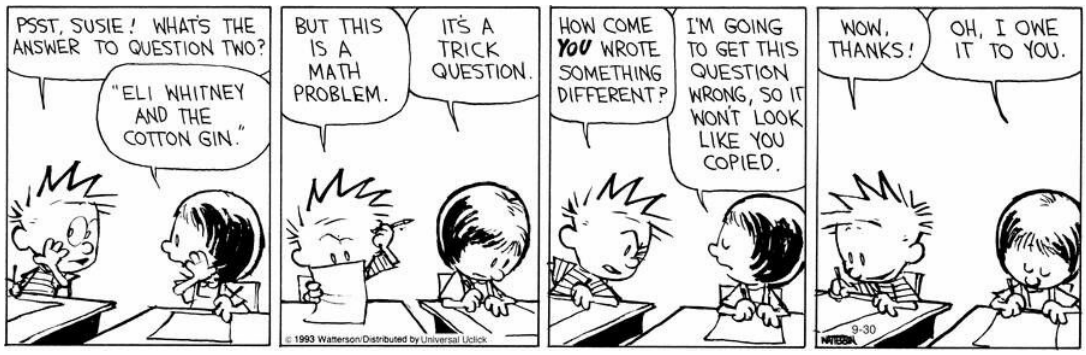
HUDM 5123 - Linear Models and Experimental Design
Mid-term Exam

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!

Note: Problems (1) - (5) are related. They deal with outcome variable Y and the following one-way ANOVA model for factor A, a four-category variable with group means μ_1, μ_2, μ_3, and μ_4:

$$Y_i = \beta_0 + \beta_1 A_{1i} + \beta_2 A_{2i} + \beta_3 A_{3i} + \epsilon_i$$

1. _____ If variables A_1, A_2, and A_3 are dummy-coded variables, with group 4 held out as the reference group, the intercept, β_0, represents the mean of all groups except the reference group; that is, $\beta_0 = (\mu_1 + \mu_2 + \mu_3)/3$.

2. _____ If variables A_1, A_2, and A_3 are deviation-coded variables, with group 4 held out as the reference group, the slope coefficient β_2 represents the deviation in means between the second and fourth groups; that is, $\beta_2 = \mu_2 - \mu_4$.

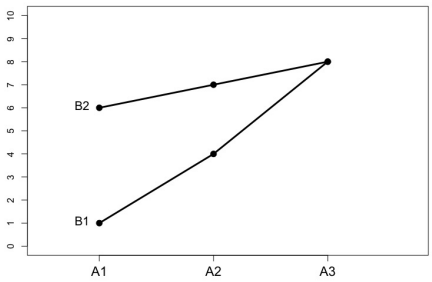
3. _____ If the sample size for the analysis is $n = 1022$, the number of degrees of freedom for the full model, given above, is 1019.

4. _____ The following reduced model could be used to test the omnibus null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ via an incremental F test:

$$Y_i = \beta_0 + \epsilon_i$$

5. _____ The predicted values based on the full model will be the same no matter whether dummy-coding or deviation-coding is used.

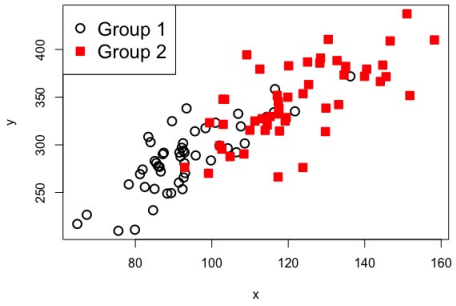
Note: Problems (6) - (11) are related. They deal with the following plot, which represents population marginal means for a two-factor design, wherein factor A has three levels and factor B has two levels.



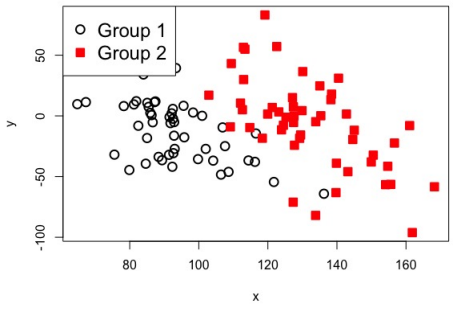
6. _____ There is an interaction between factors A and B.
7. _____ There is a main effect for factor A.
8. _____ There is a main effect for factor B.
9. _____ There is a simple main effect for factor B at level 1 of factor A.
10. _____ There is a simple main effect for factor B at level 2 of factor A.
11. _____ There is a simple main effect for factor B at level 3 of factor A.

---

12. _____ In multiple regression, multicollinearity occurs when one or more predictors can be predicted by a linear combination of other predictors in the model.
13. _____ The variance inflation factor for a particular variable, X_j, is calculated as $1/(1 - R_j^2)$, where R_j^2 is the multiple squared correlation from regressing the outcome, Y on all the predictors except X_j.
14. _____ Suppose the effect of group was estimated using ANOVA, ignoring covariate X for the data pictured below. The use of ANCOVA, adjusting for the linear relationship between X and Y, will decrease the estimate of the effect relative to ANOVA for the data plotted below.



15. _____ Suppose the effect of group was estimated using ANOVA, ignoring covariate X for the data pictured below. The use of ANCOVA, adjusting for the linear relationship between X and Y, will decrease the estimate of the effect relative to ANOVA for the data plotted below.



16. _____ In simple linear regression with outcome Y and predictor X, the total sums of squares (TSS) is the sum of squared deviations of Y from a horizontal line at the mean of X.

17. _____ The sample covariance is bounded between -1 and 1, inclusive, where a value of zero represents no linear relationship and a value of 1 or -1 represents perfect linear relationship.

18. _____ An interaction effect between factor A and factor B occurs when the effect of factor A on the outcome varies across the levels of factor B (or vice versa).

The two-way design with factors A and B, each with three levels, may be represented by cell and marginal means as follows.

	B_1	B_2	B_3	
A_1	μ_{11}	μ_{12}	μ_{13}	$\mu_{1\bullet}$
A_2	μ_{21}	μ_{22}	μ_{23}	$\mu_{2\bullet}$
A_3	μ_{31}	μ_{32}	μ_{33}	$\mu_{3\bullet}$
	$\mu_{\bullet 1}$	$\mu_{\bullet 2}$	$\mu_{\bullet 3}$	$\mu_{\bullet\bullet}$

19. _____ The following full and reduced models, if used in an incremental F test, would test the null hypothesis $H_0 : \mu_{\bullet 1} = \mu_{\bullet 2} = \mu_{\bullet 3}$ via Type 3 sums of squares.
Full model:

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 A_{1i} + \beta_2 A_{2i} + \beta_3 B_{1i} + \beta_4 B_{2i} + \\ & \beta_5 A_{1i} B_{1i} + \beta_6 A_{1i} B_{2i} + \beta_7 A_{2i} B_{1i} + \beta_8 A_{2i} B_{2i} + \epsilon_i \end{aligned}$$

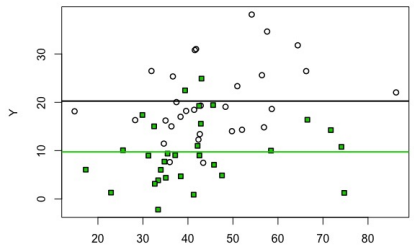
Reduced model:

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 A_{1i} + \beta_2 A_{2i} + \\ & \beta_5 A_{1i} B_{1i} + \beta_6 A_{1i} B_{2i} + \beta_7 A_{2i} B_{1i} + \beta_8 A_{2i} B_{2i} + \epsilon_i \end{aligned}$$

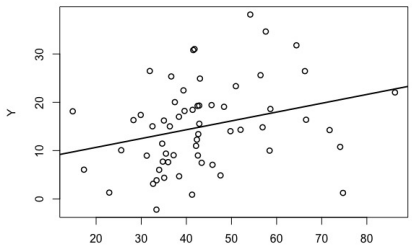
20. _____ In general, the number of degrees of freedom associated with the numerator of the incremental F test statistic is the difference in the number of parameters between the full model and the nested reduced model.

3 Matching (1 pt each)

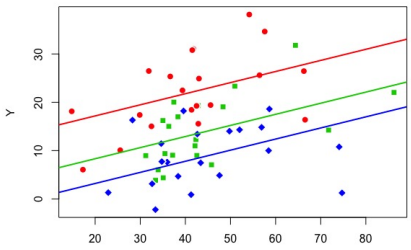
Each plot, (A)-(H), below was generated from a linear regression model fit to outcome variable Y with two categorical predictors and a continuous predictor. The categorical predictors are factor A (2 levels) and factor B (3 levels), and the continuous covariate is called X. The variable A_{1i} represents a deviation-coded predictor for level 1 of factor A with level 2 used as the reference category. The variables B_{1i} and B_{2i} represent deviation-coded predictors for levels 1 and 2, respectively, of factor B with level 3 used as the reference category. Match each plot with the model that generated the plot. Each plot will match with one and only one model.



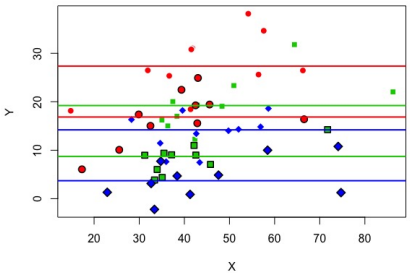
(A)



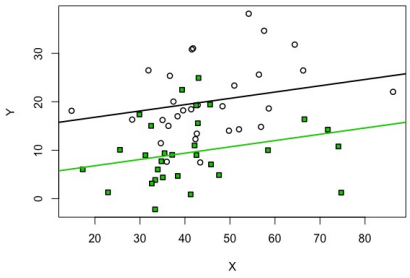
(B)



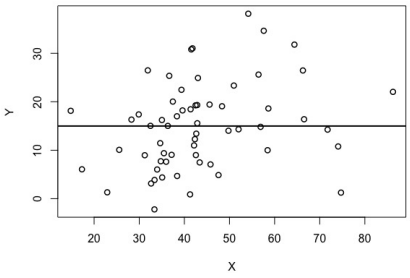
(C)



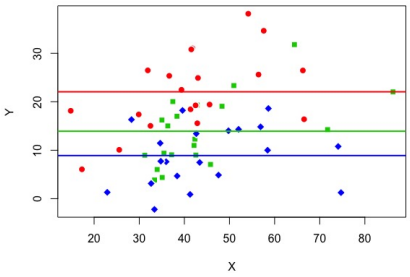
(D)



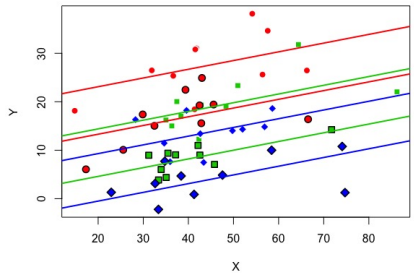
(E)



(F)



(G)



(H)

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

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

23. $Y_i = \beta_0 + \beta_1 A_{1i} + \epsilon_i$ _____

24. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_{1i} + \epsilon_i$ _____

25. $Y_i = \beta_0 + \beta_1 B_{1i} + \beta_2 B_{2i} + \epsilon_i$ _____

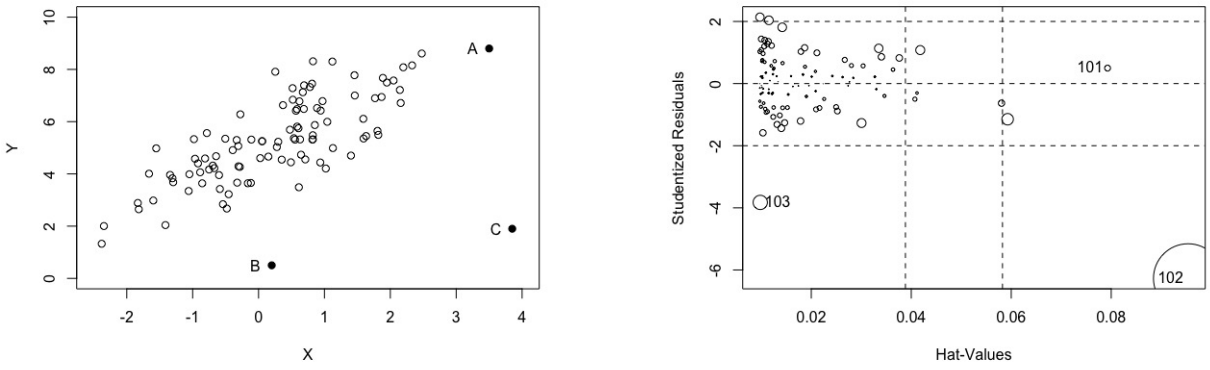
26. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 B_{1i} + \beta_3 B_{2i} + \epsilon_i$ _____

27. $Y_i = \beta_0 + \beta_1 B_{1i} + \beta_2 B_{2i} + \beta_3 A_{1i} B_{1i} + \beta_4 A_{1i} B_{2i} + \epsilon_i$ _____

28. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 B_{1i} + \beta_3 B_{2i} + \beta_4 A_{1i} B_{1i} + \beta_5 A_{1i} B_{2i} + \epsilon_i$ _____

4 Multiple Choice (1 pt each)

Note: Problems (29) - (31) are related to the plots shown below. The scatter plot in the left panel below shows the bivariate relationship between variables X and Y. The influence plot from package **car** is shown in the right panel below. Take note of the three points labeled A, B, and C in the scatterplot and the three points labeled 101, 102, and 103 in the influence plot.



29. Which statement accurately characterizes the relationships between the concepts of discrepancy, influence, and leverage with the quantitative measures of Cook's distance, hat values, and studentized residuals?

    (a) Cook's distance is a measure of discrepancy, hat values are a measure of leverage, and studentized residuals are a measure of influence.

    (b) Cook's distance is a measure of leverage, hat values are a measure of discrepancy, and studentized residuals are a measure of influence.

(c) Cook's distance is a measure of leverage, hat values are a measure of influence, and studentized residuals are a measure of discrepancy.

(d) Cook's distance is a measure of influence, hat values are a measure of discrepancy, and studentized residuals are a measure of leverage.

(e) Cook's distance is a measure of influence, hat values are a measure of leverage, and studentized residuals are a measure of discrepancy.

30. Which of the following describes the correspondence between points A, B, and C in the scatterplot and points 101, 102, and 103 in the influence plot?

(a) A = 101; B = 102; C = 103

(b) A = 101; B = 103; C = 102

(c) A = 102; B = 101; C = 103

(d) A = 103; B = 101; C = 102

(e) A = 103; B = 102; C = 101

31. Which of the three points labeled A, B, and C will be have the largest difference between its unstandardized (raw) residual and its studentized residual?

(a) point A

(b) point B

(c) point C

(d) differences are identical for points A, B, and C

(e) studentized residuals cannot be reliably determined for points A, B, and C

Note: Problems (32) - (34) are related. Transitional cell carcinoma is the most frequently observed cancer of the bladder and urinary tract in dogs. A veterinary oncologist designed a study to compare three chemotherapeutic treatment options: (1) daily oral dose of piroxicam, (2) weekly IV administration of mitoxantrone, and (3) weekly injection of vinblastine. A placebo group was also included in the study. The outcome for the study was estimated tumor volume reduction on MRI scan at termination of the chemotherapy regime. Note, group 1 is piroxicam, group 2 is mitoxantrone, group 3 is vinblastine, and group 4 is placebo. Suppose that the study enrolled 13, 20, 15, and 26 dogs, respectively, in each group.

32. Which of the following contrasts, if found to be significantly different from zero, would confirm that the chemotherapy treatments were, on average, more or less effective than placebo?

(a) $\psi = 1\mu_1 + 1\mu_2 + 1\mu_3 - 1\mu_4$

(b) $\psi = 1\mu_1 + 0\mu_2 + 0\mu_3 - 1\mu_4$

(c) $\psi = 1/3\mu_1 + 1/3\mu_2 + 1/3\mu_3 - 1\mu_4$

(d) $\psi = 0\mu_1 + 0\mu_2 + 0\mu_3 - 1\mu_4$

(e) $\psi = 1/4\mu_1 + 1/4\mu_2 + 1/4\mu_3 - 1\mu_4$

33. The formula for the t statistic for a contrast test is

$$t_0 = \hat{\psi} \Big/ \sqrt{MS_{\text{error}} \sum_{j=1}^{a} \frac{c_j^2}{n_j}}$$

For ψ_1, defined as $\psi_1 = 1\mu_1 + 0\mu_2 + 0\mu_3 - 1\mu_4$, what is the value of $\sum_{j=1}^{a} \frac{c_j^2}{n_j}$?

(a) 7/60

(b) 3/26

(c) 28/195

(d) 0

(e) 23/260

34. Consider an outcome variable Y, dichotomous predictor D, and continuous predictor X and the following regression model that involves them.

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

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.

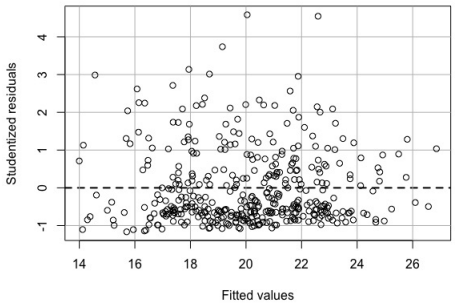
(c) The slopes may differ but intercepts are constrained to be the same.

(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.

---

35. The plot below shows the fitted values from a regression fit plotted against the studentized residuals for the same fit.These data were generated from a linear model that satisfies all assumptions for valid inferences about regression coefficients we have discussed except one, which is clearly violated. Which assumption is most clearly violated based on the graphical evidence in the plot?



(a) Constant variance of errors

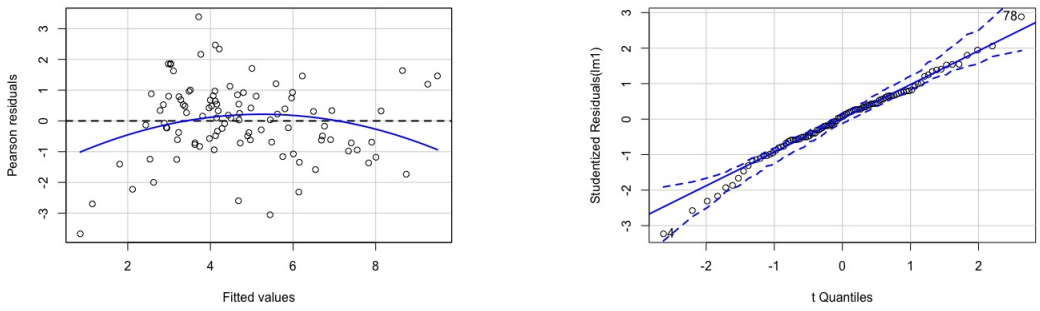
(b) Normality of errors

(c) Independence of errors

(d) Linearity

(e) Multicollinearity

36. The studentized residual vs fitted values plot (left) and qq plot (right) for a linear model fit are shown below. Which statement most accurately characterizes conclusions about assumptions for valid inferences under the linear regression model that may be drawn based on the plots?



(a) Normality of errors **is not** a concern; model mispecification **is not** a concern

(b) Normality of errors **is not** a concern; model mispecification **is** a concern

(c) Normality of errors **is** a concern; model mispecification **is not** a concern

(d) Normality of errors **is** a concern; model mispecification **is** a concern

(e) The plots do not provide enough information to make a judgement

5 Free Response

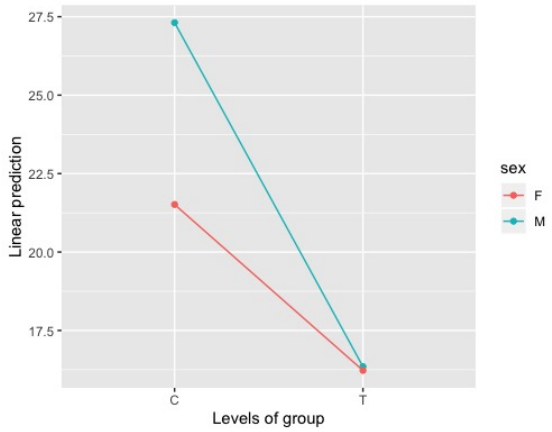
37. (5 pts) The table below shows regression analysis results for the linear regression of 5th grade math achievement score on socioeconomic status (WKSES14) and parent's education level (pg_Educ) for a sample of 573 children. Redraw the table below, correcting any violations to APA style in your version.

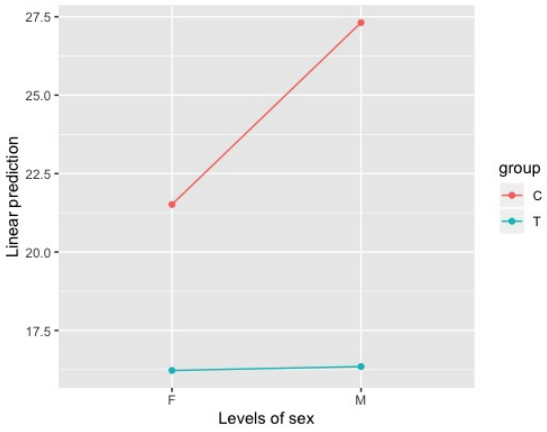
Variable	Estimate	SE	p-value
Intercept	114.29211	1.50	3.1e-14
WSES14	0.04733	.020	0.011
pg_Educ	0.03901	.033	0.475

Table 1: Regression coefficients

Writing space

The last two questions deal with the acupuncture data we have worked with in class. The outcome is headache severity rating one year after being enrolled in a randomized study to examine the efficacy of acupuncture treatment (vs a control group that did not receive acupuncture) in treating pain and discomfort due to headache disorder. The two categorical predictors for this example are treatment (1 = acupuncture; 0 = control) and sex (1 = female; 0 = male). Interaction plots are shown below, followed by results from the two-way ANOVA.






```
Anova Table (Type III tests)

Response: pk5
             Sum Sq  Df  F value    Pr(>F)
(Intercept)  65240   1 277.7699 < 2.2e-16
sex            345   1   1.4691 0.2264443
group         2601   1  11.0751 0.0009851
sex:group      317   1   1.3498 0.2462466
Residuals    69757 297
```

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 the acupuncture treatment, and would like you to work up pairwise comparisons that test whether headache severity ratings are lower for participants treated with acupuncture. Describe whether you will examine simple pairwise comparisons that condition on participant sex or pairwise comparisons that average over both sexes 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.

```
Results are averaged over the levels of: sex
> emm1 <- emmeans(object = lm1,
+                 specs = ~ group)
NOTE: Results may be misleading due to involvement in interactions
> emm1
 group emmean   SE  df lower.CL upper.CL
 C       24.4 1.85 297     20.8     28.1
 T       16.3 1.59 297     13.2     19.4

Results are averaged over the levels of: sex
Confidence level used: 0.95
> pairs(emm1)
 contrast estimate   SE  df t.ratio p.value
 C - T        8.13 2.44 297 3.328   0.0010

Results are averaged over the levels of: sex

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

> emm2 <- emmeans(object = lm1,
+                 specs = ~ group | sex)
> joint_tests(object = emm2,
+             by = "sex")
sex = F:
 model term df1 df2 F.ratio p.value
 group        1 297   7.516 0.0065

sex = M:
 model term df1 df2 F.ratio p.value
 group        1 297   5.971 0.0151

> pairs(x = emm2)
sex = F:
 contrast estimate   SE  df t.ratio p.value
 C - T        5.29 1.93 297 2.741   0.0065

sex = M:
 contrast estimate   SE  df t.ratio p.value
 C - T       10.96 4.49 297 2.444   0.0151
```

Writing space.