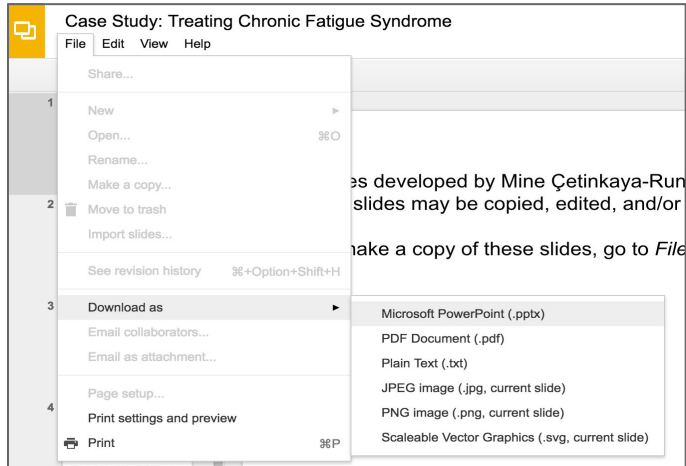


Slides developed by Mine Çetinkaya-Rundel of OpenIntro
Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.
The slides may be copied, edited, and/or shared via the [CC BY-SA license](#)

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.



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%
 - 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

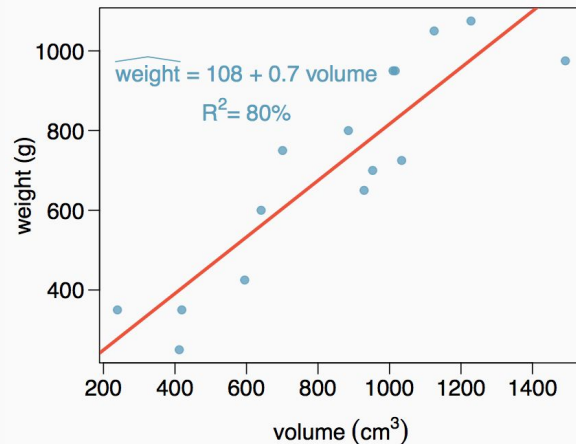
Weights of books

| | 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.)

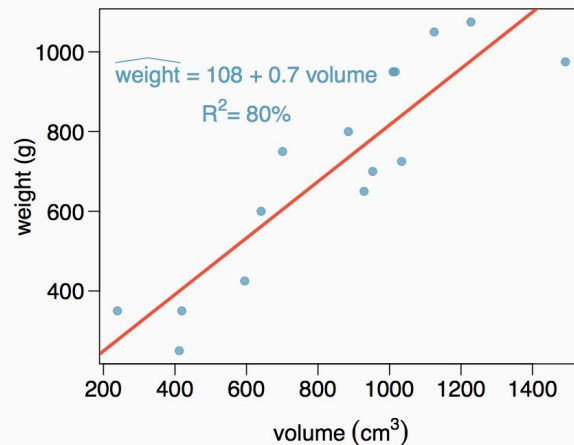
The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



- a. Weights of 80% of the books can be predicted accurately using this model.
- b. Books that are 10 cm³ over average are expected to weigh 7g 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 cm³ over average are expected to weigh 7g 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 |

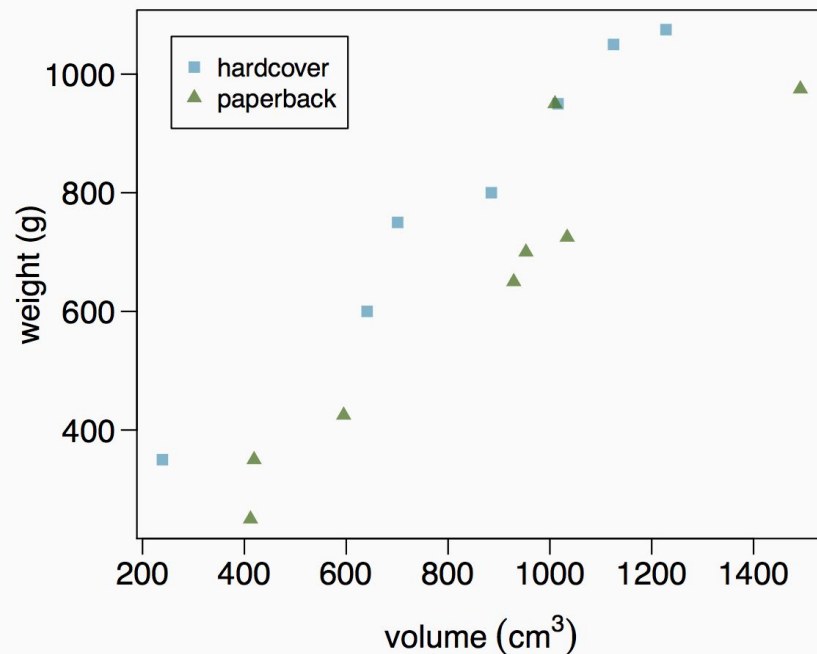
Residual standard error: 123.9 on 13 degrees of freedom

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

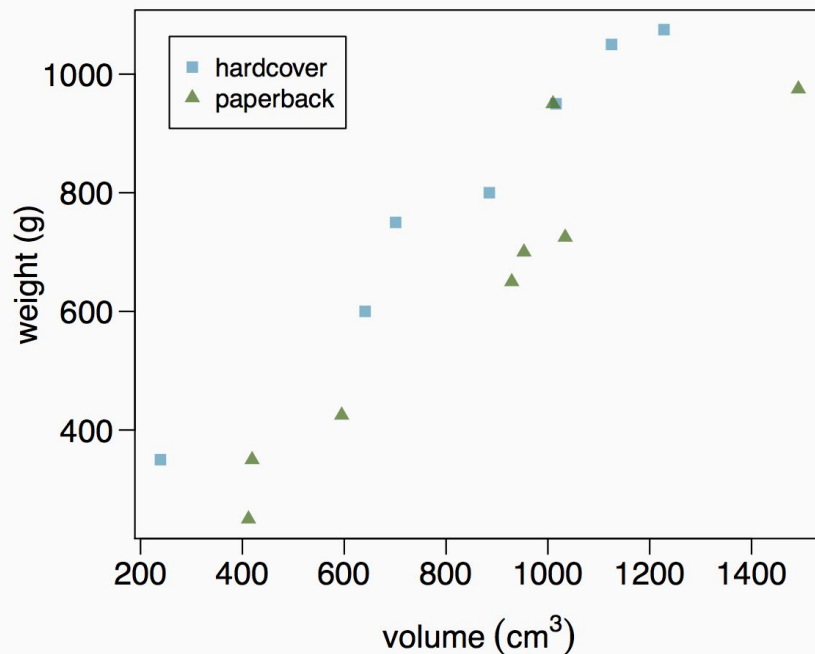
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



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

| <u>Estimate</u> | <u>Std. Error</u> | <u>t value</u> | <u>Pr(> t)</u> | |
|-----------------|-------------------|----------------|--------------------|------|
| (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 |

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

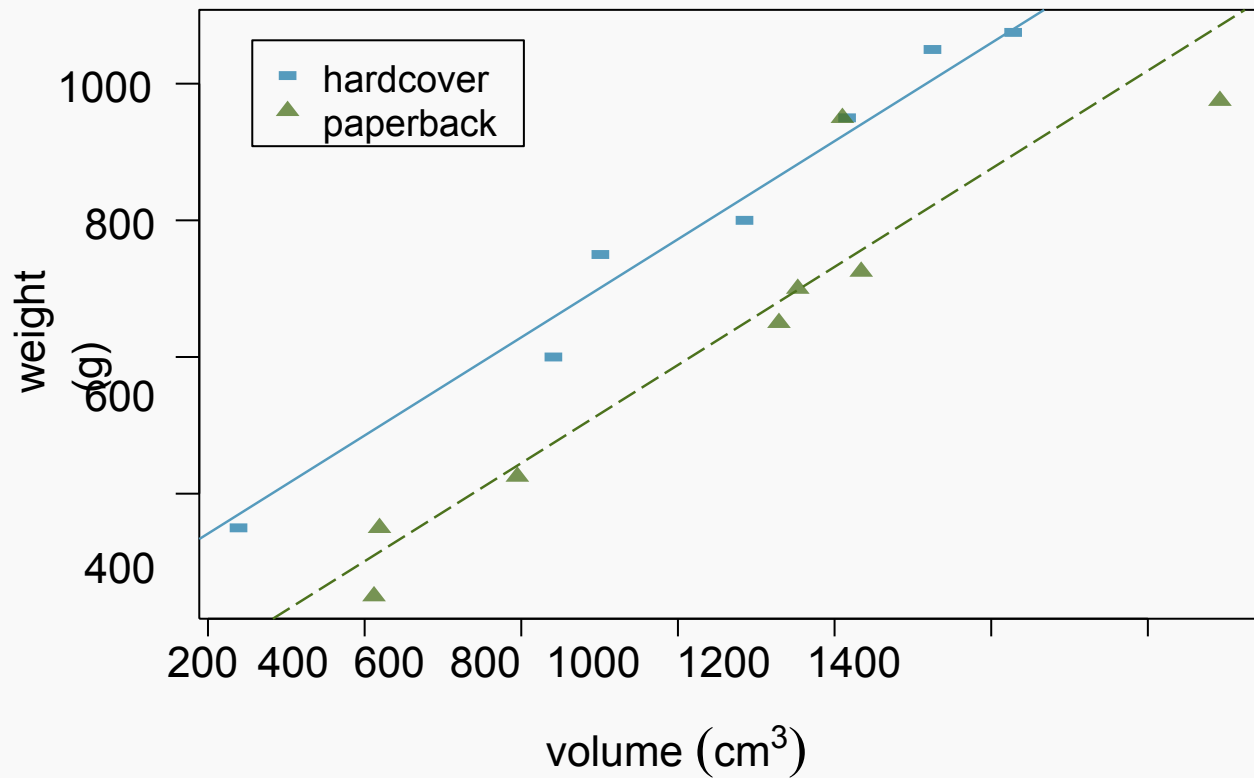
1. For **hardcover** books: plug in **0** for cover

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

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



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 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) $197.96 + 0.72 * 600 - 184.05 * 1$

(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 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 three years of the kid's life, have an IQ of 0 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 |

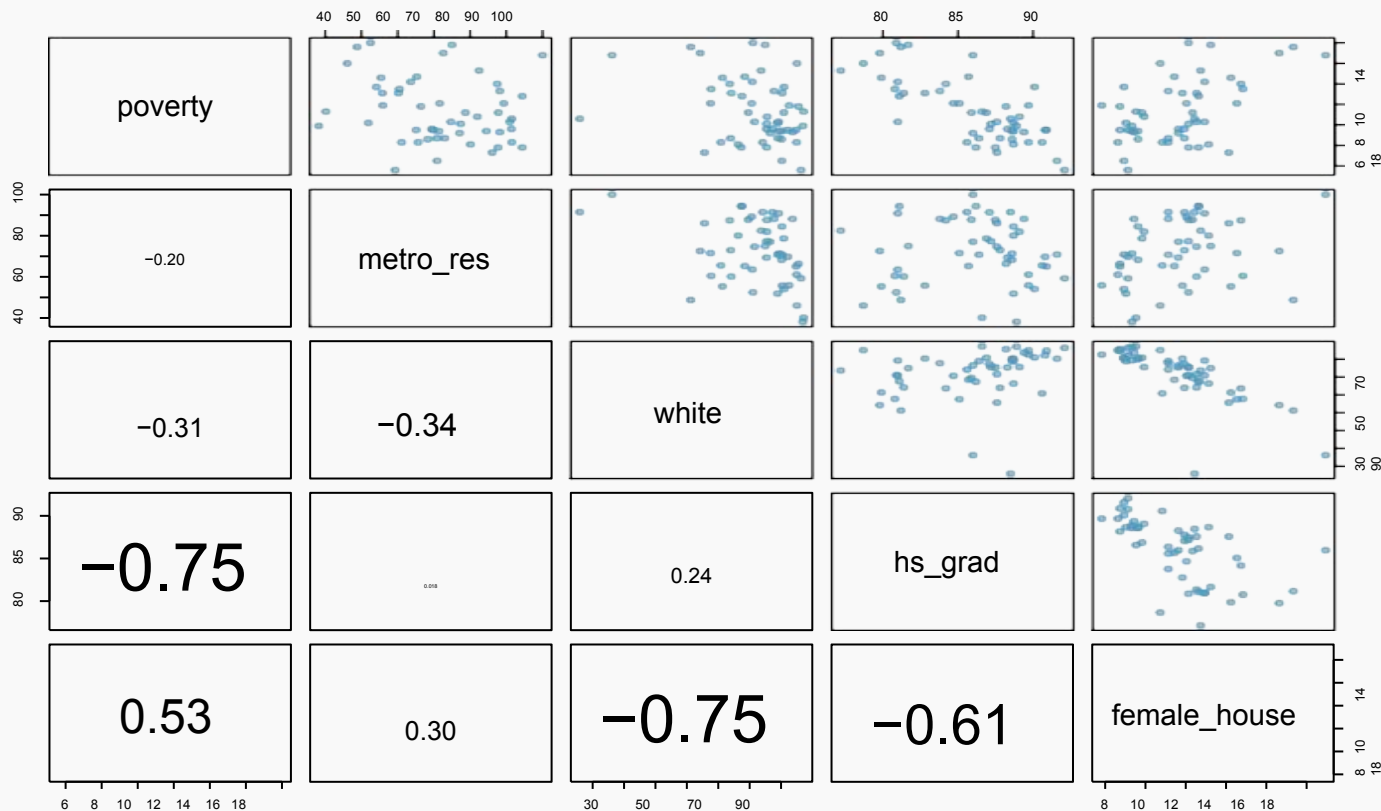
All else being equal, kids whose moms worked during the first three years of the kid's life

(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 | | | |

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 of x : $SS_{Model} = SS_{Total} - SS_{Error} \rightarrow \text{explained variability}$
 $= 480.25 - 347.68 = 132.57$

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

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 | | | |

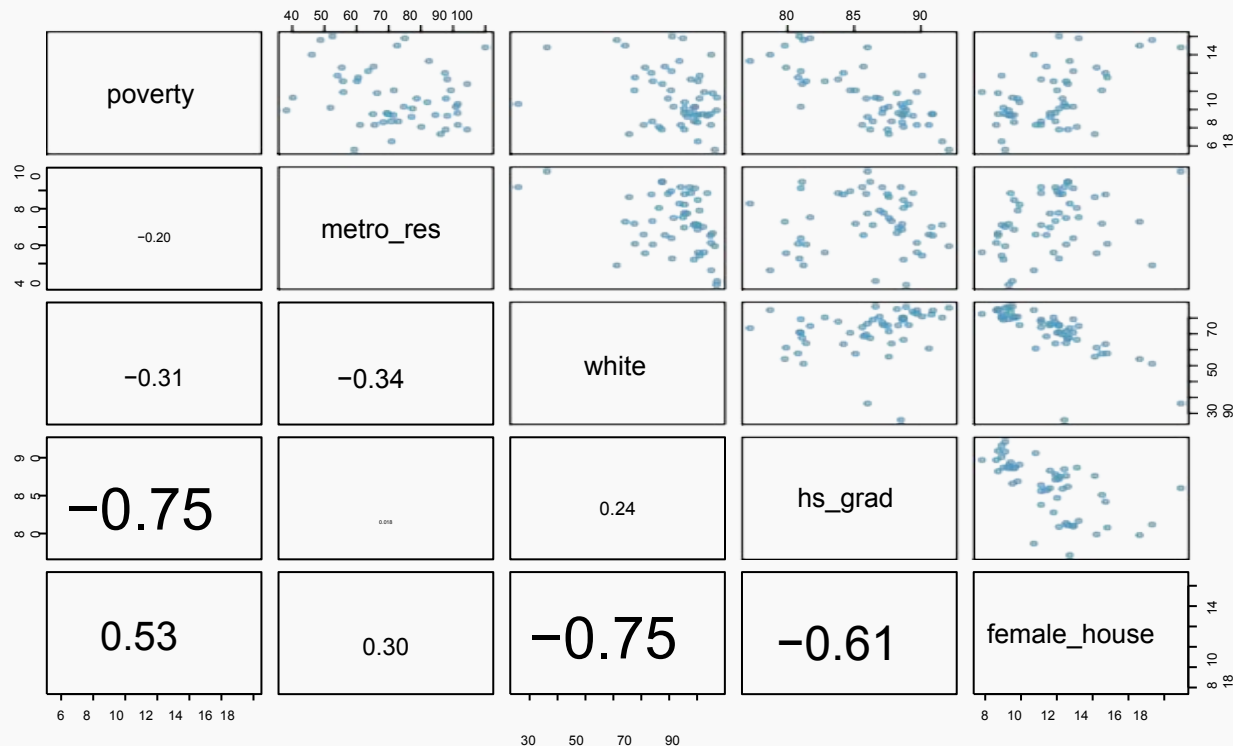
Predicting poverty using % female hh + % white

| <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 (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.

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 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

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

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_{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 | | | |

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

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](#).