Week4-Exercises-Solutions

Exercise solutions

Week 4

Continuing on with the hers_subset example. Write down the regression equation for a regression with an outcome of body mass index, and with age and physical activity (physact) as covariates. Interpret each parameter in this equation.

The regression equation is

$$BMI = \beta_0 + \beta_1 age + \beta_2 MLA + \beta_3 MMA + \beta_4 SLA + \beta_5 SMA$$

where

- β_0 is the mean BMI for someone "about as active" and aged 0
- β_1 is the amount by which the mean BMI increases for every one year increase in age for individuals with the same level of exercise.
- β_2 is the mean difference in BMI between the "much less active group" and the "about as active" group, for individuals of the same age.
- β_3 is the mean difference in BMI between the "much more active group" and the "about as active" group, for individuals of the same age.
- β_4 is the mean difference in BMI between the "somewhat less active" group and the "about as active" group, for individuals of the same age.
- β_5 is the mean difference in BMI between the "somehwat more active" group and the "about as active" group, for individuals of the same age.

To work out the linear combination for these difference, I find it helpful to use the following notation. Let μ_{group} represent the mean BMI for physical activity group *group* for those aged zero. Then

- $\mu_{MLA} = \beta_2 - \beta_0$
- $\mu_{MMA} = \beta_3 - \beta_0$
- $\mu_{SLA} = \beta_4 - \beta_0$

- $\mu_{SMA} = \beta_5 - \beta_0$

We can therefore represent the three comparisons in terms of μ, and then substitute in the betas.

Comparison 1: Mean difference $= \mu_{MMA} - \mu_{MLA} = (\beta_3 - \beta_0) - (\beta_2 - \beta_0) = \beta_3 - \beta_2$

Comparison 2: Mean difference $= \mu_{MMA} - \mu_{SMA} = (\beta_3 - \beta_0) - (\beta_5 - \beta_0) = \beta_3 - \beta_5$

Comparison 3:

$$\text{Mean difference} = \frac{\mu_{MMA} + \mu_{SMA}}{2} - \frac{\mu_{MLA} + \mu_{SLA}}{2} = \frac{\beta_3 - \beta_0 + \beta_5 - \beta_0}{2} - \frac{\beta_2 - \beta_0 + \beta_4 - \beta_0}{2} = \frac{\beta_3 + \beta_5 - \beta_2 -}{2}$$

This last comparison could potentially be more complex than this as well, depending on what we mean by combining the two groups. E.g. if by "somewhat and much more active combined" we mean the mean of a sample of equal numbers of somewhat and much more active participants combined, then the expression above would be suitable. However if we mean the mean of a sample of unequal numbers of the two groups (perhaps representative numbers from our sample), then the expression above would be incorrect. Can you think of how to adapt this so that we have a weighted mean?

Carry out this regression and report on the key findings.

Finally, express the following comparisons in terms of the regression coefficients of your equation above, and calculate these using Stata or R

The mean difference between much more active, and much less active, for individuals of the same age.

The mean difference between much more active, and somewhat more active, for individuals of the same age.

[Challenge question] The mean difference between the more active groups (somewhat and much more active combined), and the less active groups (somewhat and less active combined), for individuals of the same age.

Stata code and output

Note that some of the output below has unfortunately been truncated (but is not truncated when carried out in Stata). The reference category chosen by Stata here is "much less active", and the order of the coefficients below is: "somewhat less active", "about as active", "somewhat more active" and "much more active". So this is different to the regression equation reported above, and you may need to change these linear combinations based on the change in reference category.

Another way to work this out is to use the "label" command in Stata. In this instance `label list physact` which will show the labeling

```
1 much less active

2 somewhat less active

3 about as active

4 somewhat more active

5 much more active
```

We then use this numeric coding in our lincom statement as follows

```
use "https://www.dropbox.com/scl/fi/onx8zrpw9qoaaf3tw8hkb/hers_subset.dta?rlkey=mprmw7n6u1n8
label list physact /* Display the label encoding for physact */
reg BMI age i.physact
lincom 5.physact - 1.physact
lincom 5.physact - 4.physact
lincom (5.physact + 4.physact - 2.physact - 1.physact)/2
## physact:
##            1 much less active
##            2 somewhat less active
##            3 about as active
##            4 somewhat more active
##            5 much more active
##
##
##       Source |       SS           df       MS      Number of obs   =       276
## -------------+----------------------------------   F(5, 270)       =      5.37
##        Model |  777.159381         5  155.431876   Prob > F        =    0.0001
##     Residual |  7809.54367       270  28.9242358   R-squared       =    0.0905
## -------------+----------------------------------   Adj R-squared   =    0.0737
##        Total |  8586.70305       275  31.2243747   Root MSE        =    5.3781
##
## ------------------------------------------------------------------------------------
##                  BMI | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## ---------------------+--------------------------------------------------------------
##                  age |  -.0729604   .0515818    -1.41   0.158    -.1745141    .0285932
##                      |
##              physact |
```

```
## somewhat less active  |  -.5625767   1.437869    -0.39   0.696    -3.393437    2.268283
##      about as active  |  -2.453703   1.308869    -1.87   0.062    -5.030591    .1231843
## somewhat more active  |  -4.392431    1.31679    -3.34   0.001    -6.984914   -1.799949
##     much more active  |  -3.995889   1.495118    -2.67   0.008    -6.939459   -1.052318
##                       |
##                 _cons |   36.01086   3.487445    10.33   0.000     29.14482     42.8769
## ------------------------------------------------------------------------------------
##
##
##  ( 1)  - 1b.physact + 5.physact = 0
##
## ------------------------------------------------------------------------------
##          BMI | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          (1) |  -3.995889   1.495118    -2.67   0.008    -6.939459   -1.052318
## ------------------------------------------------------------------------------
##
##
##  ( 1)  - 4.physact + 5.physact = 0
##
## ------------------------------------------------------------------------------
##          BMI | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          (1) |   .3965426   1.076743     0.37   0.713    -1.723337    2.516422
## ------------------------------------------------------------------------------
##
##
##  ( 1)  - .5*1b.physact - .5*2.physact + .5*4.physact + .5*5.physact = 0
##
## ------------------------------------------------------------------------------
##          BMI | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          (1) |  -3.912871    .909941    -4.30   0.000    -5.704353    -2.12139
## ------------------------------------------------------------------------------
```

R code and output

```
library(multcomp, quietly=TRUE)
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
```

```r
##     geyser
hers_subset <- read.csv("https://www.dropbox.com/scl/fi/ywlbb7duvez2nyk66ojp1/hersdata.csv?rl

lm.exercise <- lm(BMI ~ age + physact, data = hers_subset)
summary(lm.exercise)
##
## Call:
## lm(formula = BMI ~ age + physact, data = hers_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.009  -3.551  -0.675   3.084  24.952
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  35.64649    1.03752  34.357  < 2e-16 ***
## age                          -0.09952    0.01539  -6.468 1.17e-10 ***
## physactmuch less active       1.67453    0.41750   4.011 6.21e-05 ***
## physactmuch more active      -2.35754    0.35068  -6.723 2.16e-11 ***
## physactsomewhat less active  1.17651    0.29478   3.991 6.75e-05 ***
## physactsomewhat more active -1.66678    0.25379  -6.567 6.10e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.296 on 2752 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.08046,    Adjusted R-squared:  0.07878
## F-statistic: 48.16 on 5 and 2752 DF,  p-value: < 2.2e-16
comparison1 <- matrix(c(0,0,-1,1,0,0), nrow=1)
comparison2 <- matrix(c(0,0,0,1,0,-1), nrow=1)
comparison3 <- matrix(c(0,0,-1,1,-1,1)/2, nrow=1)
lincom1 <- glht(lm.exercise, comparison1)
lincom2 <- glht(lm.exercise, comparison2)
lincom3 <- glht(lm.exercise, comparison3)
summary(lincom1)
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = BMI ~ age + physact, data = hers_subset)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
```

```
## 1 == 0  -4.0321     0.4873  -8.275 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
confint(lincom1)
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = BMI ~ age + physact, data = hers_subset)
##
## Quantile = 1.9608
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate lwr     upr
## 1 == 0 -4.0321  -4.9876 -3.0766
summary(lincom2)
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = BMI ~ age + physact, data = hers_subset)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0  -0.6908     0.3539  -1.952   0.0511 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
confint(lincom2)
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = BMI ~ age + physact, data = hers_subset)
##
## Quantile = 1.9608
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate  lwr       upr
## 1 == 0 -0.690755 -1.384710  0.003201
```

```r
summary(lincom3)
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = BMI ~ age + physact, data = hers_subset)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0  -3.4377     0.2883  -11.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
confint(lincom3)
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = BMI ~ age + physact, data = hers_subset)
##
## Quantile = 1.9608
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate lwr     upr
## 1 == 0 -3.4377  -4.0029 -2.8725
```