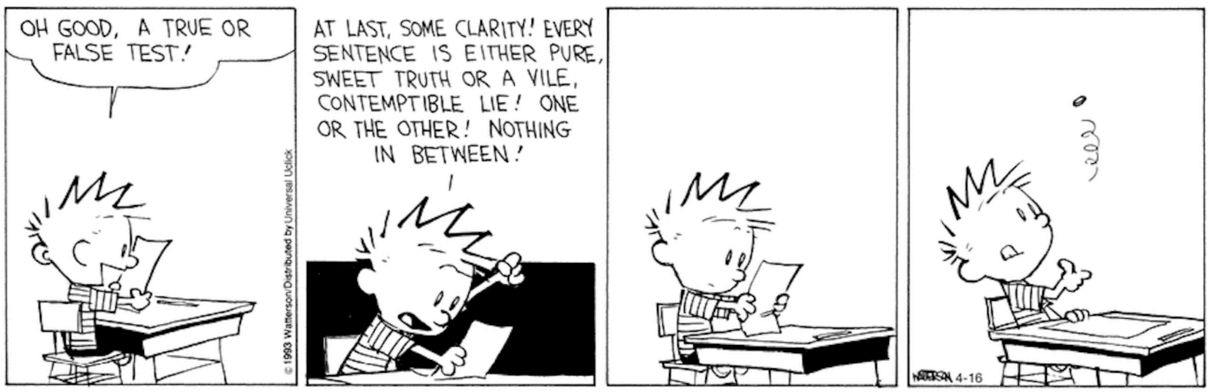
HUDM 5123 - Linear Models and Experimental Design
Mid-term Exam

1 Exam Rules

Don't cheat. If you don't know the answer, take your best guess. 50 points total.



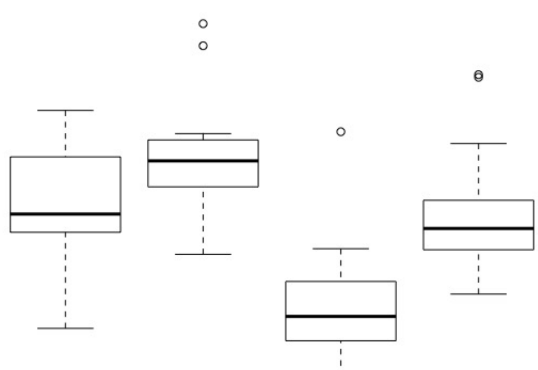
2 True/False (1 pt each) No Tiffs!

1. ____ The degrees of freedom associated with the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} \times X_{i2} + \epsilon_i$$

fit to a sample of $n = 70$ observations is 66.

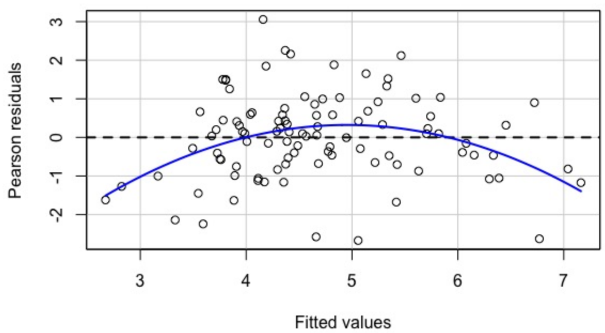
2. ____ Group sample variances for a single-factor ANOVA design for a factor with four levels are as follows: 11.6, 4.2, 14.1, and 9.7; boxplots by cell of the design are displayed in the plot below; and results of Bartlett's test of homogeneity of variance yielded a p-value of .18.



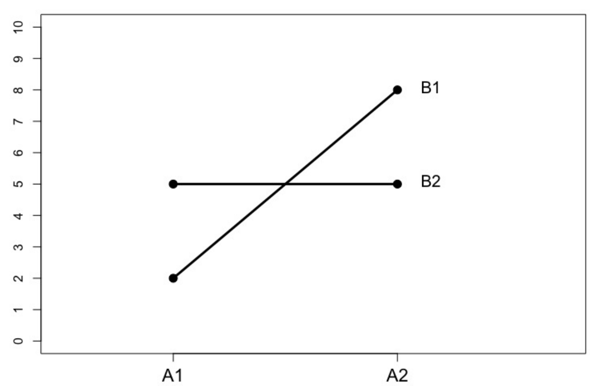
Based on the sample variances, boxplots, and p-value, it is safe to conclude that the constant residual variance assumption is tenable here.

3. ____ Consider a four group one-way ANOVA design. Suppose interest centers on a test of the null hypothesis $H_0 : \mu_1 = (\mu_2 + \mu_3 + \mu_4)/3$. If the researcher defines a contrast $\psi = c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4$ and tests $H_0 : \psi = 0$, the values $c_1 = -3, c_2 = 1, c_3 = 1, c_4 = 1$ correspond with a test of the desired hypothesis.

4. ____ The residual plot below suggests the assumption of correct model specification (linearity) is violated.



The following plot represents population marginal means for a scenario with two factors (A and B), each with two levels.



5. ____ There is an interaction between factors A and B.

6. ____ There is a main effect for factor A.

7. ____ There is a main effect for factor B.

8. ____ There is a simple main effect for factor A at level 1 of factor B.

9. ____ There is a simple main effect for factor A at level 2 of factor B.

10. ____ Suppose a linear regression model with p predictors is fit to data of size N with the assumption that the error term $\epsilon \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. An unbiased estimate for the residual variance, σ_ϵ^2, may be obtained by adding up N residuals and dividing their sum by $N - p - 1$. That is,

$$s^2 = \frac{1}{N-p-1} \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i \right).$$

11. ____ The formula for the incremental F test for comparing nested regression models is as follows:

$$F = \frac{(SSResid_R - SSResid_F)/(df_R - df_F)}{SSResid_F/df_F}.$$

12. ____ Experiments in which participants are repeatedly measured yield invalid results if analyzed with standard OLS regression because the assumption of independent errors is violated.

13. ____ The inclusion of a predictor in a multiple regression that is a linear combination of the other predictors in the model is a violation of the constant variance assumption.

14. ____ A continuous outcome was regressed on a categorical variable with three categories. Dummy codes were used to code three categorical variables, and the third category was held out as a reference. The estimates for the intercept and the coefficients on the first and second dummy-coded variables were, respectively, 31.4, 5.5, and -1.2,. The model-based estimate for the mean of the first group is 36.9.

15. ____ The null hypothesis for the one-way ANOVA omnibus test for a factor with four levels is $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.

16. ____ The null hypothesis $H_0 : \mu_1 = (\mu_2 + \mu_3)/2$ is identical to the null hypothesis $H_0 : \psi = 0$, where $\psi = 1\mu_1 + 1/2\mu_2 + 1/2\mu_3$.

17. ____ The formula for multiple R^2 is RSS/TSS, where RSS is the residual sum of squares and TSS is the total sum of squares.

18. ____ The t distribution is skewed to the right, while the F distribution is symmetric.

19. ____ The standard error of a contrast depends on the residual standard error, the group sample sizes, and the number of parameters in the model.

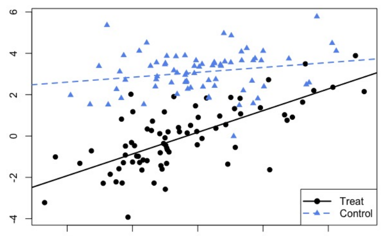
20. ____ The following full and reduced models were compared with an incremental F test:

$$\begin{aligned} Y_{i_F} &= \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_3 D_{i3} + \beta_4 D_{i4} + \epsilon_{i_F} \\ Y_{i_R} &= \beta_0 + \epsilon_{i_R}. \end{aligned}$$

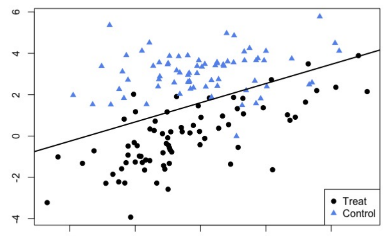
The numerator degrees of freedom for the incremental F test is equal to 5.

3 Matching (1 pt each)

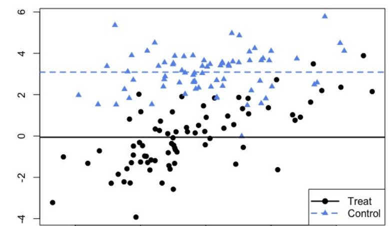
Each plot, (A)-(H), below was generated from a linear regression model fit to outcome variable Y with dummy-coded dichotomous predictor D and continuous predictor X. Match each plot with the model that generated the plot. Each plot will match with one and only one model.



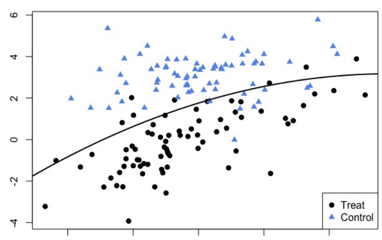
A)



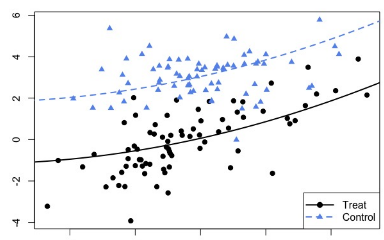
B)



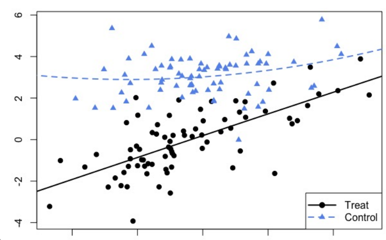
C)



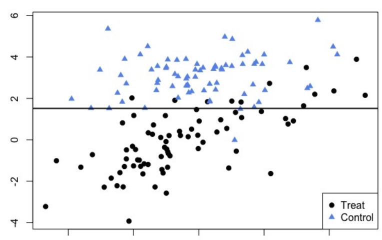
D)



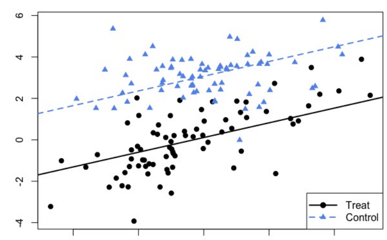
E)



F)



G)



H)

21. $Y_i = \beta_0 + \epsilon_i$ ______

22. $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ ______

23. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$ ______

24. $Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$ ______

25. $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \epsilon_i$ ______

26. $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i + \epsilon_i$ ______

27. $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 X_i^2 + \epsilon_i$ ______

28. $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 D_i X_i + \beta_5 D_i X_i^2 + \epsilon_i$ ______

4 Multiple Choice (1 pt each)

Note: Problems (29) - (31) are related. A psychologist interested in the effects of exercise on depression planned a one-factor pilot study where the outcome was measured as baseline to post-experiment improvement on a standard depression inventory. The psychologist recruited a number of participants diagnosed with major depressive disorder that had comparable levels of baseline depression and who also reported at baseline that they did not exercise regularly. Participants were randomly assigned to the three levels of the exercise factor: (a) no exercise, (b) 30 minutes every other day, (c) 60 minutes every other day.

29. Which of the following contrasts, if found to be significantly different from zero, would confirm a pairwise difference between the "no exercise" group and the "60 min" exercise group?

    (a) $\psi = 1\mu_{0 \text{ min}} + 0\mu_{30 \text{ min}} - 1\mu_{60 \text{ min}}$
    (b) $\psi = 1\mu_{0 \text{ min}} - 1\mu_{30 \text{ min}} + 0\mu_{60 \text{ min}}$
    (c) $\psi = 0\mu_{0 \text{ min}} + 1\mu_{30 \text{ min}} - 1\mu_{60 \text{ min}}$
    (d) $\psi = 1/2\mu_{0 \text{ min}} + 1/2\mu_{30 \text{ min}} - 1\mu_{60 \text{ min}}$
    (e) $\psi = 1\mu_{0 \text{ min}} - 1/2\mu_{30 \text{ min}} - 1/2\mu_{60 \text{ min}}$

A psychiatrist at the same university as the psychologist ran a pilot study to examine the efficacy of a certain SSRI drug on depression. The psychiatrist recruited similar participants as the psychologist, with the exception that participants that reported currently taking medication to treat depression were excluded. The SSRI factor consisted of four levels: (a) 0 mg (placebo pill), (b) 20 mg (considered a mild dose), (c) 40 mg (considered a moderate dose), (d) 60 mg (considered a large dose). Participants were randomly assigned to one of the four groups.

30. Which of the following contrasts, if found to be significantly different from zero, would confirm that the drug is more effective than a placebo, regardless of dose?

    (a) $\psi = 1\mu_{0 \text{ mg}} - 1/3\mu_{20 \text{ mg}} - 1/3\mu_{40 \text{ mg}} - 1/3\mu_{60 \text{ mg}}$
    (b) $\psi = 1\mu_{0 \text{ mg}} - 1\mu_{20 \text{ mg}} - 0\mu_{40 \text{ mg}} - 0\mu_{60 \text{ mg}}$
    (c) $\psi = 1\mu_{0 \text{ mg}} - 0\mu_{20 \text{ mg}} - 1/2\mu_{40 \text{ mg}} - 1/2\mu_{60 \text{ mg}}$
    (d) $\psi = 0\mu_{0 \text{ mg}} - 1/3\mu_{20 \text{ mg}} - 1/3\mu_{40 \text{ mg}} - 1/3\mu_{60 \text{ mg}}$
    (e) $\psi = 1/3\mu_{0 \text{ mg}} - 1/3\mu_{20 \text{ mg}} - 1/3\mu_{40 \text{ mg}} - 1/3\mu_{60 \text{ mg}}$

The psychologist and psychiatrist both found evidence to support the efficacy of exercise and the SSRI drug in their pilot studies. They teamed up and hypothesized that the effect of taking the SSRI at any dose would be enhanced by exercise at any level, relative to not exercising at all. They planned a two-factor experiment using the factors and levels from their respective pilot studies. Each participant was randomized to one of the twelve groups (e.g., no exercise and no drug). Suppose the researchers defined contrasts of the form:

$$\begin{aligned}\psi = {} & c_1\mu_{0 \text{ min, } 0 \text{ mg}} + c_2\mu_{0 \text{ min, } 20 \text{ mg}} + c_3\mu_{0 \text{ min, } 40 \text{ mg}} + c_4\mu_{0 \text{ min, } 60 \text{ mg}} \\ & + c_5\mu_{30 \text{ min, } 0 \text{ mg}} + c_6\mu_{30 \text{ min, } 20 \text{ mg}} + c_7\mu_{30 \text{ min, } 40 \text{ mg}} + c_8\mu_{30 \text{ min, } 60 \text{ mg}} \\ & + c_9\mu_{60 \text{ min, } 0 \text{ mg}} + c_{10}\mu_{60 \text{ min, } 20 \text{ mg}} + c_{11}\mu_{60 \text{ min, } 40 \text{ mg}} + c_{12}\mu_{60 \text{ min, } 60 \text{ mg}}.\end{aligned}$$

31. Which of the following sets of contrast coefficients, $c_1, c_2, \ldots, c_{12}$, could be used to test this hypothesis by testing $H_0 : \psi = 0$?

(a)	1	-1/3	-1/3	-1/3	-1	1/3	1/3	1/3	0	0	0	0
(b)	1	-1/3	-1/3	-1/3	-1/2	1/6	1/6	1/6	-1/2	1/6	1/6	1/6
(c)	1	0	0	-1	-1/2	0	0	1/2	-1/2	0	0	1/2
(d)	1	-1	-1	-1	-1	1	1	1	-1	1	1	1
(e)	0	1	-1/2	-1/2	0	-1	1/2	1/2	0	1	-1	0

Note: Problem (32) is not related to Problems (29) - (31).

32. Consider an outcome variable Y, dichotomous predictor D, and continuous predictor X. A model that fails to follow the principal of marginality is given by

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 D_i X_i + \epsilon_i.$$

Which statement accurately describes the modeling constraints across groups with $D = 0$ and $D = 1$ implied by the model?

(a) The intercepts may differ but the slopes are constrained to be the same.

(b) The intercepts may differ but the slope in $D = 0$ group is constrained to be zero.

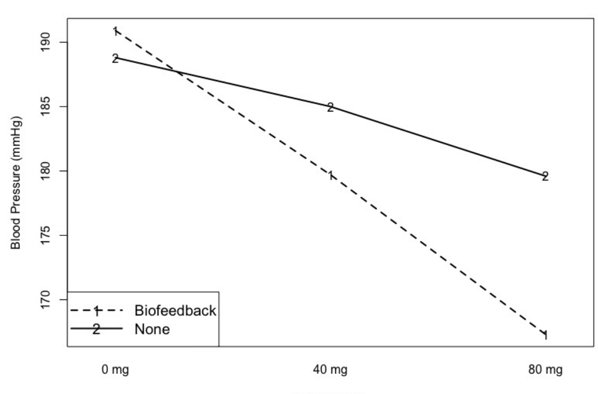
(c) The slopes may differ but intercepts are constrained to be the same.

(d) The slopes and intercepts are constrained to be the same across groups.

(e) The slopes may differ but the intercept in $D = 0$ group is constrained to be zero.

Note: Problems (33) - (36) are related. Drug therapy (characterized by the use of a pharmacological drug at a dose of either 0 mg, 40 mg, or 80 mg per day) and biofeedback (characterized by mindfulness meditation, deep breathing exercises, etc.) are two options for controlling high blood pressure. An equal number of participants were randomly assigned to each of six cells in the two-factor (2×3) design. Output from the two-way ANOVA, along with the interaction plot of cell means are given below.

```
Response: bp
             Sum Sq Df    F value    Pr(>F)
(Intercept) 1984893  1 25970.8299 < 2.2e-16
bio             400  1     5.2392   0.02602
drug           2696  2    17.6384 1.273e-06
bio:drug        519  2     3.3923   0.04095
Residuals      4127 54
```



33. What was the total sample size analyzed in this study?

(a) 60 (b) 61 (c) 48 (d) 54 (e) 59

34. Given the graphical and statistical evidence, which of the following is an appropriate next step in the analysis?

(a) Test pairwise comparisons for drug, averaging over the levels of biofeedback.

(b) Test simple pairwise comparisons for drug, conditional on the levels of biofeedback.

(c) Test main omnibus effect of drug, averaging over the levels of biofeedback.

(d) Test simple omnibus effect of drug at both levels of biofeedback.

(e) Test for an interaction between drug and biofeedback factors.

35. Let B_i represent the dummy-coded biofeedback indicator and let D_{i40} and D_{i80} represent, respectively, dummy-coded indicators for the 40 mg and 80 mg levels of the drug factor. Let BP_i represent the blood pressure outcome measure. Which of the following full and reduced model pairs may be used to test the significance of the two-way interaction via an incremental F test?

(a) $$\begin{aligned}\text{BP}_{iF} &= \beta_0 + \beta_1 B_i \times D_{i40} + \beta_2 B_i \times D_{i80} + \epsilon_{iF}\\ \text{BP}_{iR} &= \beta_0 + \epsilon_{iR}\end{aligned}$$

(b) $$\begin{aligned}\text{BP}_{iF} &= \beta_0 + \beta_1 B_i + \beta_2 B_i \times D_{i40} + \beta_3 B_i \times D_{i80} + \epsilon_{iF}\\ \text{BP}_{iR} &= \beta_0 + \beta_1 B_i + \epsilon_{iR}\end{aligned}$$

(c) $$\begin{aligned}\text{BP}_{iF} &= \beta_0 + \beta_1 B_i + \beta_2 D_{i40} + \beta_3 D_{i80} + \beta_4 B_i \times D_{i40} + \beta_5 B_i \times D_{i80} + \epsilon_{iF}\\ \text{BP}_{iR} &= \beta_0 + \beta_1 B_i + \beta_2 D_{i40} + \beta_3 D_{i80} + \epsilon_{iR}\end{aligned}$$

(d) $$\begin{aligned}\text{BP}_{iF} &= \beta_0 + \beta_1 B_i + \beta_2 D_{i40} + \beta_3 D_{i,80\text{ mg}} + \beta_4 B_i \times D_{i,40\text{ mg}} + \beta_5 B_i \times D_{i80} + \epsilon_{iF}\\ \text{BP}_{iR} &= \beta_0\end{aligned}$$

(e) $$\begin{aligned}\text{BP}_{iF} &= \beta_0 + \beta_1 B_i \times D_{i40} + \beta_2 B_i \times D_{i80} + \epsilon_{iF}\\ \text{BP}_{iR} &= \beta_0 + \beta_1 B_i + \beta_2 D_{i40} + \beta_3 D_{i80} + \epsilon_{iR}\end{aligned}$$

36. The variable names are as described in the last problem. What does the intercept parameter β_0 represent in the following model?

$$\text{BP}_i = \beta_0 + \beta_1 B_i + \beta_2 D_{i40} + \beta_3 D_{i80} + \beta_4 B_i \times D_{i40} + \beta_5 B_i \times D_{i80} + \epsilon_i.$$

(a) The mean blood pressure of the no biofeedback placebo group.

(b) The mean blood pressure of the biofeedback placebo group.

(c) The overall (grand) mean blood pressure.

(d) The overall mean blood pressure of the no biofeedback group, averaged across levels of drug.

(e) The overall mean blood pressure of the biofeedback group, averaged across levels of drug.

5 Free Response

37. (5 pts) List five properties of this table that are not aligned with APA style guidelines.

Variable	Estimate	SE	p-value
Intercept	114.29211	1.50	3.1e-14
WSES14	0.04733	.020	0.011
pg_Educ	0.03901	.033	0.475

Table 1: Regression coefficients

(a)

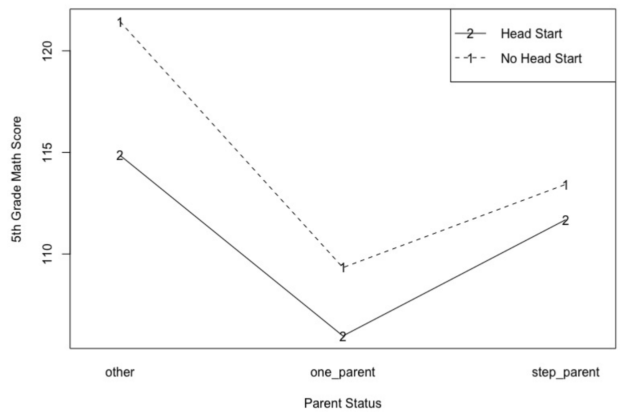
(b)

(c)

(d)

(e)

The last two question deal with the ECLS-K data we have worked with in class. The outcome is fifth grade math score on a standardized math test. The two categorical predictors for this example are household parental status (three levels) and Head Start attendance (two levels). A plot of marginal means is shown below, followed by results from the two-way ANOVA.




```
Response: C6R4MSCL
                  Sum Sq   Df    F value    Pr(>F)
(Intercept)     22260283    1 45443.8975 < 2.2e-16 ***
PAR_STATUS         71583    2    73.0681 < 2.2e-16 ***
HS                  6631    1    13.5375 0.0002355 ***
PAR_STATUS:HS       2466    2     2.5167 0.0807990 .
Residuals        3603270 7356
```

38. (4 pts) Suppose you are a statistical analyst working on this project and the principal investigator (PI) has asked you to analyze the data. In particular, the PI is most interested in the effect of parent status, and would like you to work up pairwise comparisons that test whether scores are higher for some parent status groups than others. Describe whether you will examine simple pairwise comparisons that condition on Head Start status or pairwise comparisons that average over Head Start status and justify your decisions using the graphical evidence and the output from the two-way ANOVA. Assume assumptions for valid inferences are met and do not mention them in your response.

Writing space.

39. (5 pts) Use the following output to implement the analysis plan you described above. Write a paragraph or two in which you report the relevant p-values and interpret the relevant results in context. Be sure to address the PI's research question in your response. Use APA style guidelines as appropriate.

```
> emm1 <- emmeans(lm1,
+                 specs = ~ PAR_STATUS | HS ,
+                 adjust = "none")
> joint_tests(emm1, by = "HS")
HS = HS:
 model term df1  df2 F.ratio p.value
 PAR_STATUS   2 7356  16.657 <.0001

HS = NO_HS:
 model term df1  df2 F.ratio p.value
 PAR_STATUS   2 7356 124.891 <.0001

> pairs(emm1)
HS = HS:
 contrast                 estimate    SE   df t.ratio p.value
 other - one_parent           8.89 1.544 7356  5.757  <.0001
 other - step_parent          3.16 2.586 7356  1.224  0.4392
 one_parent - step_parent    -5.72 2.613 7356 -2.190  0.0729

HS = NO_HS:
 contrast                 estimate    SE   df t.ratio p.value
 other - one_parent          12.10 0.808 7356 14.983  <.0001
 other - step_parent          8.01 1.253 7356  6.392  <.0001
 one_parent - step_parent    -4.10 1.426 7356 -2.873  0.0114

----------------------------------------------

> emm2 <- emmeans(lm1,
+                 specs = ~ PAR_STATUS,
+                 adjust = "none")
NOTE: Results may be misleading due to involvement in interactions
> emm2
 PAR_STATUS  emmean    SE   df  lower.CL upper.CL
 other          118 0.552 7356      117      119
 one_parent     108 0.674 7356      106      109
 step_parent    113 1.327 7356      110      115

Results are averaged over the levels of: HS
Confidence level used: 0.95
> pairs(emm2)
 contrast                 estimate    SE   df t.ratio p.value
 other - one_parent          10.50 0.871 7356 12.048  <.0001
 other - step_parent          5.59 1.437 7356  3.887  0.0003
 one_parent - step_parent    -4.91 1.488 7356 -3.299  0.0028

Results are averaged over the levels of: HS
```

Writing space.