References

1. Cavalieri P, Marovich P, Patetta MJ, Walsh S, Bond C, SAS Institute (2000) Statistics I: Introduction to anova, regression, and logistic regression: course notes. SAS Institute, Cary, NC
2. Daniel WW (2006) Biostatistics: a foundation for analysis in the health sciences 8th edition with SPSS software CD Rom 14.0 set (Wiley series in probability and statistics). Wiley, New York
3. Darroch J (1997) Biologic synergism and parallelism. Am J Epidemiol 145(7): 661–668
4. Delwiche LD, Slaughter SJ (2008) The little SAS book: a primer. SAS Institute, Cary, NC
5. Hennekens CH, Buring JE, Mayrent SL (1987) Epidemiology in medicine. Lippincott Williamns & Wilkins, Philadelphia
6. Hosmer DW, Lemeshow S (2000) Applied logistic regression (Wiley series in probability and statistics). Wiley, New York
7. Kelsey JL, Whittemore AS, Evans AS, Thompson WD (1996) Methods in observational epidemiology. Oxford University Press, Oxford, New York
8. Kleinbaum DG, Kupper LL, Nizam A, Muller KE (2007) Applied regression analysis and multivariable methods (Duxbury applied). Duxbury Press, North Scituate, MA
9. Patetta MJ, Amrhein J (2005) Categorical data analysis using logistic regression: course notes. SAS Institute, Cary, NC
10. Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott Williamns & Wilkins, Philadelphia
11. Schlesselman JJ (1982) Case-control studies: design, conduct, analysis. Oxford University Press, Oxford, New York
12. Susser ES, Schwartz S, Morabia A, Bromet E (2006) Psychiatric epidemiology: searching for the causes of mental disorders. Oxford University Press, Oxford, New York

C. DiMaggio, *SAS for Epidemiologists: Applications and Methods*,
DOI 10.1007/978-1-4614-4854-9,

Solutions

Chapter 2

2.1 Submitting Commands and Reading Output

```
/** compare ages upstate and nyc ami admissions with ttest **/

libname p8483 " ";

proc ttest data=p8483.demo$_1$; * identify the data
set you are working with;
class nyc; * identify the variable you are using to
identify and compare groups;
var age; /* continuous variable you are analyzing*/
title ’Comparing ages nyc and upstate ami patients’;
*title for your output;
run;


/* univariate analysis of age */

proc univariate data=p8483.demo$_1$;
var age;
histogram; *returns histogram of age variable;
probplot;
run;
```

- The mean age of upstate AMI patients is 70.
- The mean age of NYC AMI patients 69.
- The difference is statistically significant. ($p < 0.0001$)
- It is not likely to be clinically important.

2.4 Using PROC CONTENTS To View Variables

- There are 502,492 observations in the data set.
- Each observation represents an acute myocardial infarction.

C. DiMaggio, *SAS for Epidemiologists: Applications and Methods*,
DOI 10.1007/978-1-4614-4854-9,

- There are 16 variables for each observation.
- The data is not sorted.
- The DATE variable is numeric.
- DISPO is a 2-unit long character variable.
- DATE looks like numbers.
- ECODE is missing from most observations, and is therefore not very useful.

Chapter 3

3.1 Reading in Data from the Editor Window

```
data test;
input
@1  ID     $6.
@7  RESULT $8.
@15 VALUE  3.;
cards;
203769Positive486
201948Positive400
202085Positive364
201755Positive416
202092Positive373
202087Positive657
201358Negative341
201429Positive448
201549Negative320
202741Positive391
201627Positive532
202004Negative268
202052Negative334
203531Positive573
204366Negative348
204042Negative252
;
run;

proc print data=test;
run;
```

3.2 Reading in Data from an External Source

```
DATA sparcs_1;
INFILE 'your/path/here/nycsparcs.TXT'  LRECl=452 OBS=100;
INPUT
@18   DATE yymmn6.
@44   AGE        3.
@50   COUNTY     $CHAR2.
@52   ZIP        $CHAR5.
```

```
@57   SEX         $CHAR1.
@58   RACE        $CHAR2.
@60   ETHNIC      1.
@71   PDX         $CHAR6.
@343  DISPO       $CHAR2;
run;

proc print data=sparcs_1;
run;
```

3.3 Creating a SAS Library

```
libname p8483 " C:\Users\Charles DiMaggio\Desktop";
```

Chapter 4

4.1 Using PROC PRINT

```
libname ex4_1;

proc contents data=;
run;

proc print data= (obs=20) noobs ;
var age pdx charge;
where los gt 14;
sum charge;
run;


proc print data=- (obs=20) noobs ;
var age pdx charge;
where los lt 7;
sum charge;
run;
```

- It appears to be hospital discharge data.
- there are 502,492 observations with 16 variables each.
- The total charges are $914,170.00.
- The total charges for patients whose length of stay was less than seven days is $104,660.00.
- Length of stay is associated with cost of care.

4.2 From SAS to Excel

```
ods html file =  style=minimal;

proc print data=- (obs=150) noobs;
var sex age zip date dispo pdx;
title 'Patient Demographics';
run;
```

ods html close;

4.3 Creating and applying formats

```
proc print data=       (obs=20) noobs;
var charge;
format charge dollar11.2;
run;

proc format;
  value dol_range low - 500000 = 'Low'
                  500001 - 1000000 = 'Medium'
                  1000001 - high = 'High';
run;
proc print data=p8483.demo_1 (obs=20) noobs;
var charge;
format charge dol_range.;
run;

proc print data=        (obs=20) noobs;
var sex;
run;

proc format;
  value $gender "F" = "Female"
                       "M" = "Male";
run;

proc print data=       (obs=20) noobs;
var sex;
format sex $gender.;
run;
```

4.4 Using Titles and Labels

```
data test;
input
@1  ID     $6.
@7  RESULT $8.
@15 VALUE  3.;
cards;
203769Positive486
```

```
201948Positive400
202085Positive364
201755Positive416
202092Positive373
202087Positive657
201358Negative341
201429Positive448
201549Negative320
202741Positive391
201627Positive532
202004Negative268
202052Negative334
203531Positive573
204366Negative348
204042Negative252
;

proc print data=test noobs label;

label
ID = 'Patient Identifier'
RESULT = 'Final Result'
VALUE = 'Assay Level';

run;

title 'Initial Blood Test Results';
run;
```

Chapter 5

5.1 Concatenating Data Sets

```
data tot_deaths; set tot_deaths; tot_911_deaths =
female_911_deaths + male_911_deaths; run;
```

- The file is missing many observations.
- The total deaths variable is missing all non-New York City deaths.
- It solved the problem.
- You receive the warning “Missing values created from missing values”.

5.2 Merging and Performing Operations on Datasets

```
data MERGE1; set Ch5demo1  Ch5demo2; run;
```

- 185 observations were read in from Ch5demo1.
- 1,421 observations were read from Ch5demo2.
- There are 1,606 observations in the new data set.

- It appears correct.
- You could look at the original data set using PROC CONTENTS.

```
proc sort data=MERGE1; by zip; proc sort data= Ch5demo3; by zip;
run;

data MERGE2; merge MERGE1 ch5demo3; by zip; run;

data MERGE2_calcs; set MERGE2; tot_death = male_911_deaths +
female_911_deaths; death_rate = (tot_death / pop_tot)* 100000;
run;

data MERGE2_calcs; set MERGE2_calcs; work_pop = pop_2024
+ pop_2534 + pop_3544 + pop_4554; work_rate = (tot_death /
work_pop)* 100000; run;
```

- You need to sort the data sets before merging.
- The death rate for 11,566 is $16.8/100,000$; the deathrate for zip code for 10032 is $18.8/100,000$.
- ZIP Code 10032 is at increased risk perhaps because it is closer to ground zero.
- The new rates for workers are 34.6 for ZIP code 11566, and 35.1 for 10032.
- The rates went up because the denominator of working individuals is smaller. The rates based on the working population are probably closer to the “true” risk.

Chapter 6

(You may need to refer back to other chapters to complete some of these problems)

6.1 Reading in the Data Set

```
DATA NYCSPARCS; /* name the data set in work directory */
INFILE \nycsparcs.TXT’ MISSOVER LRECl=452; * OBS=15000;
/* notice how we commented out the obs=, will read all the
   observations */INPUT /* input the variables you are
      interested in */
@18   DATE yymmn6.
@44   AGE       3.
@50   COUNTY    $CHAR2.
@52   ZIP       $CHAR5.
@57   SEX       $CHAR1.
@58   RACE      $CHAR2.
@71   PDX       $CHAR6.
@349  LOS       4.
;
```

6.2 Creating New Variables

```
/**************CREATE MONTH VARIABLE***********/
month=MONTH(DATE);
/***********CREATE NUMERIC GENDER VARIABLE****************/
IF sex=’M’ then male=1;
   else male=0;
IF sex=’F’ then female=1;
           else female=0;

/**********CREATE NUMERIC VARIABLES FOR RACE ***********/
IF RACE = ’01’ then White=1;
               else White=0;
IF RACE = ’02’ then Black=1;
               else Black=0;
IF RACE = ’04’ then Asian=1;
               else Asian=0;
IF RACE = ’88’ then Other_Race=1;
               else Other_Race=0;
IF RACE = ’99’ then Unknown_Race=1;
               else Unknown_Race=0;

/************MORTALITY **************************/
IF DISPO = ’20’  then death=1;
                 else death=0;

/*********** SUBSTANCE ABUSE *****************/

if pdx in /* use ICD9 codes to create diagnoses */
(’2910’,’2911’,’2912’,’2913’,’2914’,’2915’,’29181’,’29189’,
’2919’,’2920’,’29211’,’29212’,
’2922’,’29281’,’29282’,’29283’,’29284’,’29289’,’2929’,
’30300’,’30301’,’30302’,’30303’,’30390’,’30391’,’30392’,
’30393’,’30400’
’30401’,’30402’,’30403’,’30410’,’30411’,’30412’,’30413’,’30420’,
’30421’,’30422’,’30423’,’30430’,’30431’,’30432’,’30433’,’30440’,
’30441’,’30442’,’30443’,’30450’,’30451’,’30452’,’30453’,’30460’,
’30461’,’30462’,’30463’,’30470’,’30471’,’30472’,’30473’,’30480’,
’30481’,’30482’,’30483’,’30490’,’30491’,’30492’,’30493’,’30500’,
’30501’,’30502’,’30503’,’3051’,’30520’,’30521’,’30522’,’30523’,
’30530’,’30531’,’30532’,’30533’,’30540’,’30541’,’30542’,’30543’,
’30550’,’30551’,’30552’,’30553’,’30560’,’30561’,’30562’,’30563’,
’30570’,’30571’,’30572’,’30573’,’30580’,’30581’,’30582’,’30583’,
’30590’,’30591’,’30592’,’30593’)
Then subst_ab=1;
Else subst_ab=0;

RUN;

PROC CONTENTS DATA=NYCSPARCS; /* check your file was read in */
RUN;
```

6.3 Using PROC MEANS

```
proc means data=nycsparcs;
var age;
run;

proc means data=nycsparcs;
var male;
run;
```

- There is a person aged 109 years old.

6.4 Using PROC FREQ

```
proc freq data=nycsparcs;
tables county race sex;
run;
```

- There is no real difference between numeric gender and character sex, although the numeric variable might be more useful if we need a dummy variable later on.
- Everyone came from a few counties, but since this is New York City data that makes sense. We refer to data documentation to determine that county 58 is the Bronx, 59 is Brooklyn, 60 is Manhattan, 61 is Queens, and 62 is Staten Island. The largest percentage of non-New York City residents comes from county 55 (Westchester).
- Over a quarter of the entries have race as other or unknown. This raises concern about the reliability and validity of that variable.

6.5 Using PROC TABULATE

```
proc tabulate data=nycsparcs; /*create output data set from
  tabulate procedure */
         where county in ('58' '59' '60' '61' '62');
         class month;
     var subst_ab;
     table month, subst_ab*sum subst_ab*pctsum;
ods output table=subst; /* note creating a table from output,
 will use it later to graph, can see it in explorer */
run;
```

- Brooklyn appears to need more resources.
- Staten Island may require less.
- These results may be affected by the number of hospital beds in a county, the number of beds available for substance abuse, and local medical practice in admitting or discharging substance abuse, the proportion of age groups in a county that might be more or less likely to abuse drugs.
- We would want to determine population-based age-stratified rates.

Chapter 7

7.1 PROC GCHART

```
proc gchart data=nycsparcs;
     vbar race;
     hbar race;
run;

proc gchart data=nycsparcs;
     vbar race / sumvar=los type=mean;
     hbar race / sumvar=los type=mean;
run;
quit;

proc gchart data=nycsparcs;
     pie race / sumvar=los type=mean
                 fill=x explode='02';
run;
quit;
```

One advantage of hbar is that it returns statistics.

7.2 PROC GPLOT

```
proc tabulate data=nycsparcs; /*create output data set from
 tabulate procedure */
         where county in ('58' '59' '60' '61' '62');
         class month;
     var subst_ab;
     table month, subst_ab*sum subst_ab*pctsum;
ods output table=subst; /* note creating a table from output,
        we'll use it later to graph, can see it in explorer */
run;


proc gplot data=subst; /* note using output table from tabulate*/
     plot subst_ab_Sum * month ;
 symbol value=diamond i=spline
      c=red w=5;
run;
quit;
```

- The chart is not zeroed on the vertical axis, which makes changes look more impressive.
- The spline makes it look like a continuous process when it very much may not be so.
- You can use the "vaxis" option after the plot statement to define axis. It might be better to just join the points instead of spline interpolation. (The change in colors is just for fun.)

```
proc gplot data=subst;
     plot subst_ab_Sum * month / vaxis= 0 to 5000 by 200;
 symbol value=diamond i=join
      c=blue w=2;
run;
quit;
```

Chapter 8

8.1 One-Way Frequencies

```
proc print data=car;
   var type region safety weight;
run;

proc freq data=car;
   tables region safety type;
run;

Partial PROC PRINT Output
Obs    type             region       safety    weight

  1    Medium           N America       0       3.395
  2    Sport/Utility    N America       0       4.180
  3    Medium           N America       0       3.145
  4    Small            N America       0       2.600
  5    Medium           N America       0       3.085
  6    Medium           N America       0       2.910
  7    Sport/Utility    N America       0       4.180
  8    Medium           Asia            0       3.415
  9    Medium           N America       0       3.995
 10    Small            N America       0       2.600
 11    Small            N America       1       2.765
 12    Small            Asia            0       2.665
 13    Medium           N America       0       3.100
 14    Medium           N America       0       3.455
 15    Medium           N America       0       3.055
 16    Large            N America       0       3.450
 17    Large            N America       0       3.640
 18    Large            N America       0       4.195
 19    Large            N America       0       3.985
 20    Large            N America       0       4.480
```

- Measurement scale of each variable:
  - Safety—OrdinaL
  - Type—NOMINAL
  - Region—nominal
  - Weight—CONTINUOUS

- The proportion of cars built in North America is 0.6354.
- There are no unusual data values that warrant further attention.

8.2 Cross Tabulations

```
proc format;
   value safdesc 0='Average or Above'
                 1='Below Average';
run;

proc freq data=car;
   tables region*safety / expected cellchi2;
   format safety safdesc.;
run;
```

- For the cars made in Asia, 42.86% had a below-average safety score.
- For the cars with an average or above safety score, 69.70% were made in North America.
- There seems to be an association between region and safety. A higher percentage (75.41 vs. 57.14) of cars from North America had a higher safety rating.
- The cell where region is Asia and safety is below average contributed the most to any possible association.

8.3 Chi-Square

```
proc freq data=car;
   tables region*safety / chisq;
   format safety safdesc.;
run;
```

You fail to reject the null hypothesis that there is not an association. The p-value represents the probability of observing a chi-square value at least as large as the one actually observed, given that the null hypothesis is true.

8.4 Spearman

```
data car2;
   set car;
   size=1*(type='Sports' or type='Small') +
        2*(type='Medium') +
        3*(type='Large' or type='Sport/Utility');
run;


proc format;
   value sizename 1=Small
                  2=Medium
                  3=Large;
run;
```

```
proc freq data=car2;
   tables size*safety / chisq measures cl;
   format safety safdesc.
          size sizename.;
run;
```

- You should use the Mantel-Haenszel test to detect an ordinal association between size and safety.
- You reject the null hypothesis that there is not an ordinal association.
- The Spearman correlation statistic indicates that an ordinal association of moderate strength exists (0.5425) between size and safety.
- The 95% confidence interval around that statistic is (0.6932, 0.3917).

Chapter 9

9.1 Contingency Table Analysis

- The row mean scores difference statistic can be used to measure the evidence of an association between type by safety?
- With a p-value less than 0.0001, there is statistical evidence of an association between type by safety.
- You could use the uncertainty coefficient to measure the strength of the association between type by safety.
- The proportion of variability in the response variable that is explained by the predictor variable is 0.3011.

9.2 Stratified Analysis

```
   tables type*safety region*safety / all;
run;
```

- Use the MH statistic to detect an association between type by safety controlling for region.
- There is a statistically significant association of type with safety, holding region constant.
- Because there are no observations for “large” vehicles.
- Yes. There may be cells with zero in the denominator leading to undefined results.

```
proc freq data=sasuser.safety;
   tables type*region*safety / all bdt;
   exact or comor;
run;
```

- Yes. The adjusted odds ratio for the association of region and safety controlling for type differs from the crude or unadjusted odds ratio.
- The Breslow–Day statistic is not statistically significant, indicating there is no interaction between type and region.

Chapter 10

10.1 PROC MEANS

- SAS used all observations, most of which were coded zero because not recorded for adults.
- $N = 200,000$, $\mu = 318.3755000$.
- 134,412 children less than 1 year old were discharged from NYC hospitals in this year.
- $N = 22,611$, $\mu = 2816.11$.

```
DATA infants; /* name the data set in work directory */
INFILE 'C:\Users\Charlie\Documents\Columbia\Epi\SAS COURSE\Data
 Sets\nycsparcs.TXT' MISSOVER LRECl=452; * OBS=15000; /*
  notice how we commented out the obs=,
will read all the observations */
INPUT /* input the variables you are interested in */
@18   DATE yymmn6.
@44   AGE       3.
@50   COUNTY    $CHAR2.
@52   ZIP       $CHAR5.
@57   SEX       $CHAR1.
@58   RACE      $CHAR2.
@60   ETHNIC    1.
@61   SOURCE    $CHAR1. /*source of admission used with type of
                        admission to id prematures*/
@62   TYPE      $CHAR1. /* type of admission to identify
@71   PDX       $CHAR6.   newborns=type 04 */
@294  BIRTHWT   4.
@343  DISPO     $CHAR2.
@349  LOS       4.
@434  CHARGE    12.;

/*******************MORTALITY *************************/
IF DISPO = '20'  then death=1;
                 else death=0;
/*********************** INFANT MORTALITY ****************/
IF AGE LT 1 and DISPO='20'  then infant_mort=1;
                            else infant_mort=0;
/********************* LOW BIRTH WEIGHT *************/
IF TYPE='4' and BIRTHWT LT 2500 then LBTWT=1;
                                else LBTWT=0;
/********************* PREMATURE BIRTH *************/
IF TYPE='4' and SOURCE='2'  then  preemie=1;
                             else prememie=0;
run;

proc means data=infants;
var birthwt;
run;
```

```
proc means data=infants n mean median std var clm;
var birthwt;
run;

data infants;
  set infants;
  if age lt 1;
  run;


proc means data= infants n mean median std var clm;
var birthwt;
run;
```

10.2 PROC UNIVARIATE

- The histogram and probability plot lead us to suspect the data are not normally distributed.
- The skewness statistic is −1.5, indicating the data are left skewed.
- The mean (2816) is less than the median (3200), again indicating left skewed.
- The kurtosis statistic is 1.3, indicating a high-peaked and heavy-tailed distribution.
- There are a lot of zeros, indicating either missing data or coding errors.
- Rerunning the analysis with zeros removed returns a more normal-appearing data distribution, with a mean closer to the median, less skewness, and a kurtosis value about the same as the previous run:

```
goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc univariate data infants mu0=2500 plot;
var birthwt;
histogram birthwt;
probplot birthwt / normal (mu=est sigma=est color=blue w=1) ;
run;


goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc univariate data infants mu0=2500 plot;
where birthwt ne 0;
var birthwt;
histogram birthwt;
probplot birthwt / normal (mu=est sigma=est color=blue w=1) ;
run;
```

10.3 PROC BOXPLOT

```
data infants;
set infants;
if county = '58' then borough = 1;
```

```
else if county ='59' then borough = 2;
else if county ='60' then borough = 3;
else if county = '61' then borough = 4;
else if county = '62' then borough = 5;
else borough = .;
run;

proc sort data=infants;
by borough;
run;

symbol color = salmon;
   title 'Boxplot NYC County Birthweights';
   proc boxplot data=infants04;
   where birthwt ne 0;
   plot birthwt*borough    / cframe   = vligb
                             cboxes   = dagr
                             cboxfill = ywh;
   run;
```

Chapter 11

11.1 PROC GLM

```
options ls=75 ps=45;  /* page and line spacing options*/
proc glm data = infants;
where birthwt ne 0;
       class borough;
       model birthwt=borough;
       means borough / hovtest;
       output out=check r=resid   p=pred;
run;
quit;  /* have to quit our of GLM */
```

The p-value for Levene's test is $p < 0.0001$; reject the null hypothesis of homogeneity

11.2 PROC GPLOT

```
/* now run gplot on the 'check' dataset created above*/
Proc gplot data=check;
         Plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
         axis1 w=2 major=(w=2) minor=none offset=(10pct);
         axis2 w=2 major=(w=2) minor=none;
         title 'plot residuals vs predictors';
run;
quit;
```

The residuals are evenly distributed about 0 with some evidence of outliers.

11.3 PROC UNIVARIATE

```
proc univariate data = check normal;
      var resid;
      histogram resid / normal;
      probplot;* resid / mu=est sigma=est color=blue w=1;
      title;
run;
```

- The mean is zero.
- It is consistent with the underlying assumptions for ANOVA.
- The kurtosis statistic is 3.3.
- The data are high peaked and heavy tailed.

11.4 LSMEANS

```
/* compare means */
proc glm data=infants;
   class borough;
   model birthwt=borough;
   lsmeans borough / pdiff=all adjust=tukey;
   title 'Data: Multiple Comparisons';
run;
quit;
```

- The lowest mean birth weight is 2,662 grams in borough 1. It is statistically significantly different from the other boroughs.
- Box plots would be helpful to compare the mean birth weights across boroughs.

Chapter 12

12.1 Read in the Pedestrian Injury Data Set

```
proc contents data=ch9.ch9exercise;
run;
```

12.2 Print Out the Data

```
proc print data=ch9.ch9exercise;
var numinj totpop perblack perhisp medhsinc pci;
id name;
run;
```

12.3 Create an Injury Rate Variable

```
data ch9exercise;
set ch9.ch9exercise;
injrate=(numinj/totpop)*1000;
run;
```

12.4 PROC UNIVARIATE

```
options ps=50 ls=76;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=ch9exercise;
   var injrate perblack perhisp medhsinc pci;
   histogram injrate perblack perhisp medhsinc pci/ normal;
   probplot  injrate perblack perhisp medhsinc pci
             / normal (mu=est sigma=est color=red w=2);
   id name;
   title 'Univariate Statistics of pedestrian injury data set';
run;
```

- The average number of injuries per 1,000 population for a Nassau County ZCTA community during the observation period was 2.
- Hempstead had the highest injury rate, six injuries per 1,000 population.
- Freeport and Hempstead had the highest proportion of Hispanic residents.
- Inwood and Hempstead had the the lowest median household income.
- “perblack” and “perhisp” appear to be least likely to be normally distributed.

12.5 Log Transformation

```
data ch9exercise;
set ch9exercise;
lnblack=log(perblack);
lnhisp=log(perhisp);
run;

proc univariate data=ch9exercise;
   var lnblack lnhisp;
   histogram lnblack lnhisp/ normal;
   probplot  lnblack lnhisp
             / normal (mu=est sigma=est color=red w=2);
   id name;
   title '';
run;
```

Yes.

12.6 PROC GPLOT

```
/*Use the following syntax to set up the options for
 your plots and define axes*/

Options ps=50 ls=64;
Goptions reset=all gunit=pct border
Fontres=presentation ftext=swissb;

Axis1 length=70 w=3 color=blue label=(h=3) value=(h=3);
Axis 2 length=70 w=3 color=blue label=(h=3) value=(h=3);

/*Invoke the above options and axes by
including the following line after your plot statement*/

/ vaxis=axis1 haxis=axis2

Options ps=50 ls=64;
Goptions reset=all gunit=pct border
               Fontres=presentation ftext=swissb;

Axis1 length=70 w=3 color=blue label=(h=3) value=(h=3);
Axis 2 length=70 w=3 color=blue label=(h=3) value=(h=3);

proc gplot data=ch9exercise;
   plot injrate * (lnblack lnhisp medhsinc pci)
        / vaxis=axis1 haxis=axis2;
   symbol1 v=dot h=2 w=4 color=red;
   title h=3 color=green
         'Scatter Plot of Pedestrian Injury Rate by Explanatory
          Variables';
run;
quit;
```

- “perhisp” seems to most demonstrate a relationship with injury rate.
- The plot would best be described as a direct linear relationship.

12.7 PROC CORR

```
proc corr data=ch9exercise rank;
   var lnblack lnhisp medhsinc pci;
   with injrate;
   title '';
run;
```

- “perhisp” was most strongly correlated with pedestrian injury rate.
- The relationship between income in a community and pedestrian injury rate was weakly negative.

- There is something in communities with large proportions of poor Hispanics or Blacks that put people at risk for pedestrian injuries.

Chapter 13

13.1 Infant Birth Weight and Hospital Charges: Assumptions

```
options ps=50 ls=76;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc univariate data=  infants normal;
var charge birthwt;
histogram charge birthwt;
probplot charge birthwt;
run;
```

The data do not appear normally distributed.

13.2 Infant Birth Weight and Hospital Charges: Correlation

```
proc corr data=infants rank;
   var charge;
   with birthwt;
   title 'PROC CORR: infant birthweight and charge';
run;
```

- There is a statistically significant relationship between birthweight and charges $r = -0.22, p < 0.0001$.
- The relationship is inversely proportional.
- We should look at it with a scatterplot.

```
Options ps=50 ls=64;
Goptions reset=all gunit=pct border
               Fontres=presentation ftext=swissb;

Axis1 length=70 w=3 color=blue label=(h=3) value=(h=3);
Axis 2 length=70 w=3 color=blue label=(h=3) value=(h=3);

proc gplot data= infants;
   plot charge * birthwt
        / vaxis=axis1 haxis=axis2;
   symbol1 v=dot h=2 w=4 color=red;
   title h=3 color=green
        'Plot of infant birthweight and charge';
run;
```

13.3 Infant Birth Weight and Hospital Charges: Linear Regression

```
proc reg data=infants;
   where birthwt ne 0;
   model charge=birthwt;
   title 'Simple Linear Regression of charges and birthweight';
run;
quit;
```

- The overall F test is 3497 $p < 0.0001$.
- Birthweight is a statistically significant predictor of total charges $p < 0.0001$.
- The model explains 14% of the variability $R^2 = 0.15$.
- Total explained variance is the square of Pearson's, correlation coefficient.

```
quit;
```

Chapter 14

14.2 Residuals

```
options ps=50 ls=97;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc reg data=sasuser.b_grades;
   model gpa=score;
   plot r.*(p. score);
   plot student.*obs. / vref=3 2 -2 -3
                        haxis=1 to 25 by 1;

   title 'Residual Diagnostic Plots';
run;
quit;
```

Index

C. DiMaggio, *SAS for Epidemiologists: Applications and Methods*,
DOI 10.1007/978-1-4614-4854-9,