Regression Review

Name: _____

Note: This was from last semester's final. Some of the syntax (code) was done differently this semester. We've crossed out and replaced the old stuff. Answers are in red and comments are in blue.

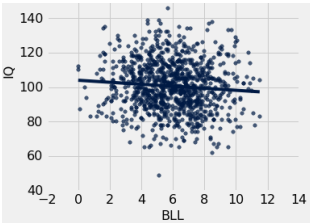
3. (26 points) Regression

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

(same as correlation(lead, 'BLL', 'IQ')

(b) (4 pt) What is the estimated average IQ of a child with a BLL of 10.0 predicted by linear regression, assuming BLL and IQ are linearly related?

BLL in standard units = (10 - 6) / 2 = 2
Predicted IQ in standard units = (-0.1)*2 = -0.2
Predicted IQ = estimated average IQ = 15*(-0.2) + 100 = 97 (the answer)

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

(Just replace 10 with x in the above.)

$y =$ 15*(-0.1)*(x - 6)/2 + 100 = -3*x + 104.5

(It's okay, or even preferred, to give the first, unsimplified, answer.)

(d) (4 pt) Complete the code below so that the last line prints out a 95% confidence interval for the IQ value at a BLL of 10.0 on the *regression line of the population* from which this sample was collected.
Hint: The `slope` and `intercept` functions appears on your midterm study guide.

Note: in this exam there were slope and intercept functions given; not in yours. slope and intercept took 3 arguments: a table, a column label for x values, and a column label for y values. They returned the slope and intercept (respectively) of the regression line.

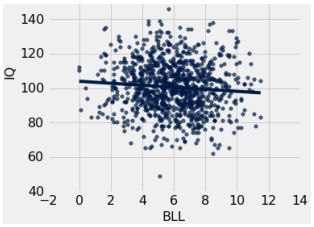
```
estimates = make_array()
for i in np.arange(400):

    r = lead.sample(with_replacement=True)

    e = slope(_____r_____, 0, 1) * ____10____ + intercept(_____r_____, 0, 1)

    estimates = np.append(estimates, e)

print(percentile( 2.5, estimates), percentile(97.5, estimates))
```

The data from the previous page are repeated here for your reference.

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



(e) (6 pt) Mark each of the following statements about **the confidence interval you computed in part (d)** as *True* or *False* based on the definition of a confidence interval and the details of your implementation. The terms "most" and "high chance" should be interpreted as around 95%.

- Circle *True* or *False*: It contains most IQ scores in the population for children with a BLL of 10.0.
  The confidence interval's width reflects how well we have estimated the average IQ for kids with BLL 10. It has nothing to do directly with individuals.
- Circle *True* or *False*: It contains most IQ scores in the sample for children with a BLL of 10.0.
  Same as above.
- Circle *True* or *False*: If `with_replacement=False` were used (line 3), the interval would have 0 width.
  All the "resamples" would be shuffled copies of each other, so the scatter plot and regression line for each one would be exactly the same.
- Circle *True* or *False*: If the study were repeated many times, most confidence intervals computed in this way would overlap.
  Most of the intervals would include the population average IQ for kids with BLL 10, so they would overlap each other.
- Circle *True* or *False*: If the study were repeated many times, the average IQ in the population would fall within most intervals computed in this way.
  This was tricky. The overall average IQ isn't what we're estimating, but rather the average IQ for kids with BLL 10.
- Circle *True* or *False*: If the study were repeated many times, the average IQ in the population for children with a BLL of 10.0 would fall within most intervals computed in this way.

(f) (2 pt) What null hypothesis would you evaluate in a statistical test to determine whether BLL and IQ are negatively correlated in the population?

The correlation between BLL and IQ in the population is 0.

(g) (2 pt) Based on the summary statistics provided, what is the minimum proportion of IQ scores in the sample that are between 70 and 130 according to Chebyshev's inequality?

70 and 130 are 2 standard deviations (15) from the mean IQ (100), so the minimum proportion is 1 - 1/(2*2), or 75%.

(h) (2 pt) Based on the scatter diagram and summary statistics, what proportion of IQ scores in the sample do you think are between 70 and 130? **Describe your reasoning.** *Hint*: You don't need to count dots.

.95. The scatter diagram looks football-shaped, so the distribution of IQ (ignoring BLL) is probably roughly normal. In a normally-distributed dataset, 95% of the elements are within 2 standard deviations of the mean.