STAT 4400 / 5400 **HOMEWORK #1** SPRING 2022

Due: Jan. 18 by midnight

(1) Let

$$y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} 2 \\ 0 \\ 4 \end{bmatrix}$$

Consider the model $y = \beta_0 + \beta_1 x + \epsilon$ where the errors ϵ_i for $i = 1, \ldots, n$, have independent normal distributions with mean 0 and a constant standard deviation denoted as σ. Assume that y accurately reflects the phenomenon in which we are interested and that x is the only relevant predictor of y and that the model generalizes to the given values of x. Assume further that y is a linear function of x.

(a) Using the method of least squares, write equations for the the estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}^2$.

(b) Solve the equations by hand (using a calculator if necessary) and find the numerical estimates of $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}^2$.

(c) Find the value of R^2.

(d) Test the hypothesis $H_0 : \beta_1 = 0$ using a level of significance of $\alpha = .05$. Write the formulas for all necessary calculations and justify your conclusion.

(e) Create a plot that shows (i) the relationship between x and y and (ii) the least squares regression line. *(Feel free to use R or any other software for this question or you may draw the plot by hand and upload the scanned pdf into your homework file.)*

(2) A test is graded from 0 to 50, with an average score of 35 and a standard deviation of 10. For comparison to other tests, it would be convenient to rescale to a mean of 100 and a standard deviation of 15.

(a) Find a linear transformation so that the test has the desired mean and standard deviation.

(b) Find a second linear transformation that gives the desired mean and standard deviation. (Use the same method as above and find another solution.)

(c) Which of these two transformations would you recommend using to compare the results of this test to other tests and why?

(3) In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other.

(a) First generate 1000 data points from a normal distribution with mean 0 and standard deviation 1, call it "var1". Generate another variable in the same way, call it "var2". Run a regression of one on the other. Is the slope coefficient statistically significant?

(b) Now run a simulation repeating this process 100 times. From each simulation, save the z-score (the estimated coefficient of the predictor divided by its standard error). If the absolute value of the z-score exceeds 1.96 then the estimated slope coefficient is statistically significant at the $\alpha = .05$ level of significance. How many of the 100 estimated slope coefficients are statistically significant?

Here is code to perform the simulation:

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
    var1 <- rnorm (1000,0,1)
    var2 <- rnorm (1000,0,1)
    fit <- lm (var2 ~ var1)
    z.scores[k] <- coef(fit)[2]/summary(fit)$coef[2,”Std. Error”]
    }
alpha = .05
cutoffn = qnorm(1-alpha/2,    lower.tail = TRUE, )
sum(abs(z.scores) > cutoffn)
```

(4) Use the dataset “child.iq.dta” which is saved in Canvas for this question.

(a) Fit a regression of child test scores on mother's age, display the data and fitted model, check assumptions, and interpret the slope coefficient. When do you recommend mothers should give birth? What are you assuming in making these recommendations?

(b) Repeat this for a regression that further includes mother's education, interpreting both slope coefficients in this model. Have your conclusions about the timing of birth changed?

(c) Now create an indicator variable reflecting whether the mother has completed high school or not. Consider interactions between the high school completion and mother's age in family. Also, create a plot that shows the separate regression lines for each high school completion status group.

(d) Finally, fit a regression of child test scores on mother's age and education level for the first 200 children and use this model to predict test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children.

(5) Use the dataset “prostate” which is included with the R package “faraway”.

(a) Fit a model with the natural logarithm of the level of prostate specific antigen (“lpsa”) as the response and the other variables in the dataset as predictors. Compute a 95% confidence interval for the coefficient associated with the natural logarithm of the level of cancer (“lcavol”).

(b) Fit a new model with only those predictors that were significant at the 5% level in model in the previous model. Compute a 95% confidence interval for the coefficient associated with the natural logarithm of the level of cancer (“lcavol”).

(c) Which of these two models do you prefer to use for predicting the level of prostate specific antigen and why?