MATH 4044 – Statistics for Data Science

Practical Week 9 Solutions

Question 1

The data for this practical is stored in a SAS data file called `hsb2.sas7bdat` located in `mydata` library on the SAS OnDemand server.

This data file contains 200 observations from a sample of high school students with demographic information about the students, such as their gender (`female`), socio-economic status (`ses`) and ethnic background (`race`). It also contains a number of scores (out of 100) on standardized tests, including tests of reading (`read`), writing (`write`), mathematics (`math`) and social studies (`socst`).

Note: All the analysis that follows is subject to the necessary conditions being satisfied, e.g. Normality, independence, equality of variance etc. Condition checking is left as an exercise.

(a) Perform and analyse a factorial ANOVA model to determine whether there is statistically significant difference in average writing scores by gender and socio-economic status. Include tests for interaction. Interpret the results.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2278.24419	455.64884	5.67	<.0001
Error	194	15600.63081	80.41562		
Corrected Total	199	17878.87500			

R-Square	Coeff Var	Root MSE	write Mean
0.127427	16.99190	8.967476	52.77500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ses	2	1063.252697	531.626349	6.61	0.0017
female	1	1334.493311	1334.493311	16.59	<.0001
ses*female	2	21.430904	10.715452	0.13	0.8753

Table 1. Results of factorial ANOVA relating write to `ses` and `female`

Based on results in Table 1, the overall model is highly significant, F(5,194) = 5.67, P-value < 0.0001. The R-squared is quite small at 0.1274 so there is considerable variability in writing scores.

There is a significant main effect due to socio-economic status, F(2,194) = 6.61, P-value = 0.0017, and gender, F(1,194) = 16.59, P-value < 0.0001. There is no evidence of interaction, P(2,194) = 0.13, P-value = 0.8753.

From the parameters estimates in Table 2, the mean writing score for females with socio-economic status classified as level three is 58.97 (t = 35.41, P-value < 0.0001). The mean score for females with socio-economic status level one is 6.47 points lower (t = -2.81, P-value = 0.0054). The mean score for females with socio-economic status level two is also lower compared to level three, by 4.72 points (t = -2.24, P-value = 0.0265). Given socio-economic status level three, the mean writing score for males is 6.10 points lower than for females (t = -2.59, P-value = 0.0103).

Differences between mean writing scores for females and males with socio-economic status one and two relative to the difference between them assuming socio-economic status three are not statistically significant (P-values 0.9558 and 0.6384, respectively).

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		58.96551724	B	1.66521846	35.41	<.0001
ses	1	-6.46551724	B	2.29911738	-2.81	0.0054
ses	2	-4.71551724	B	2.10909411	-2.24	0.0265
ses	3	0.00000000	B	.	.	.
female	0	-6.10344828	B	2.35497453	-2.59	0.0103
female	1	0.00000000	B	.	.	.
ses*female	1 0	0.20344828	B	3.66332291	0.06	0.9558
ses*female	1 1	0.00000000	B	.	.	.
ses*female	2 0	1.40663977	B	2.98867884	0.47	0.6384
ses*female	2 1	0.00000000	B	.	.	.
ses*female	3 0	0.00000000	B	.	.	.
ses*female	3 1	0.00000000	B	.	.	.

Table 2. Factorial ANOVA solution table

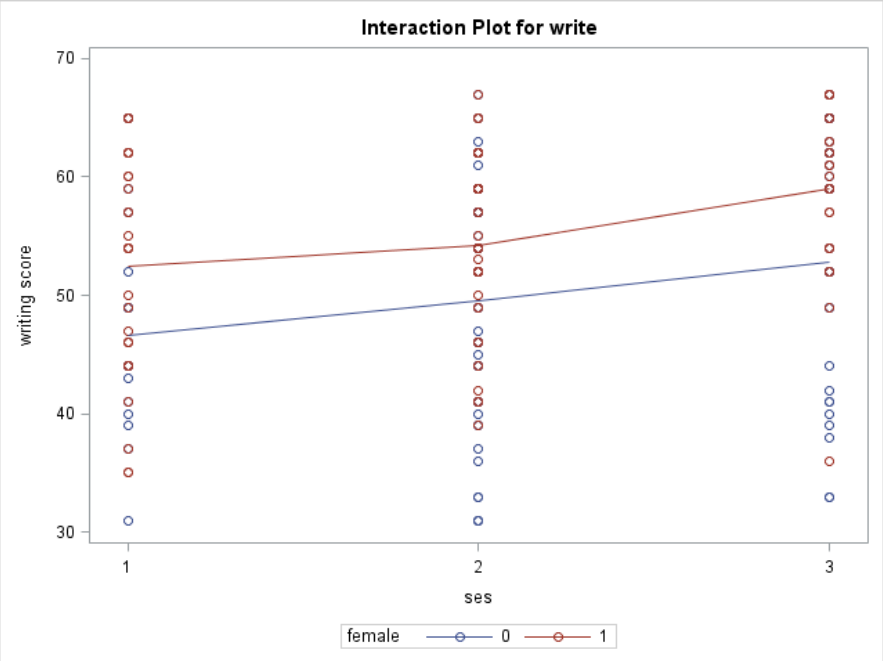

Figure 1. Interaction plot for the factorial ANOVA model in Table 1

The interaction plot confirms that while there are significant main effects for gender and socio-economic status, there is no significant interaction. Mean writing scores for females are higher than for males all levels of `ses`, by a similar amount.

ses*female Effect Sliced by female for write					
female	DF	Sum of Squares	Mean Square	F Value	Pr > F
0	2	419.005033	209.502516	2.61	0.0765
1	2	683.025308	341.512654	4.25	0.0157

Table 3. Simple effects of `ses`

Based on Table 3, differences in mean writing scores by socio-economic status are statistically significant for females (P-value = 0.0157) but not for males, P-value = 0.0765.

ses*female Effect Sliced by ses for write					
ses	DF	Sum of Squares	Mean Square	F Value	Pr > F
1	1	355.506383	355.506383	4.42	0.0368
2	1	523.867189	523.867189	6.51	0.0115
3	1	540.155172	540.155172	6.72	0.0103

Table 4. Simple effects of `female`

From Table 4, the effect of gender on writing scores is statistically significant for each socio-economic status since all P-values < 0.05.

Based on Table 5, the difference between socio-economic status one and two is not statistically significant (P-value = 0.3421). Differences between socio-economic status level three and the other two levels are however statistically significant, with P-values of 0.0018 (level three versus level one) and 0.0214 (level three vs level two).

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

ses	write LSMEAN	LSMEAN Number
1	49.5500000	1
2	51.9015957	2
3	55.9137931	3

Least Squares Means for effect ses Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: write			
i/j	1	2	3
1		0.3421	0.0018
2	0.3421		0.0214
3	0.0018	0.0214	

Table 5. Post-hoc comparisons for `ses` controlling for `female`

From Table 6, there is a statistically significant difference in means by gender (P-value < 0.0001).

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

female	write LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
0	49.6717535	<.0001
1	55.2385057	

Table 6. Post-hoc comparisons for `female` controlling for `ses`

As the interaction term between gender and socio-economic status was not statistically significant, there is no need to examine post-hoc comparisons for the corresponding means.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

ses	female	write LSMEAN	LSMEAN Number
1	0	46.6000000	1
1	1	52.5000000	2
2	0	49.5531915	3
2	1	54.2500000	4
3	0	52.8620690	5
3	1	58.9655172	6

Least Squares Means for effect ses*female
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: write

i/j	1	2	3	4	5	6
1		0.2903	0.8766	0.0493	0.2445	0.0003
2	0.2903		0.7064	0.9565	1.0000	0.0597
3	0.8766	0.7064		0.1144	0.6241	0.0002
4	0.0493	0.9565	0.1144		0.9862	0.2263
5	0.2445	1.0000	0.6241	0.9862		0.1042
6	0.0003	0.0597	0.0002	0.2263	0.1042	

Table 7. Post-hoc comparisons for interaction between `female` and `ses`

(b) Define dummy variables for the variable `ses`. Relate `write` to `ses` and `female` (which already is a dummy variable) including interaction using multiple regression. Discuss your results and compare to results from part (a).

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2278.24419	455.64884	5.67	<.0001
Error	194	15601	80.41562		
Corrected Total	199	17879			

Root MSE	8.96748	R-Square	0.1274
Dependent Mean	52.77500	Adj R-Sq	0.1049
Coeff Var	16.99190		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	52.86207	1.66522	31.74	<.0001
female		1	6.10345	2.35497	2.59	0.0103
ses1		1	-6.26207	2.85202	-2.20	0.0293
ses2		1	-3.30888	2.11753	-1.56	0.1198
ses1_female		1	-0.20345	3.66332	-0.06	0.9558
ses2_female		1	-1.40664	2.98868	-0.47	0.6384

Table 8. Regression results for `write` with `ses` and `female` as explanatory variables

Based on results in Table 8, the overall model is statistically significant, F(5,194), P-value < 0.0001. The adjusted R-squared indicates that gender, socio-economic status and interaction term together explain only 10.49% of variability in writing scores.

The base category is males with socio-economic status three, represented by the intercept. Slopes represent differences in mean writing scores from the base category. Note that parameter estimates and P-values in Table 8 here are equivalent to those in the parameter estimates table from PROC GLM shown in Table 2.

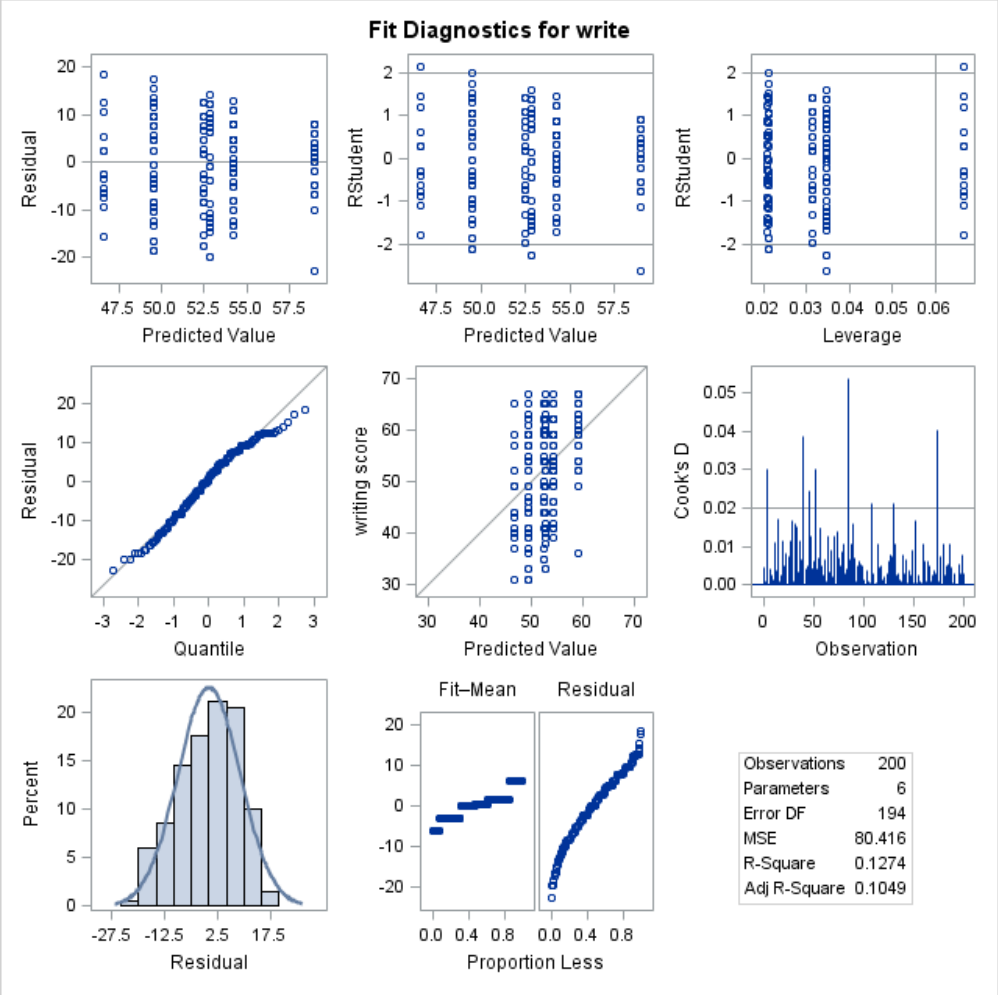

Figure 2. Diagnostics plots for the regression model in Table 6

Diagnostics plots in Figure 2 suggest some violations of assumptions necessary for linear regression, such as homoscedasticity, independence and Normality of residuals.

APPENDIX – SAS code

```
ods graphics on;

proc glm data=work.hsb2;
    class ses female;
    model write=ses | female / ss3 solution;
    lsmeans ses | female / pdiff adjust=tukey;
    /* Simple interaction effects */
    lsmeans female*ses / slice=female;
    lsmeans female*ses / slice=ses;
    run;
quit;

data work.hsb2_dummies;
    set work.hsb2;
    if ses=1 then ses1=1; else ses1=0;
    if ses=2 then ses2=1; else ses2=0;
    if ses=1 and female=1 then ses1_female = 1;
    else ses1_female = 0;
    if ses=2 and female=1 then ses2_female = 1;
    else ses2_female = 0;
    run;

proc reg data=work.hsb2_dummies plots=diagnostics;
    model write=female ses1 ses2 ses1_female ses2_female;
    run;
quit;

ods graphics off;
```