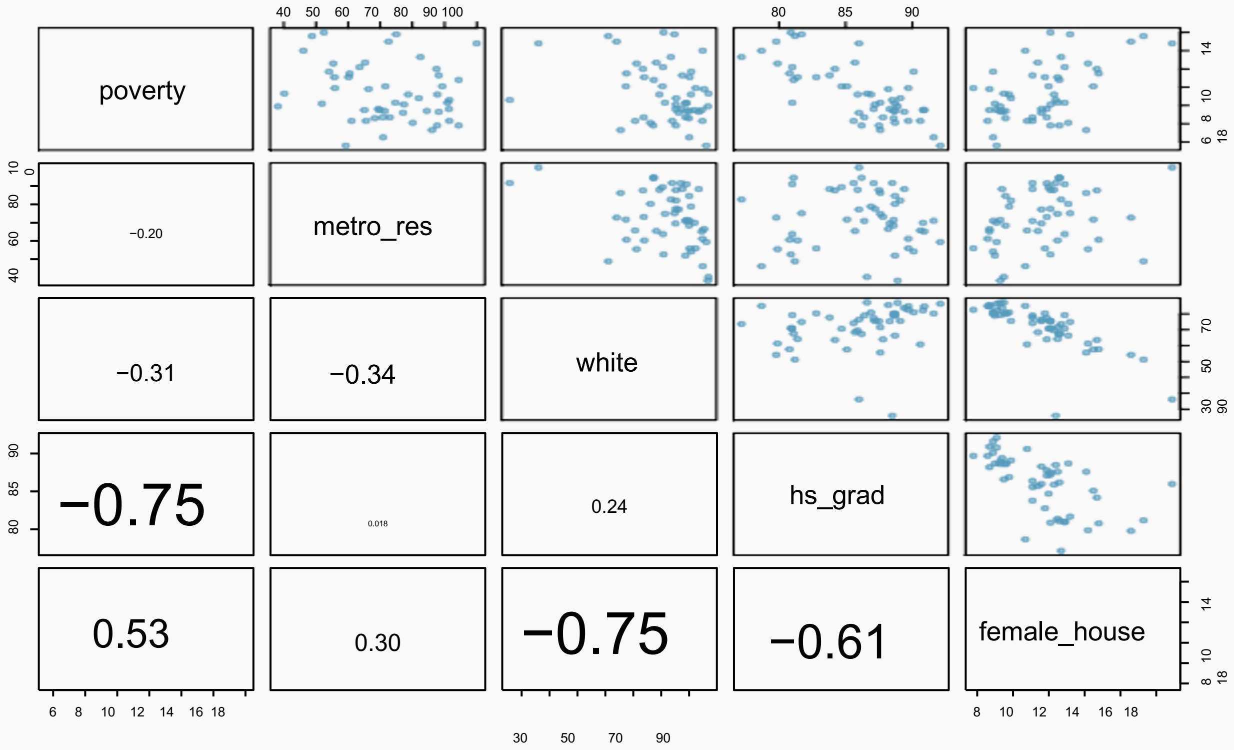
$$SS_{REG} = 339.47$$
$$SS_{TOT} = 480.25$$
$$n = 51$$
$$p = 2$$

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

$$= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1}\right)$$
$$= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48}\right)$$
$$= 1 - 0.74$$
$$= 0.26$$

Collinearity

Predicting poverty in US states: pair plot



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

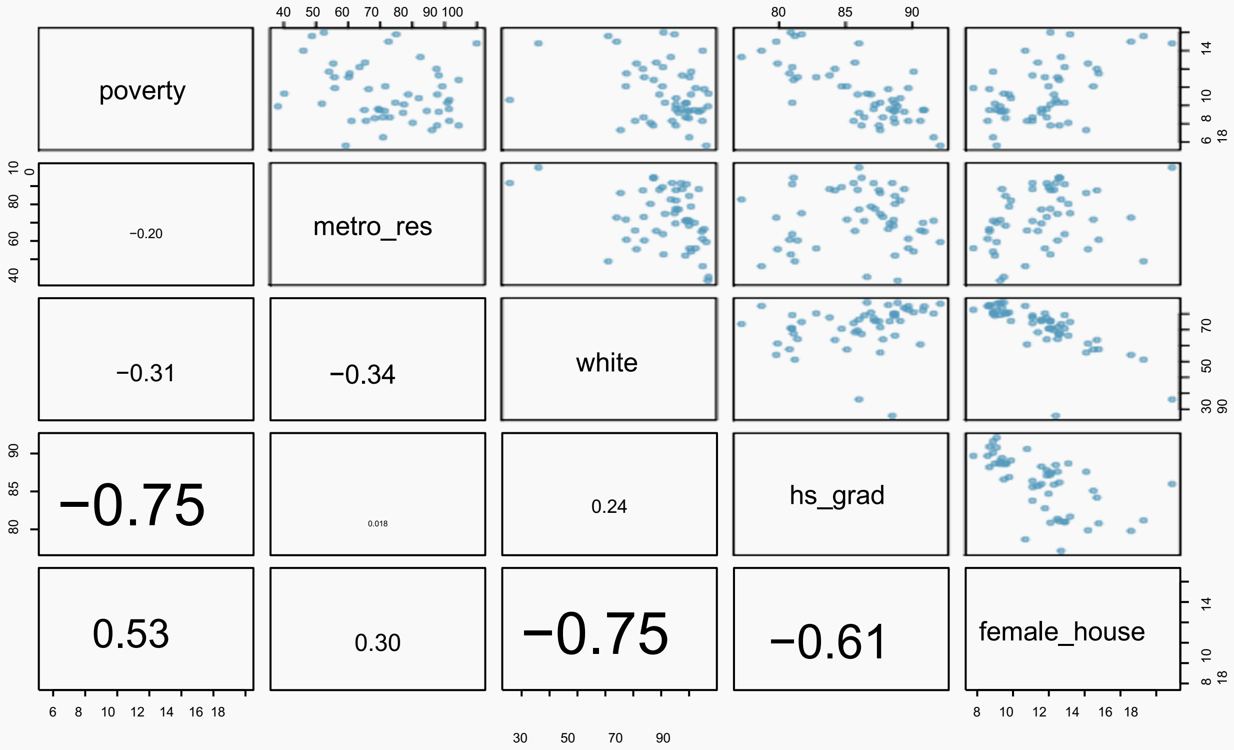
$$R^2 = 1 - \frac{SS_{REG}}{SS_{TOT}} = 0.29$$

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

$$R^2 = 1 - \frac{SS_{REG}}{SS_{TOT}} = 0.29$$

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



Collinearity between explanatory variables

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 and may lead to overfitting. Instead, we prefer the simplest best model.

When we have multiple variables, we want to select a subset of these variables to include in the model.

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.

Beauty in the classroom

beauty
gender.male
age
formal.yes [1]
lower.yes [2]
native.non english
minority.yes
students [3]
tenure.tenure track [4]
tenure.tenured

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

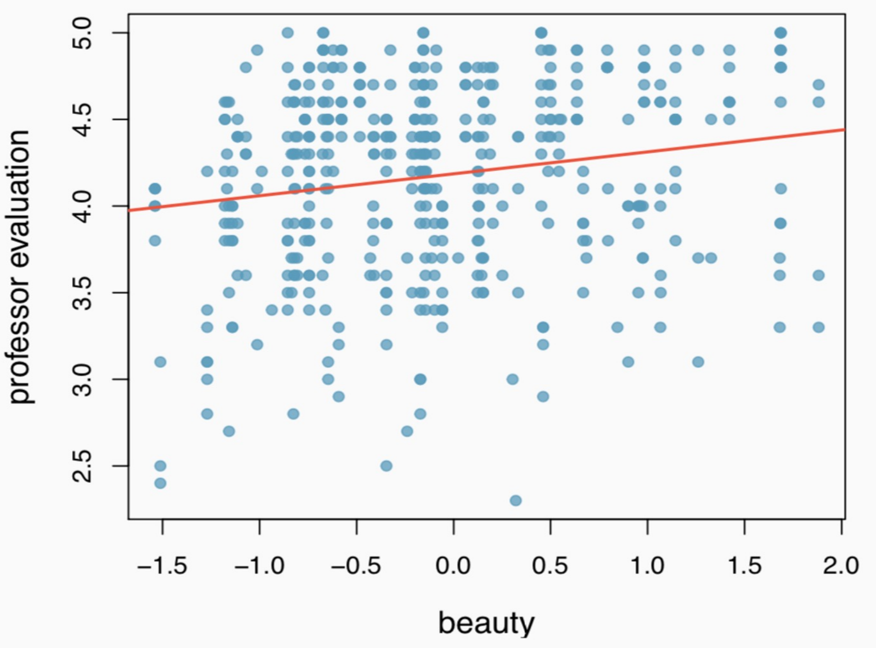
[2] `lower`: lower division course, levels: `yes`, `no`

[3] `students`: number of students

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

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.*

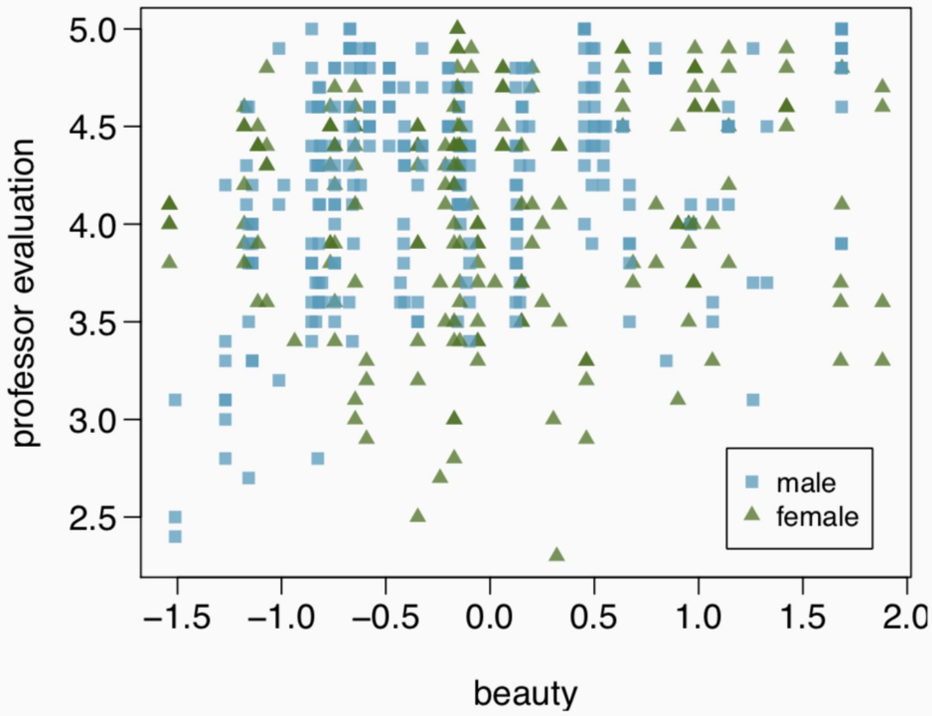
(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(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 evaluated higher, lower, or about the same as female professors?



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.

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

a. Since p-value is positive, higher the professor’s age, the higher we would expect them to be rated.
b. If we keep all other variables in the model, 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 keep all other variables in the model, 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

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*
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
Step 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806
	beauty + gender + age + formal + tenure	0.0773
	beauty + gender + age + formal + native	0.0713

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 remanning variables does not result in a higher R^2_{adj}

Backward-Elimination vs. Forward-Selection

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

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 . This is commonly the case in machine learning applications.
- When we care about understanding which variables are statistically significant predictors of the response, or if there is interest in producing a simpler model at the potential cost of a little prediction accuracy, then the p-value approach is preferred.
- Regardless of the approach we use, our job is not done after variable selection – we must still verify the model conditions are reasonable.

Checking model conditions using graphs

Modeling conditions

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

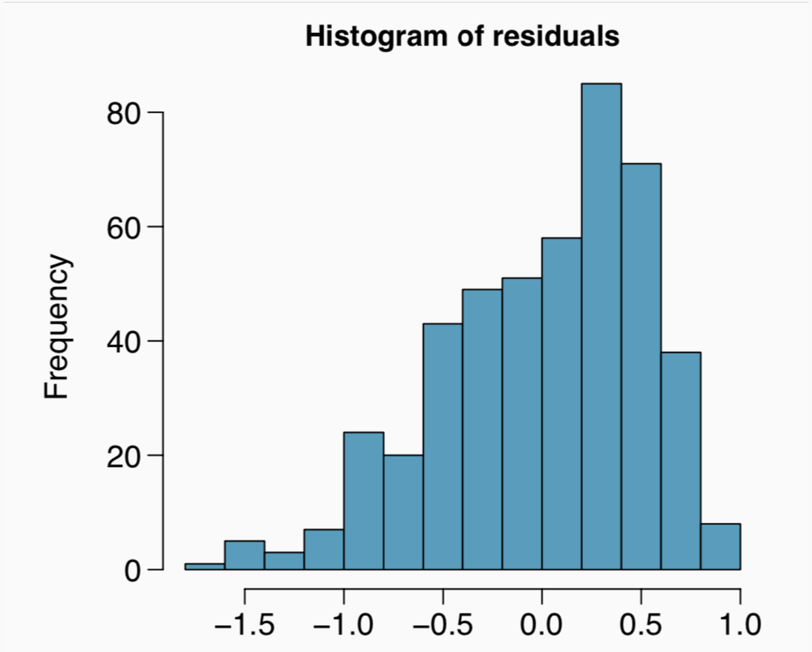
The model depends on the following conditions

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, which we will go through in detail in the following slides.

(1) nearly normal residuals

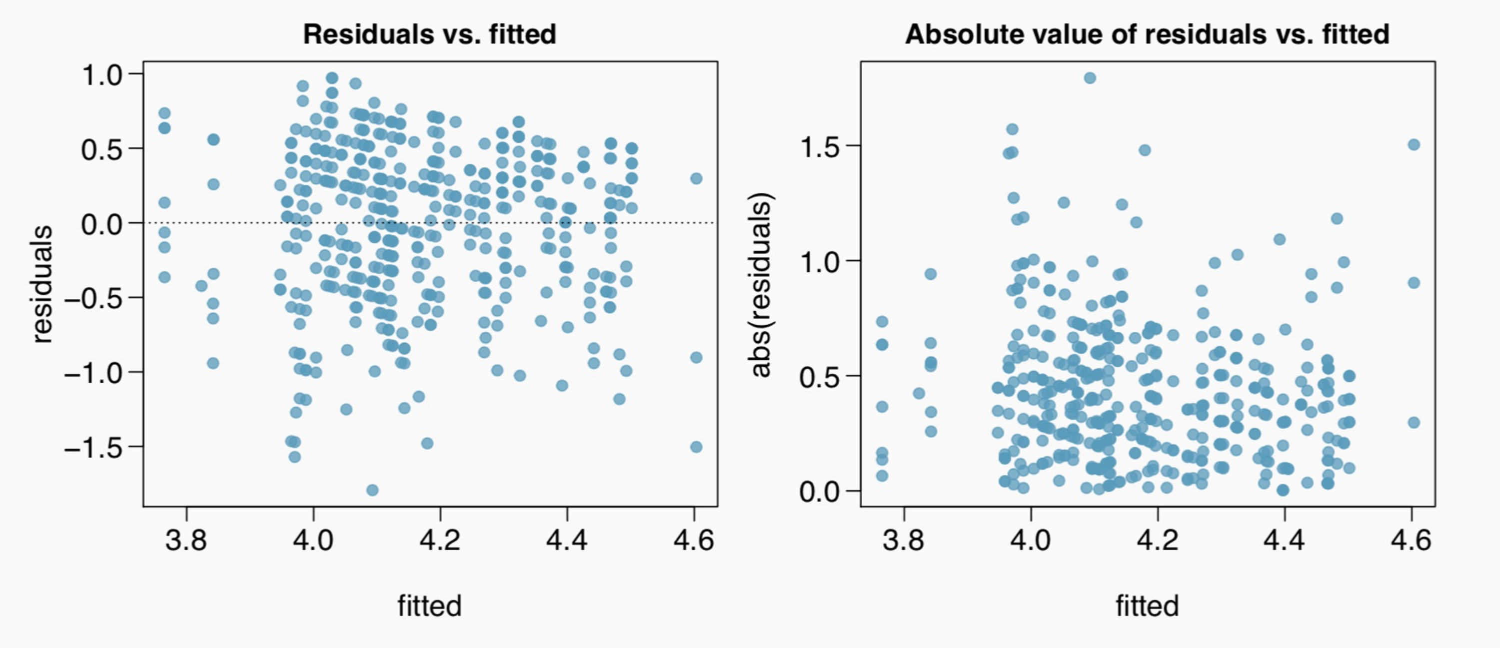
Histogram of the residuals.



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*.

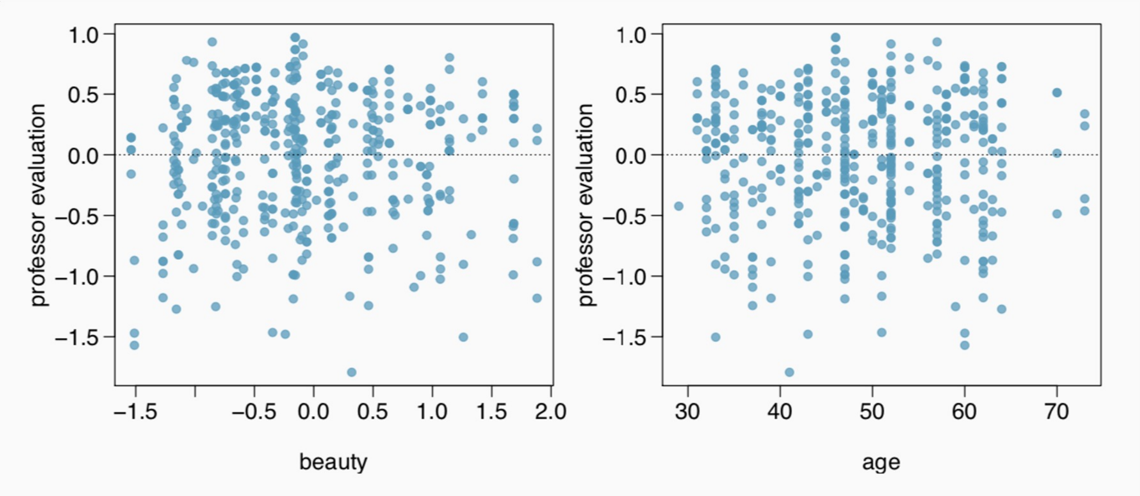
With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

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.

(4) linear relationships

Scatterplot of residuals vs. each (numerical) explanatory variable.



Does this condition appear to be satisfied?

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)

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.