Unit 5 – Regression and Correlation
Practice Problems (2 of 3)
Solutions

Before you begin. Download from the course website
hersdata_small.xlsx

Description of Dataset

Source
Hulley et al (1998) Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605-613

The Heart and Estrogen/progestin Replacement Study (HERS) was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus.

The dataset for this homework (**hersdata_small.xlsx**) is a simple random sample of n=1000. A subset of the variables are considered:

Data dictionary/Codebook *(Partial)*

Variable	Label	Type	Codings
age	Age, years	numeric	Continuous, range, [45:79]
BMI	Body Mass index (kg/m²)	numeric	Continuous, range, [15.21:54.13]
glucose	Fasting glucose (mg/dL)	numeric	Continuous, range, [29:298]
LDL	LDL cholesterol (mg/dL)	numeric	Continuous, range, [44.4:393.4]
drinkany	Any current alcohol use	numeric	1 = yes 0 = no
exercise	Exercise at least 3x/week	numeric	1 = yes 0 = no
HT	Randomization	numeric	1 = hormone therapy 0 = placebo
physact	Comparative ("compared to other women your age") physical activity	Numeric	1 = much less active 2 = somewhat less active 3 = about as active 4 = somewhat more active 5 = much more active
statins	Statin use	Numeric	1 = yes 0 = no
diabetes	Diabetes	Numeric	1 = yes 0 = no

initialize session

```
setwd("/cloud/project")                                  # Set working directory
getwd()                                                  # Check working directory

## [1] "/cloud/project"

options(scipen=999)                                      # Turn off scientific notation
rm(list = ls())                                          # Clear the workspace environment
```

import excel data

```
library(readxl)                                          # read_excel() in package {readxl}
source <- read_excel("hersdata_small.xlsx")
str(source)                                              # str() to check structure of data

## tibble [1,000 × 38] (S3: tbl_df/tbl/data.frame)
##  $ id       : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ BMI      : num [1:1000] 21.7 29.1 31.9 26.9 38.1 ...
##  $ BMI1     : num [1:1000] NA 29.9 31.9 25.1 36.9 ...
##  $ DBP      : num [1:1000] 67 91 80 61 89 70 72 61 83 71 ...
##  $ HDL      : num [1:1000] 68 59 44 52 36 42 49 46 59 48 ...
##  $ HDL1     : num [1:1000] NA 75 35 92 34 55 55 52 63 45 ...
##  $ HT       : num [1:1000] 1 1 0 0 0 1 1 0 1 1 ...
##  $ LDL      : num [1:1000] 111 148.8 84.8 183.2 75 ...
##  $ LDL1     : num [1:1000] NA 98.2 59 88 78.2 ...
##  $ SBP      : num [1:1000] 114 140 153 141 141 122 180 132 129 131 ...
##  $ TG       : num [1:1000] 165 171 146 189 295 145 126 267 113 176 ...
##  $ TG1      : num [1:1000] NA 179 110 55 244 108 139 171 105 226 ...
##  $ WHR      : num [1:1000] 0.786 0.992 0.927 0.836 0.861 ...
##  $ WHR1     : num [1:1000] NA 0.967 0.972 0.796 0.896 ...
##  $ age      : num [1:1000] 76 63 72 62 54 58 69 70 63 61 ...
##  $ age10    : num [1:1000] 7.6 6.3 7.2 6.2 5.4 ...
##  $ diabetes: num [1:1000] 0 1 1 0 0 0 0 0 0 1 ...
##  $ dmpills : num [1:1000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ drinkany: num [1:1000] 1 1 0 0 1 0 0 0 1 0 ...
##  $ exercise: num [1:1000] 0 1 0 0 0 1 1 0 1 0 ...
##  $ globrat : num [1:1000] 3 4 2 3 2 5 4 4 2 3 ...
##  $ glucose : num [1:1000] 115 185 67 96 109 108 111 90 90 132 ...
##  $ glucose1: num [1:1000] NA 101 154 103 98 93 91 97 83 121 ...
##  $ htnmeds : num [1:1000] 1 0 1 1 1 1 1 1 1 ...
##  $ insulin : num [1:1000] 0 0 1 0 0 0 0 0 0 0 ...
##  $ medcond : num [1:1000] 0 0 1 1 0 1 1 0 0 1 ...
##  $ nonwhite: num [1:1000] 0 0 0 0 0 1 0 0 1 0 ...
##  $ physact : num [1:1000] 3 2 1 4 2 3 4 4 4 3 ...
##  $ poorfair: num [1:1000] 0 0 1 0 1 0 0 0 1 0 ...
##  $ raceth  : num [1:1000] 1 1 1 1 1 2 1 1 2 1 ...
##  $ smoking : num [1:1000] 0 0 0 0 0 1 0 0 1 0 ...
##  $ statins : num [1:1000] 0 1 1 0 1 0 1 0 0 1 ...
##  $ tchol   : num [1:1000] 212 242 158 273 170 240 197 259 279 226 ...
##  $ tchol1  : num [1:1000] NA 209 116 191 161 220 215 269 257 228 ...
##  $ waist   : num [1:1000] 77 105.5 101 86.5 105 ...
##  $ waist1  : num [1:1000] NA 104 106 86 105 ...
##  $ weight  : num [1:1000] 60.4 78.7 88 65.2 94 ...
##  $ weight1 : num [1:1000] NA 80.8 88 60.9 91 ...
```

retain vars of interest

```
library(tidyverse)                                       # pipe operator %>% and select() in package {tidyverse}

hersdata <-  source %>%                                  # here we created a smaller dataframe w vars of interest
select(age,BMI,glucose,LDL,drinkany,exercise,HT,physact,statins,diabetes)

hersdata <- data.frame(hersdata)
```

```
glimpse(hersdata)                                  # using glimpse( ) in package {tidyverse}.   You could also do str( )

## Rows: 1,000
## Columns: 10
## $ age      <dbl> 76, 63, 72, 62, 54, 58, 69, 70, 63, 61, 64, 66, 65, 65, 71, 7…
## $ BMI      <dbl> 21.68, 29.12, 31.93, 26.93, 38.14, 33.70, 26.20, 26.84, 33.31…
## $ glucose  <dbl> 115, 185, 67, 96, 109, 108, 111, 90, 90, 132, 108, 107, 93, 8…
## $ LDL      <dbl> 111.0, 148.8, 84.8, 183.2, 75.0, 169.0, 122.8, 159.6, 197.4, …
## $ drinkany <dbl> 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0…
## $ exercise <dbl> 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0…
## $ HT       <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1…
## $ physact  <dbl> 3, 2, 1, 4, 2, 3, 4, 4, 4, 3, 4, 3, 2, 4, 5, 3, 2, 3, 4, 4, 2…
## $ statins  <dbl> 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0…
## $ diabetes <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1…   # Note that all vars are numeric
```

```
create factor vars as needed
library(tidyverse)

hersdata <- hersdata %>%
            mutate(drinkanyf = factor(drinkany,
                                      levels=c(0,1),
                                      labels=c("0 = no", "1 = yes")),
                                      na.rm=TRUE) %>%
            mutate(exercisef = factor(exercise,
                                      levels=c(0,1),
                                      labels=c("0 = no", "1 = yes")),
                                      na.rm=TRUE) %>%
            mutate(HTf = factor(HT,
                                      levels=c(0,1),
                                      labels=c("0 = placebo", "1 = hormone therapy")),
                                      na.rm=TRUE) %>%
            mutate(physactf = factor(physact,
                                      levels=c(1,2,3,4,5),
                                      labels = c("1 = much less active",
                                                 "2 = somewhat less active",
                                                 "3 = about as active",
                                                 "4 = somewhat more active",
                                                 "5 = much more active")),
                                       na.rm=TRUE) %>%
            mutate(statinsf = factor(statins,
                                      levels=c(0,1),
                                      labels=c("0 = no", "1 = yes")),
                                      na.rm=TRUE) %>%
            mutate(diabetesf = factor(diabetes,
                                      levels=c(0,1),
                                      labels=c("0 = no", "1 = yes")),
                                      na.rm=TRUE)


keepvars <- c("drinkanyf", "exercisef", "HTf", "physactf", "statinsf", "diabetesf")


glimpse(hersdata[keepvars])

## Rows: 1,000
## Columns: 6
## $ drinkanyf <fct> 1 = yes, 1 = yes, 0 = no, 0 = no, 1 = yes, 0 = no, 0 = no, 0…
## $ exercisef <fct> 0 = no, 1 = yes, 0 = no, 0 = no, 0 = no, 1 = yes, 1 = yes, 0…
## $ HTf       <fct> 1 = hormone therapy, 1 = hormone therapy, 0 = placebo, 0 = p…
## $ physactf  <fct> 3 = about as active, 2 = somewhat less active, 1 = much less…
## $ statinsf  <fct> 0 = no, 1 = yes, 1 = yes, 0 = no, 1 = yes, 0 = no, 1 = yes, …
## $ diabetesf <fct> 0 = no, 1 = yes, 1 = yes, 0 = no, 0 = no, 0 = no, 0 = no, 0 …
```

1.

By any means you like, obtain numerical summaries of the four continuous variables: **age, BMI, glucose,** and **LDL.**

```
Q1: Descriptives on continuous vars
library(stargazer)                                                    # stargazer( ) in package {stargazer}
library(tidyverse)

keepvars <- c("glucose", "age", "BMI", "LDL")                         # convenient for stargazer() that follows

stargazer(hersdata[keepvars],                                         # note - square brackets to index columns/vars
          type="text",
          summary.stat=c("n","mean", "sd", "min", "p25", "median", "p75","max"),
          title="Q1: Descriptives on Continuous Variables")

##
## Q1: Descriptives on Continuous Variables
## ========================================================================
## Statistic   N    Mean   St. Dev.  Min   Pctl(25) Median  Pctl(75)   Max
## ------------------------------------------------------------------------
## glucose   1,000 111.214  35.461    29      92      99      113      298
## age       1,000 66.701   6.575     45      62      67       71      79
## BMI        999  28.315   5.453   15.210  24.535  27.590   31.330  54.130
## LDL        997  144.656  36.843  44.400 121.000  141.000 165.800  393.400
## ------------------------------------------------------------------------
```

2.

By any means you like, obtain numerical summaries of the six discrete variables: **drinkany, exercise, HT, physact, statins, diabetes.**

```
Q2: Descriptives on discrete vars
library(summarytools)

# Hack!!! By creating object keepvars, I only need to issue the command freq( ) one time to obtain all the descriptives

keepvars <- c("drinkanyf", "exercisef", "HTf", "physactf", "statinsf", "diabetesf")


# Note use of square brackets when specifying which columns/variables to use
freq(hersdata[keepvars])

## Frequencies
## hersdata$drinkanyf
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------- ------ --------- -------------- --------- --------------
##       0 = no    586     58.66          58.66     58.60          58.60
##      1 = yes    413     41.34         100.00     41.30          99.90
##         <NA>      1                               0.10         100.00
##        Total   1000    100.00         100.00    100.00         100.00
##
## hersdata$exercisef
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------- ------ --------- -------------- --------- --------------
##       0 = no    607     60.70          60.70     60.70          60.70
##      1 = yes    393     39.30         100.00     39.30         100.00
##         <NA>      0                               0.00         100.00
##        Total   1000    100.00         100.00    100.00         100.00
##
##
```

```
hersdata$HTf
## Type: Factor
##
##                            Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------------------- ------ --------- -------------- --------- --------------
##              0 = placebo    492     49.20          49.20     49.20          49.20
##      1 = hormone therapy    508     50.80         100.00     50.80         100.00
##                     <NA>      0                                0.00         100.00
##                    Total   1000    100.00         100.00    100.00         100.00
##
## hersdata$physactf
## Type: Factor
##
##                                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------------------------- ------ --------- -------------- --------- --------------
##          1 = much less active     72      7.20           7.20      7.20           7.20
##      2 = somewhat less active    179     17.90          25.10     17.90          25.10
##           3 = about as active    322     32.20          57.30     32.20          57.30
##      4 = somewhat more active    312     31.20          88.50     31.20          88.50
##          5 = much more active    115     11.50         100.00     11.50         100.00
##                          <NA>      0                                0.00         100.00
##                         Total   1000    100.00         100.00    100.00         100.00
##
## hersdata$statinsf
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------- ------ --------- -------------- --------- --------------
##       0 = no    635     63.50          63.50     63.50          63.50
##      1 = yes    365     36.50         100.00     36.50         100.00
##         <NA>      0                                0.00         100.00
##        Total   1000    100.00         100.00    100.00         100.00
##
## hersdata$diabetesf
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------- ------ --------- -------------- --------- --------------
##       0 = no    749     74.90          74.90     74.90          74.90
##      1 = yes    251     25.10         100.00     25.10         100.00
##         <NA>      0                                0.00         100.00
##        Total   1000    100.00         100.00    100.00         100.00
```

Exercises #3 - #7 consider non-diabetics only (diabetes==0)

```
create dataset of NON-diabetics only
library(tidyverse)                                                    # filter() and select() in package {tidyverse}

nondiabetics <- hersdata %>%
                filter(diabetes==0) %>%                                 # note double equal sign in filter()
                select(glucose,exercisef,age,drinkanyf,BMI,physactf) %>%    # select( ) to choose vars of interest
                na.omit()                                               # complete data only


glimpse(nondiabetics)
## Rows: 748
## Columns: 6
## $ glucose   <dbl> 115, 96, 109, 108, 111, 90, 90, 108, 107, 80, 90, 92, 94, 10…
## $ exercisef <fct> 0 = no, 0 = no, 0 = no, 1 = yes, 1 = yes, 0 = no, 1 = yes, 1…
## $ age       <dbl> 76, 62, 54, 58, 69, 70, 63, 64, 66, 65, 71, 72, 76, 73, 65, …
## $ drinkanyf <fct> 1 = yes, 0 = no, 1 = yes, 0 = no, 0 = no, 0 = no, 1 = yes, 1…
## $ BMI       <dbl> 21.68, 26.93, 38.14, 33.70, 26.20, 26.84, 33.31, 22.42, 27.2…
## $ physactf  <fct> 3 = about as active, 4 = somewhat more active, 2 = somewhat …
```

#3

Fit a single predictor model of Y=**glucose** to X= **exercise** among non-diabetics ONLY.
In 1-2 sentences, report and interpret the output.

```
Q3: Fit Y=glucose to X=exercise among non-diabetics only
library(jtools)                                                    # summ( ) in {jtools}
library(broom)                                                     # tidy( ) in {broom}
library(stargazer)                                                 # stargazer( ) in {stargazer}

fitq3 <- lm(glucose ~ factor(exercisef), data=nondiabetics)         # obtain fit. save as fitq3

# Basic displays of fit (no packages required)
summary(fitq3)                                                     # summary( ) in {basic}. no library() needed

##
## Call:
## lm(formula = glucose ~ factor(exercisef), data = nondiabetics)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.7569  -6.7569  -0.7569   5.2431  29.1538
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)               97.7569     0.4520  216.28 < 0.0000000000000002 ***
## factor(exercisef)1 = yes  -1.9107     0.6999   -2.73             0.00648 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.438 on 746 degrees of freedom
## Multiple R-squared:  0.009893,   Adjusted R-squared:  0.008566
## F-statistic: 7.454 on 1 and 746 DF,  p-value: 0.00648

coef(fitq3)                                                        # coef( ) to show betas

##              (Intercept) factor(exercisef)1 = yes
##                97.756881                -1.910727

confint(fitq3)                                                     # confint( ) to show CI for betas

##                              2.5 %     97.5 %
## (Intercept)              96.869539 98.6442224
## factor(exercisef)1 = yes -3.284656 -0.5367976

cbind(coef(fitq3),confint(fitq3))                                  # cbind( ) to combine betas and CI

##                                       2.5 %     97.5 %
## (Intercept)              97.756881 96.869539 98.6442224
## factor(exercisef)1 = yes -1.910727 -3.284656 -0.5367976

anova(fitq3)                                                       # anova( ) to show anova table

## Analysis of Variance Table
##
## Response: glucose
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## factor(exercisef)   1    664  663.95  7.4538 0.00648 **
## Residuals         746  66451   89.08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Pretty displays of fit (packages required)
tidy(fitq3)                                                    # tidy( ) in package {broom} for pretty

## # A tibble: 2 × 5
##   term                    estimate std.error statistic p.value
## 1 (Intercept)                97.8      0.452    216.   0
## 2 factor(exercisef)1 = yes   -1.91     0.700     -2.73 0.00648

tidy(anova(fitq3))                                            # analysis of variance (prettier if html)

## # A tibble: 2 × 6
##   term                 df  sumsq meansq statistic  p.value
## 1 factor(exercisef)     1   664.  664.       7.45  0.00648
## 2 Residuals           746 66451.   89.1     NA    NA

summ(fitq3)                                                    # summ( ) in package {jtools}
                                                               # Disadvantage:  no anova table
## MODEL INFO:
## Observations: 748
## Dependent Variable: glucose
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,746) = 7.45, p = 0.01
## R² = 0.01
## Adj. R² = 0.01
##
## Standard errors: OLS
## ------------------------------------------------------------
##                                 Est.   S.E.   t val.      p
## ------------------------------ ------- ------ -------- ------
## (Intercept)                   97.76   0.45   216.28   0.00
## factor(exercisef)1 = yes      -1.91   0.70    -2.73   0.01
## ------------------------------------------------------------

stargazer(fitq3,type="text")                                   # stargazer( ) in package {stargazer}

##
## ===================================================
##                             Dependent variable:
##                         ---------------------------
##                                    glucose
## ---------------------------------------------------
## factor(exercisef)1 = yes          -1.911***
##                                    (0.700)
##
## Constant                          97.757***
##                                    (0.452)
##
## ---------------------------------------------------
## Observations                        748
## R2                                 0.010
## Adjusted R2                        0.009
## Residual Std. Error           9.438 (df = 746)
## F Statistic                 7.454*** (df = 1; 746)
## ===================================================
## Note:                   *p<0.1; **p<0.05; ***p<0.01
```

Putting it all together:

(1) The fitted line is Y = 97.76 – (1.91)*exercise for Y=glucose

(2) R-squared = .01 says that only 1% of the variation in glucose is explained by this model.

(3) The p-value of .006 for the overall F test (value= 7.45) tells us that this fitted model explains statistically significantly more of the variability in the outcome than is explained by the null model of using the average Y only.

#4

Next, fit a multiple predictor model of Y= **glucose** among non-diabetics ONLY.. Fit the following predictors: **exercise, age, drinkany,** and **BMI**. In 1-2 sentences, interpret the output.

```
Q4: Fit Y=glucose to X's = exercise, age, drinkany, and BMI among non-diabetics only
library(jtools)                                                    # summ( ) in {jtools}
library(broom)                                                     # tidy( ) in {broom}
library(stargazer)                                                 # stargazer( ) in {stargazer}

fitq4 <- lm(glucose ~ age + BMI + factor(exercisef) + factor(drinkanyf), data=nondiabetics)  # fit

summ(fitq4)                                                        # I rather like this display!!

## MODEL INFO:
## Observations: 748
## Dependent Variable: glucose
## Type: OLS linear regression
##
## MODEL FIT:
## F(4,743) = 14.79, p = 0.00
## R² = 0.07
## Adj. R² = 0.07
##
## Standard errors: OLS
## ------------------------------------------------------------
##                                 Est.   S.E.   t val.      p
## ------------------------------ ------- ------ -------- ------
## (Intercept)                    81.21   4.12    19.73   0.00
## age                             0.05   0.05     1.05   0.30
## BMI                             0.47   0.07     7.05   0.00
## factor(exercisef)1 = yes       -1.09   0.69    -1.59   0.11
## factor(drinkanyf)1 = yes       -0.40   0.68    -0.60   0.55
## ------------------------------------------------------------

tidy(fitq4, conf.int=TRUE)                                            # Nice - betas with CI

## # A tibble: 5 × 7
##   term                  estimate std.error statistic  p.value conf.low conf.high
## 1 (Intercept)            81.2       4.12      19.7   5.45e-70  73.1       89.3
## 2 age                     0.0533    0.0509     1.05  2.95e- 1  -0.0466     0.153
## 3 BMI                     0.466     0.0661     7.05  4.17e-12   0.336      0.596
## 4 factor(exercisef)1 =…  -1.09      0.688     -1.59  1.13e- 1  -2.44       0.260
## 5 factor(drinkanyf)1 =…  -0.402     0.676     -0.595 5.52e- 1  -1.73       0.924

tidy(anova(fitq4))                                                    # Pretty analysis of variance

## # A tibble: 5 × 6
##   term                df   sumsq meansq statistic   p.value
## 1 age                  1    22.7   22.7     0.271  6.03e- 1
## 2 BMI                  1  4690.  4690.     56.1    2.00e-13
## 3 factor(exercisef)    1   208.   208.      2.49   1.15e- 1
## 4 factor(drinkanyf)    1    29.6   29.6     0.354  5.52e- 1
## 5 Residuals          743 62165.    83.7    NA     NA
```

```
stargazer(fitq4,type="text")                                    # default display of betas and SE(beta)s
```

```
##
## ================================================
##                            Dependent variable:
##                        ---------------------------
##                                    glucose
## ------------------------------------------------
## age                                 0.053                    # beta (age) = 0.053
##                                    (0.051)                   # SE ( beta(age) ) = 0.051
##
## BMI                               0.466***
##                                    (0.066)
##
## factor(exercisef)1 = yes           -1.091
##                                    (0.688)
##
## factor(drinkanyf)1 = yes           -0.402
##                                    (0.676)
##
## Constant                          81.206***
##                                    (4.116)
##
## ------------------------------------------------
## Observations                         748
## R2                                  0.074
## Adjusted R2                         0.069
## Residual Std. Error           9.147 (df = 743)
## F Statistic                 14.792*** (df = 4; 743)
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

(1) The fitted line is now:
Y = 81.21 + (0.05)*age + (0.47)*BMI – (1.09)*exercise – (0.40)*drinkany

(2) R-squared = .074 says that 7.4% of the variation in glucose is explained by this model.

(3) The p-value of the overall F-test (value=14.79) is <<< .00001.
So again, the fitted model performs better than the null model of using the average Y only.

#5

Perform a partial F-test for the significance of **exercise** controlling for **age,drinkany,** and **BMI** among non-diabetics ONLY. Interpret.

```
Q5: Partial F-Test
library(stargazer)

fitq5_reduced <- lm(glucose ~ age + BMI + factor(drinkanyf), data=nondiabetics)
fitq5_full <- lm(glucose ~ factor(exercisef) + age + BMI + factor(drinkanyf), data=nondiabetics)

temp <- anova(fitq5_reduced,fitq5_full)
print(temp,signif.stars=FALSE)                    # option signif.stars=FALSE to to suppress printing stars

## Analysis of Variance Table
##
## Model 1: glucose ~ age + BMI + factor(drinkanyf)
## Model 2: glucose ~ factor(exercisef) + age + BMI + factor(drinkanyf)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    744 62375
## 2    743 62165  1    210.24 2.5128 0.1134
```

The partial F-test (null hypothesis: exercise is not significant in the adjusted model) has p-value = .11. Do NOT reject. Conclude that, controlling for age, drink and BMI, exercise is not statistically significant for the prediction of glucose.

```
fitq5_exercise <- lm(glucose ~ factor(exercisef), data=nondiabetics)
stargazer(fitq5_exercise,fitq5_reduced,fitq5_full,
          type="text")
```

```
##
## ================================================================================================
##                                                Dependent variable:
##                            ------------------------------------------------------------------
##                                                       glucose
##                                  (1)                   (2)                    (3)
## ------------------------------------------------------------------------------------------------
## factor(exercisef)1 = yes       -1.911***                                    -1.091
##                                 (0.700)                                     (0.688)
##
## age                                                    0.056                  0.053
##                                                       (0.051)                (0.051)
##
## BMI                                                   0.484***               0.466***
##                                                       (0.065)                (0.066)
##
## factor(drinkanyf)1 = yes                              -0.388                 -0.402
##                                                       (0.676)                (0.676)
##
## Constant                      97.757***               80.073***              81.206***
##                                (0.452)                (4.058)                (4.116)
##
## ------------------------------------------------------------------------------------------------
## Observations                     748                    748                    748
## R2                              0.010                  0.071                  0.074
## Adjusted R2                     0.009                  0.067                  0.069
## Residual Std. Error         9.438 (df = 746)       9.156 (df = 744)       9.147 (df = 743)
## F Statistic              7.454*** (df = 1; 746) 18.846*** (df = 3; 744) 14.792*** (df = 4; 743)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

#6

Create four 0/1 design variables to represent the 5 possible outcomes of **physact** among non-diabetics ONLY. By any means you like, produce a check on the creation of your design variables.

```
Q6: Create 4 design variables to represent the 5 levels of physact
library(tidyverse)                                      # mutate() to create a new variable in package {tidyverse}
library(summarytools)                                   # freq() in package {summarytools}

dataq6 <- nondiabetics %>%
          mutate(physact2 = ifelse(physactf=="2 = somewhat less active",1,0), na.rm=T,      # note double equal signs
                 physact3 = ifelse(physactf=="3 = about as active",1,0), na.rm=T,
                 physact4 = ifelse(physactf=="4 = somewhat more active",1,0), na.rm=T,
                 physact5 = ifelse(physactf=="5 = much more active",1,0), na.rm=T)

keepvars <- c("physactf", "physact2", "physact3", "physact4", "physact5")
freq(dataq6[keepvars])

## Frequencies
## dataq6$physactf
## Type: Factor
##
##                               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------------- ------ --------- -------------- --------- --------------
##          1 = much less active     44      5.88           5.88      5.88           5.88
##       2 = somewhat less active    105     14.04          19.92     14.04          19.92
##            3 = about as active    242     32.35          52.27     32.35          52.27
##       4 = somewhat more active    259     34.63          86.90     34.63          86.90
##           5 = much more active     98     13.10         100.00     13.10         100.00
##                          <NA>      0                               0.00         100.00
##                          Total    748    100.00         100.00    100.00         100.00
##
```

```
## dataq6$physact2
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0    643     85.96          85.96     85.96          85.96
##           1    105     14.04         100.00     14.04         100.00
##        <NA>      0                               0.00         100.00
##       Total    748    100.00         100.00    100.00         100.00
##
## dataq6$physact3
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0    506     67.65          67.65     67.65          67.65
##           1    242     32.35         100.00     32.35         100.00
##        <NA>      0                               0.00         100.00
##       Total    748    100.00         100.00    100.00         100.00
##
## dataq6$physact4
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0    489     65.37          65.37     65.37          65.37
##           1    259     34.63         100.00     34.63         100.00
##        <NA>      0                               0.00         100.00
##       Total    748    100.00         100.00    100.00         100.00
##
## dataq6$physact5
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0    650     86.90          86.90     86.90          86.90
##           1     98     13.10         100.00     13.10         100.00
##        <NA>      0                               0.00         100.00
##       Total    748    100.00         100.00    100.00         100.00
```

#7

Fit a multiple predictor model of Y= **glucose** among non-diabetics ONLY. Consider as the predictor ONLY the design variables for **physact**. In 1-2 sentences, interpret the output.

```
Q7: Fit Y = glucose to design variables for physact (2 ways)
library(stargazer)

# METHOD 1:  Letting R create the indicator variables "under the hood"
fitq7_R <- lm(glucose ~ factor(physactf), data=dataq6)

# METHOD 2: Modeling indicators explicity with user-created indicators You choose method that you prefer!
fitq7_explicit <- lm(glucose ~ physact2 + physact3 + physact4 + physact5, data=dataq6)
```

```
# Compare to see which type of display you prefer!
stargazer(fitq7_R, fitq7_explicit,
          type="text")
```

```
##
## ===============================================================
##                                         Dependent variable:
##                                     ---------------------------
##                                               glucose
##                                        (1)            (2)
## ---------------------------------------------------------------
## factor(physactf)2 = somewhat less active     -2.334                          # These first 4 rows are METHOD 1 model
##                                             (1.688)
##
## factor(physactf)3 = about as active          -2.983*
##                                             (1.540)
##
## factor(physactf)4 = somewhat more active    -5.178***
##                                             (1.532)
##
## factor(physactf)5 = much more active         -4.162**
##                                             (1.705)
##
## physact2                                                  -2.334             # Next 4 rows are METHOD 2 model
##                                                          (1.688)
##
## physact3                                                 -2.983*
##                                                          (1.540)
##
## physact4                                                -5.178***
##                                                          (1.532)
##
## physact5                                                 -4.162**
##                                                          (1.705)
##
## Constant                                   100.591***    100.591***
##                                             (1.417)       (1.417)
##
## ---------------------------------------------------------------
## Observations                                  748           748
## R2                                           0.022         0.022
## Adjusted R2                                  0.017         0.017
## Residual Std. Error (df = 743)               9.397         9.397
## F Statistic (df = 4; 743)                   4.263***      4.263***
## ===============================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

The multiple predictor model of glucose in relationship to physical activity using 4 design variables, and therefore no assumption of a specific relationship, was statistically significant (Overall F-Statistic = 4.26, p-value = .002). Post-hoc examination of the effects of increasing physical activity on glucose suggests a negative association and, in particular, that increasing physical activity is associated with a downward trend in glucose. However, this output does not include a formal test of trend.