Solution to exercise 3.2

We first present the

SAS-program used for solving the problems in the exercise.

```
Title 'Exercise 3.2';
Data stat;
input schoolno P A TO;
cards;
1 286 155.0 2127
2 58 85.0 442
3 75 97.5 516
4 167 196.4 1458
5 22 24.6 130
;
run;
data newstat;
set stat;
sqrtA=A**0.5;
PsqrtA=P*sqrtA;
run;

proc print data=newstat;
run;

ods graphics on;
Title 'Model IA: TO=PsqrtA P';
proc glm data=newstat;
model TO=PsqrtA P/solutions;
run;
Title 'Model IB: TO=P PsqrtA';
proc glm data=newstat;
model TO=P PsqrtA /solutions;
run;
Title 'Model II: TO=PsqrtA sqrtA';
proc glm data=newstat;
model TO=PsqrtA sqrtA/solutions;
run;
Title 'Simpler Model: TO=PsqrtA (with intercept)';
proc glm data=newstat;
model TO=PsqrtA /solutions;
run;
Title 'Simpler Model: TO=PsqrtA (without intercept)';
proc glm data=newstat;
model TO=PsqrtA /noint solutions;
run;

ods graphics off;
```

The data

Obs	schoolno	P	A	TO	sqrtA	PsqrtA
1	1	286	155.0	2127	12.4499	3560.67
2	2	58	85.0	442	9.2195	534.73
3	3	75	97.5	516	9.8742	740.57
4	4	167	196.4	1458	14.0143	2340.38
5	5	22	24.6	130	4.9598	109.12

The effect of different ordering of class variables and corresponding model terms

Model IA: TO = PsqrtA P

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2758393.917	1379196.958	1047.51	0.0010
Error	2	2633.283	1316.642		
Corrected Total	4	2761027.200			

R-Square	Coeff Var	Root MSE	TO Mean
0.999046	3.882469	36.28556	934.6000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PsqrtA	1	2758171.473	2758171.473	2094.85	0.0005
P	1	222.444	222.444	0.17	0.7209

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PsqrtA	1	37839.25719	37839.25719	28.74	0.0331
P	1	222.44421	222.44421	0.17	0.7209

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	108.8486968	34.25236314	3.18	0.0864
PsqrtA	0.6208036	0.11580217	5.36	0.0331
P	-0.6481748	1.57694079	-0.41	0.7209

Model IB: TO = P PsqrtA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2758393.917	1379196.958	1047.51	0.0010
Error	2	2633.283	1316.642		
Corrected Total	4	2761027.200			

R-Square	Coeff Var	Root MSE	TO Mean
0.999046	3.882469	36.28556	934.6000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
P	1	2720554.660	2720554.660	2066.28	0.0005
PsqrtA	1	37839.257	37839.257	28.74	0.0331

Source	DF	Type III SS	Mean Square	F Value	Pr > F
P	1	222.44421	222.44421	0.17	0.7209
PsqrtA	1	37839.25719	37839.25719	28.74	0.0331

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	108.8486968	34.25236314	3.18	0.0864
P	-0.6481748	1.57694079	-0.41	0.7209
PsqrtA	0.6208036	0.11580217	5.36	0.0331

The only difference between the two sets of output is the values for the Type I SSs. They depend on the ordering of the class variables and therefore we use the Type III SSs as default.

1. and 2. Estimating parameters in Model I

From the above output we see that

$$\begin{bmatrix}\hat{a}\\ \hat{b}\\ \hat{c}\end{bmatrix} = \begin{bmatrix}108.8486968\\ 0.6208036\\ -0.6481748\end{bmatrix}$$

$$\hat{\sigma}^2 = 1316.642$$

4. Estimating parameters in Model II

Model II: TO=PsqrtA sqrtA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2759832.631	1379916.316	2310.32	0.0004
Error	2	1194.569	597.284		
Corrected Total	4	2761027.200			

R-Square	Coeff Var	Root MSE	TO Mean
0.999567	2.614958	24.43940	934.6000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PsqrtA	1	877347.0425	877347.0425	1468.89	0.0007
sqrtA	1	1661.1587	1661.1587	2.78	0.2373

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	25.79550949	46.86575992	0.55	0.6373
PsqrtA	0.55391765	0.01445274	38.33	0.0007
sqrtA	10.06519472	6.03541532	1.67	0.2373

$$\begin{bmatrix}\hat{d}\\ \hat{f}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}25.79550949\\ 0.55391765\\ 10.06519472\end{bmatrix}$$

$$\hat{\sigma}^2 = 597.284$$

3. Testing whether c in model I is equal to 0

We must now consider Model I without the $c\cdot P$ term. We get

Simpler Model: TO=PsqrtA (with intercept)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2758171.473	2758171.473	2897.52	<.0001
Error	3	2855.727	951.909		
Corrected Total	4	2761027.200			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	98.97899480	20.76941101	4.77	0.0175
PsqrtA	0.57348451	0.01065390	53.83	<.0001

The test statistic for the hypothesis that c is =0 in Model I:"TO= P PsqrtA (with intercept)" becomes – now denoting the model without the $\boldsymbol{c}\cdot P$ –term (ie, the model TO=PsqrtA (with intercept)) "Hypothesis":

$$\frac{(SS_{res}(Hyp)-SS_{res}(Mod))/(DF_{res}(Hyp)-DF_{res}(Mod))}{SS_{res}(Mod)/DF_{res}(Mod)} = \frac{(2855.727-2633.283)/1}{2633.283/2} = \frac{222.444/1}{2633.283/2} = 0.168948$$

which is equal F-value for the P-effect in the table with the Type III SS in the larger model. Therefore the above computation of the test statistic is superfluous: We obtain the test statistic directly from the Type III table in the larger model. We see that the hypothesis is accepted with a wide margin since the probability of getting a more extreme F-value is 0.7209.

3. (continued) Testing whether a in model I is equal to 0

Simpler Model: TO=PsqrtA (without intercept)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7103938.392	7103938.392	1161.03	<.0001
Error	4	24474.608	6118.652		
Uncorrected Total	5	7128413.000			

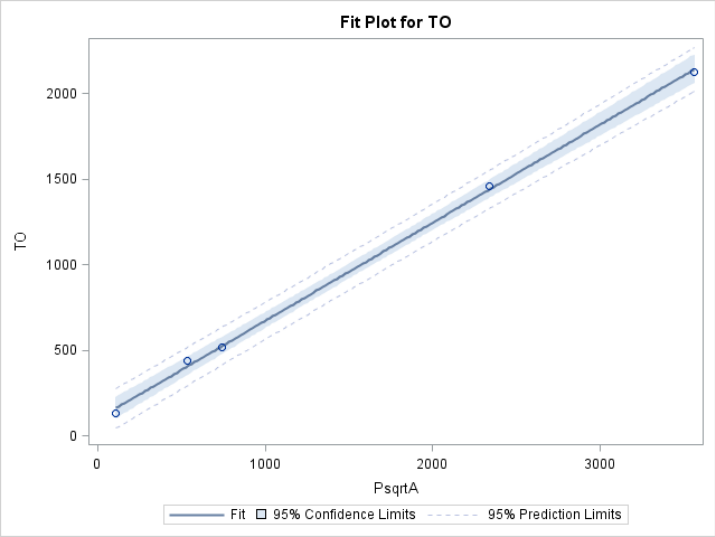
The test statistic for no intercept in the model:"TO=PsqrtA (with intercept)" becomes – now denoting the model without intercept "Hypothesis":

$$\frac{(SS_{res}(Hyp)-SS_{res}(Mod))/(DF_{res}(Hyp)-DF_{res}(Mod))}{SS_{res}(Mod)/DF_{res}(Mod)} = \frac{(24474.608-2855.727)/1}{2855.727/3} = 22.7111 = 4.77^2$$

which is equal to the squared value of the t Value for assessing whether the intercept in the model "TO=PsqrtA (with intercept)" may be assumed to be equal to 0. Like before we do not need to estimate the parameters in the last model. We automatically get the relevant test statistic from the output for the larger model with the intercept. Since the probability of getting a more extreme test statistic under the hypothesis is 0.0175, smaller than 0.05, we shall reject the hypothesis, and the resulting model becomes

$$\widehat{TO} = 108.8486968 + 0.6208036 \times P\sqrt{A}$$

The estimated line is shown below



4. (continued) Testing whether g in model II is equal to 0

From the Type III SS it follows that we may assume that g is equal to 0 (the probability of a more extreme value of the test statistic is 0.2373 > 0.05).

Accepting this, the reduced Models I and II become identical and therefore the test for the intercept will be the same.

5. Is one model better than the other?

We cannot test this since none of the models is a simplification of the other, but since they contain the same number of parameters one might prefer the one with the largest squared multiple correlation coefficient R^2, which very marginally is model I. However, the risk of over-fitting is of course imminent when we have estimated 3 parameters based on 5 observations so using R^2in this situation is not advisable.

■