Advanced Research Dataset Homeworks

Gerardo A. Casteleiro, MS, LMHC

10/29/2019

Contents

Introduction

This document provides a step-by-step comprehensive process for completing the MHS7730 Advanced Research dataset work in R. For access to the RMarkdown file, see: https://github.com/gcasteleiro/adv_research.

Dataset 1 - Multiple Regression

1. Load the data file called EX9Q1.sav. This is a demonstration data file which for many years was supplied with every copy of SPSS (and called '1991 U.S. General Social Survey.sav') but is not with recent versions.

1.a. Open the data file and familiarise yourself with the variables. This file contains more than 40 variables for each of about 1500 respondents.

```
#import the dataset
EX9Q1 <- read_sav("EX9Q1(1).sav")

#check for missing data
any(is.na(EX9Q1))
```

```
## [1] TRUE
```

```
#check the data variable names
names(EX9Q1)
```

```
##  [1] "sex"      "race"     "region"   "happy"    "life"     "sibs"
##  [7] "childs"   "age"      "educ"     "paeduc"   "maeduc"   "speduc"
## [13] "prestg80" "occcat80" "tax"      "usintl"   "obey"     "popular"
## [19] "thnkself" "workhard" "helpoth"  "hlth1"    "hlth2"    "hlth3"
## [25] "hlth4"    "hlth5"    "hlth6"    "hlth7"    "hlth8"    "hlth9"
## [31] "work1"    "work2"    "work3"    "work4"    "work5"    "work6"
## [37] "work7"    "work8"    "work9"    "prob1"    "prob2"    "prob3"
## [43] "prob4"    "filter_$"
```

```
#count missing data
length(which(is.na(EX9Q1)))
```

```
## [1] 21033
```

1.b. The variable "prestg80" is a scale variable which codes the respondent's occupational prestige score (a higher value indicates a more prestigious occupation). We are going to investigate which other variables predict the occupational prestige score. Undertake a multiple regression to determine whether occupational prestige is predicted by the variables listed below, but first check the nominal variables for whether or not they are dichotomous (for example, use Frequencies). You will find one that is not dichotomous, and you should decide how to deal with that. You could: produce dummy variables; collapse two categories into one; or, as one category has proportionately low N, just exclude that category. Justify your decision. Predictors: respondent's sex, their race, their general happiness (happy), the number of children they have (childs), the highest year of school completed (educ), and whether the respondent takes illegal drugs (hth5) or has a drinking problem (hlth4)

Below the data is checked. Then the multiple regression model with `lm` (linear model) is fitted and tested.

```r
#check nominal variables for dichotomy
table(EX9Q1$sex)
```

```
##
##   1   2
## 636 881
```

```r
table(EX9Q1$happy) #level of happiness
```

```
##
##   1   2   3
## 467 872 165
```

```r
table(EX9Q1$childs) #no. of children
```

```
##
##   0   1   2   3   4   5   6   7   8
## 419 255 375 215 127  54  24  23  17
```

```r
table(EX9Q1$educ) #highest yr hs completed
```

```
##
##   0   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
##   2   5   5   6  12  25  68  56  73  85 461 130 175  73 194  43  45  22
##  20
##  30
```

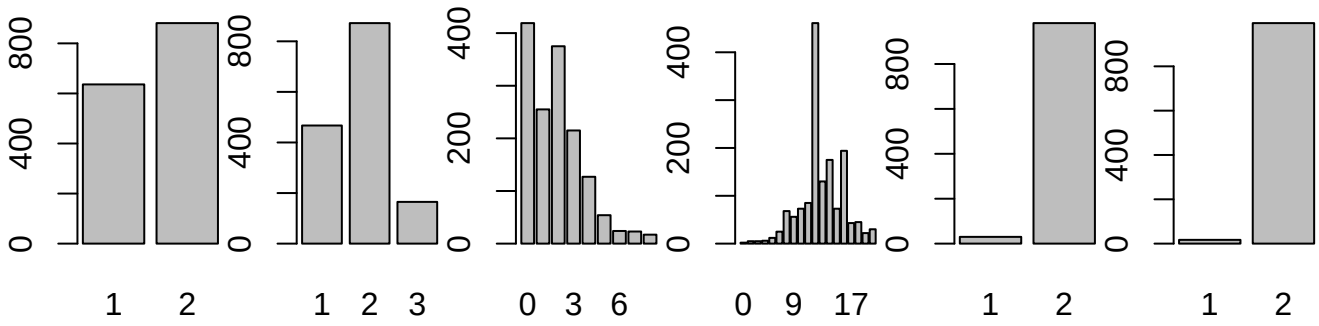
```r
table(EX9Q1$hlth5) #illegal drugs
```

```
##
##   1   2
##  30 982
```

```r
table(EX9Q1$hlth4) #drinking problem
```

```
##
##   1   2
##  17 995
```

```r
tablist <- list(
table(EX9Q1$sex),
table(EX9Q1$happy), #level of happiness
table(EX9Q1$childs), #no. of children
table(EX9Q1$educ), #highest yr hs completed
table(EX9Q1$hlth5), #illegal drugs
table(EX9Q1$hlth4) #drinking problem
)
#plots
for(i in 1:length(tablist)){
  barplot(tablist[[i]])
}
```




```
#OR: lapply(tablist,barplot)

table(EX9Q1$race) #race is nominal and not dichotomous (1 = "white" 2 = "black" 3 = "other")
```

```
##
##    1    2    3
## 1264  204   49
```

```
#for race variable, use: as.factor(EX9Q1$race)

#fit model
mr.fit <- lm(formula = prestg80 ~ sex + happy + childs + educ + hlth5 +
    hlth4 + as.factor(race), data = EX9Q1)
#check model
summary(mr.fit)
```

```
##
## Call:
## lm(formula = prestg80 ~ sex + happy + childs + educ + hlth5 +
##     hlth4 + as.factor(race), data = EX9Q1)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##     Min      1Q  Median      3Q     Max
## -33.809  -8.108  -0.080   7.720  29.685
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.9155     7.0683  -0.130  0.89697
## sex               -0.6774     0.7466  -0.907  0.36448
## happy             -0.7166     0.6020  -1.190  0.23418
## childs             0.1481     0.2247   0.659  0.51008
## educ               2.2227     0.1301  17.090  < 2e-16 ***
## hlth5              7.0864     2.1698   3.266  0.00113 **
## hlth4              2.0384     2.8497   0.715  0.47459
## as.factor(race)2  -4.1639     1.1152  -3.734  0.00020 ***
## as.factor(race)3  -2.0058     2.1801  -0.920  0.35779
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.1 on 924 degrees of freedom
##   (584 observations deleted due to missingness)
## Multiple R-squared:  0.2864, Adjusted R-squared:  0.2802
## F-statistic: 46.36 on 8 and 924 DF,  p-value: < 2.2e-16
```

We can go some steps further. Just because this model is significant doesn't mean it's the one we should use; it's definitely not the most parsimonious.

```
#backward model selection could result in a better model
#(feel free to skip this entire section)

drop1(mr.fit)
```

```
## Single term deletions
##
## Model:
## prestg80 ~ sex + happy + childs + educ + hlth5 + hlth4 + as.factor(race)
##                 Df Sum of Sq    RSS    AIC
## <none>                       113923 4500.9
## sex              1       101 114024 4499.8
## happy            1       175 114098 4500.4
## childs           1        54 113976 4499.4
## educ             1     36008 149931 4755.2
## hlth5            1      1315 115238 4509.7
## hlth4            1        63 113986 4499.5
## as.factor(race)  2      1771 115694 4511.3
```

```
##AIC is the Akaike Information Criterion, which points to the relative quality
##of statistical models in a given dataset. The lowest AIC can be taken out in the
##next model.

mr.fit2 <- lm(prestg80 ~ happy + childs + educ + hlth5 + hlth4 +
                as.factor(race), data = EX9Q1)

summary(mr.fit2)
```

```
##
## Call:
## lm(formula = prestg80 ~ happy + childs + educ + hlth5 + hlth4 +
##     as.factor(race), data = EX9Q1)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##     Min      1Q  Median      3Q     Max
## -33.425  -8.053  -0.118   7.821  30.058
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.4299     7.0449  -0.203 0.839207
## happy             -0.7515     0.6007  -1.251 0.211255
## childs             0.1351     0.2242   0.603 0.546854
## educ               2.2287     0.1299  17.160  < 2e-16 ***
## hlth5              6.8257     2.1505   3.174 0.001553 **
## hlth4              2.0251     2.8494   0.711 0.477425
## as.factor(race)2  -4.1877     1.1148  -3.757 0.000183 ***
## as.factor(race)3  -1.9602     2.1793  -0.899 0.368635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.1 on 925 degrees of freedom
##   (584 observations deleted due to missingness)
## Multiple R-squared:  0.2858, Adjusted R-squared:  0.2804
## F-statistic: 52.88 on 7 and 925 DF,  p-value: < 2.2e-16
```

```
drop1(mr.fit2)
```

```
## Single term deletions
##
## Model:
## prestg80 ~ happy + childs + educ + hlth5 + hlth4 + as.factor(race)
##                 Df Sum of Sq    RSS    AIC
## <none>                       114024 4499.8
## happy            1       193 114217 4499.4
## childs           1        45 114069 4498.1
## educ             1     36297 150321 4755.6
## hlth5            1      1242 115266 4507.9
## hlth4            1        62 114087 4498.3
## as.factor(race)  2      1788 115812 4510.3
```

```
mr.fit3 <- lm(prestg80 ~ happy + educ + hlth5 + hlth4 + as.factor(race), data = EX9Q1)

summary(mr.fit3)
```

```
##
## Call:
## lm(formula = prestg80 ~ happy + educ + hlth5 + hlth4 + as.factor(race),
##     data = EX9Q1)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##     Min      1Q  Median      3Q     Max
## -33.325  -7.954  -0.168   7.721  29.899
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.9276     7.0111  -0.132 0.894769
```

```
## happy             -0.7895     0.5991  -1.318 0.187923
## educ               2.2142     0.1258  17.594  < 2e-16 ***
## hlth5              6.8787     2.1441   3.208 0.001381 **
## hlth4              1.9591     2.8472   0.688 0.491578
## as.factor(race)2  -4.1283     1.1132  -3.708 0.000221 ***
## as.factor(race)3  -1.8633     2.1756  -0.856 0.391955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.1 on 928 degrees of freedom
##   (582 observations deleted due to missingness)
## Multiple R-squared:  0.2861, Adjusted R-squared:  0.2815
## F-statistic: 61.97 on 6 and 928 DF,  p-value: < 2.2e-16
```

```
drop1(mr.fit3)
```

```
## Single term deletions
##
## Model:
## prestg80 ~ happy + educ + hlth5 + hlth4 + as.factor(race)
##                 Df Sum of Sq    RSS    AIC
## <none>                       114281 4507.5
## happy            1       214 114495 4507.2
## educ             1     38121 152402 4774.6
## hlth5            1      1268 115548 4515.8
## hlth4            1        58 114339 4506.0
## as.factor(race)  2      1737 116017 4517.6
```

```
mr.fit4 <- lm(prestg80 ~ happy + educ + hlth5 + as.factor(race), data = EX9Q1)

summary(mr.fit4)
```

```
##
## Call:
## lm(formula = prestg80 ~ happy + educ + hlth5 + as.factor(race),
##     data = EX9Q1)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##     Min      1Q  Median      3Q     Max
## -33.283  -7.977  -0.130   7.619  29.906
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.5698     4.8542   0.529 0.596656
## happy              -0.8222     0.5971  -1.377 0.168822
## educ                2.2144     0.1257  17.610  < 2e-16 ***
## hlth5               7.1013     2.1165   3.355 0.000825 ***
## as.factor(race)2   -4.1338     1.1122  -3.717 0.000214 ***
```

```
## as.factor(race)3  -1.8191     2.1732  -0.837 0.402778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.09 on 930 degrees of freedom
##   (581 observations deleted due to missingness)
## Multiple R-squared:  0.2856, Adjusted R-squared:  0.2818
## F-statistic: 74.36 on 5 and 930 DF,  p-value: < 2.2e-16
```

```
drop1(mr.fit4)
```

```
## Single term deletions
##
## Model:
## prestg80 ~ happy + educ + hlth5 + as.factor(race)
##                 Df Sum of Sq    RSS    AIC
## <none>                       114356 4509.9
## happy            1       233 114589 4509.8
## educ             1     38135 152491 4777.3
## hlth5            1      1384 115741 4519.2
## as.factor(race)  2      1739 116095 4520.0
```

```
mr.fit5 <- lm(prestg80 ~ educ + hlth5 + as.factor(race), data = EX9Q1)

summary(mr.fit5)
```

```
##
## Call:
## lm(formula = prestg80 ~ educ + hlth5 + as.factor(race), data = EX9Q1)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##     Min      1Q  Median      3Q     Max
## -33.351  -8.304   0.144   8.150  30.604
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         0.1824     4.4728   0.041  0.967481
## educ                2.2269     0.1239  17.972   < 2e-16 ***
## hlth5               7.4725     2.0891   3.577  0.000365 ***
## as.factor(race)2   -4.3139     1.0968  -3.933 0.0000899 ***
## as.factor(race)3   -2.5420     2.0947  -1.214  0.225230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.07 on 940 degrees of freedom
##   (572 observations deleted due to missingness)
## Multiple R-squared:  0.2849, Adjusted R-squared:  0.2818
## F-statistic: 93.61 on 4 and 940 DF,  p-value: < 2.2e-16
```

```r
#test assumptions
gvlma(mr.fit5) #acceptable
```

```
##
## Call:
## lm(formula = prestg80 ~ educ + hlth5 + as.factor(race), data = EX9Q1)
##
## Coefficients:
##      (Intercept)              educ             hlth5  as.factor(race)2
##           0.1824            2.2269            7.4725           -4.3139
## as.factor(race)3
##          -2.5420
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = mr.fit5)
##
##                      Value     p-value                   Decision
## Global Stat        27.5999 0.000015032 Assumptions NOT satisfied!
## Skewness            0.3501 0.554057708    Assumptions acceptable.
## Kurtosis            2.8179 0.093215613    Assumptions acceptable.
## Link Function      23.7284 0.000001109 Assumptions NOT satisfied!
## Heteroscedasticity  0.7035 0.401621961    Assumptions acceptable.
```

```r
#there's a more streamline way to do this with udpate(), but I did it manually to
#show the iterative process

#example
test <- update(mr.fit, .~. - educ)
summary(test)
```

```
##
## Call:
## lm(formula = prestg80 ~ sex + happy + childs + hlth5 + hlth4 +
##     as.factor(race), data = EX9Q1)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##     Min      1Q  Median      3Q     Max
## -26.125  -9.580  -0.792   7.094  43.565
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
##                   Estimate Std. Error t value    Pr(>|t|)
## (Intercept)       33.2298     7.7648   4.280 0.000020674 ***
```

```
## sex               -1.3260     0.8531  -1.554     0.12044
## happy             -1.6823     0.6853  -2.455     0.01428 *
## childs            -0.7778     0.2488  -3.127     0.00182 **
## hlth5              6.7625     2.4848   2.722     0.00662 **
## hlth4              2.1934     3.2638   0.672     0.50173
## as.factor(race)2  -6.2818     1.2616  -4.979 0.000000762 ***
## as.factor(race)3  -2.7487     2.4966  -1.101     0.27119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.72 on 927 degrees of freedom
##   (582 observations deleted due to missingness)
## Multiple R-squared:  0.0623, Adjusted R-squared:  0.05522
## F-statistic: 8.798 on 7 and 927 DF,  p-value: 0.0000000001663
```

1.c. Report the results of the analysis.

In the selected model (`mr.fit5`), highest year of school completed (`educ`), illegal drugs (`hlth5`) and (`race`) significantly accounted for approximately 28% of the variance of occupational prestige (`prestg80`). The model was significant, ($F(4,940) = 93.61$, $p < .001$).

1.d. Repeat this analysis separately for men and women. Are there any major differences in

the pattern of results for these two groups?

We separate the data using pipes `%>%` and the `tidyverse` package (`dyplr`)

```
men.data <- EX9Q1 %>% filter(sex == 1)
men.fit <- lm(prestg80 ~ happy + childs + educ + hlth5 + hlth4 +
                as.factor(race), data = men.data)
summary(men.fit)
```

```
##
## Call:
## lm(formula = prestg80 ~ happy + childs + educ + hlth5 + hlth4 +
##     as.factor(race), data = men.data)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##      Min       1Q   Median       3Q      Max
## -28.6196  -7.7728  -0.3973   7.5305  28.9437
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)        5.60958    9.55430   0.587     0.557
## happy             -0.95896    0.96465  -0.994     0.321
## childs             0.06713    0.34588   0.194     0.846
```

```
## educ               1.86065    0.18267  10.186   < 2e-16 ***
## hlth5             10.50286    2.52832   4.154 0.0000401 ***
## hlth4             -2.04827    4.13409  -0.495     0.621
## as.factor(race)2 -7.32852    1.84228  -3.978 0.0000828 ***
## as.factor(race)3 -2.19857    3.20170  -0.687     0.493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.3 on 393 degrees of freedom
##   (235 observations deleted due to missingness)
## Multiple R-squared:  0.3041, Adjusted R-squared:  0.2917
## F-statistic: 24.54 on 7 and 393 DF,  p-value: < 2.2e-16
```

It seems that for men, education, drug use and race are highly predictive of occupational prestige score; they account for approximately 29% of the variance of job prestige, and the model was significant ($F(7,393) = 24.54$, $p < .001$).

```
women.data <- EX9Q1 %>% filter(sex == 2)
women.fit <- lm(prestg80 ~ happy + childs + educ + hlth5 + hlth4 + as.factor(race), data = women.data)
summary(women.fit)
```

```
##
## Call:
## lm(formula = prestg80 ~ happy + childs + educ + hlth5 + hlth4 +
##     as.factor(race), data = women.data)
##
## Residuals:
## <Labelled double>: R's Occupational Prestige Score (1980)
##     Min      1Q  Median      3Q     Max
## -31.632  -8.307   0.116   6.945  28.028
##
## Labels:
##  value     label
##      0 DK,NA,NAP
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.1826    11.8270   0.523   0.6014
## happy              -0.6977     0.7589  -0.919   0.3583
## childs              0.2777     0.2924   0.950   0.3426
## educ                2.5894     0.1851  13.987   <2e-16 ***
## hlth5              -5.0361     4.5126  -1.116   0.2649
## hlth4               7.2431     3.8945   1.860   0.0635 .
## as.factor(race)2  -1.9813     1.3789  -1.437   0.1514
## as.factor(race)3  -1.5673     2.9415  -0.533   0.5944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.79 on 524 degrees of freedom
##   (349 observations deleted due to missingness)
## Multiple R-squared:  0.2995, Adjusted R-squared:  0.2902
## F-statistic: 32.01 on 7 and 524 DF,  p-value: < 2.2e-16
```

On the other hand, for women, education alone seems to account for occupational prestige score, the model accounts for approximately 30% of the variance, and is significant ($F(7,524) = 32.01$, $p < .001$). Drug use also seems to be suggestive, but is not significant at the 0.05 alpha.

Dataset 2 - Analyses of Variance

1. Load the data file Ex8Q1.sav. This file contains the data collected during a study of the effects of time on memory for details of a crime. A total of thirty five participants watched a video of a crime. They were then interviewed either one, two or three days later. Participants' memory for the details of the crime was scored out of a total of 50.

1.a. First, to check whether the allocation of participants to conditions was random, compare the age of the participants in the three groups. Is the result good news for the experimenters?

Compare age of participants in groups:

```
na.omit(Ex8Q1) %>%
  group_by(Delay)%>%
  dplyr::select(Age)%>%
  summarize_all(mean)
```

```
## Adding missing grouping variables: `Delay`
```

```
## # A tibble: 3 x 2
##        Delay   Age
##    <dbl+lbl> <dbl>
## 1 1 [1 day]   20.6
## 2 2 [2 days]  20.2
## 3 3 [3 days]  20
```

```
#good news for experiments. Mean age is about the same for all three grps
```

```
#confirm using base R
mean(na.omit(Ex8Q1[Ex8Q1$Delay == 1,]$Age))
```

```
## [1] 20.58333
```

```
mean(na.omit(Ex8Q1[Ex8Q1$Delay == 2,]$Age))
```

```
## [1] 20.25
```

```
mean(na.omit(Ex8Q1[Ex8Q1$Delay == 3,]$Age))
```

```
## [1] 20
```

1.b. Now test the hypothesis that the duration of the delay between encoding and recall will affect the accuracy of recall. Report the result of your analysis.

Fit ANOVA model

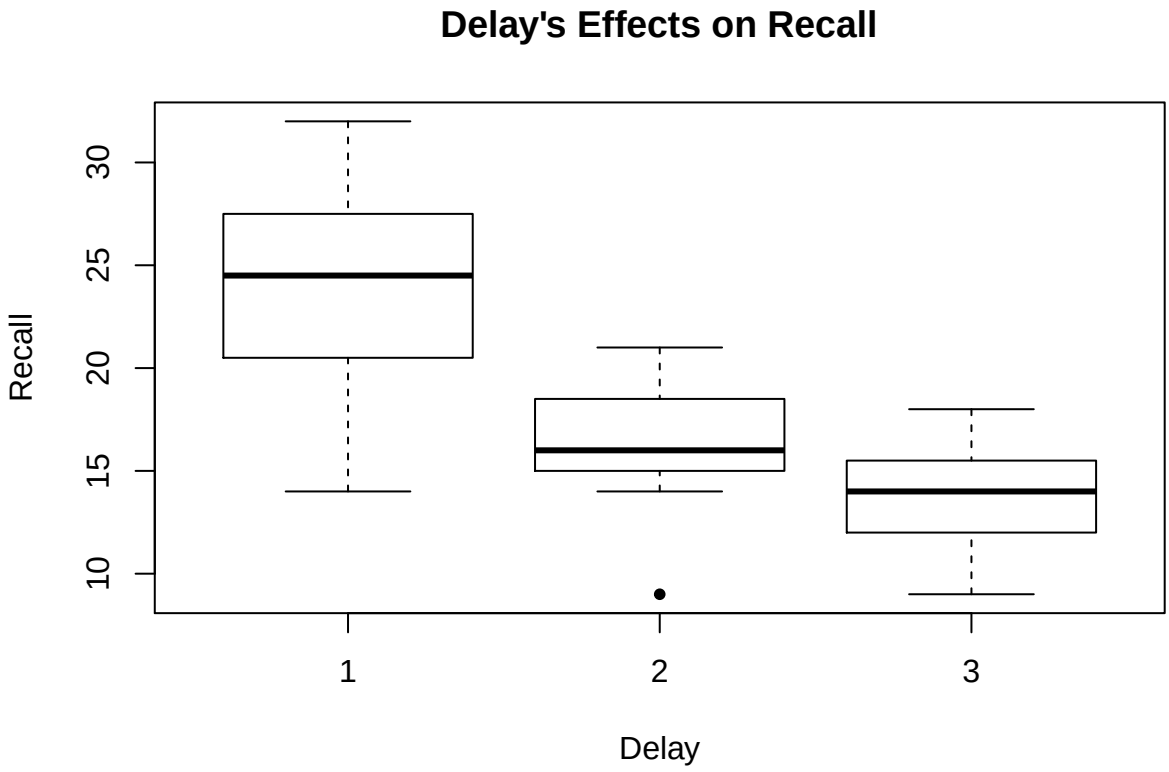


Figure 1: Graph depicting the effects of delay on recall

```
aov.fit <- aov(Recall~as.factor(Delay), data = Ex8Q1)

#summary
summary(aov.fit)
```

```
##                  Df Sum Sq Mean Sq F value     Pr(>F)
## as.factor(Delay)  2  625.0  312.52    21.5 0.00000121 ***
## Residuals        32  465.2   14.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

The duration of delay between encoding and recall will affect the accuracy of recall ($F(2,32) = 21.5$, $p < .001$).

1.c. Produce a graph showing recall at each delay, and perform post-hoc tests to determine whether each increase in delay results in a significant decrease in recall.

Graph

```
boxplot(Recall~Delay, data = Ex8Q1, pch = 20, main = "Delay's Effects on Recall")
```

Now we confirm with a post-hoc test:

```r
#Tukey Honest Significant Differences test
TukeyHSD(aov.fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Recall ~ as.factor(Delay), data = Ex8Q1)
##
## $`as.factor(Delay)`
##          diff        lwr       upr     p adj
## 2-1 -7.500000 -11.325250 -3.674750 0.0000978
## 3-1 -9.924242 -13.835464 -6.013021 0.0000016
## 3-2 -2.424242  -6.335464  1.486979 0.2937774
```

```r
cohen.d(Ex8Q1[,4:5],"Delay")
```

```
## Call: cohen.d(x = Ex8Q1[, 4:5], group = "Delay")
## Cohen d statistic of difference between two means
##        lower effect upper
## Recall  -2.9   -1.8 -0.72
##
## Multivariate (Mahalanobis) distance between groups
## [1] 1.8
## r equivalent of difference between two means
## Recall
##  -0.68
```

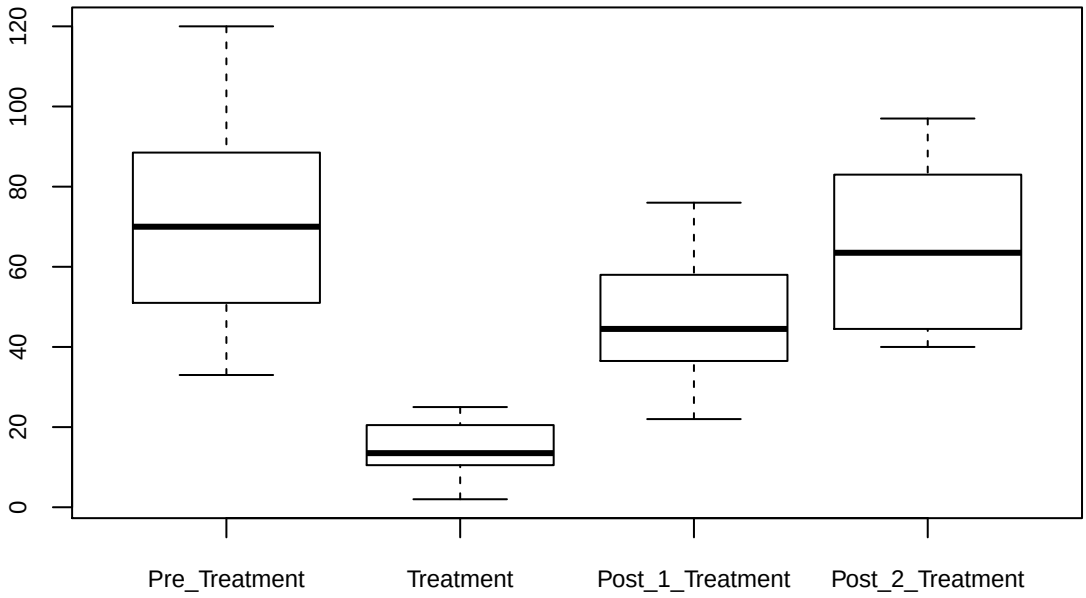
Differences in 1-2 and 1-3 are significant, but 2-3 is not. Overall, delay has a major detrimental effect on recall (Cohen's d = -1.8).

2. A psychologist is evaluating a prison-based treatment program for violent offenders serving long sentences. The psychologist was given access to the prison records which included information about any official reprimands received, or misdemeanours committed by the offenders. She used this information to calculate a behaviour score for each offender (high value indicates poor behaviour). The psychologist calculated this score for the year immediately prior to treatment, for the year of treatment and for each of the first and second years following treatment. The file Ex8Q2.sav contains the data from 12 prisoners.

2.a. Does the behaviour score change significantly across the four years?

Probably the easiest way to see the change is with a boxplot (HOW we should find out was not specified).

```r
boxplot(Ex8Q2[,-1], cex.axis = 0.75, main = "Prisoner Bx Change Over Time")
```




```
#1st column is ID so we take it out with subsetting [,-1].
```

In order to use the ezANOVA function in "ez" package, we need "long" data this means that each kind of treatment for each prisoner needs to have its own column. We use the `reshape` package to `melt` the data.

```
ldata <- melt(data.frame(Ex8Q2),
                id = "Prisoner_ID",
                tx = c("Pre_Treatment","Treatment","Post_1_Treatment","Post_2_Treatment"))
colnames(ldata) <- c("id","tx","score") #name the columns to variable names (iv/dv)
head(ldata) #the head() function prints the 1st six rows of the data.frame
```

```
##   id            tx score
## 1  1 Pre_Treatment    46
## 2  2 Pre_Treatment    66
## 3  3 Pre_Treatment    44
## 4  4 Pre_Treatment    89
## 5  5 Pre_Treatment    88
## 6  6 Pre_Treatment    62
```

Run the ANOVA:

```
ezANOVA(
  data = ldata,
  wid = id, #this asks for the participant's ID
  within = tx, #the "within-subjects" condition
  dv = score, #the dependent variable
  type =3 #this asks for a certain type of math - type 3 SPSS-comparable results
)
```

```
## $ANOVA
```

```
##   Effect DFn DFd        F                  p p<.05       ges
## 2     tx   3  33 34.85599 0.0000000002427302     * 0.6208088
##
## $`Mauchly's Test for Sphericity`
##   Effect         W         p p<.05
## 2     tx 0.4511836 0.1732339
##
## $`Sphericity Corrections`
##   Effect       GGe           p[GG] p[GG]<.05       HFe            p[HF]
## 2     tx 0.6534505 0.0000001955834         * 0.7936968 0.00000001294157
##   p[HF]<.05
## 2         *
```

```
#Mauchly's test of sphericity is not significant, assumptions are met without need
#of corrections.
```

2.b. During treatment the prisoners are housed in a special unit, and as a result the behaviour score for the year of treatment may not be comparable to the measures taken before and after treatment. Reanalyse the data excluding this data point.

Now we're going to run it again minus the "Treatment" condition (2 ways to do this):

```
ezANOVA(
  data = ldata[which(ldata$tx!="Treatment"),], #subsetting out the condition
  wid = id,
  within = tx,
  dv = score,
  type =3
)
```

```
## $ANOVA
##   Effect DFn DFd        F           p p<.05       ges
## 2     tx   2  22 8.064098 0.002360156     * 0.2253756
##
## $`Mauchly's Test for Sphericity`
##   Effect         W         p p<.05
## 2     tx 0.7659468 0.2636289
##
## $`Sphericity Corrections`
##   Effect       GGe       p[GG] p[GG]<.05       HFe       p[HF] p[HF]<.05
## 2     tx 0.8103378 0.004856575         * 0.9301367 0.003076188         *
```

```
#or re-fitting the data in the 'melting' process...
ldata2 <- melt(data.frame(Ex8Q2[,-3]),
               id = "Prisoner_ID",
               tx = c("Pre_Treatment","Post_1_Treatment","Post_2_Treatment"))
colnames(ldata2) <- c("id","tx","score") #name the columns to variable names (iv/dv)

ezANOVA(
  data = ldata2,
```

```
  wid = id,
  within = tx,
  dv = score,
  type =3
)
```

```
## $ANOVA
##   Effect DFn DFd        F           p p<.05       ges
## 2     tx   2  22 8.064098 0.002360156     * 0.2253756
##
## $`Mauchly's Test for Sphericity`
##   Effect         W         p p<.05
## 2     tx 0.7659468 0.2636289
##
## $`Sphericity Corrections`
##   Effect       GGe       p[GG] p[GG]<.05       HFe       p[HF] p[HF]<.05
## 2     tx 0.8103378 0.004856575         * 0.9301367 0.003076188         *
```

The result is still significant.

2.c. The psychologist planned to compare the behaviour at one year and two years post treatment with the behaviour prior to treatment. Undertake a contrast which includes these comparisons.

Given that I'm using another package to calculate the repeated measures ANOVA, I'm not able to conduct post-hoc tests because I'm not getting so much an "object" as a "list" with the results. Instead, I complete comparisons using paired samples t-tests, using my "wide" data (original or raw data)

```
t.test(Ex8Q2$Pre_Treatment,Ex8Q2$Post_1_Treatment, paired = TRUE) #significant
```

```
##
##  Paired t-test
##
## data:  Ex8Q2$Pre_Treatment and Ex8Q2$Post_1_Treatment
## t = 3.1311, df = 11, p-value = 0.009559
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.302683 41.863984
## sample estimates:
## mean of the differences
##                24.58333
```

```
t.test(Ex8Q2$Pre_Treatment,Ex8Q2$Post_2_Treatment, paired = TRUE) #not significant
```

```
##
##  Paired t-test
##
## data:  Ex8Q2$Pre_Treatment and Ex8Q2$Post_2_Treatment
## t = 0.91939, df = 11, p-value = 0.3776
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -7.318368 17.818368
## sample estimates:
## mean of the differences
##                    5.25
```

The other option is using `pairwise.t.test()`, which takes the values of a vector (x) by groups (g)... for which I can use the "long data" that I created for the repeated measures ANOVA. This is important because if I were doing a lot of these I wouldn't want to keep switching between the two. It helps for preventing errors and increasing reproducibility. It also lets me use p-value adjustments (e.g., "Bonferroni" correction).

```
pairwise.t.test(ldata2$score, ldata2$tx, p.adjust.method = "bonf", paired = TRUE)
```

```
##
##  Pairwise comparisons using paired t tests
##
## data:  ldata2$score and ldata2$tx
##
##                  Pre_Treatment Post_1_Treatment
## Post_1_Treatment 0.029         -
## Post_2_Treatment 1.000         0.015
##
## P value adjustment method: bonferroni
```

```
#Using Bonferroni correction
```

Two paired-samples t-tests were conducted to compare the means between pre-treatment behavior scores and one- and two-year post-treatment scores. There was a significant difference in the scores for pre-treatment (M = 70.9, SD = 24.6) and one-year post-treatment (M = 46.3, SD = 15.9) behavior scores; t(11) = 0.92, p = 0.009. On the other hand, there is no significant difference between pre-treatment (M = 70.9, SD = 24.6) and two-year post-treatment (M = 65.6, SD = 19.9) behavior scores; t(11) = 0.92, p = .378.

2.d. What would you conclude about the effectiveness of this treatment?

The treatment seems to be effective but only for the short term. There is a large effect size (η^2 = 0.385, CI_{95} = -0.035, 0.810).

3. It has been suggested that some dyslexic children are affected by the colour of the paper and text when trying to read, possibly because they find the "glare" of white paper off-putting. To investigate this 15 dyslexic and 15 non-dyslexic male year 7 school pupils were tested. All participants were asked to read 3 matched passages as quickly and accurately as possible. The time taken to read each passage was recorded and a 5 second penalty was added for each error made. One passage was presented in white on black paper (W/B), one printed in black on white paper (B/W) and one in black on yellow (B/Y). Each participant saw each colour combination once, and the order in which the passages were read and the pairing of passage to colour combination were determined at random for each child. The data from this study are tabulated in Table 1 below.

3.a. Describe the design of this study.

This study is a 2x3 mixed factorial design.

3.b. Prepare a data file for this study.

```
df <- data.frame(id = 1:20,
                 type = factor(c(rep("dys",10),rep("non",10))),
                 wb = c(40,48,39,40,46,52,61,41,53,42,28,21,23,30,26,20,26,30,23,20),
                 bw = c(45,50,38,38,43,50,58,40,55,38,30,25,24,28,32,24,25,27,20,21),
                 by = c(44,51,40,37,44,53,56,39,57,45,32,24,23,27,29,22,25,26,32,22)
                 )
head(df)
```

```
##   id type wb bw by
## 1  1  dys 40 45 44
## 2  2  dys 48 50 51
## 3  3  dys 39 38 40
## 4  4  dys 40 38 37
## 5  5  dys 46 43 44
## 6  6  dys 52 50 53
```

3.c. Analyse the data to determine whether there is any evidence that text/paper colour combination affects reading speed in dyslexic children.

```
df %>%
  filter(type == "dys")%>%
  summarise_at(.vars = c("wb","bw","by"),mean)
```

```
##     wb   bw   by
## 1 46.2 45.5 46.6
```

```
(46.2 + 46.6)/2 - 45.5 #0.9 second slower for black on white paper. but is it sig?
```

```
## [1] 0.9
```

3.d. Write the results section of a report describing the outcome of this study.

```
str(ldata3)
```

```
## 'data.frame':    60 obs. of  4 variables:
##  $ id        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ink_pap   : Factor w/ 3 levels "wb","bw","by": 1 1 1 1 1 1 1 1 1 1 ...
##  $ read_speed: num  40 48 39 40 46 52 61 41 53 42 ...
##  $ type      : Factor w/ 2 levels "dys","non": 1 1 1 1 1 1 1 1 1 1 ...
```

```
#str shows the "structure" of the data, helps fit the model below

aov.fit4 <- aov(read_speed ~ ink_pap, data = ldata3[ldata3$type == "dys",])

#summarize model
summary(aov.fit4)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## ink_pap      2    6.2    3.10   0.058  0.943
## Residuals   27 1434.5   53.13
```

```
#model not significant

#post-hoc just for kicks
TukeyHSD(aov.fit4)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = read_speed ~ ink_pap, data = ldata3[ldata3$type == "dys", ])
##
## $ink_pap
##       diff       lwr      upr     p adj
## bw-wb -0.7 -8.782265 7.382265 0.9749226
## by-wb  0.4 -7.682265 8.482265 0.9917355
## by-bw  1.1 -6.982265 9.182265 0.9393007
```

```
#also goose-egg
```

For individuals with dyslexia, there was no significant effect of color of ink/paper on reading speed at the $p < .05$ for three conditions ($F(2,77) = 0.058$, $p = .943$).

```
aov.fit5 <- aov(read_speed ~ type, data = ldata3)
summary(aov.fit5)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         1   6365    6365   200.6 <2e-16 ***
## Residuals   58   1840      32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(aov.fit5)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = read_speed ~ type, data = ldata3)
##
## $type
##          diff       lwr       upr p adj
## non-dys -20.6 -23.51122 -17.68878     0
```

```
ldata3 %>%
  group_by(type) %>%
  dplyr::select(read_speed)%>%
  summarize_all(mean)
```

```
## Adding missing grouping variables: `type`
```

```
## # A tibble: 2 x 2
##   type  read_speed
##   <fct>      <dbl>
## 1 dys         46.1
## 2 non         25.5
```

```
46.1 - 25.5
```

```
## [1] 20.6
```

```
ezANOVA(
  data = ldata3,
  wid = id,
  between = type,
  dv = read_speed,
  type =3
)
```

```
## Coefficient covariances computed by hccm()
```

```
## $ANOVA
##   Effect DFn DFd        F               p p<.05       ges
## 2   type   1  18 70.68885 0.0000001198907     * 0.7970433
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd      SSn      SSd        F          p p<.05
## 1   1  18 49.08889 209.8333 4.210961 0.05500009
```

However, the 2x3 ANOVA model including the two groups (dyslexic/non-dyslexic) revealed that, as expected, dyslexia had a significant main effect on reading time $F(1, 18) = 70.68$, $p < .001$, $\eta^2 = 0.79$ [1]

[1] Cohen's $f = 1.939563$; to make sure my scientific notation skills are on par -> 1.198907e-07 < .001 = TRUE

Dataset 3 - ANCOVA & MANOVA

1. A psychologist who is interested in aggression has devised an experimental paradigm in which participants play a computer game with an opponent. When the opponent makes an error the participant is invited to "punish" their opponent by exposing him to a blast of loud noise. The duration and volume of the noise blast are combined to give a measure of aggression. A total of 60 participants were tested using this procedure before being given feedback about their performance in the game. One third of the participants received negative feedback, one third received positive feedback and the remaining third received neutral feedback. Finally, the participants played the game again and their level of aggression was measured as before. The data from this study can be found in the file Ex10Q1.sav.

```
Ex10Q1 <- read_sav("Ex10Q1.sav")
str(Ex10Q1)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    62 obs. of  4 variables:
##  $ Part_Num : num  1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "format.spss")= chr "F8.0"
##  $ Condition: 'haven_labelled' num  1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "format.spss")= chr "F8.0"
##   ..- attr(*, "display_width")= int 10
##   ..- attr(*, "labels")= Named num  1 2 3
##   .. ..- attr(*, "names")= chr  "Positive Interaction" "Neutral Intraction" "Negative Interaction"
##  $ Pre_test : num  133 114 126 133 118 ...
##   ..- attr(*, "format.spss")= chr "F8.2"
##  $ Post_Test: num  80 114.1 99 119.8 97.1 ...
##   ..- attr(*, "format.spss")= chr "F8.2"
```

1.a. Undertake an ANOVA to determine whether the post-feedback levels of aggression are affected by feedback.

Fit ANOVA

```
aov.fit6 <- aov(Post_Test ~ as.factor(Condition), data = Ex10Q1)
summary(aov.fit6)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Condition)  2   3510  1755.0   3.881 0.0263 *
## Residuals            57  25777   452.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

Post-test levels of aggression are affected by feedback.

1.b. Participants were randomly assigned to each of the three feedback conditions, and as a result the pre-test scores for these three groups should not differ. Test whether this is the case.

Test the pre-test for differences

```
summary(aov(Pre_test ~ as.factor(Condition), data = Ex10Q1))
```

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(Condition)  2   2700  1350.0    5.59 0.00607 **
## Residuals            57  13765   241.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

```
#nope - problem!
```

There is clearly a difference in the pretest.

1.c. In light of the answer to the previous question, use the pre-test scores as a covariate and re-examine the effect of feedback on aggression.

To run the ANCOVA, include the "covariant" variable(s) `Pre_test` *in front* of the IV in the formula.

```
summary(aov(Post_Test ~ Pre_test + as.factor(Condition), data = Ex10Q1))
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## Pre_test              1   2087  2086.7   4.686 0.0347 *
## as.factor(Condition)  2   2263  1131.3   2.540 0.0879 .
## Residuals            56  24938   445.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

This shows that when the `Pre_test` is included in the formula, the main effect for `Condition` is not significant.

Option 2 - create a gain score - subtract pre_test scores from post_test.

```
Ex10Q1$gain <- Ex10Q1$Post_Test - Ex10Q1$Pre_test
#create new variable gain score

summary(aov(gain ~ as.factor(Condition), data = Ex10Q1))
```

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(Condition)  2  11136    5568   6.849 0.00216 **
## Residuals            57  46340     813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

```
#run one-way ANOVA
```

1.d. How does the inclusion of pre-test aggression as a covariate change the outcome of the analysis?

With the inclusion of pre-test as a covariate, the p value decreases and the F statistic increases.

```
TukeyHSD(aov(Post_Test ~ as.factor(Condition) * Pre_test, data = Ex10Q1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Post_Test ~ as.factor(Condition) * Pre_test, data = Ex10Q1)
##
## $`as.factor(Condition)`
##          diff        lwr       upr     p adj
## 2-1 18.534264   2.566008 34.502520 0.0192146
## 3-1 11.635607  -4.332649 27.603863 0.1941648
## 3-2 -6.898657 -22.866913  9.069599 0.5544382
```

1.e. What should the psychologist conclude regarding the effect of feedback on aggression?

The psychologist should conclude that *some* of the difference in the `post_test` is essentially accounted for in the `pre_test`. Additionally, the kind of feedback has an effect on aggression. The biggest, and only significant, difference is noted between "positive" and "neutral" feedback.

2. Geiselman and colleagues 1 developed the Cognitive Interview (CI) to help police officers obtain accurate information from witnesses. Research has demonstrated that the use of the CI results in an increase in recall for the details of an event, however, there is less evidence that the CI results in more accurate descriptions of the people involved in the event. A psychologist has developed a new interview, which she calls the "Visual Interview" (VI) which is specifically designed to help witness describe the people they saw. A group of 20 participants watched a video of two actors performing a number of actions. After a delay of 24 hours the participants were interviewed using the CI or the VI. Each participant's description of the actions was scored out of 100 and their description of the appearance of the actors was scored out of 60. The data are contained in file Ex10Q2.sav.

```
Ex10q2 <- read_sav("Ex10q2.sav")
head(Ex10q2)
```

```
## # A tibble: 6 x 4
##   Part_Number Interview Memory_Events Memory_People
##         <dbl> <dbl+lbl>         <dbl>         <dbl>
```

```
## 1           1    1 [CI]          80.9          42.5
## 2           2    1 [CI]          56.7          31.0
## 3           3    1 [CI]          62.3          13.4
## 4           4    1 [CI]          41            20.0
## 5           5    1 [CI]          70.4          20.2
## 6           6    1 [CI]          43.6          30.0
```

2.a. Describe the design of this study

2x2 mixed design

2.b. Check your data to determine whether it is appropriate for analysis using MANOVA.

Test data to ensure it is appropriate.

Correlations are checked between the dependent variables to check for correlations above .8, which can be problematic:

```
#Pearson's product-moment correlation
cor.test(Ex10q2$Memory_Events,Ex10q2$Memory_People)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Ex10q2$Memory_Events and Ex10q2$Memory_People
## t = 1.0762, df = 18, p-value = 0.2961
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.220663  0.620839
## sample estimates:
##       cor
## 0.2458661
```

```
#0.24 - no problem found
```

Test for normality:

```
#Shapiro-Wilk test
shapiro.test(Ex10q2$Memory_Events)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ex10q2$Memory_Events
## W = 0.97122, p-value = 0.7804
```

```
#normal
shapiro.test(Ex10q2$Memory_People)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ex10q2$Memory_People
## W = 0.98082, p-value = 0.9443
```

```
#normal
leveneTest(Ex10q2$Memory_People, as.factor(Ex10q2$Interview))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.5008 0.4882
##       18
```

```
leveneTest(Ex10q2$Memory_Events, as.factor(Ex10q2$Interview))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0203 0.8883
##       18
```

```
#homogeneity of variance
```

Data is appropriate.

2.c. The psychologists hypothesised that the VI would result in better memory for the appearance of the actors, but that there would be no difference between the VI and CI groups for recall of the events. She predicted that this relationship would hold for both short and long delays. Analyse the data to test these hypotheses.

There are no observations provided in this dataset for delay.

2.d. Assuming these results are reliable, what are the implications with regard to how police should interview witnesses?

Fit the model:

```
manova.fit <- manova(as.matrix(Ex10q2)[,3:4] ~ Ex10q2$Interview)
manova.fit
```

```
## Call:
##    manova(as.matrix(Ex10q2)[, 3:4] ~ Ex10q2$Interview)
##
## Terms:
##                  Ex10q2$Interview Residuals
## Memory_Events              87.726  3757.102
## Memory_People             447.720  1145.321
## Deg. of Freedom                 1        18
```

```
##
## Residual standard errors: 14.44742 7.976775
## Estimated effects may be unbalanced
```

```
summary(manova.fit, test = "Pillai")
```

```
##                  Df  Pillai approx F num Df den Df  Pr(>F)
## Ex10q2$Interview  1 0.36533   4.8927      2     17 0.02097 *
## Residuals        18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Pillai's Trace is the most robust test
summary(aov(Memory_Events ~ Interview, Ex10q2))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Interview    1     88   87.73    0.42  0.525
## Residuals   18   3757  208.73
```

```
#not sig
summary(aov(Memory_People ~ Interview, Ex10q2))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Interview    1  447.7   447.7   7.036 0.0162 *
## Residuals   18 1145.3    63.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#significant
eta_sq(aov(Memory_People ~ Interview, Ex10q2),partial = T, ci.lvl = 0.95)
```

```
##        term partial.etasq conf.low conf.high
## 1 Interview         0.281    0.009     0.531
```

```
#effect size and confidence intervals
```

Assumptions of homogeneity of variance-covariance matrices and equality of variance were confirmed, and small correlations were found among the dependent variables. There was a significant difference between the two interview types, $F(2,17) = 4.892$, $p = .021$, Pillai's Trace = .365. Analyses of the independent variables (Bonferroni adjusted $\alpha = .05/2$), showed that the 'memory events' condition did not significantly differ between interviews $F(1,18) = 0.42$, $p = .525$. Significant differences in interview were found for the 'memory people' condition $F(1,18) = 7.04$, $p = .016$. Individuals who had the Cognitive Interview (CI) had a lower mean score on 'memory people' condition ($M = 28.6$) than the individuals who had the Visual Interview (VI) ($M = 38.1$). There was a large effect size, $\eta_p^2 = .281$, 95% CI [0.009, 0.531].

Police should use the Visual Interview (VI) to interview witnesses about people given that there is a statistically significant difference of approximately 9.5 points in accuracy.

Dataset 4 - Correlation

These exercises have been prepared for use in conjunction with Chapter 6 of the 5th edition of "SPSS for Psychologists" by Brace, Kemp and Snelgar (2012). This exercise uses the data file Employee data.sav which we corrected as part of Exercise 5. Load the corrected file now

1. Is there any evidence of a correlation between starting salary and current salary? What is the magnitude and direction of the correlation and is it statistically significant? What does this tell us?

Correlation between starting and current salary

```
head(Employee_data)
```

```
## # A tibble: 6 x 10
##      id gender bdate         educ  jobcat salary salbegin jobtime prevexp
##   <dbl>  <dbl> <date>     <dbl+l> <dbl+l> <dbl+> <dbl+lb> <dbl+l> <dbl+l>
## 1     1      2 1952-02-03 15 [15] 3 [Man~  57000    27000      98     144
## 2     2      2 1958-05-23 16 [16] 1 [Cle~  40200    18750      98      36
## 3     3      1 1929-07-26 12 [12] 1 [Cle~  21450    12000      98     381
## 4     4      1 1947-04-15  8 [8]  1 [Cle~  21900    13200      98     190
## 5     5      2 1955-02-09 15 [15] 1 [Cle~  45000    21000      98     138
## 6     6      2 1958-08-22 15 [15] 1 [Cle~  32100    13500      98      67
## # ... with 1 more variable: minority <dbl+lbl>
```

```
cor.test(Employee_data$salbegin, Employee_data$salary)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Employee_data$salbegin and Employee_data$salary
## t = 40.276, df = 472, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8580696 0.8989267
## sample estimates:
##       cor
## 0.8801175
```

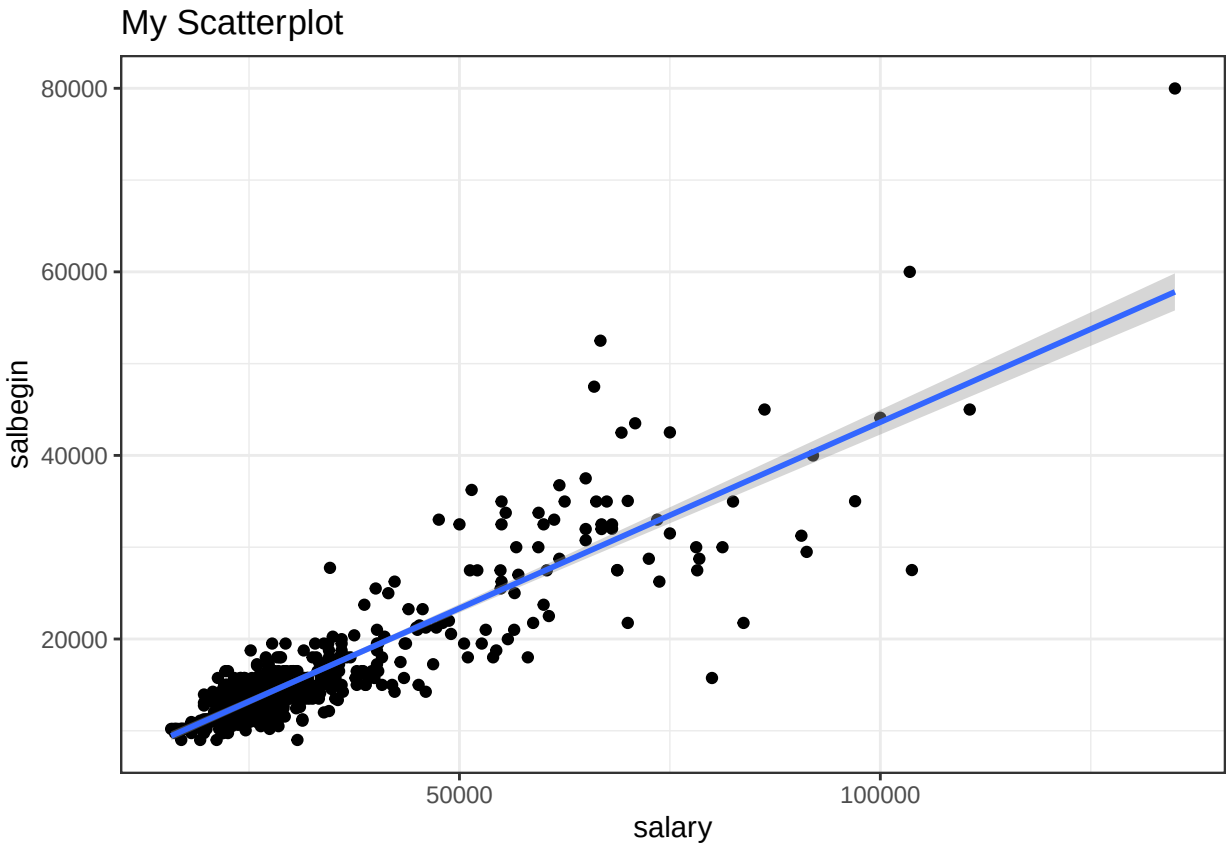
This relationship is highly correlated and significant (r = .88, N = 474, $p < .001$).

2. Draw a scattergram to illustrate the relationship between starting salary and current salary. Add a regression line to the scattergram.

"Scatterplot"

```
Employee_data%>%
  dplyr::select(salbegin,salary)%>%
  ggplot(aes(x = salary , y = salbegin)) + geom_point() +
  geom_smooth(method = "lm") + theme_bw() + ggtitle("My Scatterplot")
```

```
## Don't know how to automatically pick scale for object of type haven_labelled. Defaulting to continuou
## Don't know how to automatically pick scale for object of type haven_labelled. Defaulting to continuou
```



3. What percentage of the variance in current salary is explained by starting salary?

Percentage of variance (regression)

```
summary(lm(salbegin~salary, data = Employee_data))
```

```
##
## Call:
## lm(formula = salbegin ~ salary, data = Employee_data)
##
## Residuals:
## <Labelled double>: Beginning Salary
##      Min       1Q   Median       3Q      Max
## -19756.7  -1604.1    -61.8   1277.9  22368.4
##
## Labels:
##  value   label
##      0 missing
##
## Coefficients:
##               Estimate Std. Error t value           Pr(>|t|)
```

```
## (Intercept) 3053.09547  386.92469   7.891 0.0000000000000211 ***
## salary         0.40567    0.01007  40.276            < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3741 on 472 degrees of freedom
## Multiple R-squared:  0.7746, Adjusted R-squared:  0.7741
## F-statistic:  1622 on 1 and 472 DF,  p-value: < 2.2e-16
```

Approximately 77% of the variance ($F(1, 472) = 1622$, $p < .001$).

4. Produce a correlation matrix showing the correlations between the following variables: beginning salary, current salary, time in the job, previous experience.

Correlation matrix

```
cor(Employee_data[,6:9]) #accomplished with subsetting [ ]
```

```
##               salary    salbegin      jobtime      prevexp
## salary    1.00000000  0.88011747  0.084092267 -0.097466926
## salbegin  0.88011747  1.00000000 -0.019753475  0.045135627
## jobtime   0.08409227 -0.01975347  1.000000000  0.002978134
## prevexp  -0.09746693  0.04513563  0.002978134  1.000000000
```

5. Examine the correlation matrix and identify which of these correlations are statistically significant.

Significant correlations

```
cor.test(Employee_data$prevexp,Employee_data$salary) #yes
```

```
##
##  Pearson's product-moment correlation
##
## data:  Employee_data$prevexp and Employee_data$salary
## t = -2.1277, df = 472, p-value = 0.03388
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.185900660 -0.007466824
## sample estimates:
##         cor
## -0.09746693
```

```
cor.test(Employee_data$jobtime,Employee_data$salary) #no (p = .07)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Employee_data$jobtime and Employee_data$salary
## t = 1.8334, df = 472, p-value = 0.06737
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006018969  0.172848789
## sample estimates:
##       cor
## 0.08409227
```

6. Which two variables are significantly negatively correlated? Can you suggest possible reasons for this relationship?

There is a weak, but significant, negative correlation is between previous experience and current salary (r = -.1, p = .03). Maybe previous experience led to some bad habits that are holding people back?

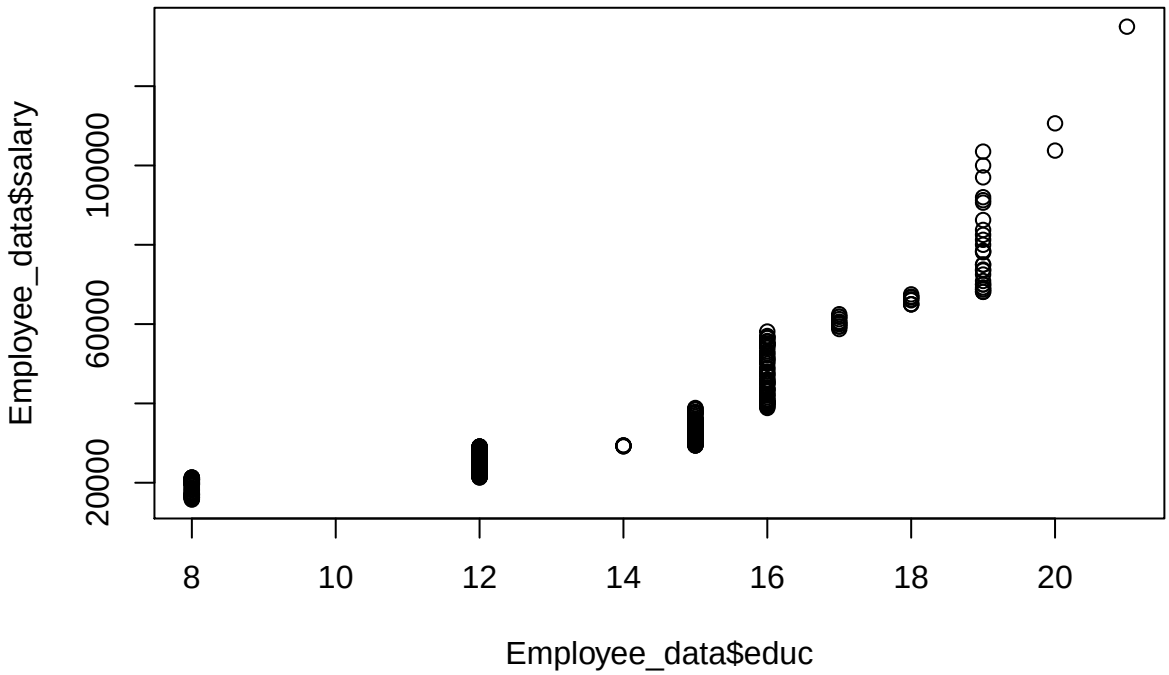
7. An organisational Psychologist wants to know whether there is a relationship between education and current salary. What is the most appropriate statistical test to use for this analysis? Justify your answer. Is the correlation significant?

Relationship between education and current salary.

```
gvlma(lm(educ ~ salary, data = Employee_data)) #assumptions not acceptable for skewness
```

```
##
## Call:
## lm(formula = educ ~ salary, data = Employee_data)
##
## Coefficients:
## (Intercept)       salary
##   9.6504036    0.0001116
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lm(educ ~ salary, data = Employee_data))
##
##                      Value          p-value                   Decision
## Global Stat        84.1150 0.00000000000000 Assumptions NOT satisfied!
## Skewness           35.0813 0.00000000316219 Assumptions NOT satisfied!
## Kurtosis            0.3246 0.56883418928367    Assumptions acceptable.
## Link Function      47.9918 0.00000000000428 Assumptions NOT satisfied!
## Heteroscedasticity  0.7173 0.39703182424550    Assumptions acceptable.
```

```
qqplot(Employee_data$educ,Employee_data$salary) #also seen in q-q plot
```



The best statistical test is Spearman's Rho - the data is not continuous, nor normally distributed.

```
cor.test(Employee_data$educ,Employee_data$salary, method = "spearman")
```

```
## Warning in cor.test.default(Employee_data$educ, Employee_data$salary,
## method = "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Employee_data$educ and Employee_data$salary
## S = 5529541, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.6884647
```

There was a significant difference between education and salary ($r_s = 0.69$, $N = 474$, $p < .001$).

Dataset 5 - Binary Logistic Regression

These exercises have been prepared for use in conjunction with Chapter 11 of the 5th edition of "SPSS for Psychologists" by Brace, Kemp and Snelgar (2012)

1. A psychologist is interested in how radiographers learn to interpret ambiguous x-ray images. He recruited a number of trainee radiographers. Each was shown an x-ray and asked to determine whether or not it showed a fracture. The psychologist recorded the number of hours of training the radiographer had completed, whether the x-ray showed a fracture or not, and whether the radiographer's decision was correct. The data from this study are coded in the file Ex11Q1.sav.

1.a. Identify the outcome and predictor variables. Which of the predictor variables are categorical?

```
head(logdata)
```

```
## # A tibble: 6 x 4
##   Radiog_ID Training           X_ray   correct
##       <dbl>    <dbl>       <dbl+lbl> <dbl+lbl>
## 1         1        2 1 [Fracture]    0 [Wrong]
## 2         5        2 1 [Fracture]    0 [Wrong]
## 3        24        2 1 [Fracture]    0 [Wrong]
## 4        25        4 0 [no fracture] 0 [Wrong]
## 5        28        4 0 [no fracture] 0 [Wrong]
## 6        33        5 1 [Fracture]    0 [Wrong]
```

```
logdata$Training
```

```
##  [1]  2  2  2  4  4  5  5  6  6  6  6  7  7  8  8  8 10 12 12  1  2  2  3
## [24]  3  3  3  4  4  4  5  5  6  6 12 14
## attr(,"label")
## [1] "N of months in training"
## attr(,"format.spss")
## [1] "F8.2"
```

```
#predictor - cont.
logdata$X_ray
```

```
## <Labelled double>: Does x-ray show fracture
##  [1] 1 1 1 0 0 1 0 1 1 1 0 0 1 0 0 1 1 1 0 0 0 0 1 0 0 0 1 0 1 1 0 0 1 0 1
##
## Labels:
##  value       label
##      0 no fracture
##      1    Fracture
```

```
#predictor - cat.

logdata$correct
```

```
## <Labelled double>: Was decision correct
##  [1] 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1
##
## Labels:
##  value   label
##      0   Wrong
##      1 Correct
```

```
#outcome - cat.
```

1.b. Carry out the appropriate analysis to determine which of the predictor variables significantly predict the radiographer's interpretation of the x-ray.

```
model <- glm(correct ~ Training, data = logdata, family = "binomial")
summary(model) #z value is "Wald"
```

```
##
## Call:
## glm(formula = correct ~ Training, family = "binomial", data = logdata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.57737  -0.19249  -0.01683   0.07825   1.86132
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -13.746      6.102  -2.253   0.0243 *
## Training       2.442      1.073   2.275   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47.804  on 34  degrees of freedom
## Residual deviance: 12.685  on 33  degrees of freedom
## AIC: 16.685
##
## Number of Fisher Scoring iterations: 8
```

```
lrm(model) # "Model Likelihood Ratio" to get X^2 results
```

```
## Logistic Regression Model
##
##  lrm(formula = model)
##
```

```
##                       Model Likelihood     Discrimination    Rank Discrim.
##                          Ratio Test           Indexes          Indexes
##  Obs            35    LR chi2      35.12    R2       0.850    C       0.975
##   0             20    d.f.             1    g        8.798    Dxy     0.950
##   1             15    Pr(> chi2) <0.0001    gr    6620.359    gamma   0.986
##  max |deriv| 9e-05                          gp       0.480    tau-a   0.479
##                                             Brier    0.061
##
##            Coef     S.E.   Wald Z Pr(>|Z|)
##  Intercept -13.7454 6.1018 -2.25  0.0243
##  Training    2.4416 1.0732  2.28  0.0229
##
```

```
confusion_matrix(model) #to produce confusion matrix
```

```
##          Predicted 0 Predicted 1 Total
## Actual 0          18           2    20
## Actual 1           1          14    15
## Total             19          16    35
```

```
18/20 #percentage hit for correct
```

```
## [1] 0.9
```

```
14/15 #percentage hit for incorrect
```

```
## [1] 0.9333333
```

```
(18/20 + 14/15)/2 #0.916 overall accuracy
```

```
## [1] 0.9166667
```

```
confint(model) #to get confidence intervals
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %    97.5 %
## (Intercept) -31.4286241 -5.639071
## Training      0.9946912  5.479247
```

```
visualize_model(model) #plot the model
points(logdata$correct, pch=20) #put in the points
```

1.c. Report the results of your analysis.

Results

A logistic regression was performed with the correct result of x-rays as the dependent variable and training in months as well as x-ray as predictor variables. A total of 35 observations were analyzed and the full model

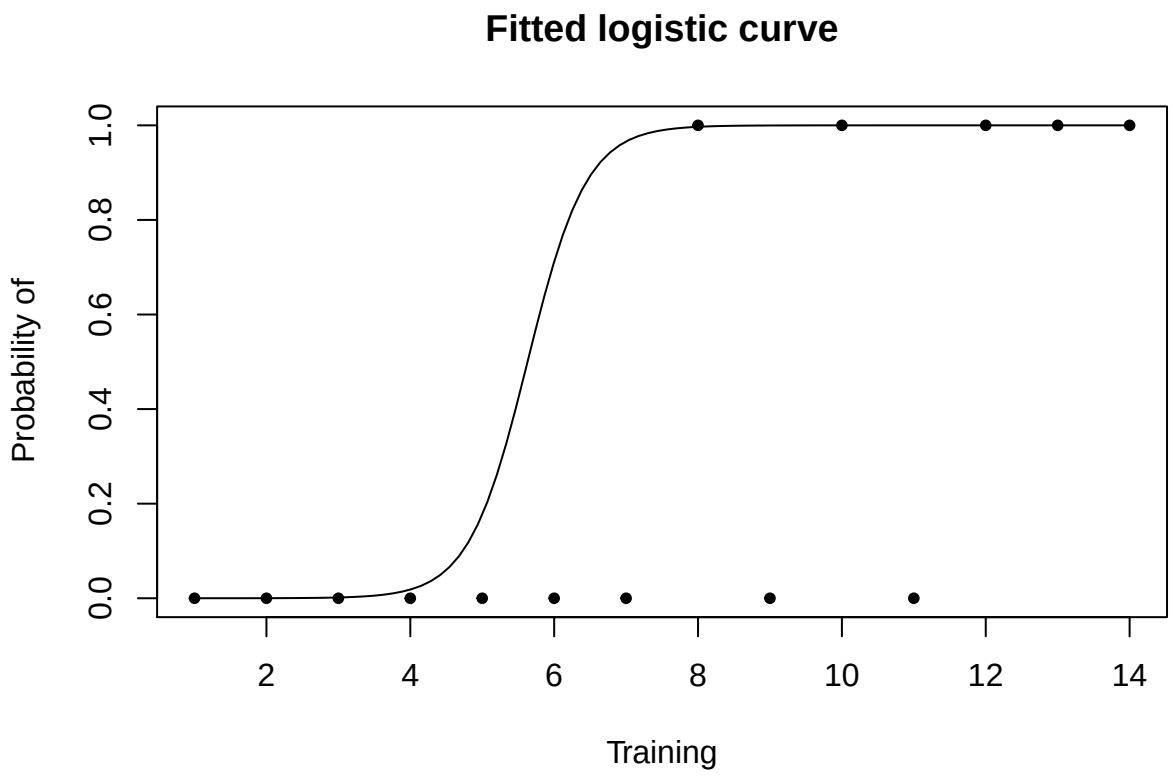


Figure 2: Visualization of Logistic Regression Model

significantly predicted rates of correct outcomes (omnibus $\chi^2 = 35.12$, df $= 1$, $p < .001$). The model accounts for 85% of the variance of correct outcomes, with approximately a 90% accuracy for correct outcomes and 93% accuracy for incorrect outcomes. Overall, the model was 91.6% accurate. The values of the coefficients reveal that the increase of one month of training increases the odds of getting a correct outcome by 2.28 ($CI_{95} = 0.99, 5.48$).

2. Load the data file called EX11Q2.sav. This is a demonstration data file which for many years was supplied with every copy of SPSS (and called '1991 U.S. General Social Survey.sav') but is not with recent versions. Open the data file and familiarise yourself with the variables. This file contains more than 40 variables for each of about 1500 respondents.

2.a We are interested in which factors predict happiness. The fourth variable in the file is called "Happy" and codes the respondents' general level of happiness using a 3 point scale (1=Very Happy, 2= Pretty Happy, 3=Not too Happy), with the values 0, 8 and 9 set as missing values. Recode this variable so that it codes whether or not the respondent is Very Happy (Very Happy= 1, Pretty Happy or Not too Happy = 0). Make sure that that the missing values are still set to 0, 8 and 9.

```
any(logdata2[,"happy"] == 8)|any(logdata2[,"happy"] == 0)|any(logdata2[,"happy"] == 9)
```

```
## [1] NA
```

```r
#confirmed NA's

logdata2$recode <- rep(NA, length(logdata2$happy))
ncol(logdata2)
```

```
## [1] 44
```

```r
logdata2[which(logdata2$happy == 1),44] <- 1
logdata2[which(logdata2$happy == 2|logdata2$happy == 3),44] <- 0

head(as.vector(logdata2$recode),20)
```

```
##  [1]  1  0  1 NA  0  0  1  0  0  0  0  1  0  0  0  1  0  1  0  1
```

2.b. Is the secret to being very happy having a large family, a good education or is happiness something that comes with age? To discover the secret to happiness undertake a Binary Logistic regression using the four variables age, education, number of siblings and number of children as predictor variables, and your recoded happiness variable as the dependent variable.

```r
model2 <- glm(recode ~ childs + age + educ + sibs, data = logdata2, family = "binomial")
summary(model2) #z value is "Wald"
```

```
##
## Call:
## glm(formula = recode ~ childs + age + educ + sibs, family = "binomial",
##     data = logdata2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2761  -0.8808  -0.7849   1.3870   1.9863
##
## Coefficients:
##             Estimate Std. Error z value          Pr(>|z|)
## (Intercept) -2.730493   0.369745  -7.385 0.000000000000153 ***
## childs       0.047252   0.035229   1.341           0.17983
## age          0.009030   0.003456   2.613           0.00899 **
## educ         0.104631   0.020862   5.015 0.000000529474574 ***
## sibs         0.019167   0.019445   0.986           0.32429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1843.8  on 1483  degrees of freedom
## Residual deviance: 1813.2  on 1479  degrees of freedom
##   (33 observations deleted due to missingness)
## AIC: 1823.2
##
## Number of Fisher Scoring iterations: 4
```

```
lrm(model2) # "Model Likelihood Ratio" to get X^2 results
```

```
## Frequencies of Missing Values Due to Each Variable
## recode childs    age   educ   sibs
##     13      8      3      7     12
##
## Logistic Regression Model
##
##  lrm(formula = model2)
##
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
##  Obs          1484    LR chi2      30.61    R2       0.029    C       0.588
##   0           1020    d.f.             4    g        0.350    Dxy     0.177
##   1            464    Pr(> chi2) <0.0001    gr       1.420    gamma   0.177
##  max |deriv| 1e-12                          gp       0.074    tau-a   0.076
##                                             Brier    0.210
##
##            Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept -2.7305 0.3697 -7.38  <0.0001
##  childs     0.0473 0.0352  1.34  0.1798
##  age        0.0090 0.0035  2.61  0.0090
##  educ       0.1046 0.0209  5.02  <0.0001
##  sibs       0.0192 0.0194  0.99  0.3243
##
```

2.c. Which of these factors significantly predict Happiness?

Age and education.

2.d. Does your model really hold the secret of great happiness? Just how good a model is it?

```
confusion_matrix(model2) #to produce confusion matrix
```

```
##          Predicted 0 Predicted 1 Total
## Actual 0        1013           7  1020
## Actual 1         458           6   464
## Total           1471          13  1484
```

```
1013/1020 #percentage hit for unhappy
```

```
## [1] 0.9931373
```

```
6/464 #percentage hit for happy
```

```
## [1] 0.01293103
```

```r
(1013/1020 + 6/464)/2 #0.50 overall accuracy
```

```
## [1] 0.5030341
```

```r
confint(model2) #to get confidence intervals
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) -3.461938383 -2.01167804
## childs      -0.022049324  0.11618262
## age          0.002243123  0.01580178
## educ         0.063990281  0.14582307
## sibs        -0.019282422  0.05707391
```

Unfortunately no. It's about 50% accurate - almost exactly as accurate as a coin toss.

Dataset 6 - Factor Analysis

These exercises have been prepared for use in conjunction with Chapter 12 of the 5th edition of "SPSS for Psychologists" by Brace, Kemp and Snelgar (2012)

1. A psychologist was interested in whether a mindfulness questionnaire measured a single dimension, or whether it had more than one dimension. The data from this study are recorded in the file Ex12.sav. The questionnaire contained 15 items each requiring a response in the range 1 to 6. The responses are coded in the variables s1q1 to s1q15.

1.a. Carry out a principal component analysis with direct oblimin rotation.

```
which(is.na(Ex12)) #found missing data = problem
```

```
## [1]  358 1107
```

```
#rectify
pcdata <- na.omit(Ex12)
head(pcdata)
```

```
## # A tibble: 6 x 15
##    s1q1  s1q2  s1q3  s1q4  s1q5  s1q6  s1q7  s1q8  s1q9 s1q10 s1q11 s1q12
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1     1     1     2     5     5     1     6     4     6     1     1
## 2     3     3     4     5     3     4     4     4     5     4     3     5
## 3     6     1     5     1     5     6     2     6     2     2     6     6
## 4     3     3     3     2     3     4     3     3     4     3     2     5
## 5     3     1     3     2     5     1     5     6     5     6     3     6
## 6     3     4     4     5     3     5     4     5     5     5     5     4
## # ... with 3 more variables: s1q13 <dbl>, s1q14 <dbl>, s1q15 <dbl>
```

```
#done

#fit the model
(pca <- psych::pca(pcdata, rotate = "oblimin"))
```

```
## Principal Components Analysis
## Call: principal(r = r, nfactors = nfactors, residuals = residuals,
##     rotate = rotate, n.obs = n.obs, covar = covar, scores = scores,
##     missing = missing, impute = impute, oblique.scores = oblique.scores,
##     method = method)
## Standardized loadings (pattern matrix) based upon correlation matrix
##        PC1   h2   u2 com
## s1q1  0.44 0.19 0.81   1
## s1q2  0.47 0.22 0.78   1
## s1q3  0.63 0.39 0.61   1
## s1q4  0.45 0.21 0.79   1
## s1q5  0.39 0.15 0.85   1
```

```
## s1q6  0.45 0.20 0.80   1
## s1q7  0.77 0.59 0.41   1
## s1q8  0.75 0.57 0.43   1
## s1q9  0.68 0.46 0.54   1
## s1q10 0.69 0.47 0.53   1
## s1q11 0.53 0.28 0.72   1
## s1q12 0.76 0.57 0.43   1
## s1q13 0.48 0.23 0.77   1
## s1q14 0.79 0.62 0.38   1
## s1q15 0.64 0.40 0.60   1
##
##                 PC1
## SS loadings    5.56
## Proportion Var 0.37
##
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient.
##
## The root mean square of the residuals (RMSR) is  0.09
##  with the empirical chi square  142.03  with prob <  0.00039
##
## Fit based upon off diagonal values = 0.94
```

```
#here I have to specify psych:: package because sjstats 'masks' pca function
summary(pca)
```

```
##
## Factor analysis with Call: principal(r = r, nfactors = nfactors, residuals = residuals,
##     rotate = rotate, n.obs = n.obs, covar = covar, scores = scores,
##     missing = missing, impute = impute, oblique.scores = oblique.scores,
##     method = method)
##
## Test of the hypothesis that 1 factor is sufficient.
## The degrees of freedom for the model is 90  and the objective function was  1.39
## The number of observations was  93  with Chi Square =  119.15  with prob <  0.022
##
## The root mean square of the residuals (RMSA) is  0.09
```

This principal components analysis confirms that 1 factor is sufficient.

1.b. State as many of the indicators of factorability as you can. For each, check and report what they indicate about the factorability of this data set. NB for those without a test of significance, simply give an impression; as in the book, you don't need to give counts.

??? Need more information about the data.

1.c. How many components have eigenvalue greater than one? Write a brief results section, with suitable table, to report which items load on each component.

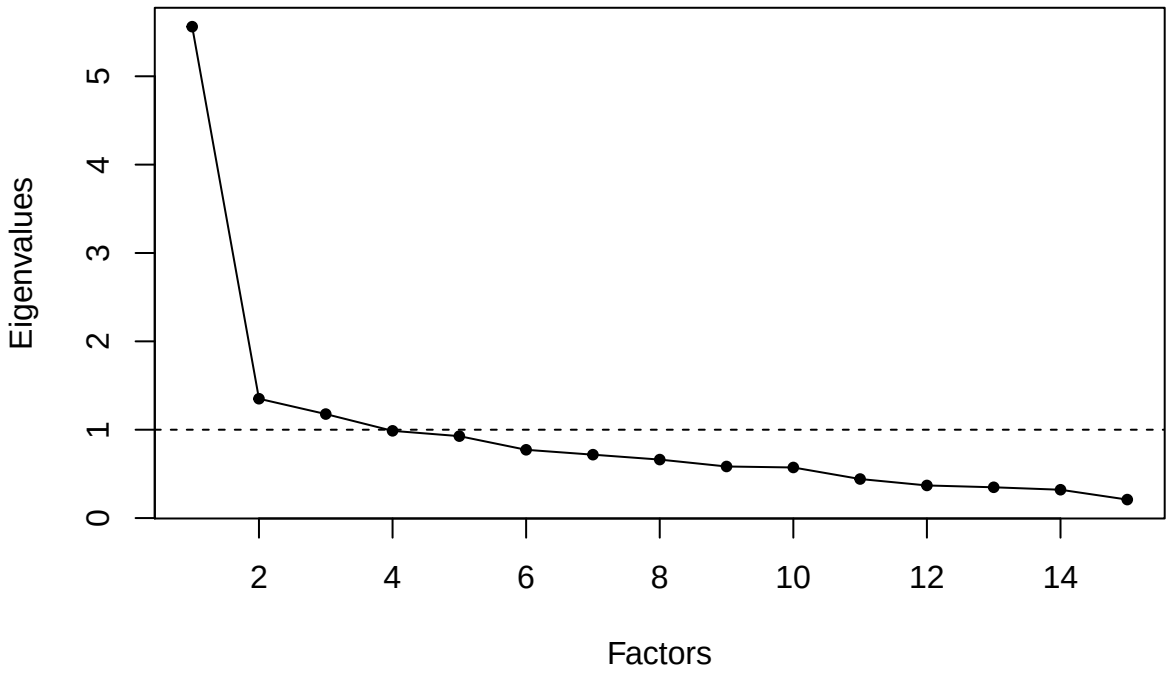


Figure 3: Scree Plot

```
mat <- cor(pcdata)

eigen(mat)$values
```

```
##  [1] 5.5612434 1.3507022 1.1770218 0.9870781 0.9272446 0.7725799 0.7172660
##  [8] 0.6616967 0.5837894 0.5725700 0.4418724 0.3689357 0.3483938 0.3203822
## [15] 0.2092238
```

```
#The first three components have an Eigenvalue greater than 1.
```

I really wouldn't keep any more than one component. Items 14, 7, 12, 8, 10, and 3 load strongest on PC1.

1.d. Consider the scree plot: how many components does that suggest?

Scree Plot

```
plot(eigen(mat)$values, type = "l", ylab = "Eigenvalues",
     xlab = "Factors")
points(eigen(mat)$values, pch = 20)
abline(h = 1, lty = 2)
```

The biggest "drop-off" takes place between factors 1 and 2. I would not retain more than one factor. This is confirmed by `1.f`.

1.e. How would you alter the analysis to assess which items load onto a single component? Do that, and report the results.

??? Alter the analysis? Why?

1.f. What other analysis/es might you conduct when considering the questionnaire?

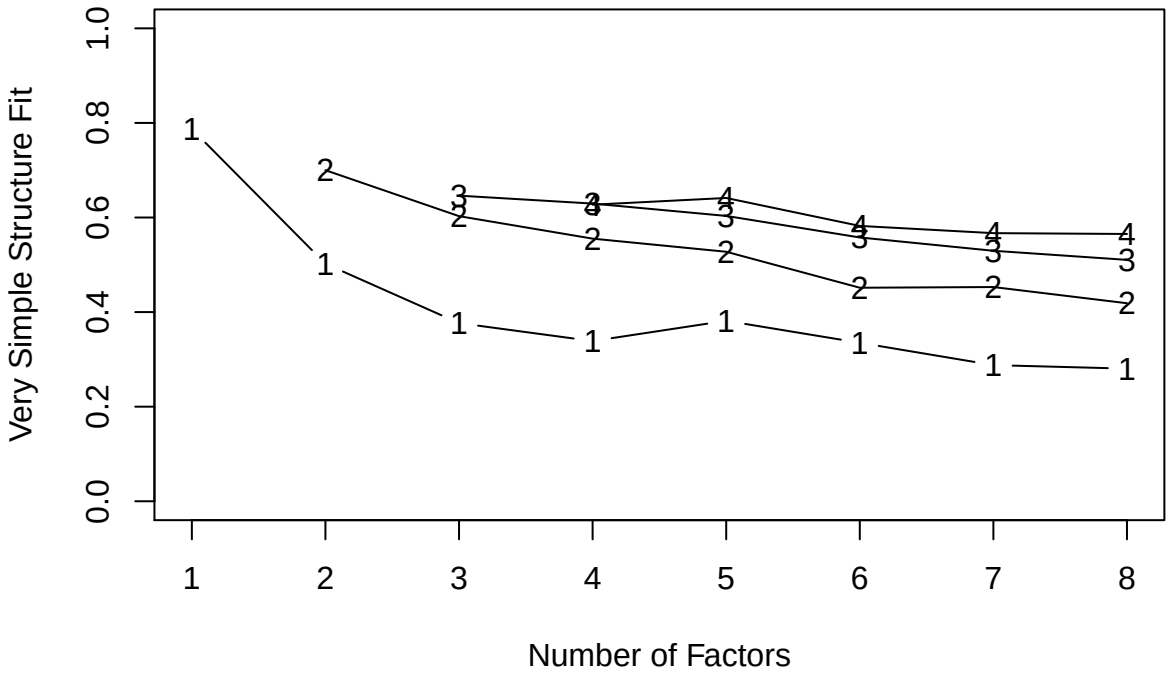
Alternative methods - none in SPSS.

```
factanal(pcdata, factors = 1)
```

```
##
## Call:
## factanal(x = pcdata, factors = 1)
##
## Uniquenesses:
##  s1q1  s1q2  s1q3  s1q4  s1q5  s1q6  s1q7  s1q8  s1q9 s1q10 s1q11 s1q12
## 0.835 0.820 0.655 0.828 0.881 0.843 0.448 0.482 0.563 0.566 0.777 0.464
## s1q13 s1q14 s1q15
## 0.828 0.402 0.639
##
## Loadings:
##       Factor1
## s1q1  0.406
## s1q2  0.424
## s1q3  0.587
## s1q4  0.415
## s1q5  0.345
## s1q6  0.396
## s1q7  0.743
## s1q8  0.720
## s1q9  0.661
## s1q10 0.659
## s1q11 0.472
## s1q12 0.732
## s1q13 0.415
## s1q14 0.773
## s1q15 0.601
##
##                Factor1
## SS loadings      4.970
## Proportion Var   0.331
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 115.49 on 90 degrees of freedom.
## The p-value is 0.0364
```

```
#this is part of the base r stats package and confirms the prior "pca" test

vss(pcdata, rotate = "oblimin")
```




```
##
## Very Simple Structure
## Call: vss(x = pcdata, rotate = "oblimin")
## VSS complexity 1 achieves a maximimum of 0.79  with  1  factors
## VSS complexity 2 achieves a maximimum of 0.7  with  2  factors
##
## The Velicer MAP achieves a minimum of 0.02  with  1  factors
## BIC achieves a minimum of  -292.26  with  1  factors
## Sample Size adjusted BIC achieves a minimum of  -23.48  with  4  factors
##
## Statistics by number of factors
##   vss1 vss2   map dof chisq  prob sqresid  fit  RMSEA  BIC SABIC complex
## 1 0.79 0.00 0.019  90   116 0.035     8.3 0.79 0.0645 -292  -8.2     1.0
## 2 0.50 0.70 0.024  76    87 0.187    11.6 0.70 0.0509 -258 -17.8     1.4
## 3 0.38 0.60 0.029  63    65 0.392    13.7 0.65 0.0384 -220 -21.2     1.5
## 4 0.34 0.56 0.040  51    47 0.645    14.5 0.63 0.0097 -184 -23.5     1.9
## 5 0.38 0.53 0.048  40    32 0.796    13.7 0.65 0.0000 -149 -22.6     1.8
## 6 0.34 0.45 0.064  30    25 0.738    14.7 0.62 0.0000 -111 -16.5     1.9
## 7 0.29 0.45 0.086  21    18 0.635    14.7 0.62 0.0000  -77 -10.7     2.2
## 8 0.28 0.42 0.116  13    12 0.568    16.7 0.57 0.0092  -47  -6.4     2.0
##   eChisq  SRMR eCRMS eBIC
## 1  106.6 0.074 0.080 -301
## 2   66.0 0.058 0.068 -279
## 3   40.7 0.046 0.059 -245
## 4   27.4 0.037 0.054 -204
## 5   15.0 0.028 0.045 -166
## 6   10.3 0.023 0.043 -126
## 7    6.5 0.018 0.041  -89
## 8    4.2 0.015 0.042  -55
```

```r
#VSS stands for Very Simple Structure, which is a method included in the `psych` package.
#It provides a plot depicting the "factors" and explicitly states what complexity is achieved
#with what number of factors.

paran(pcdata, iterations = 5000)
```

```
##
## Using eigendecomposition of correlation matrix.
## Computing: 10%  20%  30%  40%  50%  60%  70%  80%  90%  100%
##
##
## Results of Horn's Parallel Analysis for component retention
## 5000 iterations, using the mean estimate
##
## ----------------------------------------------------
## Component   Adjusted    Unadjusted    Estimated
##             Eigenvalue  Eigenvalue    Bias
## ----------------------------------------------------
## 1           4.806514    5.561243      0.754728
## ----------------------------------------------------
##
## Adjusted eigenvalues > 1 indicate dimensions to retain.
## (1 components retained)
```

```r
#This method is by far the most robust - it follows Horn's Parallel Analysis and employs
#bootstrapping (shown here with 5000 iterations).
```