PBIO 504 Introductory biostaTISTICs

**HW5: Correlation and Linear Regression**

**Reading Assignment**

• Chapters 17 and 18 in the textbook*.*

# The Data Sets

**1) Birth weight and gestational age study** (note: for some gestational ages there are multiple observations, but all of these data points are independent):

gestage (wks) bwt (grams)

1. 1820
2. 1900
3. 2210
4. 2250
5. 2500
6. 2775
7. 2901
8. 3090, 3200
9. 3270, 3347
10. 3400, 3424, 3500
11. 3493, 3600, 3616
12. 3680, 3720
13. 3800

**2) Lead and iron exercise (percent of dose absorbed): each value is the amount of the metal (mg) measured in one ml of blood**

***Study population 1*** ***Study population 2***

iron lead iron lead

17 8 11 10

22 17 25 22

35 18 30 33

43 25 44 40

80 58 50 50

85 59 65 52

91 41 71 54

72 30 82 55

96 43 90 57

100 58 100 60

11 3 80 53

INSTRUCTIONS:

1. Make a scatter-plot of the relationship of gestational age (in weeks, on the x-axis) to birth weight (in grams, on the y-axis).

2. Perform a linear regression analysis for the model: ***bwt* = *gestage***. Add the fitted regression line to the plot. Answer questions 1-6 on the worksheet about the data plot and regression results

3. Make separate scatter-plots of the iron and lead data from the 2 different study populations (it does not matter which variables are plotted on the x- and y-axes). Compute the correlation coefficients and test the null hypothesis that ***ρ*** =0 for each population separately. Answer questions 7-10 on the worksheet.

NOTE: do not have to include graphs.

**Homework 5** Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Questions about the analysis of birth weight and gestational age:***

1. What is your impression from the scatter plot with the fitted regression line, in terms of how well the linear regression model appears to fit the data? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Report the estimated slope\_\_\_\_\_\_ and estimated intercept\_\_\_\_\_\_

Interpretation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Report and interpret the R2 value.

***R2*** =\_\_\_\_\_\_ Interpretation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(4) Report the test statistic and the p-value for the test that the slope equals 0

***t*** = \_\_\_\_\_ ***p*** =\_\_\_\_\_ Interpretation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(5) Use the regression equation for a baby born at 38 weeks. What is the predicted birth weight for such an infant? \_\_\_\_\_\_\_\_\_\_\_\_\_\_

(6) Suppose we want to predict the birth weight of a child born at 28 weeks. How would you do it, or why would you not? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Correlation results of the analysis of lead and iron data.***

(7) Based on the scatter plots of each of the 2 study populations, is it appropriate to perform a correlation analysis? Describe your impressions of the data.

Population 1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Population 2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(8) Report the correlation coefficients:

***r*** =\_\_\_\_\_\_ in population 1 and ***r*** =\_\_\_\_\_\_ in population 2

(9) Report the test statistic and the p-value from the test that ***ρ*** =0 in population 1; interpret the results: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(10) “Outliers” are said to be a potential problem for correlation and regression analyses – why do you think this is so? How would an outlier data point affect the results of a correlation analysis and how might it affect the results of a linear regression?