Homework #11

Question #1 (all parts are scored)

The University of Sheffield (United Kingdom) conducted a nutritional study comparing the weight loss efficacy of three popular diets. The study recruited 78 overweight individuals and randomly assigned them to one of the three diet protocols for 6 weeks, measuring their weight before and after dieting. The *first few rows* of the data are shown below:

Person	gender	Age	Height	pre.weight	Diet	weight6weeks
1	0	22	159	58	1	54.2
2	0	46	192	60	1	54.0
3	0	55	170	64	1	63.3
4	0	33	171	64	1	61.1
5	0	50	170	65	1	62.2
6	0	50	201	66	1	64.0

A):

Suppose a researcher proposes the following analysis plan: "To assess whether each diet is effective, we should stratify the data by the variable"Diet" and then use seperate two-sample t-tests within each strata to determine how different that diet's average before and after weights are. We can then assess which diet was most effective based upon which of the three p-values is the smallest". Briefly explain two (2) flaws with the researcher's proposal.

B):

Suppose we analyze these data using an ANOVA model that uses the categorical variable "Diet" to predict weight loss (a new variable derived by taking "pre.weight" - "weight6weeks"). The sum of squares of the residuals for this model is $\sum_i r_i^2 = 430.2$; for comparison, the model which predicts the overall average weight loss for everyone, regardless of their diet, has a sum of squares of $\sum_i r_i^2 = 501.3$. Use this information to test whether there is a statistically significant association between diet and weight loss.

As you know, ANOVA is a special case of regression modeling where a single categorical explanatory variable is used to predict a quantitative outcome. The following table summarizes the regression coefficients (slopes) of the model described in Part B.

Variable	Coefficient	t-statistic	p-value
Intercept	3.3	6.75	0.000
Diet = 2	-0.27	0.41	0.684
Diet = 3	1.845	2.75	0.007

C):

Interpret the coefficient of the variable: "Diet = 3", your answer should directly address the idea of a "reference category".

D):

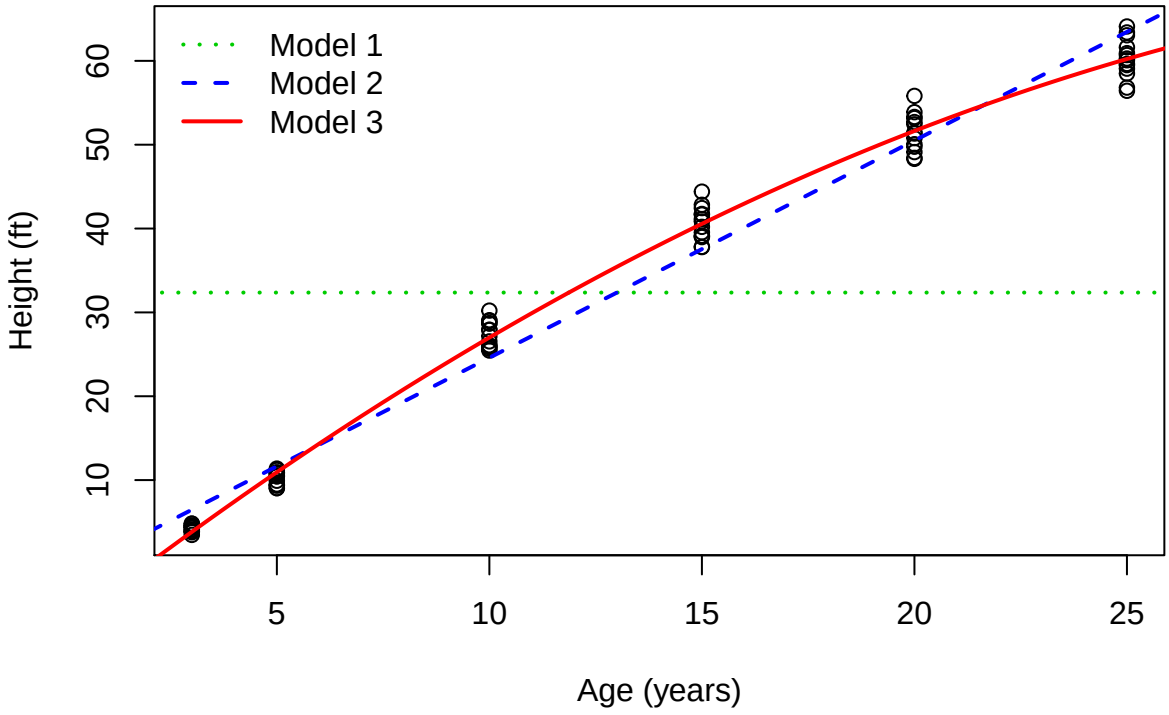
Use the information in the regression coefficient table to determine which of the three diets resulted in the largest weight loss.

E):

Based upon the design of this study, is it likely that age is a confounding variable in the relationship between "Diet" and weight loss?

Question #2 (optional)

Loblolly pine trees are one of several pine trees native to southeastern portions of the United States, ranging from central Texas to eastern Florida. The trees were given the name "Loblolly" because of their prevalence in lowlands and swampy areas. In this study, researchers tracked the growth rate of 14 Loblolly pine trees over a span of 25 years, resulting in a total of 84 height measurements (in feet). The heights of these trees over time, along with three different regression models, are shown in the plot below. A summary of these models is also provided.



Model	Sum of Squares	R^2	Adjusted R^2
$\widehat{\text{Height}} = \beta_0$	$\sum_i r_i^2 = 35474$	-	-
$\widehat{\text{Height}} = \beta_0 + \beta_1 \text{Age}$	$\sum_i r_i^2 = 712$	0.980	0.979
$\widehat{\text{Height}} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Age}^2$	$\sum_i r_i^2 = 235$	0.993	0.992

A): (extra credit)

We can evaluate the assumptions of a regression model by studying its residuals. Based upon the fitted line plot of these models, do you believe these assumptions are satisfied for Model 2? Are they satisfied for Model 3? State your conclusion and provide a brief explanation for each model.

B): (extra credit)

ANOVA can be used to compare *nested models*. Are model 1 and model 3 nested? Are model 2 and model 3 nested? Briefly explain.

C): (extra credit)

Use ANOVA to determine whether the inclusion of a quadratic term provides a significant improvement over the simple linear model that uses "Age" to predict "Height". State your null hypothesis (in words is okay), organize your test in an ANOVA table, and state your conclusion.

Question #3 (all parts are scored)

The Tips Data Set contains data on tips collected by a server working in a suburban restaurant. For this question you will analyze these data in Minitab.

A):

Test whether the amount tipped differs by day of the week. State your hypotheses and include any software output relavent to the test. Provide a one sentence conclusion.

B):

Use a residual plot to check the assumptions of the test you performed in part A.

C):

Repeat the test you performed in part A using the transformed outcome variable "log(Tip)". Include any software output relavent to the test and provide a one sentence conclusion.

D):

Explain why the p-value from the test in part C was smaller than the p-value from the test in part D