FOUNDATIONS OF STATISTICAL DECISION MAKING

Relationships and Prediction

June 12, 2018

Aaron R. Baggett, Ph.D.

Department of Physical Therapy
University of Mary Hardin-Baylor
PHTH 7147: Critical Inquiry I

OUTLINE

Outline

- Correlation
- Predicting outcomes (Regression)

Recap

- Statistical variables
- Multiple group comparisons (ANOVA)

Resources

- Slides, data, and handouts available at:

`bit.ly/umhb_dpt`

Data

- Today's example data are from the 2002-2004 National Education Longitudinal Study (NELS)
- Nationally representative, longitudinal study of U.S. high school students
- Surveys of students, their parents, math and English teachers, and school administrators
- Student assessments in math (10th & 12th grades) and English (10th grade)

Data

- Variables:
  1. `grades`: GPA of student in 2002
  2. `pared`: Highest education of parent (in years)
  3. `hwork`: Amount of time spent doing homeworkd during the week (in hours)

Data

- Let’s look at the NELS data

	grades	pared	hwork
1	78	13	2
2	79	14	6
3	79	13	1
4	89	13	5
5	82	16	3
6	77	13	4
...	...	...	...
100	74	12	4

Data

Variable correlations

	Grades	Parent Education	Homework
Grades	1.00	–	–
Parent Education	0.29 (0.08)	1.00	–
Homework	0.33 (0.11)	0.28 (0.08)	1.00

CORRELATION

Correlation

- Statistical technique used to determine the degree to which two variables are related
- Two numerical variables: Pearson's r
- The degree of relationship between two variables can vary from -1.0 to 1.0
- This is sometimes referred to as magnitude
- The closer the relationship is to -1.0 or 1.0, the stronger the magnitude or degree of relation between two variables

Correlation

- Correlation coefficients describe two characteristics:
  1. The degree to which two variables are related
  2. The direction, or type of effect one variable has on the other (i.e., positive or negative)

Correlation

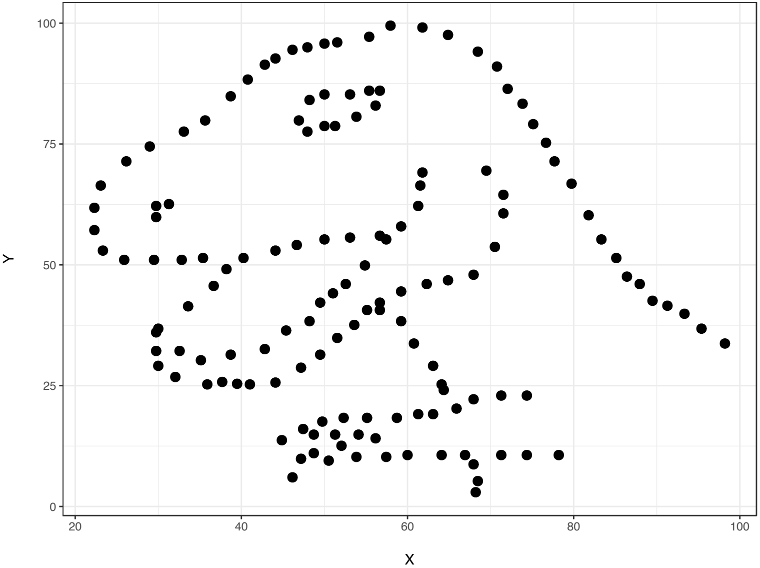
- Two types of correlation:
  1. Positive Correlation:
     - Higher scores on one variable associated with higher scores on a second variable
  2. Negative Correlation:
     - Higher scores on one variable associated with lower scores on a second variable

Correlation

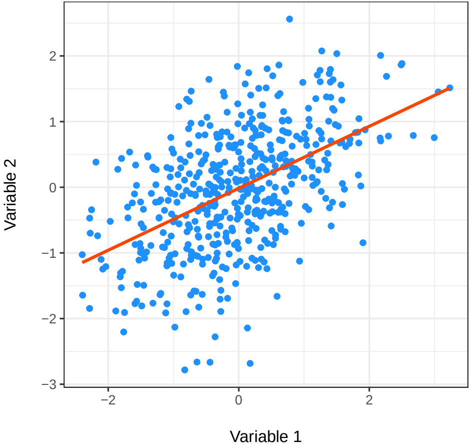
- It’s always recommended that you visualize your correlational data first
- The results may yield more information than the r alone
- For example, imagine we have the following data with $r = -0.064$

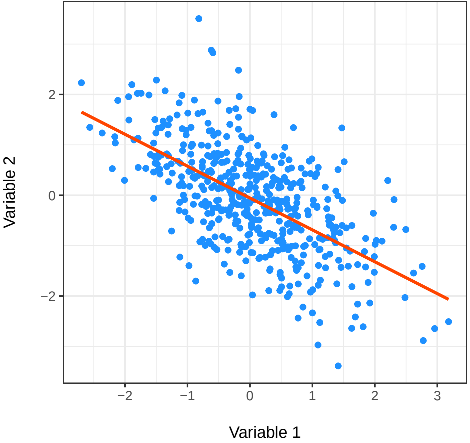
	x	y
1	55.38	97.18
2	51.54	96.03
3	46.15	94.49
—	—	—
142	44.10	92.69

Correlation



Correlation


Positive Correlation


Negative Correlation

Correlation

- What determines a strong, medium, and small correlation?
  - Cohen (1988) suggested the following:
    - $r \leq$ 0.10 = small
    - $>$ 0.10 $r \leq$ 0.30 = medium
    - $r \geq$ 0.50 = large

Correlation

- Once calculated, r can be squared (r^2)
- This is called a coefficient of determination
- Proportion of variability in one variable that can be accounted for (or explained) by variability in the other variable
- The remaining proportion can be explained by factors other than your variables
  - **Ex.:** r = 0.50 $\rightarrow$ r^2 = 0.25

Correlation

- We often examine correlations visually using a scatterplot
- Graphically depicts the relationship between 2 variables
- Typically, the predictor is on the X-axis and the outcome is on the Y-axis

Correlation

	Quantitative X	**Ordinal X**	**Nominal X**
Quantitative Y	Pearson's r	—	—
Ordinal Y	Biserial r_b	Spearman ρ	—
Nominal Y	Point Biserial r_{pb}	Rank Biserial r_{rb}	Phi (ϕ)

Calkins (2005)

PREDICTION AND REGRESSION

Prediction and Regression

- Regression is a statistical procedure used to predict values of one variable from values of another variable
- It is a hypothetical model of the relationship between at least two variables
- The model used is a linear one
- Therefore, we describe the relationship using the equation of a straight line

Prediction and Regression

- Imagine we suspect parents' education and time spent doing homework combine to predict students' grades

Prediction and Regression

- Regression model equation:

$$Y = a + bX_1 + bX_2 + e$$

- a = Intercept
    - Point where regression line crosses Y axis
- b = Slope of the line

Prediction and Regression

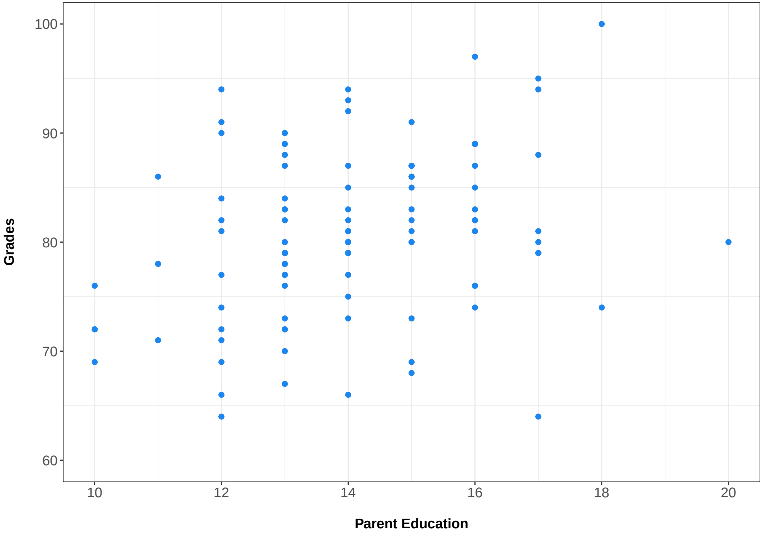
$$Y = a + bX_1 + bX_2 + e$$

- Y = Criterion or dependent variable
  - Variable being measured and predicted
  - Y = students' grades
- X = Predictor or independent variable
  - Variable we use to predict the outcome
  - X_1 = parents' education
  - X_2 = homework

NELS Data

pared	n	mean_grades	sd_grades	mean_hw	sd_hw
10	4	72.25	2.87	3.75	1.26
11	3	78.33	7.51	3.33	1.15
12	13	78.08	9.80	4.46	1.81
13	23	79.22	6.07	4.78	2.19
14	19	81.37	6.85	5.32	1.57
15	15	81.67	6.82	5.53	2.45
16	12	83.42	6.54	5.75	2.30
17	8	82.50	9.96	5.62	2.39
18	2	87.00	18.38	6.00	1.41
20	1	80.00		6.00	

NELS Data



NELS Regression

- Let's regress students' grades on parent education and time spent doing homework
- Notice the intercept term and coefficients for `pared` and `hwork`
- Interpretation can be tricky

NELS Regression

```
## 
##  LINEAR REGRESSION
##
##  Model Fit Measures
##  ───────────────────────────
##    Model    R        R²      
##  ───────────────────────────
##        1    0.390    0.152   
##  ───────────────────────────
##
##
##  MODEL SPECIFIC RESULTS
##
##  MODEL 1
##
##  Model Coefficients
##  ─────────────────────────────────────────────────────
##    Predictor    Estimate    SE       t        p        
##  ─────────────────────────────────────────────────────
##    Intercept      63.227    5.240    12.07    < .001   
##    pared           0.871    0.384     2.27     0.026   
##    hwork           0.988    0.361     2.74     0.007   
##  ─────────────────────────────────────────────────────
```

NELS Regression

```
##
##  MODEL 1
##
##  Model Coefficients
##  ──────────────────────────────────────────────────────
##    Predictor    Estimate    SE       t        p
##  ──────────────────────────────────────────────────────
##    Intercept      63.227    5.240    12.07    < .001
##    pared           0.871    0.384     2.27     0.026
##    hwork           0.988    0.361     2.74     0.007
##  ──────────────────────────────────────────────────────
```

Interpretation

For a student who spends 0 hours weekly doing homework and whose parent has 0 years of education, we would predict his/her GPA to be approximately 63.23.

NELS Regression

```
##
##  MODEL 1
##
##  Model Coefficients
##  —————————————————————————————————————————————————————
##    Predictor    Estimate    SE        t        p
##  —————————————————————————————————————————————————————
##    Intercept      63.227    5.240    12.07    < .001
##    pared           0.871    0.384     2.27     0.026
##    hwork           0.988    0.361     2.74     0.007
##  —————————————————————————————————————————————————————
```

Interpretation, contd.

For every 1 unit increase in parent education and time spent weekly doing homework, we would expect this students’ GPA to increase by 0.871 and 0.988 points, respectively.

NELS Regression

What’s wrong here?

NELS Regression

- We need to mean center both `pared` (*M* = 14.03, *SD* = 1.93) and `hwork` (*M* = 5.09, *SD* = 2.06)
- This will allow more realistic interpretation

NELS Regression

```
## 
##  LINEAR REGRESSION
## 
##  Model Fit Measures
##  ───────────────────────────
##    Model    R        R²
##  ───────────────────────────
##        1    0.390    0.152
##  ───────────────────────────
## 
## 
##  MODEL SPECIFIC RESULTS
## 
##  MODEL 1
## 
##  Model Coefficients
##  ─────────────────────────────────────────────────────────
##    Predictor       Estimate    SE       t         p
##  ─────────────────────────────────────────────────────────
##    Intercept          80.47    0.709    113.47    < .001
##    pared_center        1.68    0.742      2.27     0.026
##    hwork_center        2.03    0.742      2.74     0.007
##  ─────────────────────────────────────────────────────────
```

NELS Regression

```
##
##  MODEL 1
##
##  Model Coefficients
##  ───────────────────────────────────────────────────────────
##    Predictor        Estimate    SE       t         p
##  ───────────────────────────────────────────────────────────
##    Intercept           80.47    0.709    113.47    < .001
##    pared_center         1.68    0.742      2.27     0.026
##    hwork_center         2.03    0.742      2.74     0.007
##  ───────────────────────────────────────────────────────────
```

Interpretation

For a student who spends M = 5.09 hours weekly doing homework and whose parent has M = 14.03 years of education, we would predict his/her GPA to be approximately 80.47.

NELS Regression

```
##
##  MODEL 1
##
##  Model Coefficients
##  ──────────────────────────────────────────────────────────
##    Predictor       Estimate    SE       t         p
##  ──────────────────────────────────────────────────────────
##    Intercept          80.47    0.709    113.47    < .001
##    pared_center        1.68    0.742      2.27     0.026
##    hwork_center        2.03    0.742      2.74     0.007
##  ──────────────────────────────────────────────────────────
```

Interpretation, contd.

For every 1 unit change in parent education and time spent weekly doing homework, we would expect a students' GPA to change by 1.68 and 2.03 points, respectively.

NELS Regression

- Overall model interpretation
- In regression, we typically use R^2 as a meausre of effect size
- Proportion of variance explained by the model

```
##
##  Model Fit Measures
##  ───────────────────────────
##    Model     R        R²
##  ───────────────────────────
##        1     0.390    0.152
##  ───────────────────────────
```

NELS Regression

```
##
##  Model Fit Measures
##  ───────────────────────────
##    Model    R        R²
##  ───────────────────────────
##        1    0.390    0.152
##  ───────────────────────────
```

Interpretation

Parents’ education and the time spent doing homework combine to explain approximately 0.152 → 15.20% of the variability in determining students’ grades.

RECAP

Recap

- Correlation and regression are used to predict outcomes using past data
- Interpretation can be tricky
- Causation cannot be assumed

QUESTIONS?