Advanced Statistics
DS2003 (BDS-4A)
Lecture 24

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

19 May, 2022

Previous Lecture

- Testing for Multicollinearity with Variance Inflation Factors (VIF)
  - What is VIF?
  - How did we calculate R^2?
- Practical Example (using Python)
  - First we calculated the various VIF values of the explanatory variables
  - Then we looked at the correlation matrix for further proof
  - We then decided to remove two variables, and then recheck the VIF values based on the new regression model with 2 fewer explanatory variables

Example of managing multicollinearity

- https://www.reneshbedre.com/blog/variance-inflation-factor.html

Code

- https://colab.research.google.com/drive/1gOL8WEbL6t5C8eHZsLtdEwB1rEwH8RdB?usp=sharing

- Excel:
  - BP (CSV): https://1drv.ms/u/s!Apc0G8okxWJ1zHS32OEeh8tVtIIE?e=aT32t4
  - BP/Qty Demanded (Excel): https://1drv.ms/x/s!Apc0G8okxWJ1zHe7dOiArhibNXFd?e=yfo9Kw

By Hand

- We may find R^2 for each explanatory variable X_i by setting X_i as the response variable in place of Y, and the remaining explanatory variables as part of a separate new regression model
  - We would have to repeat this for each X_i
  - Knowing R^2 means we can easily calculate VIF

Another example: Modeling kid's test scores

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

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

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

Interpreting the slope

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

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

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 a job in the first three years of the kid's life, have an IQ of 0 (and ???) 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 `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 `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

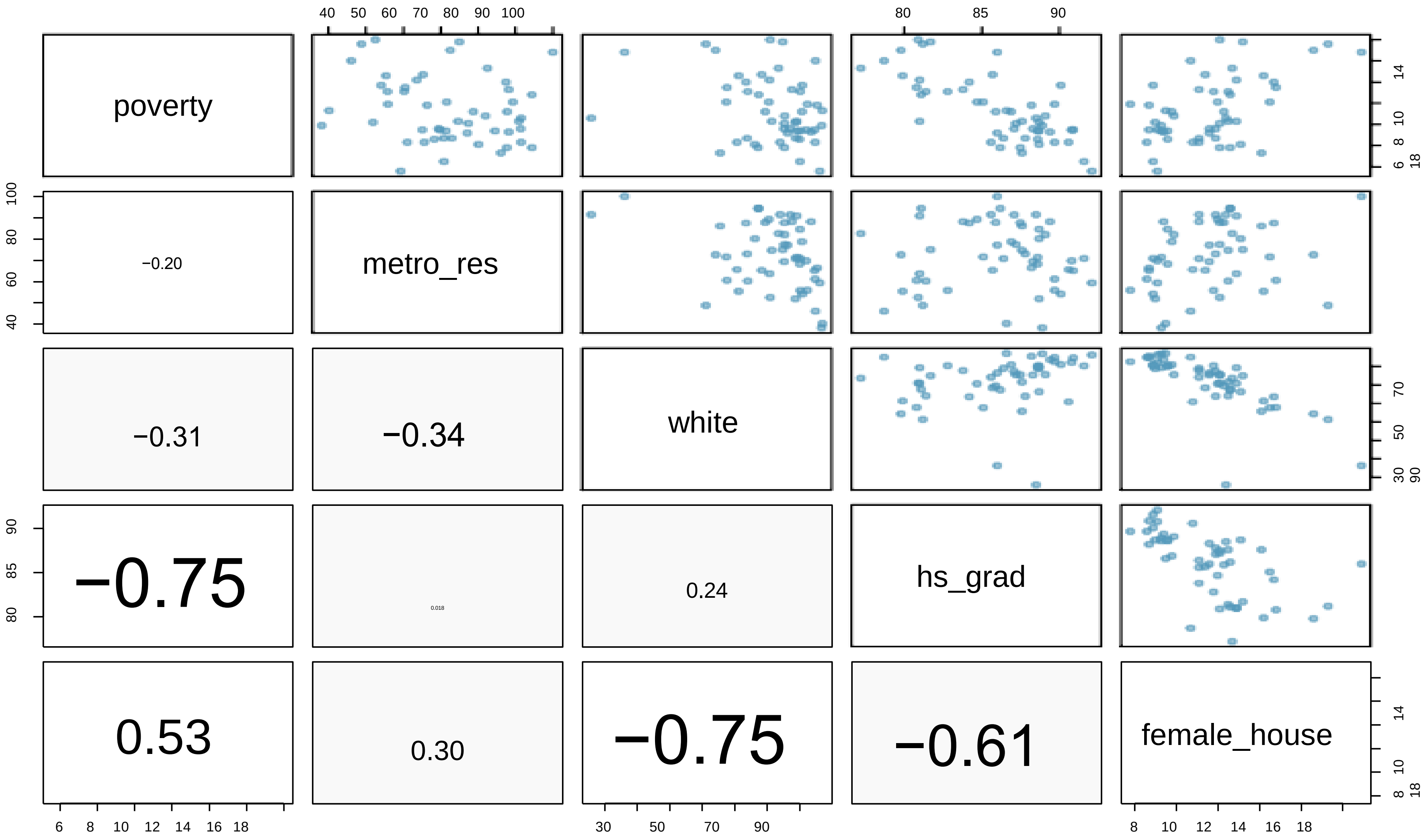
All else being equal, kids whose moms worked during the first three years of the kid's life

(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



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			

$$
\begin{aligned}
\text{Sum of squares of } y\text{: } SS_{Total} &= \sum (y-\bar{y})^2 = 480.25 \rightarrow \textit{total variability} \\
\text{Sum of squares of residuals: } SS_{Error} &= \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability} \\
\text{Sum of squares of } x\text{: } SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \textit{explained variability} \\
&= 480.25 - 347.68 = 132.57
\end{aligned}
$$

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

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 xs.*
- *And next we'll learn another measure of explained variability, adjusted R^2, that requires the use of the third approach, ratio of explained and unexplained variability.*

Predicting poverty using % female hh + % white

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

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

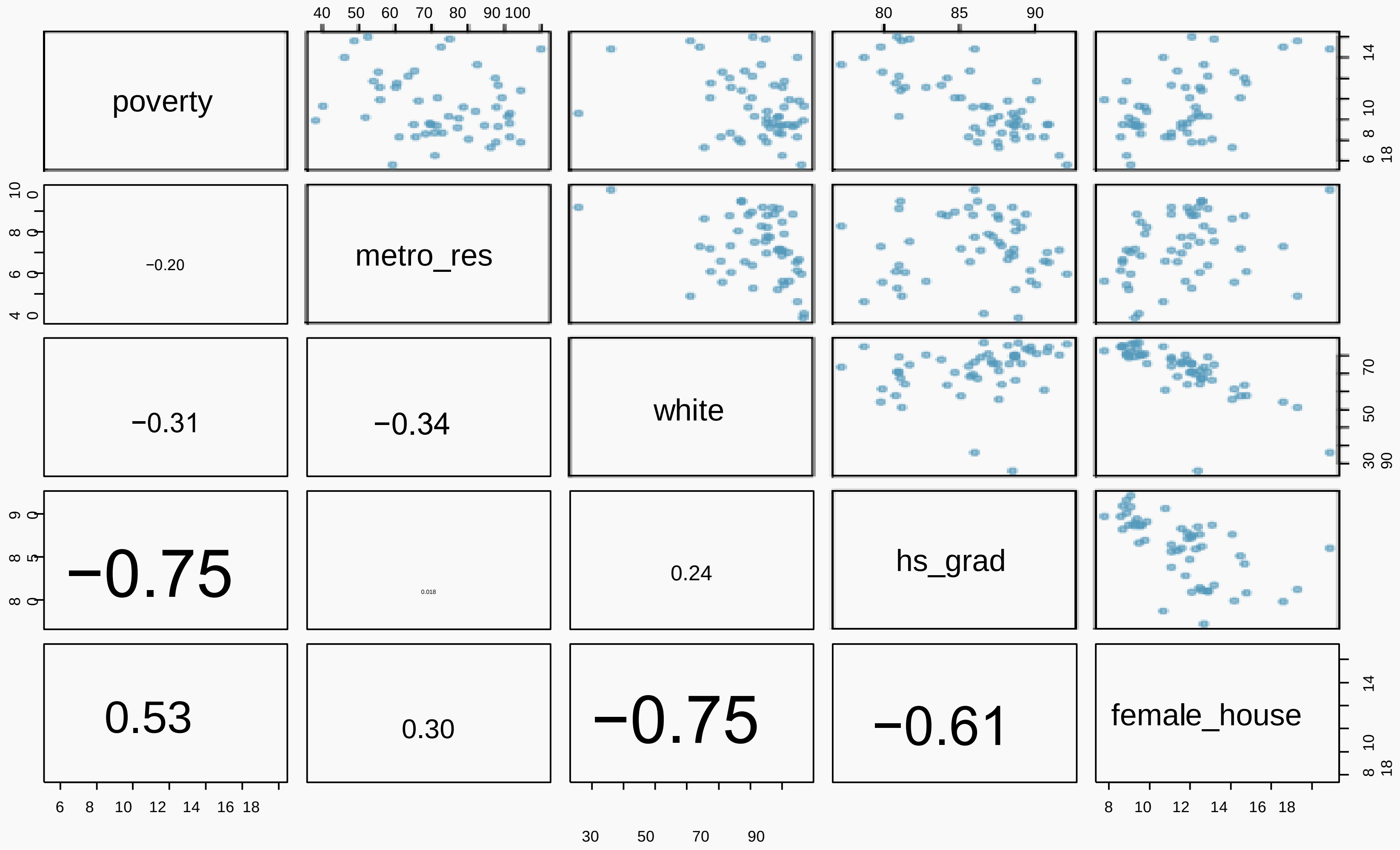
Predicting poverty using % female hh + % white

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

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

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

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



Collinearity between explanatory variables

poverty vs. % female head of household

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

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

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

Collinearity between explanatory variables (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 <u>independent</u> 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 <u>independent</u> 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

When <u>any</u> 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

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

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

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

Calculate adjusted R^2

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

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

Model selection

Beauty in the classroom

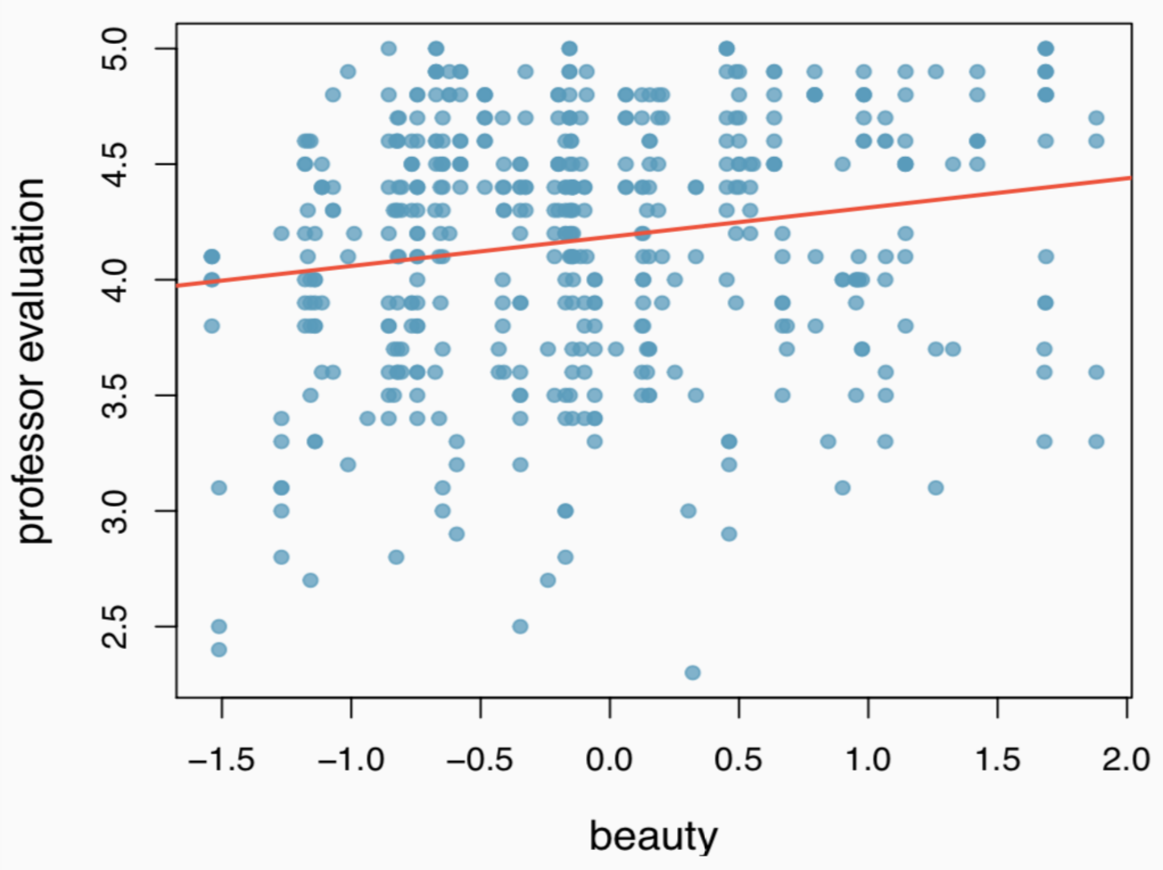
- Data: Student evaluations of instructors' beauty and teaching quality for 463 courses at the University of Texas.
- Evaluations conducted at the end of semester, and the beauty judgements were made later, by six students who had not attended the classes and were not aware of the course evaluations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

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

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.

Useful Links & Resources

- **Reference**:
  - openintro.org/os (Chapter 9, Section 9.1, 9.2)