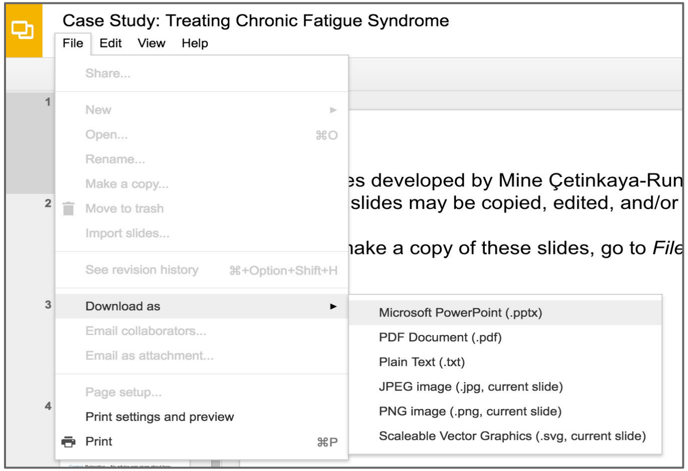
Slides developed by Mine Çetinkaya-Rundel of OpenIntro
Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.

To make a copy of these slides, go to *File > Download as > [option]*, as shown below. Or if you are logged into a Google account, you can choose *Make a copy...* to create your own version in Google Drive.



Model selection

Beauty in the classroom

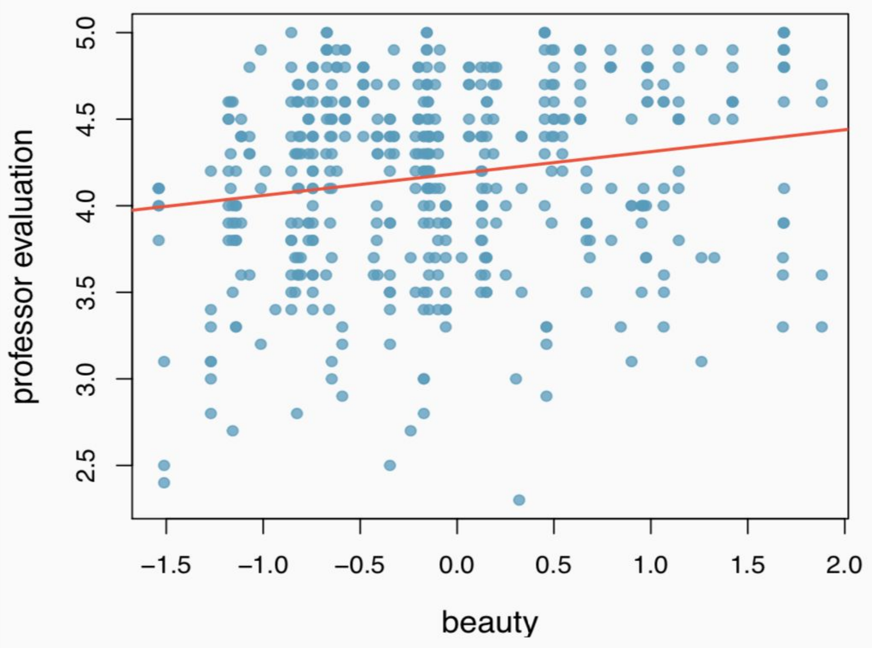
- Data: Student evaluations of instructors’ beauty and teaching quality for 463 courses at the University of Texas.
- Evaluations conducted at the end of semester, and the beauty judgements were made later, by six students who had not attended the classes and were not aware of the course evaluations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

Hamermesh & Parker. (2004)“Beauty in the classroom: instructorsÕ pulchritude and putative pedagogical productivityÓ

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 <u>correct</u> 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.*

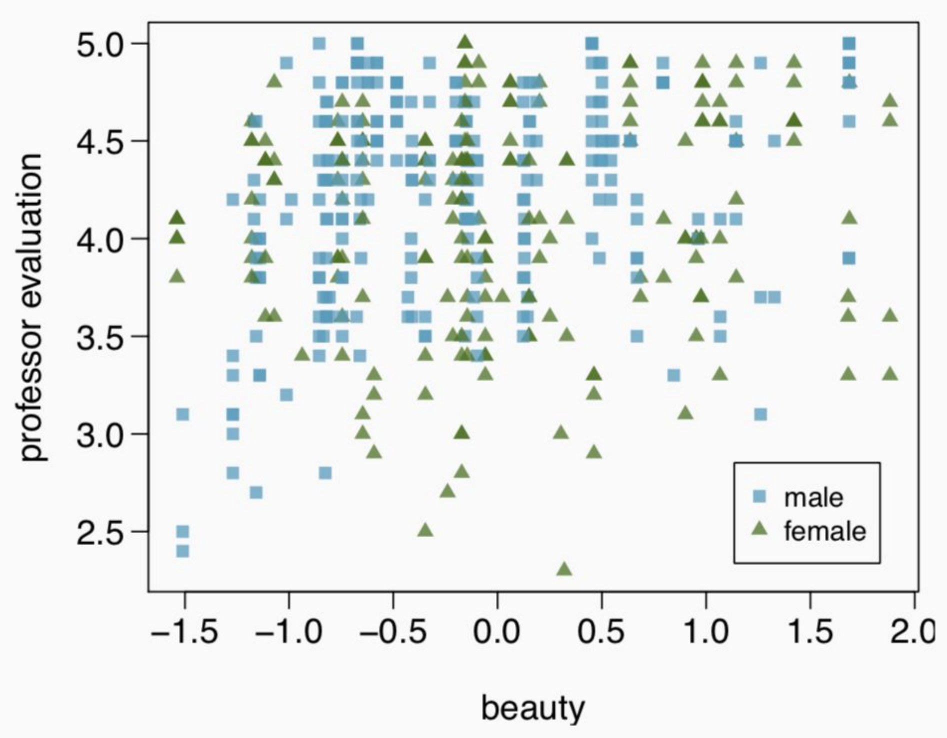
(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(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?

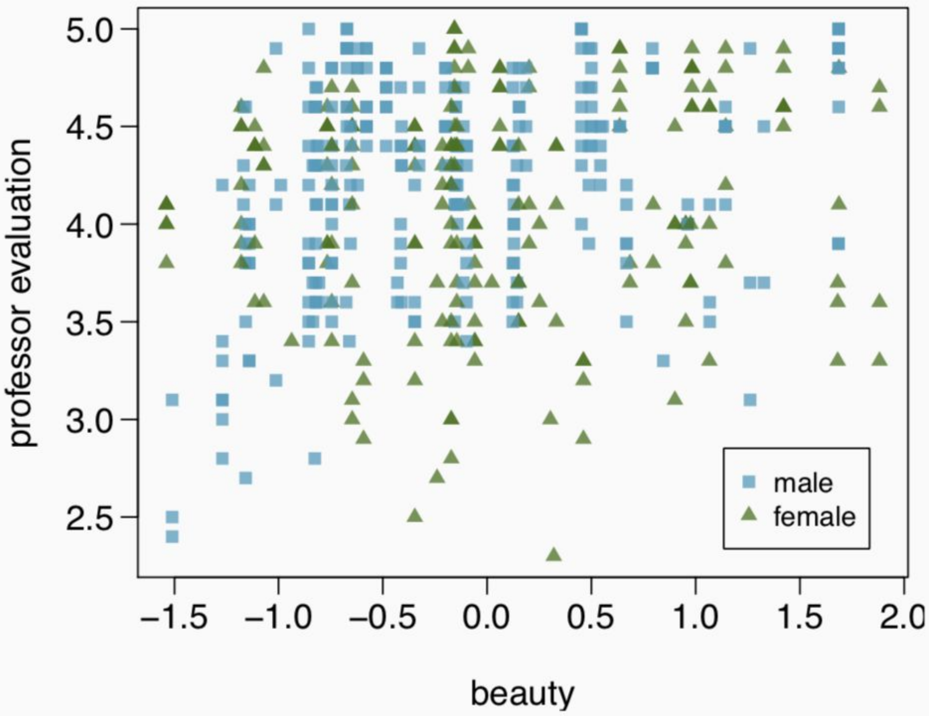


Exploratory analysis

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?



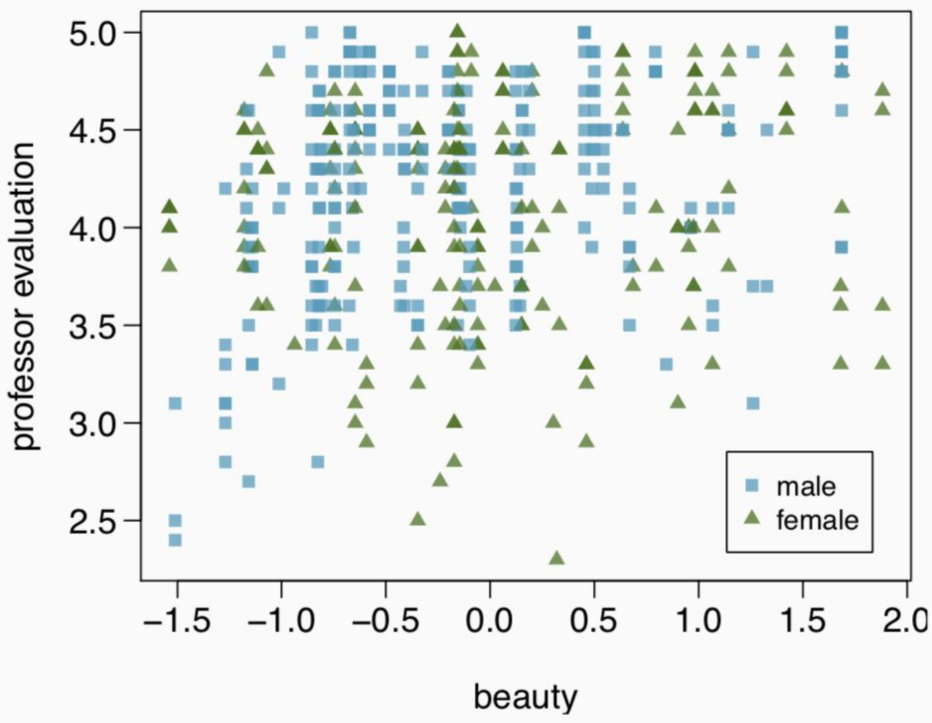
Exploratory analysis

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?

Few females with very low beauty scores.



Professor rating vs. beauty + gender

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

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

(a) higher

(b) lower

(c) about the same

Professor rating vs. beauty + gender

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

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

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

(b) lower

(c) about the same

Full Model

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

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

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

[3] `students`: number of students

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

Hypotheses

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

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

Assessing significance: numerical variables

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

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

a. Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.
b. If we 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: categorical variables

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

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

a. Reference level is non tenure track.
b. All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.
c. All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.
d. All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.

Assessing significance: categorical variables

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

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

a. Reference level is non tenure track.
b. All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.
c. All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.
d. *All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.*

Assessing significance

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

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

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

`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  nativenon english
        -0.008844           0.132422           -0.243003
tenuretenure track    tenuretenured
        -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 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

Backward elimination 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 R2 approach tending to include more predictors in the final model.
- When the sole goal is to improve prediction accuracy, use R2 . 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.

Find more resources at openintro.org/os, including

- Slides
- Videos
- Statistical Software Labs
- Discussion Forums (free support for students and teachers)
- Learning Objectives

Teachers only content is also available for Verified Teachers, including

- Exercise solutions
- Sample exams
- Ability to request a free desk copy for a course
- Statistics Teachers email group

Questions? Contact us.