Week1-Exercises-Solutions

Exercise solutions

Week 1

The following exercise will allow you to test yourself against what you have learned so far. The solutions will be released at the end of the week.

Using the dataset hers_subset.csv dataset, use simple linear regression in R or Stata to measure the association between diastolic blood pressure (DBP - the outcome) and body mass index (BMI - the exposure).

a) Summarise the important findings by interpreting the relevant parameter values, associated P-values and confidence intervals, and R^2 value. Three to four sentences is usually enough here. & b) From your regression output, calculate by how much the mean DBP changes for a 5kgm^{-2} increase in BMI? Can you verify this by modifying your data and re-running your regression?

Stata code and output

```
/* Part a */
import delimited "https://www.dropbox.com/s/t0ml83xesaaazd0/hers_subset.csv?dl=1"
reg dbp bmi
/* Part b*/
gen bmi5 = bmi / 5
reg dbp bmi5
## . /* Part a. import delimited "https://www.dropbox.com/s/t0ml83xesaaazd0/hers_subset.cs
## (encoding automatically selected: ISO-8859-1)
## (38 vars, 276 obs)
##
## . reg dbp bmi
##
##       Source |       SS           df       MS      Number of obs   =       276
## -------------+----------------------------------   F(1, 274)       =      4.84
##        Model |  423.883938         1  423.883938   Prob > F        =    0.0286
##     Residual |  23988.8842       274  87.5506722   R-squared       =    0.0174
## -------------+----------------------------------   Adj R-squared   =    0.0138
```

```
##        Total |  24412.7681       275  88.7737022   Root MSE        =    9.3569
##
## ------------------------------------------------------------------------------
##          dbp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          bmi |   .2221827   .1009756     2.20   0.029     .0233961    .4209693
##        _cons |   67.82592   2.923282    23.20   0.000     62.07097    73.58087
## ------------------------------------------------------------------------------
##
## . /* Part b*/
## . gen bmi5 = bmi / 5
##
## . reg dbp bmi5
##
##       Source |       SS           df       MS      Number of obs   =       276
## -------------+----------------------------------   F(1, 274)       =      4.84
##        Model |  423.883889         1  423.883889   Prob > F        =    0.0286
##     Residual |  23988.8842       274  87.5506724   R-squared       =    0.0174
## -------------+----------------------------------   Adj R-squared   =    0.0138
##        Total |  24412.7681       275  88.7737022   Root MSE        =    9.3569
##
## ------------------------------------------------------------------------------
##          dbp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##         bmi5 |   1.110913   .5048781     2.20   0.029     .1169804    2.104847
##        _cons |   67.82592   2.923282    23.20   0.000     62.07097    73.58087
## ------------------------------------------------------------------------------
```

R code and output

```
# Part a
hers_subset <- read.csv("https://www.dropbox.com/s/t0ml83xesaaazd0/hers_subset.csv?dl=1")
lm.hers <- lm(DBP ~ BMI, data = hers_subset)
summary(lm.hers)
##
## Call:
## lm(formula = DBP ~ BMI, data = hers_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.6420  -6.4584  -0.7538   5.8199  27.0639
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.8259     2.9233    23.2   <2e-16 ***
## BMI           0.2222     0.1010     2.2   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.357 on 274 degrees of freedom
## Multiple R-squared:  0.01736,    Adjusted R-squared:  0.01378
## F-statistic: 4.842 on 1 and 274 DF,  p-value: 0.02862
confint(lm.hers)
##                   2.5 %     97.5 %
## (Intercept) 62.07097446 73.5808680
## BMI          0.02339609  0.4209693

# Part b
hers_subset$BMI5 <- hers_subset$BMI / 5
lm.hers <- lm(DBP ~ BMI5, data = hers_subset)
summary(lm.hers)
##
## Call:
## lm(formula = DBP ~ BMI5, data = hers_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.6420  -6.4584  -0.7538   5.8199  27.0639
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.8259     2.9233    23.2   <2e-16 ***
## BMI5          1.1109     0.5049     2.2   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.357 on 274 degrees of freedom
## Multiple R-squared:  0.01736,    Adjusted R-squared:  0.01378
## F-statistic: 4.842 on 1 and 274 DF,  p-value: 0.02862
```

We find evidence that diastolic blood pressure increases as body mass index increases (P = 0.029). For every one kg/m^{-2} increase in BMI, the mean diastolic blood pressure increases by 0.22mmHg, and we are 95% confident the true increase lies between 0.023 and 0.42mmHg. BMI accounts for 1.7% of the overall variability in diastolic blood pressure.

If a one kg/m^{-2} increase in BMI accounts for a 0.22mmHg increase in DBP, then a 5kg/m^{-2} increase in BMI accounts for a `5x0.22 = 1.1`mmHg increase in DBP. We can confirm this in Stata or R by creating a new covariate `BMI5` which is BMI scaled by a factor of 1/5 (so that a 1 increase in BMI5 corresponds to a 5 increase in BMI).

c) Manually calculate the β_1 standard error, the t-value, p-value and R^2

From 3.3.7 of the textbook, the standard error of the regression coefficient is as follows:

$$\hat{\mathrm{se}}(\hat{\beta}_1) = \frac{\text{Root mean squared error}}{\hat{\sigma}_x\sqrt{(n-1)}}$$

We can use R or Stata to calculate $\hat{\sigma}_x = 5.5879$. Substituting this in we obtain $\hat{\mathrm{se}}(\hat{\beta}_1) = 9.357/(5.5879\sqrt{275}) = 0.101$ in agreement with the Stata and R output.

The t-value is the regression coefficient divided by it's standard error $t = 0.222/0.101 = 2.2$, and the P-value can be calculated by looking up a corresponding t-table with $n-2 = 276-2 = 274$ degrees of freedom:

Stata code and output for the p-value

```
/*Compute*/
  disp tprob(274,2.2)
## . /*Comput.   disp tprob(274,2.2)
## .0286418
```

R code and output

```
(1-pt(2.2,274))*2
## [1] 0.0286418
```

R^2 is the fraction of the total variance explained by the model so is equal to $R^2 = 423.88/24412.77 = 0.017$. These two variances are default output in Stata. In R the model sum of squares and residual sum of squares can be obtained using `anova(lm.hers)`, after which the R^{2} can be calculated.

d) Based on your regression, make a prediction for the mean diastolic blood pressure of people with a BMI of 28kgm^{-2}.& e) Calculate and interpret a confidence interval for this prediction. & f) Calculate and interpret a prediction interval for this prediction.

Stata code and output

```
import delimited "https://www.dropbox.com/s/t0ml83xesaaazd0/hers_subset.csv?dl=1"
reg dbp bmi

lincom _cons + 28*bmi

set obs 277
replace bmi = 28 in 277
predict fitDBP
predict seprDBP, stdf
gen upper = fitDBP + 1.96*seprDBP in 277
gen lower = fitDBP -1.96*seprDBP in 277

list bmi fitDBP lower upper in 277
```

```
## . import delimited "https://www.dropbox.com/s/t0ml83xesaaazd0/hers_subset.csv?dl(encodi
## (38 vars, 276 obs)
##
## . reg dbp bmi
##
##       Source |       SS           df       MS      Number of obs   =       276
## -------------+----------------------------------   F(1, 274)       =      4.84
##        Model |  423.883938         1  423.883938   Prob > F        =    0.0286
##     Residual |  23988.8842       274  87.5506722   R-squared       =    0.0174
## -------------+----------------------------------   Adj R-squared   =    0.0138
##        Total |  24412.7681       275  88.7737022   Root MSE        =    9.3569
##
## ------------------------------------------------------------------------------
##          dbp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          bmi |   .2221827   .1009756     2.20   0.029     .0233961    .4209693
##        _cons |   67.82592   2.923282    23.20   0.000     62.07097    73.58087
## ------------------------------------------------------------------------------
##
## .
## . lincom _cons + 28*bmi
##
##  ( 1)  28*bmi + _cons = 0
##
## ------------------------------------------------------------------------------
##          dbp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
```

```
##           (1) |   74.04704   .5647208   131.12   0.000     72.93529    75.15878
## ------------------------------------------------------------------------------
##
## .
## . set obs 277
## Number of observations (_N) was 276, now 277.
##
## . replace bmi = 28 in 277
## (1 real change made)
##
## . predict fitDBP
## (option xb assumed; fitted values)
##
## . predict seprDBP, stdf
##
## . gen upper = fitDBP + 1.96*seprDBP in 277
## (276 missing values generated)
##
## . gen lower = fitDBP -1.96*seprDBP in 277
## (276 missing values generated)
##
## .
## . list bmi fitDBP lower upper in 277
##
##      +--------------------------------------+
##      | bmi     fitDBP      lower      upper |
##      |--------------------------------------|
## 277. |  28   74.04704   55.67424   92.41984 |
##      +--------------------------------------+
##
## .
```

R code and output

```r
hers_subset <- read.csv("https://www.dropbox.com/s/t0ml83xesaaazd0/hers_subset.csv?dl=1")
lm.hers <- lm(DBP ~ BMI, data = hers_subset)

new_observation <- data.frame(BMI = 28)
predict(lm.hers, newdata = new_observation, interval="confidence")
##        fit      lwr      upr
## 1 74.04704 72.93529 75.15878
predict(lm.hers, newdata = new_observation, interval="prediction")
```

```
##          fit      lwr      upr
## 1 74.04704 55.59306 92.50101
```

We predict that the mean diastolic blood pressure for those with a BMI of 28kgm^{-2} to be 74mmHg. We are 95% confident the true mean lies between 72.9mmHg and 75.2mmHg. We expect that 95% of women with that BMI will have a diastolic blood pressure between 55.6mmHg and 92.5mmHg.