The first problem in this lab will make use of output from the following program:

get file='c:\recall.sav'.
compute NRECALL = (1/sqrt(n))*arsin(sqrt(recall)).
compute d1 = 0.
if (event eq 1) d1=1.
compute d2 = 0.
if (event eq 2) d2=1.
.
compute d25 = 0.
if (event eq 25) d25=1.
compute d26 = 0.
if (event eq 26) d26=1.
regression vars=NRECALL,EVENTYR,BIRTHYR,d1 to d26 / dep=NRECALL
   / enter EVENTYR,BIRTHYR / stepwise d1 to d26
   / save=resid(RECALADJ).
compute AGEEV=EVENTYR-BIRTHYR.
compute AGEEV2=(AGEEV-21)**2.
compute AGEEV3=(AGEEV-21)**3.
pearson corr AGEEV AGEEV2 AGEEV3.
regression vars=RECALADJ,AGEEV / dep=RECALADJ / enter.
regression vars=AGEEV,EVENTYR,BIRTHYR / dep=AGEEV / enter.
examine vars=AGEEV/percentiles(5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90,95)/plot none/stat none.
temporary.
recode AGEEV(lo thru 7=7)(7.1 thru 9.8=9)(9.81 thru 12=12)
   (12.1 thru 14=14)(14.1 thru 15=15)(15.1 thru 17=17)
   (17.1 thru 18=18)(18.1 thru 20=20)(20.1 thru 21=21)
   (21.1 thru 22=22)(22.1 thru 23=23)(23.1 thru 24=24)
   (24.1 thru 26=26)(26.1 thru 27=27)(27.1 thru 28=28)
   (28.1 thru 30=30)(30.1 thru 31=31)(31.1 thru 33.2=33)
   (33.3 thru 36=36)(36.1 thru hi=40).
examine vars=RECALADJ by AGEEV/plot boxplot/stat none/nototal.
regression vars=RECALADJ,AGEEV,AGEEV2,AGEEV3 / dep=RECALADJ
   / enter AGEEV / enter AGEEV2 / enter AGEEV3
   / save pred(RECALEST).
graph scatterplot(bivar)=AGEEV with RECALEST.
(Note: You do not need to print up the entire output from this program to be passed in with your lab.)
1. In this lab we complete our analysis of the recall data.

   a. Your dependent variable must be a recall measure that is homoscedastic. You will remember that this is why we transformed the RECALL measure to NRECALL by taking its arcsin square root and by then weighting it by the number of subjects on which it is based. Please do this again.

   b. To approximate a random sample of "event-cohorts," we must have measures of recall on comparable events that are remembered by comparable age cohorts. On the one hand, we want to remove differences in recall due both to the recency of events as well as to the disproportionately stronger or weaker effects of particular events. On the other hand, we want to remove "maturity effects" (i.e., the tendency at the time of the survey for older cohorts to remember more than younger ones) from the data. For these reasons NRECALL must be adjusted for the linear effects of BIRTHYR and EVENTYR. In addition NRECALL must be adjusted for the greater social importance of some events than others by creating dummy variables for each event, and by removing variance from NRECALL that is due to events that are significantly (at the .05 level) more or less socially important than the others. For example, a dummy variable for the event, "Hiroshima bombing" is created as follows:

   ```
   compute d1 = 0.
   if (event eq 1) d1=1.
   ```

   Be sure you understand what these two SPSS commands do.

   c. In part b you were asked to create "dummy variables for each event." Although you may be able to create these variables, you would not be able to enter (i.e., force) dummy variables for all 26 events into your regression model. Why is this?

   d. The regression requested in part b yields residuals (RECALADJ) that measure "recall adjusted for maturity and event effects." Karl Mannheim argued that people are most influenced by events that they experience between the ages of 17 and 25. This suggests that people will be more likely to recall events that occurred around their 21st birthday than that occurred when they were younger or older. (Note: 21 is the midpoint between 17 and 25.) Accordingly one might hypothesize a curvilinear relation between one's recall of an event and one's age when the event took place. To estimate this relation, create three variables to respectively measure the linear, quadratic, and cubic effects of age-at-time-of-event (AGEEV). To reduce their collinearity and to render their slopes more meaningful, subtract the theoretically meaningful number, 21 (namely, the age at which events have greatest influence), from AGEEV before obtaining quadratic (AGEEV2) and cubic (AGEEV3)
measures. Explain how you can tell that the three polynomials are not orthogonal.

e. Show that AGEEV explains no variance in RECALADJ, and that EVENTYR and BIRTHYR explain all variance in AGEEV. Explain how these findings are related. (Hint: First, explain why the latter finding is necessary given how AGEEV was computed. Second, explain how the former finding necessarily results from the latter.)

f. Obtain a boxplot of RECALADJ by AGEEV. Using this plot, demonstrate that there is a nonlinear relation between RECALADJ and AGEEV.

g. Obtain three hierarchical linear models by regressing RECALADJ first on AGEEV, then on AGEEV and AGEEV2, and then on AGEEV, AGEEV2, and AGEEV3. How much of the variance in RECALADJ is explained by its relation with a quadratic function of AGEEV? Does a cubic reexpression of AGEEV explain a significant additional amount of variance in RECALADJ? Which of these three models best fits the data? (Use the .05 significance level in obtaining your answer.)

h. From the best-fitting regression model estimated in part g, plot the estimated recall values (RECALEST) by AGEEV. Describe in words the relation that your regression model depicts as existing between RECALADJ and AGEEV. (i.e., how does it appear visually?) Do your findings provide support for Mannheim's theory of generations?

2. In writing up her results the researcher concludes, "The children's average IQ score was 100. The children were found to exhibit a decrease of 10 points in IQ for each unit of increase in TV viewing. Independent of this linear trend, there was a slight tendency (2 IQ
points above what one would estimate given this trend) for children with medium TV viewing to have higher IQs than children with either extremely high or extremely low viewing. Assume that the numbers in this description refer to parameter estimates from an ordinary least squares (OLS) regression equation.

a. Write down the regression equation that the researcher estimated. (Be sure to include the estimated unstandardized regression coefficients in the equation.)

b. Show specifically how she coded her independent variable(s). (I.e., indicate for each independent variable, what numerical value was assigned to each of the four groups.)

3. Consider the following interpretation of the same data: "The children's average IQ score was 100. On average, the combined high and medium-high viewers had IQs that were 10 points below this average. Independent of this, the combined low and medium-high viewers were found to have IQs on average that were 5 points above the overall average." Again assume that the numbers in this description refer to parameter estimates from an ordinary least squares (OLS) regression equation.

a. What regression equation was estimated here? (Again, be sure to include the estimated unstandardized regression coefficients in the equation.)

b. State how independent variable(s) were coded prior to estimating the equation. (I.e., again indicate for each independent variable, what numerical value was assigned to each of the four groups.)

4. The y-hats referred to in problems 2 and 3 are not the same for children in any of the four groups. How is this possible, given that each problem deals with an analysis of the same data? (Hint: Recall that when working with the "memorization data" in class, we ended up with the same y-hats no matter if we used dummy variables, effect codes, orthogonal contrasts, or orthogonal polynomial contrasts. Why is this not the case here?)

Below please find R and SAS code for the first problem:

# R

# Code:
recall <- read.table("C:/recall.txt")
names(recall) <- c("birthyr","event","eventyr","recall")
attach(recall)

# New variable "n"
recall[1:5] = c(44,47,67,51,52)
recall[6:5] = c(63,63,63,78,73)
recall[7:5] = c(80,87,78,73,80)
recall[8:5] = c(87,94,87,86,87)
recall[9:5] = c(66,97,86,65,48)
names(recall)[5] <- "n"

recall$nrecall <- (1/sqrt(recall$n))*asin(sqrt(recall$recall))
recall[,7:32] <- 0
names(recall)[7:32] <- c("d1","d2","d3","d4","d5","d6","d7","d8","d9",
  "d10","d11","d12","d13","d14","d15","d16","d17","d18",
  "d19","d20","d21","d22","d23","d24","d25","d26")
recall[event==1,7] = 1
recall[event==2,8] = 1
recall[event==3,9] = 1
recall[event==4,10] = 1
recall[event==5,11] = 1
recall[event==6,12] = 1
recall[event==7,13] = 1
recall[event==8,14] = 1
recall[event==9,15] = 1
recall[event==10,16] = 1
recall[event==11,17] = 1
recall[event==12,18] = 1
recall[event==13,19] = 1
recall[event==14,20] = 1
recall[event==15,21] = 1
recall[event==16,22] = 1
recall[ event==17,23] = 1
recall[ event==18,24] = 1
recall[ event==19,25] = 1
recall[ event==20,26] = 1
recall[ event==21,27] = 1
recall[ event==22,28] = 1
recall[ event==23,29] = 1
recall[ event==24,30] = 1
recall[ event==25,31] = 1
recall[ event==26,32] = 1

reg1 <- lm(nrecall ~ eventyr + birthyr + d1+d2+d3+d4+d5+d6+d7+d8+d9+
    d10+d11+d12+d13+d14+d15+d16+d17+d18+d19+d20+d21+d22+d23+d24+d25+d26,
data=recall)
reg1.step <- step(reg1)
summary(reg1.step)

reg2 <- lm(nrecall ~ birthyr + eventyr +
    d2+d3+d4+d5+d8+d9+d11+d12+d13+d14+d17+d18+d19+d25, data=recall)
summary(reg2)
recall$recalladj <- reg2$residuals

recall$ageev <- recall$eventyr-recall$birthyr
recall$ageev2 <- (recall$ageev-21)^2
recall$ageev3 <- (recall$ageev-21)^3
cor(recall[,34:36])

reg3 <- lm(recalladj ~ ageev, data=recall)
summary(reg3)

reg4 <- lm(ageev ~ eventyr + birthyr, data=recall)
summary(reg4)

quantile(recall$ageev,
c(.05,.10,.15,.20,.25,.30,.35,.40,.45,.50,.55,.60,.65,.70,.75,.80,.85,
    .90,.95))

recall[ recall$ageev <= 7, 37] = 7
recall[ recall$ageev > 7 & recall$ageev <= 9.8, 37] = 9
recall[ recall$ageev > 9.8 & recall$ageev <= 12, 37] = 12
recall[ recall$ageev > 12 & recall$ageev <= 14, 37] = 14
recall[ recall$ageev > 14 & recall$ageev <= 15, 37] = 15
recall[ recall$ageev > 15 & recall$ageev <= 17, 37] = 17
recall[ recall$ageev > 17 & recall$ageev <= 18, 37] = 18
recall[ recall$ageev > 18 & recall$ageev <= 20, 37] = 20
recall[recall$ageev > 20 & recall$ageev <= 21, 37] = 21
recall[recall$ageev > 21 & recall$ageev <= 22, 37] = 22
recall[recall$ageev > 22 & recall$ageev <= 23, 37] = 23
recall[recall$ageev > 23 & recall$ageev <= 24, 37] = 24
recall[recall$ageev > 24 & recall$ageev <= 26, 37] = 26
recall[recall$ageev > 26 & recall$ageev <= 27, 37] = 27
recall[recall$ageev > 27 & recall$ageev <= 28, 37] = 28
recall[recall$ageev > 28 & recall$ageev <= 30, 37] = 30
recall[recall$ageev > 30 & recall$ageev <= 31, 37] = 31
recall[recall$ageev > 31 & recall$ageev <= 33.2, 37] = 33
recall[recall$ageev > 33.2 & recall$ageev <= 36, 37] = 36
recall[recall$ageev > 36, 37] = 40
names(recall)[37] <- "ageev.collapsed"

boxplot(recall$recalladj ~ recall$ageev.collapsed, xlab="AGEEV",
        ylab="Unstandardized residual")
abline(h=0, col="grey")
x11()

reg5 <- lm(recalladj ~ ageev, data=recall)
summary(reg5)
reg6 <- lm(recalladj ~ ageev + ageev2, data=recall)
summary(reg6)
reg7 <- lm(recalladj ~ ageev + ageev2 + ageev3, data=recall)
summary(reg7)

recall$reccallest5 <- reg5$fitted.values
plot(recall$ageev, recall$reccallest5, xlab="AGEEV", ylab="Predicted value from reg5")
x11()
recall$reccallest6 <- reg6$fitted.values
plot(recall$ageev, recall$reccallest6, xlab="AGEEV", ylab="Predicted value from reg6")
x11()
recall$reccallest7 <- reg7$fitted.values
plot(recall$ageev, recall$reccallest7, xlab="AGEEV", ylab="Predicted value from reg7")

* SAS
* Code;

DATA recall;
  INFILE 'C:\recall.txt';
  INPUT birthyr event eventyr recall;
d1=0; d2=0; d3=0; d4=0; d5=0; d6=0; d7=0; d8=0; d9=0;
d10=0;
d11=0; d12=0; d13=0; d14=0; d15=0; d16=0; d17=0; d18=0; d19=0;
d20=0;
d21=0; d22=0; d23=0; d24=0; d25=0; d26=0;
IF event = 1 THEN d1=1; IF event = 2 THEN d2=1;
IF event = 3 THEN d3=1; IF event = 4 THEN d4=1;
IF event = 5 THEN d5=1; IF event = 6 THEN d6=1;
IF event = 7 THEN d7=1; IF event = 8 THEN d8=1;
IF event = 9 THEN d9=1; IF event = 10 THEN d10=1;
IF event = 11 THEN d11=1; IF event = 12 THEN d12=1;
IF event = 13 THEN d13=1; IF event = 14 THEN d14=1;
IF event = 15 THEN d15=1; IF event = 16 THEN d16=1;
IF event = 17 THEN d17=1; IF event = 18 THEN d18=1;
IF event = 19 THEN d19=1; IF event = 20 THEN d20=1;
IF event = 21 THEN d21=1; IF event = 22 THEN d22=1;
IF event = 23 THEN d23=1; IF event = 24 THEN d24=1;
IF event = 25 THEN d25=1; IF event = 26 THEN d26=1;
IF birthyr=28 THEN n=44; IF birthyr=32 THEN n=47;
IF birthyr=34 THEN n=67; IF birthyr=35 THEN n=51;
IF birthyr=36 THEN n=52; IF birthyr=37 THEN n=63;
IF birthyr=38 THEN n=63; IF birthyr=39 THEN n=78;
IF birthyr=40 THEN n=73; IF birthyr=41 THEN n=80;
IF birthyr=42 THEN n=87; IF birthyr=43 THEN n=94;
IF birthyr=44 THEN n=87; IF birthyr=45 THEN n=66;
IF birthyr=46 THEN n=97; IF birthyr=47 THEN n=86;
IF birthyr=48 THEN n=65; IF birthyr=49 THEN n=48;
nrecall = (1/sqrt(n))*arsin(sqrt(recall));
RUN;

PROC REG data=recall;
   MODEL nrecall = eventyr birthyr d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12
d13 d14 d15 d16
d17 d18 d19 d20 d21 d22 d23 d24 d25 d26 / selection=forward SLENTRY= .05 SLSTAY=.05;
   OUTPUT OUT=recall r=recalladj;
RUN;

DATA recall; SET recall;
   ageev = eventyr - birthyr;
ageev2 = (ageev-21)**2;
ageev3 = (ageev-21)**3;
RUN;

PROC CORR data=recall;
   VAR AGEEV AGEEV2 AGEEV3;
RUN;

PROC REG data=recall;
  MODEL recalladj = ageev;
RUN;

PROC REG data=recall;
  MODEL ageev = eventyr birthyr;
RUN;

PROC UNIVARIATE data=recall noprint;
  VAR ageev;
  OUTPUT out=percentiles
    PCTLPTS=5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95
    PCTLPRE=p;
PROC PRINT data=percentiles;
  TITLE "Percentiles of AGEEV";
RUN;

DATA recall;
SET recall;
  IF ageev <= 7 THEN ageevc=7;
  IF ageev > 7 AND ageev <= 9.8 THEN ageevc=9;
  IF ageev > 9.8 AND ageev <= 12 THEN ageevc=12;
  IF ageev > 12 AND ageev <= 14 THEN ageevc=14;
  IF ageev > 14 AND ageev <= 15 THEN ageevc=15;
  IF ageev > 15 AND ageev <= 17 THEN ageevc=17;
  IF ageev > 17 AND ageev <= 18 THEN ageevc=18;
  IF ageev > 18 AND ageev <= 20 THEN ageevc=20;
  IF ageev > 20 AND ageev <= 21 THEN ageevc=21;
  IF ageev > 21 AND ageev <= 22 THEN ageevc=22;
  IF ageev > 22 AND ageev <= 23 THEN ageevc=23;
  IF ageev > 23 AND ageev <= 24 THEN ageevc=24;
  IF ageev > 24 AND ageev <= 26 THEN ageevc=26;
  IF ageev > 26 AND ageev <= 27 THEN ageevc=27;
  IF ageev > 27 AND ageev <= 28 THEN ageevc=28;
  IF ageev > 28 AND ageev <= 30 THEN ageevc=30;
  IF ageev > 30 AND ageev <= 31 THEN ageevc=31;
  IF ageev > 31 AND ageev <= 33 THEN ageevc=33;
  IF ageev > 33 AND ageev <= 36 THEN ageevc=36;
  IF ageev > 36 THEN ageevc=40;
RUN;

PROC SORT data=recall;
  BY ageevc;
RUN;
PROC BOXPLOT data=recall;
PLOT recalladj*ageevc /BOXSTYLE=SCHEMATIC;
RUN;

PROC REG data=recall;
  MODEL recalladj=ageev;
  OUTPUT OUT=recall p=recallest;
RUN;

PROC REG data=recall;
  MODEL recalladj=ageev ageev2;
  OUTPUT OUT=recall p=recallest2;
RUN;

PROC REG data=recall;
  MODEL recalladj=ageev ageev2 ageev3;
  OUTPUT OUT=recall p=recallest3;
RUN;

PROC GPLOT;
  PLOT recallest*ageev;
  TITLE "From model with just AGEEV";
RUN;

PROC GPLOT;
  PLOT recallest2*ageev;
  TITLE "From model with AGEEV and AGEEV2";
RUN;

PROC GPLOT;
  PLOT recallest3*ageev;
  TITLE "From model with AGEEV, AGEEV2 and AGEEV3";
RUN;