Programming Exercises

24. *Hint:* Think of a way to group the countries without using a DATA step. Use a CLASS statement to tell SAS to compute the descriptive statistics by United States versus non-United States.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

\*\* Create a format to group non-US countries;

PROC FORMAT;

VALUE $Cgroup 'United States' = 'US'

OTHER = 'non-US';

RUN;

\*\* PROC MEANS needs to have the desired statistics

specified;

PROC MEANS DATA = sasdata.bigcomp MEAN STDDEV MEDIAN

MIN MAX N;

CLASS Country;

VAR Assets;

FORMAT Country $Cgroup.;

TITLE 'Descriptive Statistics of Company Assets';

RUN;

\*\* Part b);

\*\* PROC UNIVARIATE will calculate the desired statistics

and more;

PROC UNIVARIATE DATA = sasdata.bigcomp;

CLASS Country;

VAR Assets;

FORMAT Country $Cgroup.;

TITLE 'Descriptive Statistics of Company Assets';

RUN;

(sections 9.1, 9.3, 9.4)

25. *Hint:* Which procedure will help you flip the layout of the data set so that you can conduct the correct hypothesis tests on plaque*?* Using the appropriate BY and/or CLASS statements will allow you to carry out the requested hypothesis tests with one procedure for each part.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

PROC TTEST DATA = sasdata.vite HO = 140 SIDES = U;

WHERE Visit = 0;

VAR Sbp;

\*\* The CLASS statement is not allowed for a one sample

test, however, using a BY statement will break

analysis into strata groups;

BY Strata;

TITLE 'Test of Mean Plaque Greater Than 140 mm/Mg';

RUN;

\*\* Part b);

\*\* An inspection of the layout of the data set shows

that plaque values for each visit appear on a

different row. In order to carry out a paired t test

for before and after differences, the data must be

flipped so that the plaque data values appear in

before and after columns with corresponding plaque

measurements on the same row;

\*\* Part c);

PROC TRANSPOSE DATA = sasdata.vite out = tvite;

WHERE Visit IN (0,2);

VAR Plaque;

BY ID Strata Treatment;

ID Visit;

RUN;

PROC SORT DATA = tvite;

BY Strata Treatment;

RUN;

PROC TTEST DATA = tvite;

PAIRED \_0 \* \_2;

\*\* Using BY to break analysis up into strata groups;

BY Strata Treatment;

TITLE1 'Tests of Difference in Plaque at Baseline';

TITLE2 'and Two Year Visit, by Strata and Treatment';

RUN;

\*\* Part d);

\*\* Calculate a difference variable;

DATA tvitediff;

SET tvite;

Diff = \_0-\_2;

RUN;

PROC TTEST DATA = tvitediff;

VAR Diff;

CLASS Treatment;

BY Strata;

TITLE1 'Tests of Difference in Plaque';

TITLE2 'Across Treatment Groups, by Strata';

RUN;

\*\* Part e);

\*\* PROC UNIVARIATE is included for the normality test

p-value. The graphics are included in the output of

parts c) and d);

PROC UNIVARIATE DATA = tvitediff NORMAL;

VAR Diff;

CLASS Strata Treatment;

TITLE 'Normality of Differences';

RUN;

\*\* Part f);

\*\* The individual paired tests in part c) suggest that

there is a true mean difference in plaque before and

after for the placebo group (p = 0.0471) and

treatment group (p < 0.0001) of strata one. In strata

two there is no significant difference in mean plaque

for either of the treatment groups. The independent

test in part d) suggests that the difference in

change of plaque thickness between the placebo and

treatment groups of strata one is significantly

different (p = 0.0411), while there is no significant

difference in strata two. This vitamin appears to

work well for people with 0.60mm+ baseline plaque

measurements;

(sections 9.1, 9.4, 9.5)

26. *Hint:* The code for these tests can be written with one PROC step.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Parts a), b), c), and d);

PROC FREQ DATA = sasdata.lefties;

TABLE Hand \* (Foot Mouse Gender) /

CHISQ PLOTS = FREQPLOT;

TITLE1 'Associations between Hand Preference and';

TITLE2 'Foot and Mouse Preference, and Gender';

RUN;

\*\* Part e);

\*\* The test for hand versus foot does not provide

evidence that there is a statistically significant

association between writing hand and kicking foot

preference (chi-square p = 0.3261);

\*\* The test for hand versus mouse does provide evidence

that there is a statistically significant association

between writing hand and mouse hand preference. The

warning below the statistics table tells us that the

chi-square test may not be valid due to small

expected cell counts, therefore Fisher's exact test

should be reported (Fisher's p < 0.0001);

\*\* The test for hand versus gender does not provide

evidence that there is a statistically significant

association between writing hand and gender (chi-

square p = 0.1640);

(sections 9.6, 9.7)

27. *Hint:* Using PROC CONTENTS or viewing the data set interactively will help you identify the order of the variables. You do not need a DATA step to limit an analysis to certain observations.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

PROC CORR DATA = sasdata.aptest;

WHERE Total > 0;

VAR TotalPassedPct Spending;

TITLE1 'Correlation Between Total Passing';

TITLE2 'and Total Spending';

RUN;

\*\* Part b);

\*\* There is a significant positive correlation

between overall passing percentage and per pupil

total spending (p = 0.0021). The sample size for the

correlation n = 47 is due to one state (Wyoming)

having no students take the AP exam, and two states

(Mississippi and South Dakota) having missing data

for total passing percentage;

\*\* Part c);

PROC CONTENTS DATA = sasdata.aptest;

RUN;

PROC CORR DATA = sasdata.aptest;

WHERE Total > 0;

VAR Spending SalaryWages EmployeeBenefits Instruct

InstructSalaryWages InstructBenefits Support

SupportPupil SupportInstruct SupportGenAdmin

SupportSchoolAdmin;

WITH NumSchools Total TotalPassedPct

FemalePassedPct AfricanAmPassedPct

FemaleAAPassedPct HispanicPassedPct

FemaleHispPassedPct;

TITLE 'Correlation of AP Exam Results by Spending';

RUN;

\*\* Part d);

PROC REG DATA = sasdata.aptest;

WHERE Total > 0;

MODEL TotalPassedPct = EmployeeBenefits;

TITLE1 'Simple Linear Regression of Total Passing';

TITLE2 'on Employee Benefits';

RUN;

\*\* Part e);

\*\* Alaska appears to have extreme values with Total

employee benefit payments (per pupil) equal to 7330

and Percent of students passed equal to 100.0;

\*\* Part f);

PROC REG DATA = sasdata.aptest;

WHERE Total > 0 and State ~= 'Alaska';

MODEL TotalPassedPct = EmployeeBenefits;

TITLE1 'Simple Linear Regression of Total Passing';

TITLE2 'on Employee Benefits Excluding Alaska';

RUN;

\*\* In part d) the slope of the regression line

is 0.00458, and the R-squared value is 0.2574.

In part f) the slope of the regression line

is 0.00321, and the R-squared value is 0.1019;

(sections 9.8, 9.9, 9.10, 9.11)

28. *Hint:* Use a function to perform the base 10 log transformation. Use an ODS TRACE statement to examine the output objects for the analysis.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part g);

%MACRO regr(indep=);

\*\* Part a);

PROC REG DATA = sasdata.patents;

WHERE Patents >= 100;

MODEL Patents = &indep;

TITLE1 'Simple Linear Regression of Patents';

TITLE2 "on &indep";

RUN;

\*\* Part b);

\*\* The slope for the regression analysis is

statistically significant (p = 0.0094), however,

there appear to be some issues with the assumptions

required for the test. For example, the residual

plots cast doubt on a constant variance and

normality. There is one observation, Santa Clara

county, that has an extreme Y value;

\*\* Part c);

DATA log;

SET sasdata.patents;

WHERE Patents >= 100;

Logy = LOG10(Patents);

Logx = LOG10(&indep);

RUN;

\*\* Parts c);

ODS TRACE ON;

PROC REG DATA = log;

\*\* Part e);

ODS OUTPUT ResidualPlot = ResidualPlot&indep;

MODEL Logy = Logx;

TITLE1 'Simple Linear Regression of Log Patents';

TITLE2 "on Log &indep";

RUN;

ODS TRACE OFF;

\*\* Part d);

\*\* The slope for the logs regression analysis is

statistically significant (p < 0.0001) with a

weak R squared value. Log of education alone may not

be the only predictor of log patents. The assumptions

seem to be somewhat improved with the residuals

following a more random pattern and the distribution

looking less skewed. Another transformation may be

more appropriate;

\*\* Part e);

PROC UNIVARIATE DATA = ResidualPlot&indep NORMAL;

VAR Residual;

TITLE 'Normality of Residuals for Log Model';

RUN;

\*\* Part f);

\*\* All of the normality tests of the residuals from PROC

UNIVARIATE confirm that the residuals are still not

normally distributed. This model may not be the most

appropriate for these data;

%MEND;

\*\* Part h);

ODS PDF FILE =

'c:\MyPDFFiles\PatentsAnalysis.pdf';

%regr(indep = Education)

%regr(indep = Income)

%regr(indep = Age)

ODS PDF CLOSE;

(sections 9.1, 9.10, 9.11)

29. *Hint:* Consider modifying the layout of the data set so that you can conduct the appropriate hypothesis tests on route time for the three plans.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

PROC MEANS DATA=sasdata.bus MEAN MEDIAN STDDEV MAXDEC=1;

VAR Plan1 Plan2 Plan3;

TITLE 'Descriptive Statistics of Bus Plans';

RUN;

\*\* Part b);

DATA buslong;

SET sasdata.bus (IN = p1 RENAME = (Plan1 = Time)

KEEP = Plan1 Day)

sasdata.bus (IN = p2 RENAME = (Plan2 = Time)

KEEP = Plan2 Day)

sasdata.bus (IN = p3 RENAME = (Plan3 = Time)

KEEP = Plan3 Day);

IF p1 THEN Plan = 1;

IF p2 THEN Plan = 2;

IF p3 THEN Plan = 3;

RUN;

PROC ANOVA DATA = buslong;

CLASS Plan;

MODEL Time = Plan;

\*\* Part c);

MEANS Plan / SCHEFFE;

TITLE 'Mean Differences of Bus Plans';

RUN;

\*\* Part d);

\*\* The overall F test suggests that there is a true mean

difference in at least two of the plans (p < 0.0001).

Based on the multiple comparisons we would recommend

plan one as it has the lowest mean (20.3 minutes)

route time and is significantly different from the

other two plans;

(sections 9.12, 9.13)