Unit 5: The General Linear Model

2. Model Selection

(Chapter 6.2)

4/4/2022

Recap from last time

1. In multiple regression, every variable is conditional on every other variable
2. For inference, we care about both the whole model and the individual variables
3. We use adjusted R^2 to account to penalize additional variables

Key ideas

1. For many real-world problems, a lot of variables contribute a little bit
2. Stepwise approaches try to correct for this, but there is no “one true way”
3. We can check assumptions for multiple regression using plots

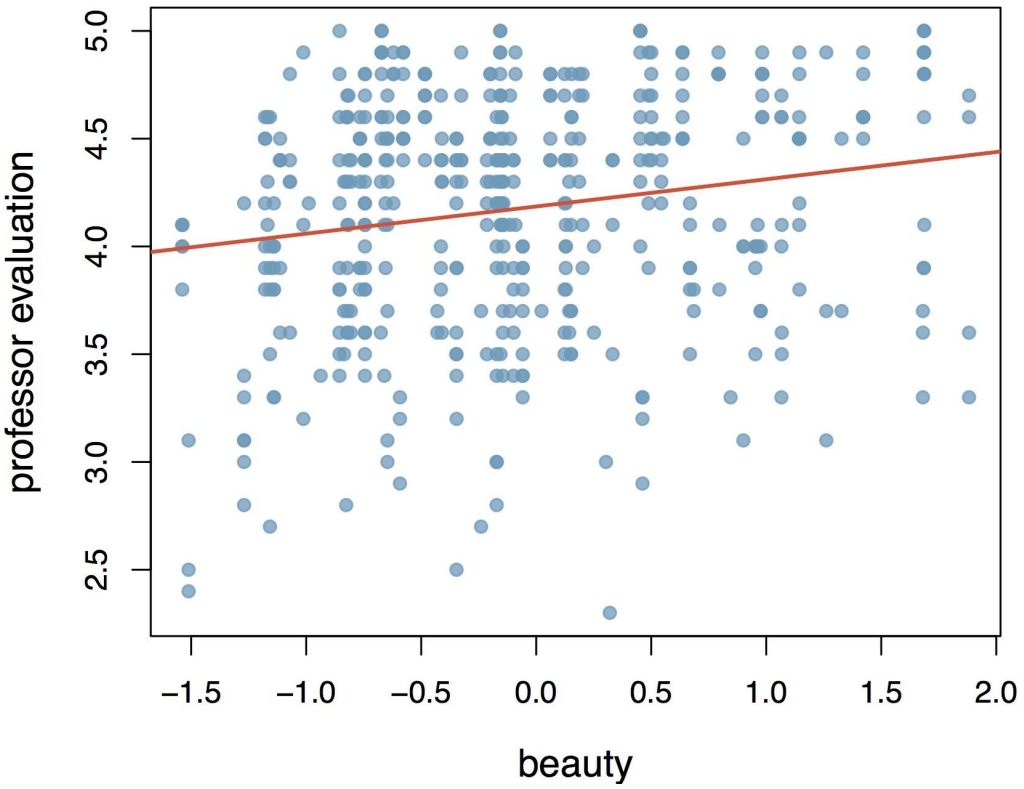
How we judge our professors

Data: Student evaluations of instructors' beauty and teaching quality for 463 courses at the University of Texas.

Evaluations conducted at the end of semester. Also judgments of the professors "beauty" made later, by six students who had not attended the classes and were not aware of the course evaluations

Hamermesh & Parker (2004). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics Education Review*

Predicting professor evaluation



Practice Question 1: Understanding regression models

	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

$R^2 = 0.0336$

Which of these correctly describes the model output?

(a) The 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 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.

Practice Question 1: Understanding regression models

	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

$R^2 = 0.0336$

Which of these correctly describes the model output?

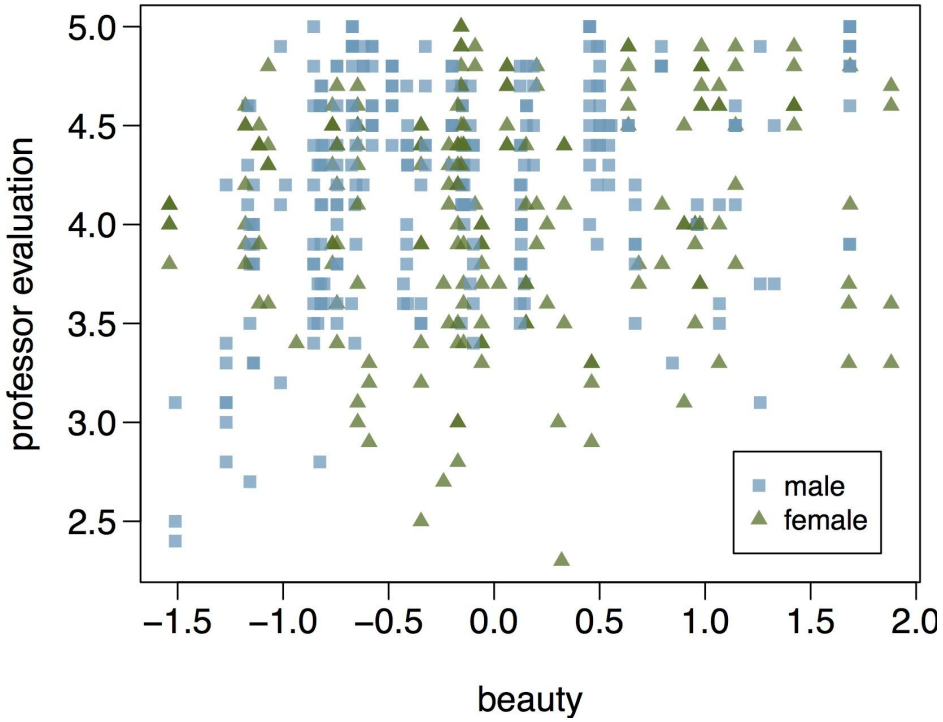
(a) The model predicts 3.36% of professor ratings correctly

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

(c) Professors who score 1 point above average in their beauty score 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.

Adding a second variable



For a given beauty score, are males rated higher or lower?

	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

For a given beauty score, are males rated higher or lower?

	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! Holding beauty constant, men are rated .17 points higher

(b) lower

(c) about the same

Let's look at the 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`

Testing 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 predictors are included in the model

H_A: $\beta_i \neq 0$ when other predictors are included in the model

Practice Question 2: Assessing significance of numerical variables

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

Which of these correctly describes the model output?

(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.

Practice Question 2: Assessing significance of numerical variables

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

Which of these correctly describes the model output?

(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.

Practice Question 3: Assessing significance of categorical variables

	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

Which of these descriptions of the model output is <u>false</u>?

(a) The reference level is non tenure track.

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

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

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

Practice Question 3: Assessing significance of categorical variables

	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

Which of these descriptions of the model output is <u>false</u>?

(a) The reference level is non tenure track.

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

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

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

Which explanatory variables do not look like reliable predictors?

	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

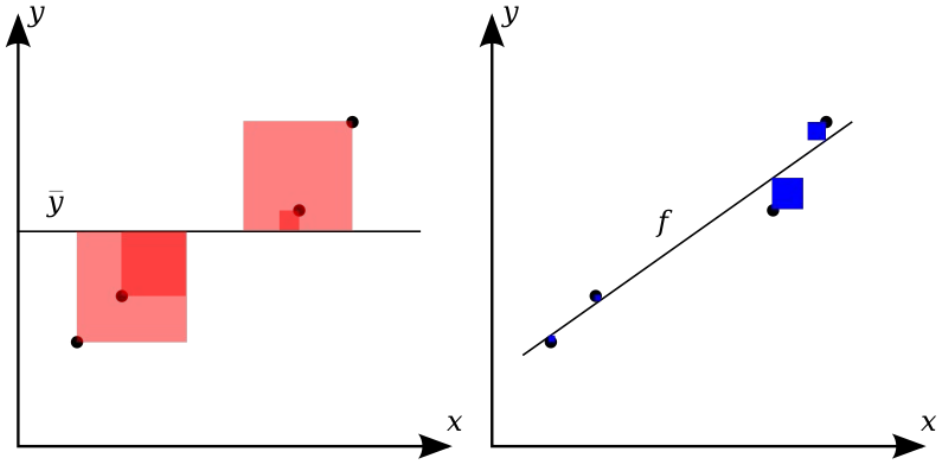
Two approaches to model selection

1. **Forward-selection with some criterion (e.g. R^2 *adjusted)*:**
   a. Start with regressions of response vs. each explanatory variable
   b. Pick the model with the highest R^2_{adj}
   c. Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
   d. Repeat until no remaining variables increase R^2_{adj}

2. **Backward-selection with some criterion (e.g. R^2 *adjusted)*:**
   a. Start with the full model
   b. Drop one variable at a time and record R^2_{adj} of each smaller model
   c. Pick the model with the highest increase in R^2_{adj}
   d. Repeat until none of the models yield an increase in R^2_{adj}

Model selection criteria: adjusted R^2

$$R^2 = 1 - \frac{SS_{resid}}{SS_{total}}$$



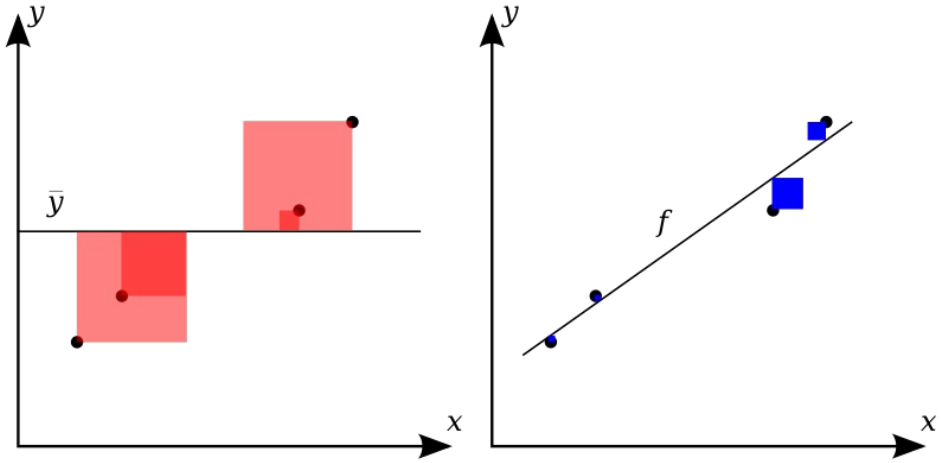
$$R^2_{adj} = 1 - \frac{SS_{resid}/(n-1)}{SS_{total}/(n-p-1)}$$

n data points

p parameters

Model selection criteria: Akaike Information Criterion (AIC)

$$R^2 = 1 - \frac{SS_{resid}}{SS_{total}}$$



$$AIC = 2p - 2\log_e(\hat{L})$$

$$= 2p + n\log_e(SS_{resid})$$

n data points

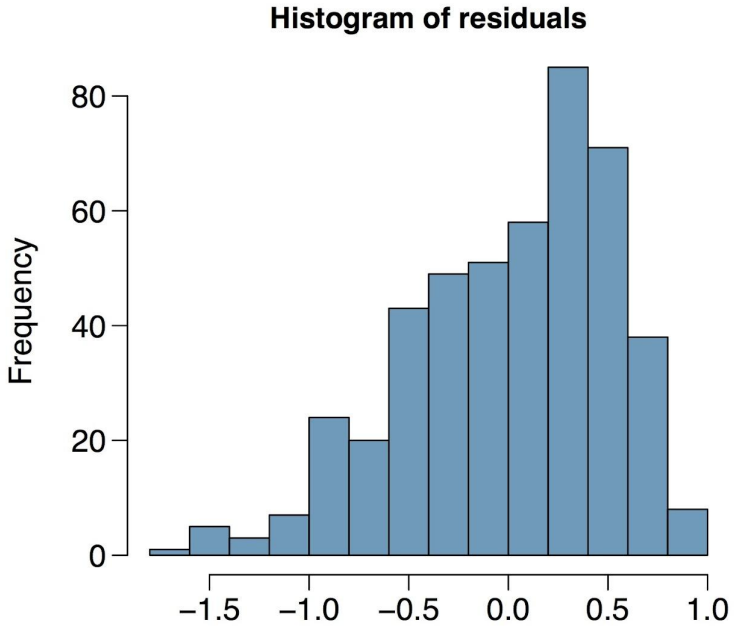
p parameters

Conditions for using multiple regression

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

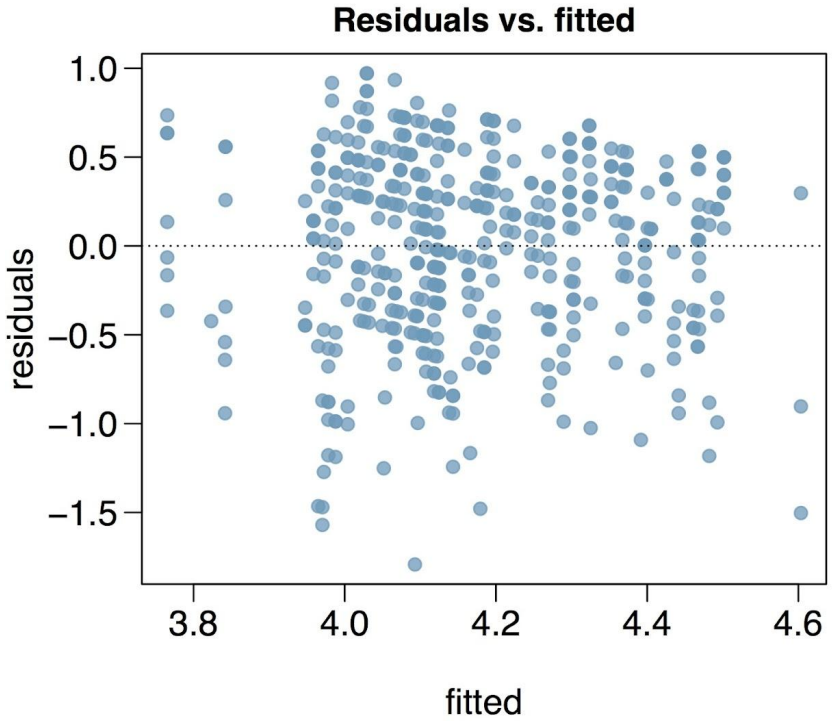
1. Residuals are nearly normal (primary concern is outliers)
2. Residuals have constant variability
3. Each variable is linearly related to the outcome

Nearly normal residuals?



Does the normal residuals condition appear to be satisfied?

Constant variability?



Does constant variability appear to be satisfied?

Checking constant variance recap

Simple linear regression: We check constant variability by plotting residuals vs. x

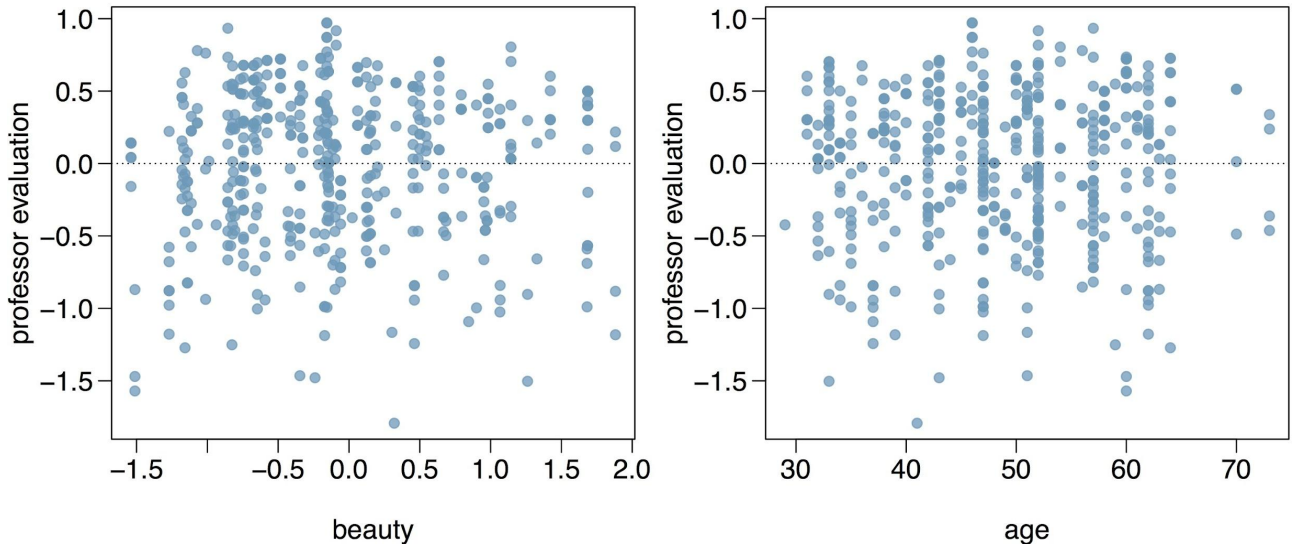
Multiple linear regression: We check constant variability by plotting residuals vs. fitted value.

Why are we using different plots?

In Multiple linear regression, there are multiple explanatory variables, so a plot of residuals vs. one of them wouldn't tell us about the whole model

Linearity?

Residuals vs. each (numerical) explanatory variable



Does it look like these predictors and the evaluation are linearly related?

(Note: One virtue of using residuals instead of the predictors on the y-axis:
We can still check for linearity without worrying about other possible violations like collinearity between the predictors.)

Key ideas

1. For many real-world problems, a lot of variables contribute a little bit
2. Stepwise approaches try to correct for this, but there is no “one true way”
3. We can check assumptions for multiple regression using plots