MATH 4044 – Statistics for Data Science

Practical Week 8 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) Carry out a one-way analysis of variance relating `write` to `prog` (program type). Check the necessary conditions and discuss the results.

Recall that the conditions to check for ANOVA are (1) independence of samples (2) Normality of the distributions for the underlying populations and (3) equality of variance for the distributions of the underlying populations. Assumption checking is left as an exercise.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3175.69786	1587.84893	21.27	<.0001
Error	197	14703.17714	74.63542		
Corrected Total	199	17878.87500			

R-Square	Coeff Var	Root MSE	write Mean
0.177623	16.36983	8.639179	52.77500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
prog	2	3175.697857	1587.848929	21.27	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
prog	2	3175.697857	1587.848929	21.27	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		46.76000000	B	1.22176444	38.27	<.0001
prog	1	4.57333333	B	1.77518257	2.58	0.0107
prog	2	9.49714286	B	1.48442643	6.40	<.0001
prog	3	0.00000000	B	.	.	.

Table 1. Results of one-way ANOVA relating `write` to `prog`

Based on results in Table 1, the overall model is highly significant, $F(2,197) = 21.27$, P-value < 0.0001.

Tukey's Studentized Range (HSD) Test for write

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	197
Error Mean Square	74.63542
Critical Value of Studentized Range	3.33976

Comparisons significant at the 0.05 level are indicated by ***.

prog Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
2 - 1	4.924	1.289	8.559	***
2 - 3	9.497	5.992	13.003	***
1 - 2	-4.924	-8.559	-1.289	***
1 - 3	4.573	0.381	8.766	***
3 - 2	-9.497	-13.003	-5.992	***
3 - 1	-4.573	-8.766	-0.381	***

Table 2. Post-hoc comparisons for `write` by `prog`

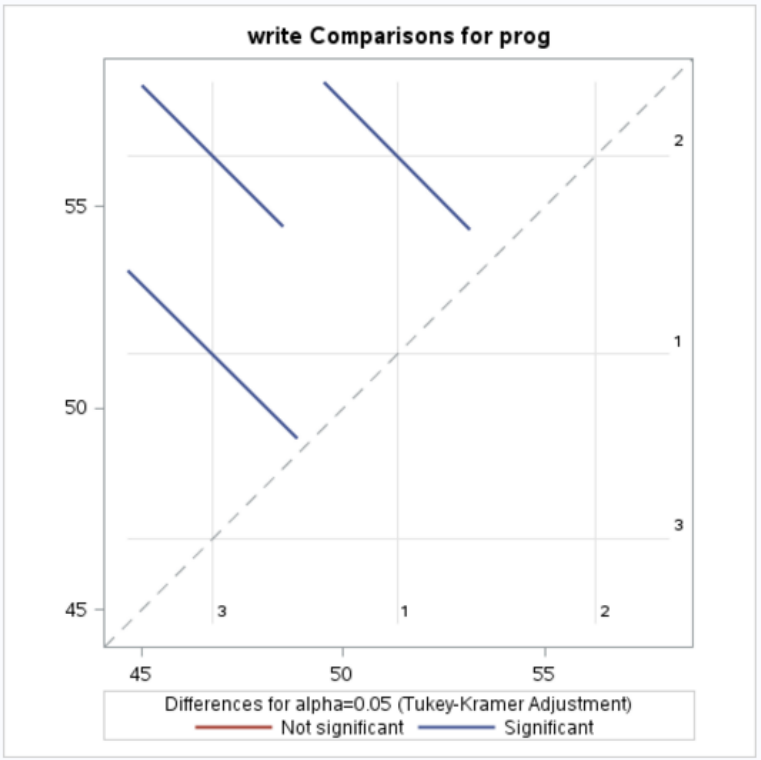

Figure 1. Diffogram corresponding to Post-hoc comparisons for `write` by `prog`

From the post-hoc comparison results shown in Table 2 or diffogram in Figure 1, mean writing scores by program are all statistically different from each other.

(b) Check the necessary assumptions and carry out an analysis of covariance relating writing scores to program type controlling for reading scores. Discuss your results.

Pearson Correlation Coefficients, N = 200 Prob > |r| under H0: Rho=0		
	read	**write**
read	1.00000	0.59678
reading score		<.0001
write	0.59678	1.00000
writing score	<.0001	

Table 3. Correlation analysis results

Figure 2. Scatterplot to illustrate the relationship between writing and reading scores

As reported in Table 3, the correlation between reading and writing scores is r = 0.5968. Since the P-value < 0.0001, this correlation is statistically significant. There is therefore a positive linear relationship between reading and writing scores, confirmed by the scatterplot in Figure 2, so it makes sense to include reading scores as a covariate.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3716.86127	1858.43063	21.28	<.0001
Error	197	17202.55873	87.32263		
Corrected Total	199	20919.42000			

Table 4. Results of ANOVA relating `read` and `prog`

From Table 4, the model relating reading scores to program type overall is statistically significant (P-value < 0.0001), which means that the independence assumption has been violated. We are not able to argue reduction in error variance due to reading scores as a covariate since it is actually confounding the outcome, at least partially.

Results of the ANCOVA model assuming equality of slopes are shown in Table 5 and Figure 3. Based on results shown in Table 5, the model overall is statistically significant, F(5,194) = 42.21 and P-value < 0.0001.

Using partial sums of squares, program type is still statistically significant, F(2,194) = 5.87, P-value = 0.0034, although the effect has been diminished (compare sums of squares and F-value). Reading scores as a covariate are highly statistically significant, F(1,194) = 69.33,

P-value < 0.0001. All else equal, the writing score increases by 0.4726 points for one point increase in the reading score. Fitted relationships between reading and writing scores by program type, without allowing for interaction, are shown in Figure 4.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7017.68123	2339.22708	42.21	<.0001
Error	196	10861.19377	55.41425		
Corrected Total	199	17878.87500			

R-Square	Coeff Var	Root MSE	write Mean
0.392512	14.10531	7.444075	52.77500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
prog	2	650.259965	325.129983	5.87	0.0034
read	1	3841.983376	3841.983376	69.33	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	24.92650849	B	2.82558196	8.82	<.0001
prog 1	2.89302615	B	1.54286626	1.88	0.0623
prog 2	4.78928219	B	1.39846996	3.42	0.0007
prog 3	0.00000000	B	.	.	.
read	0.47258640		0.05675632	8.33	<.0001

Table 5. ANCOVA results relating `write` to `read` and `prog`, including interaction

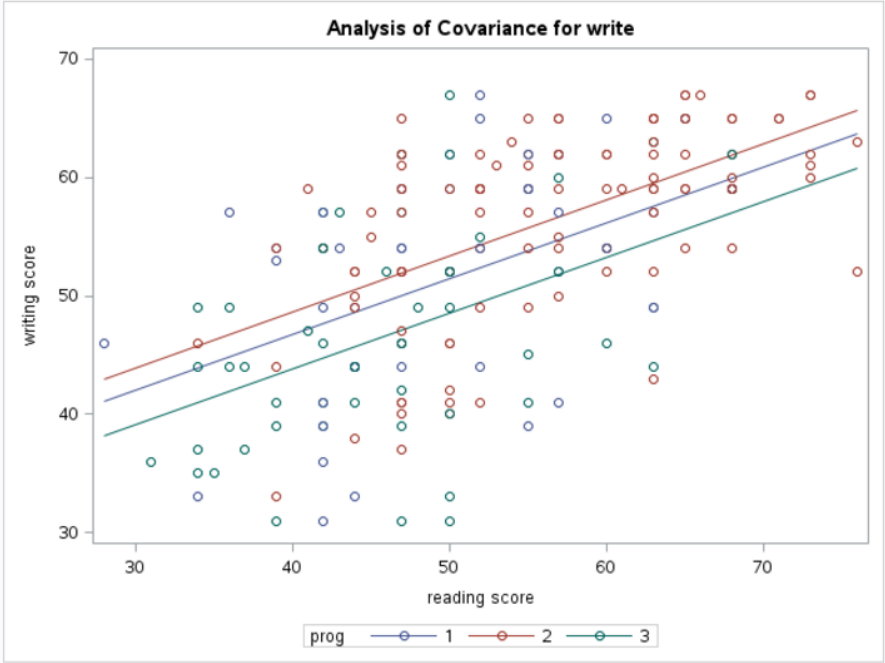

Figure 3. Fitted lines assuming equality of slopes

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7019.05483	1403.81097	25.08	<.0001
Error	194	10859.82017	55.97845		
Corrected Total	199	17878.87500			

R-Square	Coeff Var	Root MSE	write Mean
0.392589	14.17693	7.481875	52.77500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
prog	2	36.465910	18.232955	0.33	0.7224
read	1	3253.618935	3253.618935	58.12	<.0001
read*prog	2	1.373596	0.686798	0.01	0.9878

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	24.45138889	B	5.64364279	4.33	<.0001
prog 1	2.88580774	B	8.36820611	0.34	0.7306
prog 2	5.71288411	B	7.13082258	0.80	0.4240
prog 3	0.00000000	B	.	.	.
read	0.48287037	B	0.11999063	4.02	<.0001
read*prog 1	-0.00058982	B	0.17121958	-0.00	0.9973
read*prog 2	-0.01826950	B	0.14230903	-0.13	0.8980
read*prog 3	0.00000000	B	.	.	.

Table 6. ANCOVA results relating `write` to `read` and `prog`, including interaction

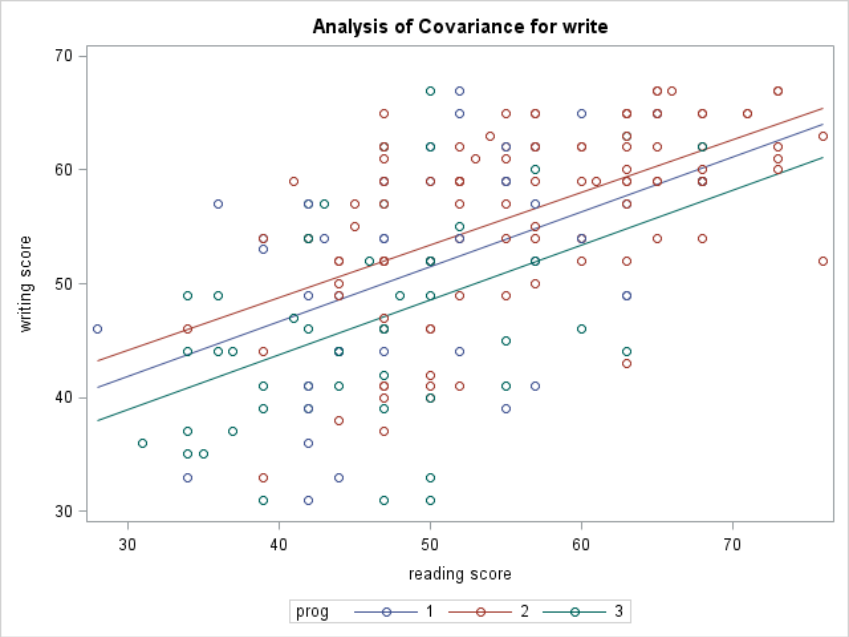

Figure 4. Fitted lines without assuming equality of slopes

We now proceed to assess the appropriateness of the equality of slopes assumption implicitly made in fitting the previous model. An ANCOVA model with an interaction term is fitted next, with results shown in Table 6. Based on results shown in Table 6, the model overall is statistically significant, $F(5,194) = 25.08$ and P-value < 0.0001.

Using partial sums of squares, program type is not statistically significant, $F(2,194) = 0.33$, P-value = 0.7224. On the other hand, reading scores as a covariate are highly statistically significant, $F(1,194) = 58.12$, P-value < 0.0001. All else equal, the writing score increases by 0.4829 points for one point increase in the reading score.

The interaction term is not statistically significant, $F(2,194) = 0.01$, P-value = 0.9878, so the assumption of homogeneity of slopes has been satisfied. Fitted relationships between reading and writing scores by program type, allowing for interaction, are shown

in Figure 4. Note that the lines are virtually parallel in this case, and we should revert to Figure 3 as a more appropriate graphical representation of the relationships.

Question 2

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

Suppose we have collected the following data:

- Individual's income (`cash`);
- Importance of charity to the individual (`import`);
- Amount given to charity (`given`);
- Gender (`gender`), where 0 represents females and 1 represents males.

Is there a difference in the amount given to charity by men and women?

(a) Carry out a one-way analysis of variance relating `given` to `gender`. Check the necessary conditions and discuss the results.

In this scenario we are dealing with two independent samples and since for all Normality tests in Tables 7 and 8 the P-values are greater than 0.05, samples for both genders can be assumed to have come from Normal populations.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.958471	Pr < W	0.4590
Kolmogorov-Smirnov	D	0.126321	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.052171	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.341369	Pr > A-Sq	>0.2500

Table 7. Results of Normality tests for females

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.967716	Pr < W	0.7538
Kolmogorov-Smirnov	D	0.103241	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.031645	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.239197	Pr > A-Sq	>0.2500

Table 8. Results of Normality tests for males

Results of Levine test of homogeneity of variance are shown in Table 10. There is no statistically significant difference in variance for the two groups; $F(1,38) = 0.09$, P-value = 0.7608 > 0.05. Therefore, all conditions for ANOVA are satisfied.

Based on results in Table 9, the overall ANOVA model is statistically significant; $F(1,38) = 17.40$, P-value = 0.0002 < 0.05.Therefore, there is a significant difference in the amount given to charity by males and females. Based on the estimates in the solution section of

Table 9, we can conclude further that that females donate more than males. The mean amount given for males is 40.72, while for females it is 40.72 + 13.32 = 54.05.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1757.334343	1757.334343	17.40	0.0002
Error	38	3838.565657	101.014886		
Corrected Total	39	5595.900000			

R-Square	Coeff Var	Root MSE	Given Mean
0.314040	20.91700	10.05062	48.05000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gender	1	1757.334343	1757.334343	17.40	0.0002

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	40.72222222	B	2.36895295	17.19	<.0001
Gender Female	13.32323232	B	3.19429551	4.17	0.0002
Gender Male	0.00000000	B	.	.	.

Table 9. Results of ANOVA relating `given` to `gender`

Levene's Test for Homogeneity of Given Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Gender	1	1559.6	1559.6	0.11	0.7390
Error	38	526189	13847.1		

Table 10. Levine's test of homogeneity of variance for `given`

(b) Check the necessary assumptions and carry out an analysis of covariance relating the amount given to gender controlling for importance (`import`). Discuss your results.

We first check whether there is a relationship between importance and the amount given. From Table 11, the correlation between the amount given and importance is r = 0.38. Since the P-value = 0.0149 < 0.05, this correlation is statistically significant. The scatterplot in Figure 5 shows a positive relationship between the amount given and importance, so it makes sense to include importance as a covariate.

Pearson Correlation Coefficients, N = 40 Prob > |r| under H0: Rho=0		
	Given	Import
Given	1.00000	0.38237 0.0149
Import	0.38237 0.0149	1.00000

Table 11. Correlation table for `given` and `import`

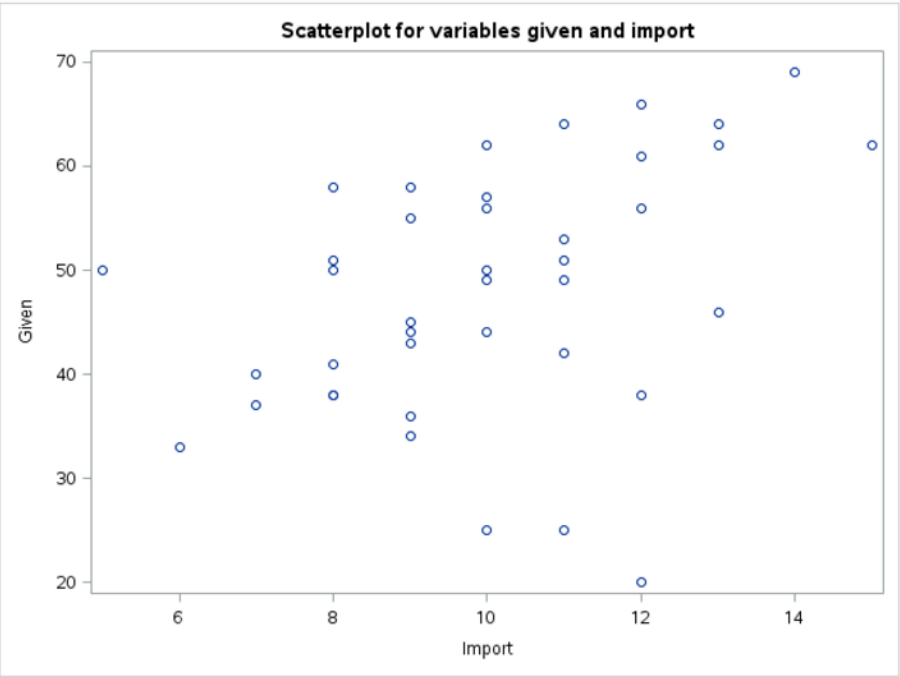

Figure 5. Scatterplot of `import` against `given`.

From Table 12, there is no statistically significant difference in mean importance between males and females (assuming unequal variances, P-value = 0.1453 > 0.05), which means that the independence assumption cannot be rejected. We may therefore be able to argue reduction in error variance due to importance as a covariate.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10.1010101	10.1010101	2.21	0.1456
Error	38	173.8989899	4.5762892		
Corrected Total	39	184.0000000			

R-Square	Coeff Var	Root MSE	Import Mean
0.054897	21.39226	2.139226	10.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gender	1	10.10101010	10.10101010	2.21	0.1456

Table 12. Results of ANOVA for `import` by `gender`

Results of an ANCOVA model with import as a covariate are shown in Table 13. Based on results shown in Table 13, the ANCOVA model overall is statistically significant, F(2,37) = 11.37 and P-value = 0.0001.

Using partial sums of squares (Type III SS), `gender` (F(1,38) = 14.01, P-value = 0.0006) is statistically significant while `import` (F(1,38) = 3.99, P-value = 0.0533) is not, however only marginally. All else equal, the amount given increases by 1.46 for one point increase in the importance score. Fitted regression lines are shown in Figure 6.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2130.577277	1065.288638	11.37	0.0001
Error	37	3465.322723	93.657371		
Corrected Total	39	5595.900000			

R-Square	Coeff Var	Root MSE	Given Mean
0.380739	20.14084	9.677674	48.05000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Import	1	373.242934	373.242934	3.99	0.0533
Gender	1	1312.403364	1312.403364	14.01	0.0006

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	26.88580390	B	7.29675305	3.68	0.0007
Import	1.46503253		0.73387561	2.00	0.0533
Gender Female	11.84340149	B	3.16383456	3.74	0.0006
Gender Male	0.00000000	B	.	.	.

Table 13. Results of ANCOVA relating `given` to `import` and `gender`

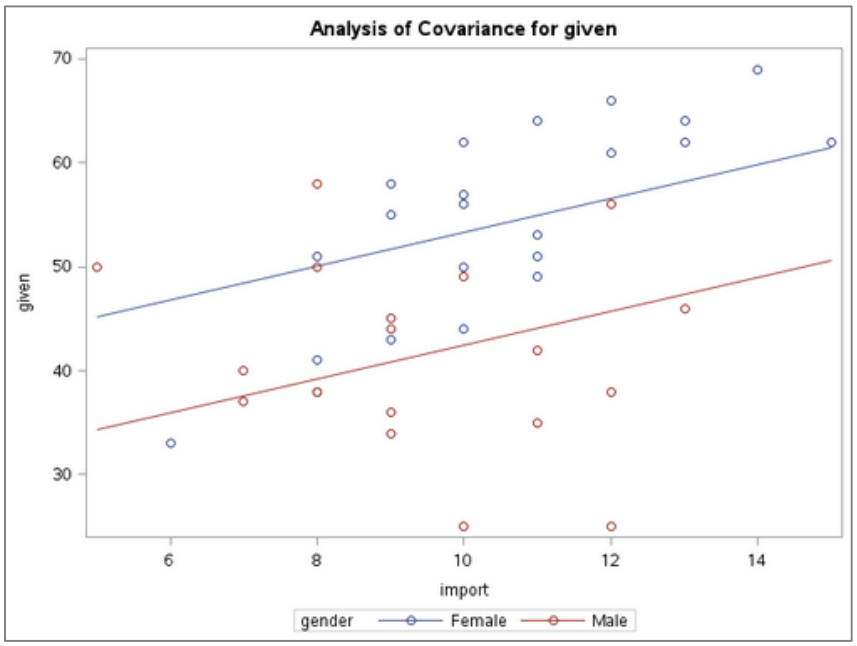

Figure 6. Fitted lines assuming equality of slopes

Results of ANCOVA with an interaction term are shown in Table 14. The interaction term is statistically significant, F(1,38) = 13.98, P-value = 0.0006 < 0.05, so the assumption of homogeneity of slopes is violated in this scenario.

Parameter estimates in the solution section of Table 14 indicate that for males, the amount given is actually negatively related to the amount given, with the latter decreasing by 1.20 per unit increase in importance. For females, the amount given increases by 3.56 (= 4.76 – 1.20) per unit of importance. When the importance score is zero, males donate on average 52.06, while females on average give 16.87 (= 52.06 –

35.19). Fitted relationships between importance and amount given by gender are shown in Figure 7.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3099.826328	1033.275443	14.90	<.0001
Error	36	2496.073672	69.335380		
Corrected Total	39	5595.900000			

R-Square	Coeff Var	Root MSE	Given Mean
0.553946	17.32942	8.326787	48.05000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Import	1	237.6709838	237.6709838	3.43	0.0723
Gender	1	518.3191663	518.3191663	7.48	0.0096
Import*Gender	1	969.2490507	969.2490507	13.98	0.0006

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	52.06104651	B	9.20621744	5.65	<.0001
Import	-1.20058140	B	0.95236732	-1.26	0.2156
Gender Female	-35.19164353	B	12.87118003	-2.73	0.0096
Gender Male	0.00000000	B	.	.	.
Import*Gender Female	4.75655154	B	1.27218996	3.74	0.0006
Import*Gender Male	0.00000000	B	.	.	.

Table 14. Results of ANCOVA relating `given` to `import` and `gender` allowing for interaction

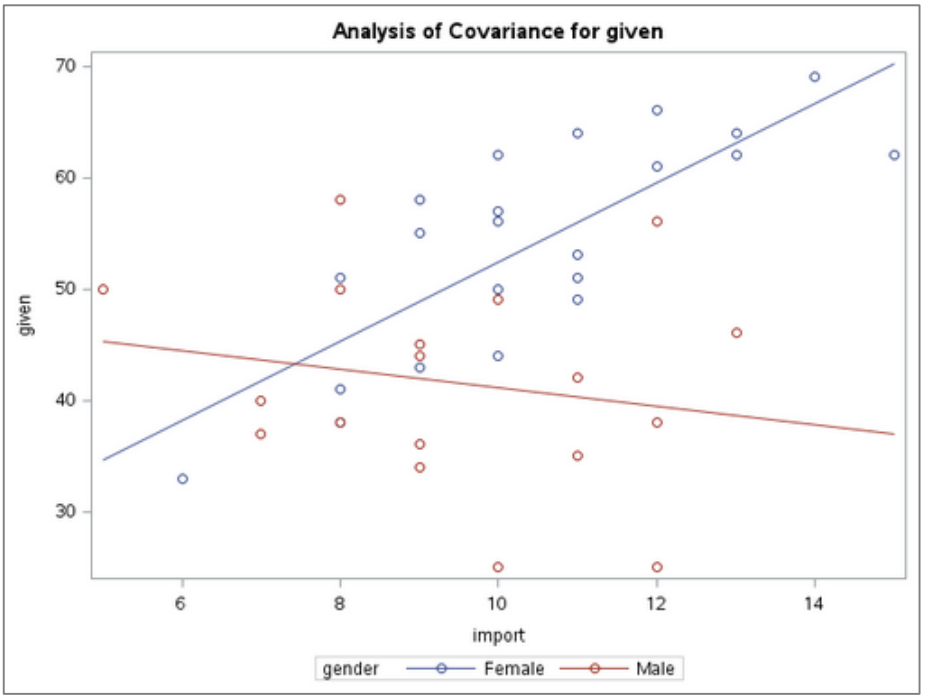

Figure 7. Fitted regression lines without assuming equal slopes

Note: If you find these results somewhat counter-intuitive, you may find it useful to re-run the analysis with variable `cash` as a covariate to better understand this data and what it appears to say about differences between males and females in terms of the amount given to charity. This is left as an exercise. The relevant code is given in the Appendix.

APPENDIX – SAS code

```
ods graphics on;

/* Question 1 */

title "ANOVA for write vs prog";

proc glm data=mydata.hsb2;
      class prog;
      model write=prog / solution;
      means prog / hovtest welch tukey;
      lsmeans prog / pdiff adjust=tukey;
      run;
quit;

title "Correlation analysis for variables read and write";

proc corr data=mydata.hsb2;
      var read write;
run;

title "Scatterplot for variables read and write";

proc sgplot data=mydata.hsb2;
      scatter x=read y=write;
run;

title "ANOVA for read vs prog (to test independence of covariate
from treatment
      effect)";

proc glm data=mydata.hsb2;
      class prog;
      model read=prog / solution;
      run;
quit;

title "ANCOVA for write vs prog and read, without interaction";

proc glm data=mydata.hsb2;
      class prog;
      model write=prog read / solution ss3;
      run;
quit;

title "ANCOVA for write vs prog and read, including interaction (to
test for
```

```
      homogeneity of slopes)";

proc glm data=mydata.hsb2;
      class prog;
      model write=prog read prog*read / solution ss3;
      run;
quit;
title;

/* Question 2 */

proc format ;
      value genderF 0='Female' 1='Male';
run;

title "Checking Normality";

proc univariate data=mydata.charity normal;
      var given;
      class gender;
      ods select testsfornormality;
      format gender genderF.;
run;

title "ANOVA for given vs gender";

proc glm data=mydata.charity;
      class gender;
      model given=gender / solution ss1;
      means gender / hovtest welch tukey;
      lsmeans gender / pdiff adjust=tukey;
      format gender genderF.;
      run;
quit;

title "Correlation analysis for variables given and import";

proc corr data=mydata.charity;
      var given import;
run;

title "Scatterplot for variables given and import";

proc sgplot data=mydata.charity;
      scatter x=import y=given;
run;

title "ANOVA for import vs gender (to check independence of
covariate from treatment effect)";

proc glm data=mydata.charity;
      class gender;
      model import=gender / ss1;
      means gender / hovtest welch;
      format gender genderF.;
      run;
```

```
quit;

title "ANCOVA for given vs gender with import as a covariate";

proc glm data=mydata.charity;
      class gender;
      model given=import gender / solution ss3;
      lsmeans gender / pdiff adjust=tukey;
      format gender genderF.;
      run;
quit;

title "ANCOVA for given vs gender and import with an interaction
term (to check homogeneity of slopes assumption)";

proc glm data=mydata.charity;
      class gender;
      model given=import gender import*gender / solution ss3;
      lsmeans gender / pdiff adjust=tukey;
      format gender genderF.;
      run;
quit;

title "The same analysis repeated with cash as a covariate";

proc sgplot data=mydata.charity;
      scatter x=cash y=given / group=gender;
      format gender genderF.;
run;

proc corr data=mydata.charity nosimple;
      var given cash;
      format gender genderF.;
run;

proc glm data=mydata.charity;
      class gender;
      model given=cash gender / solution ss3;
      lsmeans gender / pdiff adjust=tukey;
      format gender genderF.;
      run;
quit;

proc glm data=mydata.charity;
      class gender;
      model given=cash gender cash*gender / solution ss3;
      lsmeans gender / pdiff adjust=tukey;
      format gender genderF.;
      run;
quit;

ods graphics off;
```