Data 8 Tutoring: Week 9 Solution

Standard Units, Regression, Least Squares
April 8, 2017

1. Standard units

1. The average score in last year's Data 8 final exam was 53 points, with an SD of 18 points. Ian scored 48 pts and Michelle scored 74. What were their scores in standard units?

Ian: (48 - 53)/18
Michelle: (74-53)/18

2. What score corresponds to -2.5 standard units?

To convert from SUs to original units just add to the AVE the given SUs times the SD

Score = 53 + (-2.5) * 18

3. What are the benefits of measuring in standard units?

Many possible answers:
- Allows for easier comparison between different units of data
- Can switch axis without changing correlation coefficient

2. Regression

It may be helpful to review the following formulas before proceeding to the next question.

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

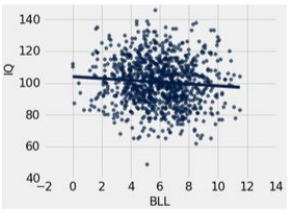
3. (25 points) Regression

The `lead` table (left) contains one row per child in a study of 1000 children's Blood Lead Levels (BLL) measured in micrograms per deciliter and their intelligence quotients (IQ). Assume that the data were collected by sampling children at random from a very large population. Summary statistics (middle) and a scatter diagram (right) are shown below. All BLLs are measured to one decimal place, and all IQ scores are integers.

BLL	IQ
7.9	90
6.2	78
3.2	110
4.1	128
7.3	88

(995 rows omitted)

Expression	Value
`np.average(lead.column('BLL'))`	6
`np.std(lead.column('BLL'))`	2
`np.average(lead.column('IQ'))`	100
`np.std(lead.column('IQ'))`	15
`correlation(lead, 'BLL', 'IQ')`	-0.1



(a) (2 pt) What is the value of `correlation(lead, 'IQ', 'BLL')`?
Hint: The `correlation` function appears on your midterm study guide.

-0.1

(b) (3 pt) What is the estimated average IQ of a child with a BLL that is 2 standard deviations above the mean BLL? Use the regression line to find this estimate, assuming BLL and IQ are linearly related.

$2 * -0.1 * 15 + 100 = 97$

(c) (4 pt) Write the equation of the regression line through this sample for the IQ y in terms of the BLL x.

$y = -0.1 \cdot \frac{15}{2} \cdot x + 100 - 6 \cdot (-0.1 \cdot \frac{15}{2}) = -0.75x + 104.5$

4. (8 points) Predictions

The `ball` table contains player data for some of the Golden State Warriors. Only the first five rows are shown.

Player	Minutes per Game	Points per Game
Klay Thompson	34	21
Andrew Bogut	20	5
Stephen Curry	34	29
James McAdoo	4	3
Andre Iguodala	28	7

You have computed the following summary statistics from the full `ball` table.

Expression	Value
`np.average(ball.column(1))`	24
`np.std(ball.column(1))`	10
`np.average(ball.column(2))`	13
`np.std(ball.column(2))`	8
`correlation(ball, 1, 2)`	0.75

(a) (6 pt) For each question below, answer with a number. You may show your work for partial credit.

- What is the value of Stephen Curry's points per game in standard units?

```
(29-13) / 8 == 2
```

- What is the slope of the regression line when the points per game are plotted on the vertical axis, the minutes per game are plotted on the horizontal axis, and a regression line is fit to the data? That's the slope of the regression line computed by `slope(ball, 1, 2)` in original units, $\frac{\text{points}}{\text{minute}}$.
The ***`slope`*** *function is defined on the last page of the midterm study guide.*

```
0.75 * 8/10 == 0.6
```

- What is the fitted value for Stephen Curry using this regression line to estimate his points per game from his minutes per game?

```
(34-24)/10 * 0.75 * 8 + 13 == 19
```

(b) (2 pt) How would the fitted value of points per game for a player who played 34 minutes per game change if Stephen Curry were removed from the table and the regression line recomputed? *Circle one.*

(a) Increase (b) Decrease (c) Stay the same (d) Not enough information

Since the regression line minimizes the sum of squared errors, and Stephen Curry's squared error is positive at x = 34 minutes, removing this term from the objective function would result in a smaller fitted value.

3. Least Squares

1. What quantity does the least squares method minimize in order to generate the “best” line?

The root mean squared error

2. Why is it unfeasible to minimize error?

To avoid cancellation when measuring the rough size of the errors, we will take the mean of the squared errors rather than the mean of the errors themselves.

4. True/False

T / **F** The least squares method is a way to generate a regression line.

T / **F** Causation implies linear correlation

T / **F** The regression line is the only line that minimizes mean squared error.

T / **F** No matter what the shape of the scatter plot, there is a unique line that minimizes the mean squared error of estimation.

T / **F** The sign of the correlation coefficient is the same as the sign of the slope

T / **F** Quadratic regression does not exist