True/False

For each of the following questions indicate true or false, then **explain** your answer. You may use examples to illustrate your answer.

(I) The p-value for any hypothesis test is calculated under the assumption that the null hypothesis is true.

(II) If the p-value for testing if $H_0 : \beta_1 \leq 0$ is less than 0.01, we would conclude there is a negative linear relationship between Y and the corresponding X.

(III) In general, if we compare a "larger" model to a "smaller" sub-model (which has a subset of the variables in the "larger" model), the "larger" model will have a lower value of SSE.

(IV) A categorical variable in linear regression always results in multiple regression lines with different slopes.

Name: _____

Full Detail

Work out the following problems. **Show your work.**

1. Athletes hemoglobin levels (blood cells that carry oxygen to the body - typically more hemoglobin is better) in `g/deciliter` were measured as the response variable, with explanatory variables of gender (`m` or `f`) as X_1, % body fat (the percentage of body fat in their body) as X_2, and the lean body mass (amount of body mass that is not body fat) in kg as X_3.
   Two models were fit, and some information about them follows:

Model	SSE
Model 1: $\hat{y} = 13.240 + 1.025X_1 - 0.031X_2 + 0.016X_3 + 0.042X_1X_2$	167.03
Model 2: $\hat{y} = 12.977 + 1.470X_1 - 0.021X_2 + 0.017X_3$	168.36

   In addition, 202 athletes were measured total.

   (a) State the null and alternative for testing if the interaction term should remain in the model.

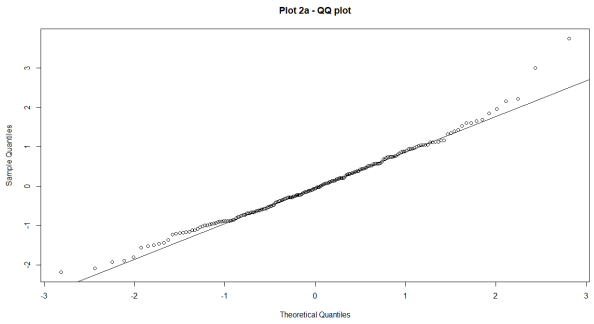
   (b) Calculate the test-statistic for testing if the interaction term should remain in the model.

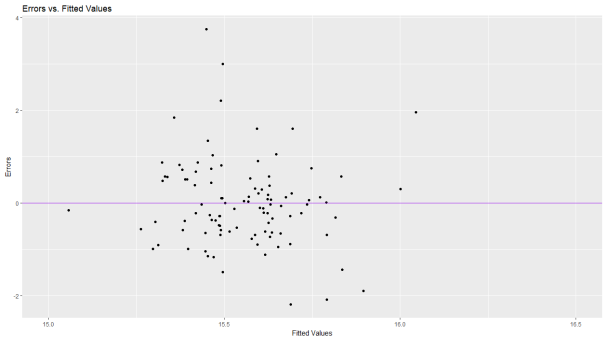
   (c) Assuming the p-value for the test in (a) was 0.2119, state your conclusion in terms of the problem, and interpret the p-value in terms of the problem. Use $\alpha = 0.05$.

   (d) Using the model implied by the result of (c), which X_i's tend to increase hemoglobin, and which tend to decrease hemoglobin? Explain.

Name: _____

2. Continuing with the previous problem, and regardless of your conclusion from 1(d), some plots for Model 2 follow:





(a) Based on the above plots, do you see any outliers present? If so, circle the observations on the plots, and explain if they are for an underestimated, or overestimated Y value.

(b) The p-values for the Fligner-Killeen test and the Shapiro-Wilks test are (respectively): 0.015432, and 0.59432. Using these, state your conclusion for each of their hypothesis tests if $\alpha = 0.05$.

(c) Based on your observations in (a), what would be a next step in the modeling process? What would you recommend, and why?

(d) List two of the assumptions of linear regression.

Name: _____

3. Continuing with the same problem, and using model 2, some information about the fitted β's follow:

	Estimate ($\hat{\beta}_i$)	p-value for $H_0 : \beta_i = 0$	Confidence Interval (95%)
(Intercept)	12.977	0.000	(12.098, 13.855)
Gender: X_1(m)	1.470	0.000	(0.925, 2.015)
% Body Fat: X_2	-0.021	0.185	(-0.052, 0.010)
Lean Body Mass: X_3	0.017	0.034	(0.001, 0.033)

For the rest of the 3(a) - 3(d), assume this is your final model.

(a) Based on the above, which explanatory variables should remain in the model, and why?

(b) Interpret the confidence interval for β_1 in terms of the problem.

(c) Interpret the estimate for β_3 in terms of the problem.

(d) Predict the hemoglobin for a female athlete that has 20% body fat, and lean muscle mass of 40kg.