Biostatistics 515/518, Winter 2019
Homework 3 (three questions)

1. For the WCGS dataset, dichotomize age into old (55 years or older) and young (younger than 55 years).

(a) Summarize height for old and young men using this dichotomization. Use the results to predict the results from a simple linear regression of height on dichotomized age. Will model-based standard error estimates be conservative or anti-conservative? Explain.

(b) Summarize systolic blood pressure (sbp) for old and young men using this dichotomization. Use the results to predict the results from a simple linear regression of systolic blood pressure on dichotomized age. Will model-based standard error estimates be conservative or anti-conservative? Explain.

For both (a) and (b), perform regression analyses to evaluate your predictions.

Note: "predict the results from a simple linear regression" – the intention is to use the output to anticipate what the fitted intercept and slope will be.

2. You are an investigator interested in whether education leads to greater economic benefit. You have access to some data collected by a colleague on monozygotic (“identical”) twins. See the “Twins” data on the course website.

The outcome variable you are interested in is hourly wage. Knowing that the distribution of incomes tends to be right-skewed, you decide to transform the hourly wage variable by taking the log. (You may either interpret results in terms of the geometric mean, or simply as arithmetic means of the log wage.)

(a) Fit a simple linear regression model to investigate the association between years of education and income. Choose the methodology that is best-suited to the question. Write a brief paragraph (3-6 sentences) that summarizes the results. Your summary should discuss the real-life importance or lack of importance of the analysis. In particular, your summary should discuss whether there are other factors that the simple regression does not account for that a more appropriate analysis would account for. In addition, comment also on the methodology you chose and why.

(b) Your cousin works for a local hardware store. She knows the topic you are researching, and wants to know how much her wage will go up if she returns to school to complete an associate's degree (2 year program). Respond to her query.

(c) Using your regression model, estimate the mean wage for those with a high-school education (12 years of school). [Note: when you are asked for an estimate, your habit should be to provide a confidence interval in addition to a point estimate.]

(d) How else might you estimate the mean wage for those with a high-school education, other than using a regression model? (Go ahead and do this.)

(e) Using a regression model to predict the wage for an individual with a high-school education (12 years of school) and give a 95% prediction interval. Comment on the methodology and appropriateness of the approach.

(f) How else might you construct a 95% prediction interval for an individual with a high-school education, other than using a regression model? (Go ahead and do this.)

3. The concept of ‘regression to the mean’ gives the following formula for predicting a response value for a variable Y from based on a predictor x: $E[Y|x] = \bar{y} + r_{xy} \times S_y \times \frac{x-\bar{x}}{S_x}$

Use this expression to derive the formulas for the regression slope and intercept presented in class.