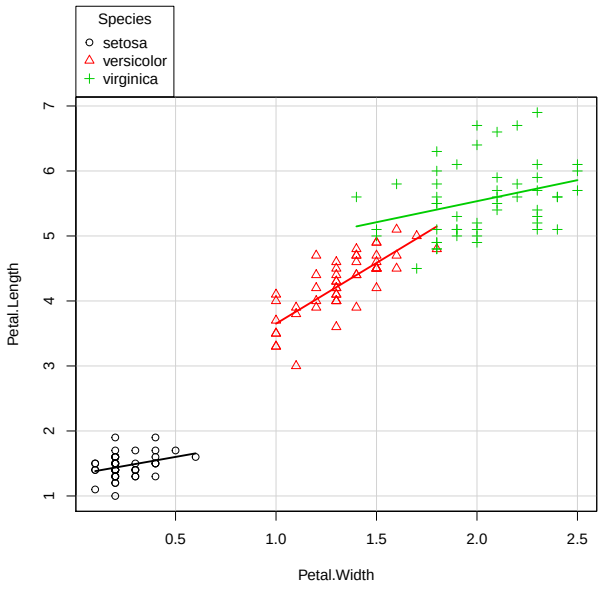
STAT 217: Homework 4

Due Wednesday, April 29th at the beginning of class

Name: _____

1. This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.. Below is the output from the model with an interaction term along with a plot:




```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width * Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84099 -0.19343 -0.03686  0.16314  1.17065
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.3276     0.1309  10.139  < 2e-16
## Petal.Width                      0.5465     0.4900   1.115   0.2666
## Speciesversicolor                0.4537     0.3737   1.214   0.2267
## Speciesvirginica                 2.9131     0.4060   7.175 3.53e-11
## Petal.Width:Speciesversicolor    1.3228     0.5552   2.382   0.0185
## Petal.Width:Speciesvirginica     0.1008     0.5248   0.192   0.8480
##
## Residual standard error: 0.3615 on 144 degrees of freedom
## Multiple R-squared:  0.9595,Adjusted R-squared:  0.9581
## F-statistic: 681.9 on 5 and 144 DF,  p-value: < 2.2e-16
```

(a) Based on the plot, does there appear to be an interaction in these data?

A. No, because the lines for each species are mostly parallel
B. Yes, because the lines for each species are mostly parallel
C. No, because the lines for each species are not parallel
D. Yes, because the lines for each species are not parallel

(b) *The R^2_{adj} for the additive model was 0.9542.* What happened to the adjusted R^2 when we went from the additive (parallel lines) model to the interaction (non-parallel lines) model and what does this tell us about the importance of allowing varying slopes?

A. The R^2_{adj} increased, suggesting that the parallel lines model, with same slopes, is adequate
B. The R^2_{adj} increased, suggesting that varying slopes are not needed
C. The R^2_{adj} increased, suggesting that the interaction model, with varying slopes, is preferred over the parallel lines model
D. The R^2_{adj} increased, suggesting that the parallel lines model, with varying slopes, is preferred over the interaction model

(c) What are the explanatory and response variables in this model?

A. Response: Petal Length; Explanatory: Petal Width and Species
B. Response: Petal Length; Explanatory: Petal Width and setosa, versicolor, and virginica
C. Response: Petal Width; Explanatory: Petal Length and setosa, versicolor, and virginica
D. Response: Petal Width; Explanatory: Petal Length and Species

(d) Below is the anova table for the interaction model.

```
## Analysis of Variance Table
##
## Response: Petal.Length
##                      Df Sum Sq Mean Sq   F value    Pr(>F)
## Petal.Width           1 430.48  430.48 3294.5561 < 2.2e-16
## Species               2  13.01    6.51   49.7891 < 2.2e-16
## Petal.Width:Species   2   2.02    1.01    7.7213 0.0006525
## Residuals           144  18.82    0.13
```

i. What are the null and alternative hypotheses for testing whether or not the slope adjustments are necessary?

ii. What is the value of the test statistic and what distribution does it follow under the null hypothesis?

iii. What is the p-value and your decision?

iv. Write your conclusion in the context of the problem.

A. There is strong evidence that the slope for petal width is the same across species in the population (p-value= 0.00065 from F-stat= 7.72 on 2 and 144 df).

B. There is strong evidence that the slope for petal width in at least one species differs the others in the population (p-value= 0.00065 from F-stat= 7.72 on 2 and 144 df).

C. There is strong evidence that the slope for petal width is the same across species in the population (p-value< 0.0001 from F-stat= 49.79 on 2 and 144 df).

D. There is strong evidence that the slope for petal width in at least one species differs the others in the population (p-value< 0.0001 from F-stat= 49.79 on 2 and 144 df).

(e) Write out the estimated regression equation.

i. Write out the estimated regression line for species *versicolor*.

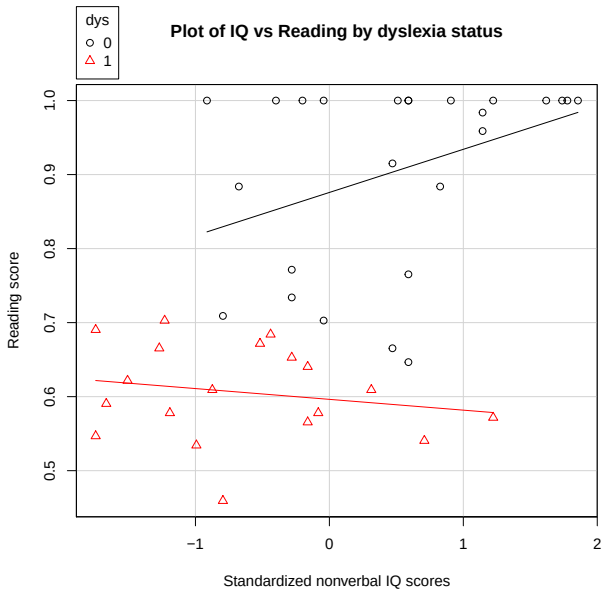
ii. Write out the estimated regression line for species *setosa*.

iii. Write out the estimated regression line for species *virginica*.

(f) Below are the AIC values for the three models fit: SLR(`fit.SLR2`), Additive model (`fit.add2`), interaction model (`fit.int2`). Which one would you prefer based on AIC?

```
##          df      AIC
## fit.SLR2  3 208.3539
## fit.add2  5 139.5702
## fit.int2  7 128.2896
```

2. The folowing dataset shows the nonverbal standardized IQ's of dyslexic and non-dyslexic children. The childrens' scores on a reading accuracy test are also recorded.




```
##
## Call:
## lm(formula = score ~ ziq + dys, data = dyslexic3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26062 -0.05565  0.02932  0.07577  0.13217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.89178    0.02312  38.580  < 2e-16
## ziq          0.02620    0.01957   1.339    0.188
## dys1        -0.26879    0.03905  -6.883 2.41e-08
##
## Residual standard error: 0.1049 on 41 degrees of freedom
## Multiple R-squared:  0.6858,Adjusted R-squared:  0.6705
## F-statistic: 44.75 on 2 and 41 DF,  p-value: 4.917e-11
```

(a) How many levels does the 'dyslexia' variable have?

(b) Calculate a 95% confidence interval for the slope coefficient on `ziq`. Use $t^* = -2.02$.

(c) Interpret your interval in the context of the problem.

3. Information was recorded about the number of steps taken and the amount of calories consumed per day in a random sample of 100 Montanans. Assumptions for the simple linear regression model were not met, so Calories Consumed was log transformed. The estimated regression equation, after the log transformation, is $ln(Cal)_i = 2000 + 0.042 * Steps_i$.

   (a) Predict the number of calories consumed for someone who takes 10000 steps in a day.

   (b) Interpret the coefficient on Steps on the original scale.

   A. For each additional step, the true median number of calories consumed is estimated to increase by 4%.

   B. For each additional step, the true mean number of calories consumed is estimated to increase by 0.042.

   C. For each additional step, the true median number of calories consumed is estimated to change by a multiplicative factor of 2.92.

   D. For each additional step, the true mean number of calories consumed is estimated to increase by 2000.