PSYCGR01: Statistics

Regression worksheet

Maarten Speekenbrink *UCL Experimental Psychology*
October, 2016

1 General Exercises

1. As part of an experiment investigating rats' speed of learning, Bond (1979) placed rats in a "shuttlebox" (a box divided into two compartments by a wall with a small door). On each trial, a tone was played, and 5 seconds later an electrical shock was delivered to the floor of the compartment in which the rat was standing. Rats could avoid the shocks by going through the door to the other compartment within 5 seconds after the tone was played. Bond recorded the average time (Y) the rats took to reach the other compartment, and the number of previous trials (X) in which they received a shock. X ranged from 0 to 15, in increments of 1. There were a total of $n = 16$ observations of X and Y.

   The mean of X was $\overline{X} = 7.5$ and the variance $S^2_X = 22.67$. The mean of Y was $\overline{Y} = 5.89$ and the variance $S^2_Y = 13.27$. The covariance between X and Y was $S_{XY} = -13.89$.

   (a) The researcher wants to know whether rats learn to avoid the shocks. What do you expect of the relation between X and Y if the rats did learn to avoid the shocks?

   Answer: I would expect the times to decrease with the number of shocks previously received.

   (b) Estimate the parameters of MODEL A:

   $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

   Answer:

   $$b_1 = \frac{S_{XY}}{S^2_X} = \frac{-13.89}{22.67} = -0.61$$

   $$b_0 = \overline{Y} - b_1 \overline{X} = 5.89 - (-0.61) \times 7.5 = 10.47$$

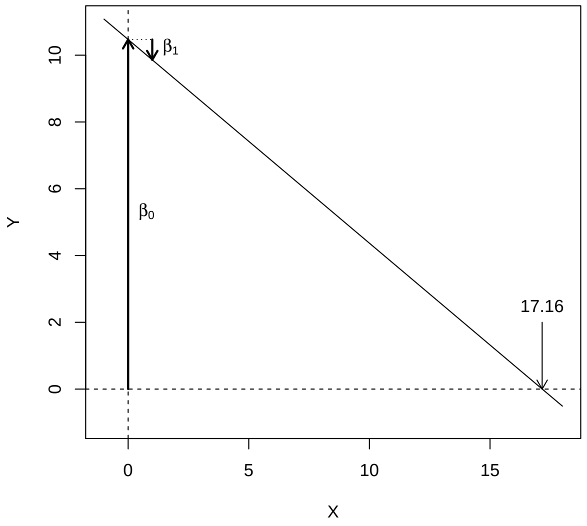
   The estimate b_1 confirms my (and your?) expectation.

   (c) Sketch a graph of the relation between Y and X according to model A (i.e., a graph of the predicted escape time as a function of the number of previous shocks).

   Answer: You need to sketch a graph with a straight line which crosses the y-axis 10.47 (the intercept) and goes down by 0.61 for each 1-unit increase in X. The line will cross the x-axis at

$$\begin{aligned} 0 &= 10.47 - .61 \times X \\ 0.61 \times X &= 10.47 \\ X &= \frac{10.47}{.61} \\ X &= 17.16 \end{aligned}$$

The graph could look as follows:



(d) The Sum of Squared Error (SSE) of model A is SSE(A) = 71.32. Compute the Proportional Reduction in Error (PRE)[1] of MODEL A over MODEL C:

$$Y_i = \beta_0 + \epsilon_i$$

and assess whether it is significant.

Answer: We're missing the value of SSE(C). You should remember that, for this simple MODEL C (with only an intercept), it is related to the (sample) variance

[1] Remember, the PRE is computed as

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}}$$

$S_Y^2 = \frac{\sum_{i=1}(Y_i - \overline{Y})^2}{n-1} = \frac{\text{SSE(C)}}{n-1}$. In more detail:

$$\begin{aligned}\text{SSE(C)} &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \sum_i (Y_i - \overline{Y})^2 \\ &= (n-1) \times S_Y^2 \\ &= (16-1) \times 13.27 \\ &= 199.05\end{aligned}$$

So the PRE is $\frac{199.05-71.32}{199.05} = 0.64$. To test for significance, we compute an F value as $\frac{(199.05-71.32)/(2-1)}{71.32/(16-2)} = 25.07$. You can also compute it from the PRE, as $F = \frac{0.64/(2-1)}{(1-0.64)/(16-2)} = 24.89$, but note that this results in a different value, due to additional rounding error. The critical value of the F is $F_{1,14;.05} = 4.60$, so clearly, the test is significant. In fact, $P(F_{1,14} \geq 25.07) < .001$. Thus, mean escape time appears to be reliably related to the number of previously received shocks.

(e) In a simple (bivariate) regression, you can compute the confidence interval for the slope as

$$b_1 \pm \sqrt{\frac{F_{1,n-2;\alpha}\text{MSE}}{(n-1)S_X^2}}.$$

Compute the 95% confidence interval for the slope in model MODEL A. How do the results compare to the analysis above?

Answer: The confidence interval is

$$b_1 \pm \sqrt{\frac{F_{1,n-2;\alpha}\text{MSE}}{(n-1)S_X^2}} = -0.61 \pm \sqrt{\frac{4.60 \times (71.32/(16-2))}{(16-1) \times 22.67}} = -0.61 \pm 0.26$$

This interval does not include the value 0, so the effect is significant, as also shown in the analysis above.

(f) The correlation between X and Y, denoted as r_{XY}, is related to the PRE. You can compute this correlation as

$$r_{XY} = \frac{S_{XY}}{S_X \times S_Y}$$

where S_X is the standard deviation of X (and similarly for S_Y) and S_{XY} the covariance between X and Y. Compute this correlation. Now take the square of the correlation, r_{XY}^2. How does this compare to the value of the PRE?

Answer: The correlation is

$$r_{XY} = \frac{-13.89}{\sqrt{22.67}\sqrt{13.27}} = -0.80.$$

The square is $r_{XY}^2 = (-.80)^2 = 0.64$. This is equal to the PRE. Indeed, for this simple (bivariate) regression model, the PRE is always equal to the squared correlation. To compute the correlation from the PRE, you can take its square root, but you will also have to add the sign of the slope b_1. In this case, the slope is negative, so the correlation is $r_{XY} = -\sqrt{\text{PRE}} = -\sqrt{0.64} = -.80$.

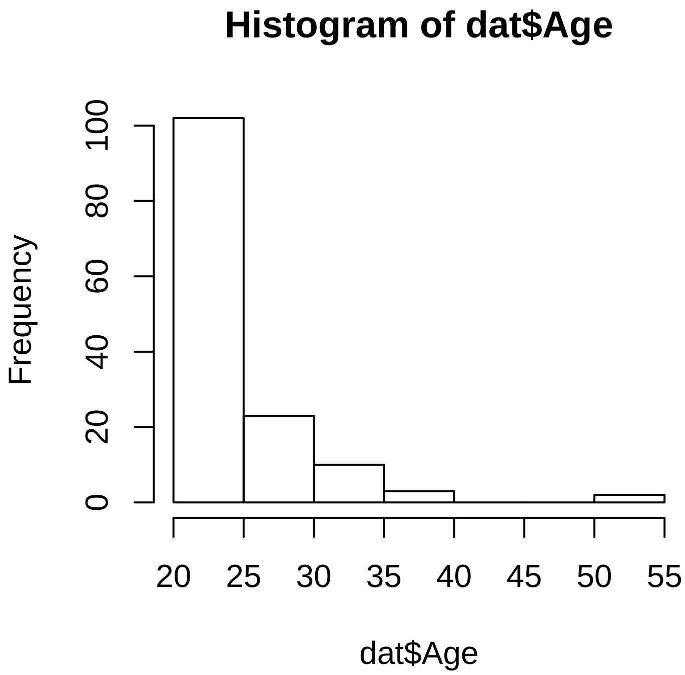
2 Computer Exercises

1. Open the datafile `2016_Questionnaire_1_cleaned.sav`. This dataset contains the `Age` (in years), `Height` (in centimetres), and `Weight` (in kilograms) of those who filled in the Moodle questionnaire for week 1 and didn't make any obvious mistakes.

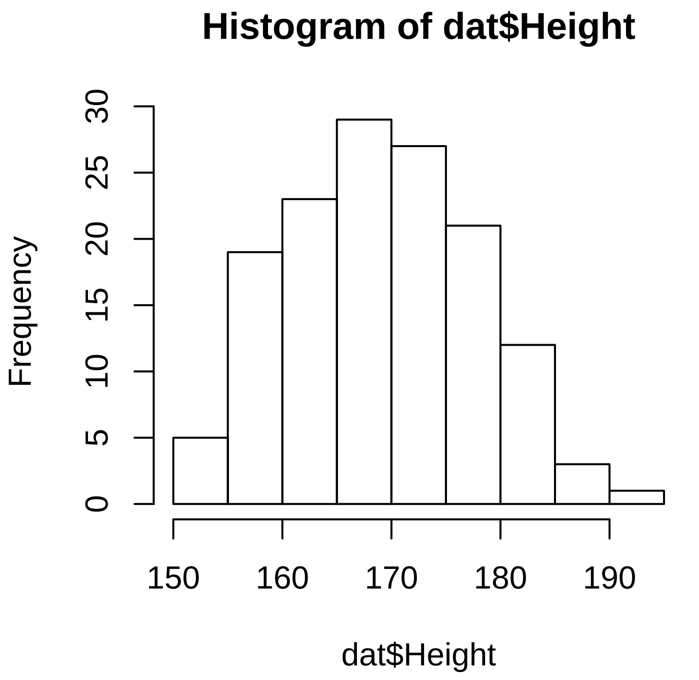
   (a) Before analysing the data, it is good practice to get a "feel" for (explore) the data. As a minimum, obtain descriptives (mean, standard deviation, minimum and maximum), as well as histograms for `Age`, `Height`, `Weight`. Do these variables look like what you would expect?

```
> # read the data
> library(foreign)
> dat <- as.data.frame(read.spss("2016_Questionnaire_1_cleaned.sav"))

Warning in read.spss("2016_Questionnaire_1_cleaned.sav"):  2016_Questionnaire
File-indicated value is different from internal value for at least one
of the three system values.  SYSMIS: indicated -1.79769e+308, expected
-1.79769e+308; HIGHEST: 1.79769e+308, 1.79769e+308; LOWEST: -1.79769e+308,
-1.79769e+308
Warning in read.spss("2016_Questionnaire_1_cleaned.sav"):  2016_Questionnaire
Unrecognized record type 7, subtype 18 encountered in system file

> # get means and ranges
> summary(dat[,c("Age","Height","Weight")])

      Age            Height          Weight      
 Min.   :20.00   Min.   :151.0   Min.   : 43.00  
 1st Qu.:22.00   1st Qu.:163.0   1st Qu.: 54.00  
 Median :23.00   Median :170.0   Median : 60.00  
 Mean   :24.91   Mean   :170.1   Mean   : 64.54  
 3rd Qu.:26.00   3rd Qu.:176.0   3rd Qu.: 70.00  
 Max.   :52.00   Max.   :192.0   Max.   :181.00  

> # standard deviations
> sd(dat$Age)

[1] 4.847155

> sd(dat$Height)
```

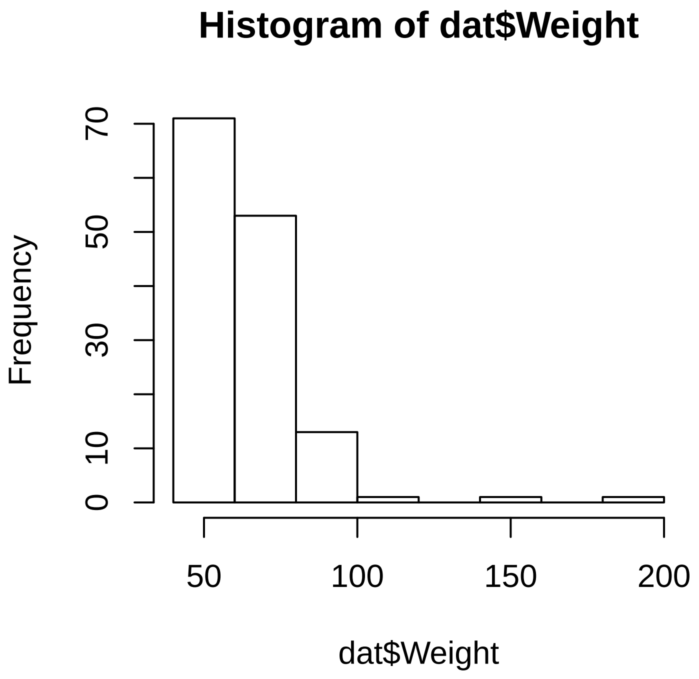
```
[1] 8.626949

> sd(dat$Weight)

[1] 17.3585

> # you can do this with one command using apply
> apply(dat[,c("Age","Height","Weight")],2,sd)

      Age    Height    Weight
 4.847155  8.626949 17.358501

> # histograms
> hist(dat$Age)
```



```
> hist(dat$Height)
```



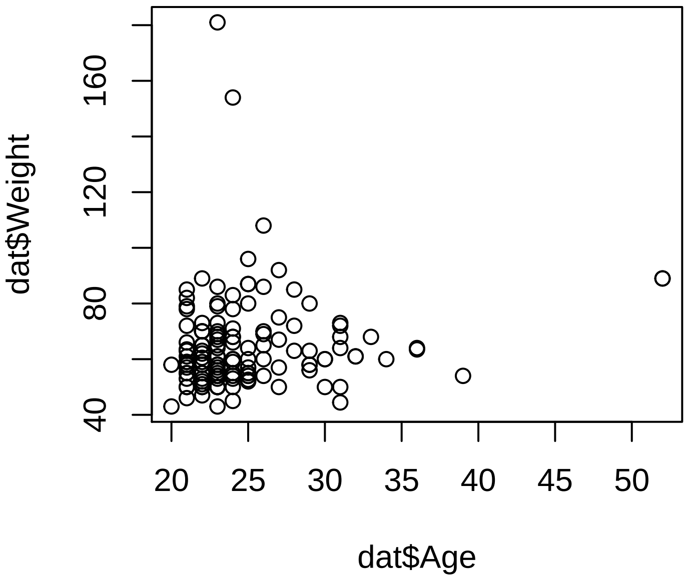
```
> hist(dat$Weight)
```



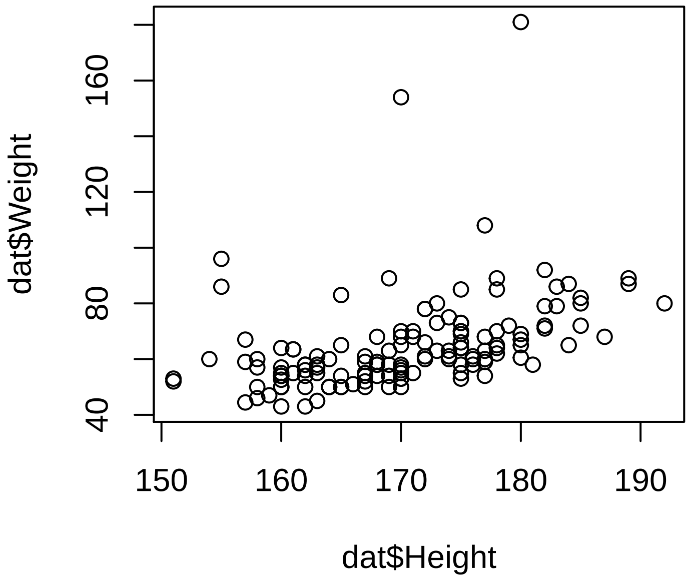
Answer: Height and Age seem fine here in terms of ranges and means. The distribution of Height seems roughly Normal, while you can see that the distribution of age is clearly skewed, but this is not necessarily problematic (and what I would expect from this population of MSc students). The distribution of Weight is also clearly skewed, with some relatively quite high weights. As we will discuss in next lecture, these look like outliers which can have a substantial influence of the model estimates and tests. For now, let's keep these cases in.

(b) We will now attempt to predict students' weight from their height and age. First, create scatterplots to investigate the relations between these variables.

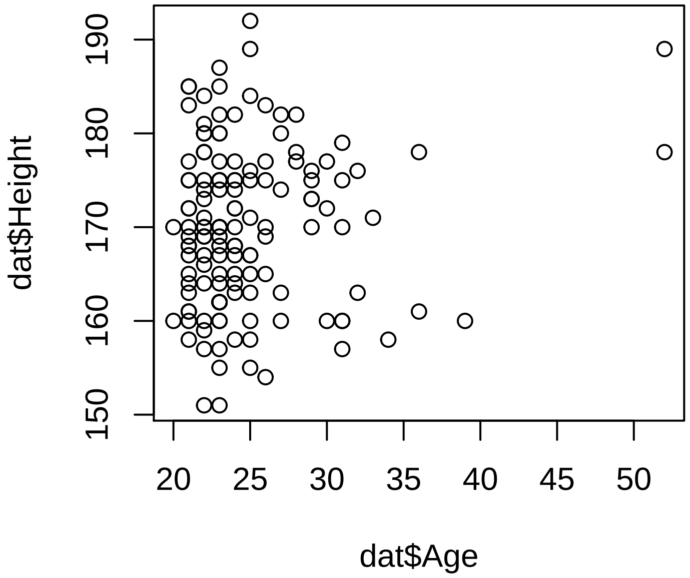
```
> # histograms
> plot(dat$Age,dat$Weight)
```



```
> plot(dat$Height,dat$Weight)
```



```
> plot(dat$Age, dat$Height)
```



(c) Estimate the parameters of the regression model

$$\texttt{Weight}_i = \beta_0 + \beta_1 \texttt{Height}_i + \epsilon_i$$

and interpret the results. Is height a good predictor of weight?

```
> # estimate the regression model
> mod <- lm(Weight~Height,data=dat)
> summary(mod)


Call:
lm(formula = Weight ~ Height, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.516  -8.512  -3.448   4.175 107.891 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -82.4739    26.3415  -3.131  0.00213 ** 
Height        0.8643     0.1547   5.588 1.18e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.73 on 138 degrees of freedom
Multiple R-squared:  0.1845,Adjusted R-squared:  0.1786 
F-statistic: 31.23 on 1 and 138 DF,  p-value: 1.182e-07
```

Answer: The estimated model is

$$\texttt{Weight}_i = -82.474 + 0.864 \times \texttt{Height}_i + e_i$$

The slope for height is significant, which you can see in the t test for the slope, $t_{138} = 5.588$, $p < .001$, and, because there is only one predictor in this model, also in the whole model test, $F_{1,138} = 31.228$, $p < .001$ (if you raise the value of the t-statistic to the power 2, you get the value of the F-statistic). So the null hypothesis that the slope is 0 in the population (height is unrelated to weight) can be rejected. For every increase of 1 cm in someone's height, the predicted weight increases by 0.864 kg. Note that the intercept is negative, and the predicted weight of someone with a height of 0 cm is -82.474. This prediction makes little sense, but when the intercept is extrapolated far from the range of the data observed, such implausible predictions can happen. Height does seem to be a good predictor. The $R^2 = 0.185$ so height explains about 18% of the variance in weight.

(d) Estimate the parameters of the regression model

$$\texttt{Weight}_i = \beta_0 + \beta_1 \texttt{Age}_i + \epsilon_i$$

and interpret the results. Is age a good predictor of weight?

```
> # estimate the regression model
> mod <- lm(Weight~Age,data=dat)
> summary(mod)


Call:
lm(formula = Weight ~ Age, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.944  -9.646  -4.210   5.050 117.354 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  52.8730     7.6679   6.895 1.77e-10 ***
Age           0.4684     0.3022   1.550    0.123    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.27 on 138 degrees of freedom
Multiple R-squared:  0.01711,Adjusted R-squared:  0.009985 
F-statistic: 2.402 on 1 and 138 DF,  p-value: 0.1235
```

Answer: The estimated model is

$$\texttt{Weight}_i = 52.873 + 0.468 \times \texttt{Age}_i + e_i$$

For every increase of 1 year in someone's age, the predicted weight increases by 0.468 kg. However, the slope for age is not significant, $t_{138} = 1.55$, $p = 0.123$, or equivalently $F_{1,138} = 2.402$, $p = 0.123$. So the null hypothesis that the slope is 0 in the population is not rejected. The $R^2 = 0.017$ so age explains only about 2% of the variance in weight.

(e) Estimate the parameters of the model

$$\texttt{Weight}_i = \beta_0 + \beta_1\texttt{Height}_i + \beta_2\texttt{Age}_i + \epsilon_i$$

and interpret the results. Are height and age good *unique* predictors of weight? How do the parameters in this model differ from the bivariate models estimated previously? If so, why would this be?

```
> # estimate the multiple regression model
> mod <- lm(Weight~Height+Age,data=dat)
> summary(mod)
```

```
Call:
lm(formula = Weight ~ Height + Age, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-15.803  -8.380  -3.340   3.288 108.660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -86.9755    26.6302  -3.266  0.00138 **
Height        0.8459     0.1554   5.442 2.36e-07 ***
Age           0.3070     0.2767   1.110  0.26905
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.72 on 137 degrees of freedom
Multiple R-squared:  0.1918,Adjusted R-squared:   0.18
F-statistic: 16.26 on 2 and 137 DF,  p-value: 4.626e-07
```

Answer: The estimated model is

$$\texttt{Weight}_i = -86.975 + 0.846 \times \texttt{Height}_i + 0.307 \times \texttt{Age}_i + e_i$$

The slope for height is significant, $t_{136} = 5.442$, $p < .001$. The slope for age is not significant, $t_{136} = 1.11$, $p = 0.269$. Holding age constant, for every increase of 1 cm in someone's height, the predicted weight increases by 0.846 kg. Holding height constant, for every increase of 1 year in someone's age, the predicted weight increases by 0.307 kg. The $R^2 = 0.192$ so together, height and age explain about 19% of the variance in weight. This is mostly due to height, as age seems to have little effect on weight in this sample. Compared to the simple regression models estimated previously, the values of the slopes have changed a little. This is because the slopes in a multiple regression model reflect *unique* effects (e.g., after accounting for the effect of age on both height and weight, what is the relation between height and weight? Is someone who is relatively tall for their age also relatively heavy for their age?). In addition, the test results are somewhat different. This is because we are now testing whether the *unqiue* effects are 0, and we are using different MODEL A and MODEL C for this then in the simple regression models.

(f) In this question, you can check for yourself what is meant by "partial regression coefficients". First, re-estimate the parameters of the regression model in Exercise 1d and save the residuals (errors) as `WtAge`. Then estimate the parameters of the model

$$\texttt{Height}_i = \beta_0 + \beta_1 \texttt{Age}_i + \epsilon_i$$

and save the residuals (errors) as `HtAge`. Finally, estimate the parameters of the model

$$\texttt{WtAge}_i = \beta_0 + \beta_1 \texttt{HtAge}_i + \epsilon_i$$

and compare the parameter estimate b_1 to the estimate b_1 obtained for the model in Exercise 1e.

If you feel like it, you can repeat this analysis, but now “partialling out” the effect of `Height` from both `Weight` and `Age` and saving the residuals as `WtHt` and `AgeHt` respectively. The estimate of β_1 in the model

$$\texttt{WtHt}_i = \beta_0 + \beta_1 \texttt{AgeHt}_i + \epsilon_i$$

should be identical to the estimate b_2 obtained for the model in Exercise 1e.

```
> # estimate the regression model
> mod <- lm(Weight~Age,data=dat)
> #save the residuals
> WtAge <- residuals(mod)
> # estimate the regression model
> mod <- lm(Height~Age,data=dat)
> #save the residuals
> HtAge <- residuals(mod)
> # estimate the 'partialled-out' regression model
> mod <- lm(WtAge~HtAge)
> summary(mod)


Call:
lm(formula = WtAge ~ HtAge)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.803  -8.380  -3.340   3.288 108.660 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.889e-16  1.324e+00   0.000        1    
HtAge       8.459e-01  1.549e-01   5.461 2.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.66 on 138 degrees of freedom
Multiple R-squared:  0.1777,Adjusted R-squared:  0.1718 
F-statistic: 29.83 on 1 and 138 DF,  p-value: 2.133e-07

> # partialling height out of age and weight
> WtHt <- residuals(lm(Weight~Height,data=dat))
> AgeHt <- residuals(lm(Age~Height,data=dat))
> summary(lm(WtHt~AgeHt))


Call:
lm(formula = WtHt ~ AgeHt)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.803  -8.380  -3.340   3.288 108.660 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.099e-16  1.324e+00   0.000    1.000
AgeHt       3.070e-01  2.756e-01   1.114    0.267

Residual standard error: 15.66 on 138 degrees of freedom
Multiple R-squared:  0.008909,Adjusted R-squared:  0.001727 
F-statistic: 1.241 on 1 and 138 DF,  p-value: 0.2673
```

Answer:

$$\texttt{WtAge}_i = 0 + 0.846 \times \texttt{HtAge}_i + e_i$$

The estimate $b_1 = 0.846$ is of course the same as the estimate $b_1 = 0.846$ in the model of Exercise 1e. In both cases, we are estimated unique effects, or partial regression coefficients. Note however that the test results (e.g., t statistics and p-value are slightly different. This is because in the multiple regression model, you take into account that you estimate three parameters. When you first partial-out the effect of age from weight and height, and then form a model where you predict the residual weight from the residual height, you do not take into account that to compute these residuals, you had to estimate parameters for age. So while the procedure of first partialling-out the effect of age from both weight and height and then predicting residual weight from residual height gives identical parameter estimates to multiple regression, which tells you something about how to interpret the slopes in a multiple regression model, the results are not identical in terms of the hypothesis tests that they produce. For testing whether the ”unique” slopes are different from 0 in the population, the hypothesis tests that the multiple regression analysis produces are better.

Doing it the other way round gives

$$\texttt{WtHt}_i = 0 + 0.307 \times \texttt{AgeHt}_i + e_i$$

and the slope of `AgeHt` is identical to the slope of `Age` in the multiple regression model.

2. Open the datafile Geller.sav. This data was simulated to replicate the results of the study by Geller, Johnston, and Madsen (1997). Geller et al. were interested in the extent to which eating disorders are related to the extent to which people base their self-esteem on bodily shape and weight. They obtained measures of the following variables:

   - SAWBS: This measure involved presenting participants with a list of possible variables related to how they feel about themselves. Subjects selected each of the attributes that they felt relevant, and then constructed a pie-chart, where the size of the wedge was an indication of the relative importance of that variable in the subject's self-perception. The angle formed by the shape-and-weight wedge was the dependent variable. To quote the author's description of this measure, "It is important to note that shape- and weight-based self-esteem is not a measure of positive or negative feelings about the body, but rather of the extent to which feelings of self-worth are based on the body."
   - WtPercep: A rating of how participants perceived their weight, from extremely overweight to extremely underweight. (Note that the optimal value is presumably near the center of this scale.)
   - ShPercep: A rating of how participants perceived their shape, from not at all attractive to extremely attractive.
   - HIQ: The Health Inventory Questionnaire measures the presence and severity of disturbed eating practices. It is based on the DSM-IV eating disorders scale. Higher scores indicate more problems.
   - EDIcomp: The Eating Disorders Composite index in the sum of three scales on the Eating Disorders Inventory. These are the Drive for Thinness, Body Dissatisfaction, and Bulimia subscales.
   - RSES: The Rosenberg Self-Esteem Scale is a ten-item Likert scale. Higher scores correspond to higher self-esteem.
   - BDI: The Beck Depression Inventory is a common measure of depression, with higher scores corresponding to more depressed mood.
   - BMI: The body mass index is a measure of body mass based on reported height and weight.
   - SES: Socio-economic status, taken from the general information sheet.
   - SocDesir: The social desirability scale is a 10 item true/false inventory that assess the subject's tendency to respond in socially desirable ways.

   Note that as the data for this exercise is simulated in accordance to the results of Geller et al (1997), and values were not rounded, the dataset contains values that could not have been obtained in the actual study (e.g., negative and non-integer values). But this does not affect the results of your analyses much.

   The researchers propose that the SAWBS will predict eating disorders over and above more traditional predictors such as depression and general self-esteem. Also, the idea is that it is not so much how people perceive their shape and weight that is important in developing eating-disorders, but how much these perceptions are determinants of their self-esteem.

(a) Estimate a model in which you predict the level of eating disorder, as measured by the `EDIcomp`, from `RSES`, `BDI` and `BMI`. Write down the estimated model. Calculate the Sum of Squared Error for this model, and the Proportional Reduction in Error (PRE) from a model with only an intercept.

```
> # read the data
> geller <- as.data.frame(read.spss("Geller.sav"))
> mod <- lm(edicomp~rses+bdi+bmi,data=geller)
> summary(mod)


Call:
lm(formula = edicomp ~ rses + bdi + bmi, data = geller)

Residuals:
     Min       1Q   Median       3Q      Max
-31.4775  -7.5257  -0.3898   5.5416  25.8363

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.5664    14.6948   0.923 0.358674
rses         -0.6801     0.2563  -2.654 0.009596 **
bdi           1.0759     0.2924   3.680 0.000422 ***
bmi           1.0194     0.4131   2.467 0.015741 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.62 on 80 degrees of freedom
Multiple R-squared:  0.5747,Adjusted R-squared:  0.5588
F-statistic: 36.04 on 3 and 80 DF,  p-value: 7.716e-15

> # calculate SSE(A)
> SSEA <- sum(residuals(mod)^2)
> SSEA

[1] 10809.62

> # calculate SSE(C)
> SSEC <- sum(residuals(lm(edicomp~1,data=geller))^2)
> SSEC

[1] 25418.75
```

```
> # calculate PRE
> (SSEC - SSEA)/SSEC

[1] 0.5747384
```

Answer:

$$\texttt{EDIcomp}_i = 13.57 - 0.68 \times \texttt{RSES}_i + 1.08 \times \texttt{BDI}_i + 1.02 \times \texttt{BMI}_i + e_i$$

For this model, SSE $= 10809.62$, and PRE $= R^2 = 0.57$, or by hand PRE $= \frac{25418.75-10809.62}{25418.75} = 0.57$.

(b) Call the previous model with the "standard predictors" MODEL C. Now estimate MODEL A1 which includes `SAWBS` as an additional predictor. Write down the estimated model. Calculate the Sum of Squared Error (SSE) for this model, and the Proportional Reduction in Error (PRE) compared to a model with only an intercept.

```
> # rename the previous model to model C
> modC <- mod
> # add the new predictor to the previous model
> modA1 <- update(modC,.~.+sawbs)
> summary(modA1)


Call:
lm(formula = edicomp ~ rses + bdi + bmi + sawbs, data = geller)

Residuals:
     Min       1Q   Median       3Q      Max 
-28.9073  -6.1283  -0.6431   6.4555  21.6653 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.38826   13.02097   0.875   0.3844    
rses        -0.59457    0.22763  -2.612   0.0108 *  
bdi          0.80572    0.26498   3.041   0.0032 ** 
bmi          0.76236    0.36974   2.062   0.0425 *  
sawbs        0.13044    0.02719   4.797 7.46e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.29 on 79 degrees of freedom
Multiple R-squared:  0.6707,Adjusted R-squared:  0.654 
F-statistic: 40.22 on 4 and 79 DF,  p-value: < 2.2e-16

> # calculate SSE(A1)
> SSEA1 <- sum(residuals(modA1)^2)
> SSEA1

[1] 8371.055

> # calculate PRE
> (SSEC - SSEA1)/SSEC

[1] 0.670674
```

Answer:

$$\texttt{EDIcomp}_i = 11.39 - 0.59 \times \texttt{RSES}_i + 0.81 \times \texttt{BDI}_i + 0.76 \times \texttt{BMI}_i + 0.13 \times \texttt{SAWBS}_i + e_i$$

And SSE(A1) $= 8371.06$, and PRE $= R^2 = 0.67$.

(c) Which null hypothesis will be tested by comparing MODEL C to MODEL A1? Test this null hypothesis.

```
> anova(modC,modA1,test="F")

Analysis of Variance Table

Model 1: edicomp ~ rses + bdi + bmi
Model 2: edicomp ~ rses + bdi + bmi + sawbs
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     80 10809.6
2     79  8371.1  1    2438.6 23.013 7.457e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Remember that MODEL C is now the model with three predictors, while MODEL A has one additional predictor (SAWBS), so the null hypothesis is

$$H_0 : \beta_{\texttt{SAWBS}} = 0$$

We can test this null hypothesis as usual

$$\begin{aligned} F &= \frac{\text{SSR}/(\text{PA} - \text{PC})}{\text{SSE(A)}/(n - \text{PA})} \\ &= \frac{(10809.62 - 8371.06)/(5 - 4)}{8371.06/(84 - 5)} \\ &= 23.01 \end{aligned}$$

As $P(F_{1,79} \geq 23.01) < .001$, the test is significant, and we reject $H_0 : \beta_{\texttt{SAWBS}} = 0$. We can get SPSS to give us the required information by ENTERing variables in blocks. You can then ask for an "R-squared change" statistic. This is the difference between the R^2 for the model in the previous block and the R^2 of the model in the current block. Remember that each R^2 is the PRE of that model compared to a model with only an intercept. So the difference between two R^2 terms is not the PRE obtained by comparing the models for which each R^2 was computed. But you can compute the required PRE of MODEL A1 compared to MODEL C from the R^2 terms (denoted as R^2_{A1} and R^2_C for MODEL A1 and MODEL C respectively) as

$$\text{PRE} = \frac{R^2_{A1} - R^2_C}{1 - R^2_C}$$

In any case, the F value and test result given for the "R-squared change" is the one we want.

(d) Now estimate MODEL A2 which in addition to the predictors in MODEL A1, also includes `WtPercep` and `ShPercep`. Write down the estimated model. Calculate the SSE for this model, and the PRE compared to a model with only an intercept.

```
> # add the new predictor to the previous model
> modA2 <- update(modA1,.~.+wtpercep + shpercep)
> summary(modA2)


Call:
lm(formula = edicomp ~ rses + bdi + bmi + sawbs + wtpercep +
    shpercep, data = geller)

Residuals:
     Min       1Q   Median       3Q      Max
-23.0896  -5.4791  -0.2726   4.1718  18.9319

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.91133   17.57033   3.922 0.000189 ***
rses        -0.50991    0.19889  -2.564 0.012299 *
bdi          0.46639    0.24067   1.938 0.056304 .
bmi         -0.27248    0.41930  -0.650 0.517727
sawbs        0.10021    0.02442   4.104    1e-04 ***
wtpercep    -4.90699    2.20818  -2.222 0.029207 *
shpercep    -4.07347    1.05030  -3.878 0.000220 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.964 on 77 degrees of freedom
Multiple R-squared:  0.7566,Adjusted R-squared:  0.7376
F-statistic: 39.89 on 6 and 77 DF,  p-value: < 2.2e-16
```

```
> # calculate SSE(A2)
> SSEA2 <- sum(residuals(modA2)^2)
> SSEA2

[1] 6186.892

> # calculate PRE
> (SSEC - SSEA2)/SSEC

[1] 0.7566013
```

Answer:

$$\begin{aligned}\texttt{EDICOMP}_i &= 68.91 - 0.51 \times \texttt{RSES}_i + 0.47 \times \texttt{BDI}_i - 0.27 \times \texttt{BMI}_i + 0.10 \times \texttt{SAWBS}_i \\ &\quad -4.91 \times \texttt{WtPerpep}_i - 4.07 \times \texttt{ShPercep}_i + e_i\end{aligned}$$

SSE(A2) $= 6186.89$, PRE $= R^2 = 0.757$.

(e) Compare MODEL A2 to MODEL A1, and determine whether inclusion of `WtPercep` and `ShPercep` reduces the SSE significantly.

```
> # model comparison
> anova(modA1,modA2,test="F")

Analysis of Variance Table

Model 1: edicomp ~ rses + bdi + bmi + sawbs
Model 2: edicomp ~ rses + bdi + bmi + sawbs + wtpercep + shpercep
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     79 8371.1
2     77 6186.9  2    2184.2 13.592 8.803e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: $F_{2,77} = 13.59$, $P(F_{1,77} \geq 13.59) < .001$, so the reduction in error is significant. PRE $= \frac{8371.06-6186.89}{8371.06} = 0.261$, so by including the perception variables, the error is reduced by 26.1%.

(f) Compare the test results for MODEL A2 to those obtained for MODEL A1. What are the differences? Can you explain these?

Answer: In MODEL A1, all predictors (but not the intercept) had significant effects on `EDICOMP`. In MODEL A2, `BMI` is no longer significant, and `BDI` is marginally (almost) significant, and the intercept is significant. It seems that `ShPercep` and `WtPercep` are taking over the role of `BMI` and `BDI`. In the case of `BMI`, this makes immediate sense. The reduction for `BDI` takes a little bit more thinking, but if you think hard you might be able to come up with a reasonable explanation.

(g) Let's try some of the automatic model building routines implemented in SPSS. To make it a bit more exciting, use `HIQ` as the dependent variable and consider all the other variables in the dataset (apart from `EDIcomp`) as potential predictors. Try the "stepwise", "forward" and "backward" methods. In each case, write down the variables included in the final model. Are there differences? If so, can you explain these?

```
> library(MASS)
> mod0 <- lm(hiq~1,data=geller) # intercept-only model
> forward_mod <- stepAIC(mod0,scope=~wtpercep + shpercep + rses + bdi +
+                          bmi + ses + socdesir,direction="forward")

Start:  AIC=371.98
hiq ~ 1

           Df Sum of Sq    RSS    AIC
+ bdi       1   2959.05 3914.2 326.69
+ rses      1   2550.62 4322.6 335.03
+ wtpercep  1   2192.00 4681.2 341.72
+ shpercep  1   2113.53 4759.7 343.12
+ socdesir  1    873.06 6000.2 362.57
+ bmi       1    329.04 6544.2 369.86
+ ses       1    219.72 6653.5 371.25
<none>                  6873.2 371.98

Step:  AIC=326.69
hiq ~ bdi

           Df Sum of Sq    RSS    AIC
+ wtpercep  1    622.73 3291.4 314.13
+ shpercep  1    527.43 3386.8 316.53
+ bmi       1    183.80 3730.4 324.65
+ socdesir  1    150.92 3763.3 325.39
+ rses      1    143.89 3770.3 325.54
<none>                  3914.2 326.69
+ ses       1      1.50 3912.7 328.66

Step:  AIC=314.13
hiq ~ bdi + wtpercep

           Df Sum of Sq    RSS    AIC
+ shpercep  1   178.342 3113.1 311.45
+ rses      1   117.364 3174.1 313.08
+ socdesir  1    98.934 3192.5 313.57
<none>                  3291.4 314.13
```

```
+ ses       1    19.473 3272.0 315.64
+ bmi       1    11.482 3280.0 315.84

Step:  AIC=311.45
hiq ~ bdi + wtpercep + shpercep

           Df Sum of Sq    RSS    AIC
+ socdesir  1   242.253 2870.8 306.65
+ rses      1   100.228 3012.9 310.71
<none>                  3113.1 311.45
+ bmi       1    17.211 3095.9 312.99
+ ses       1     9.867 3103.2 313.19

Step:  AIC=306.65
hiq ~ bdi + wtpercep + shpercep + socdesir

       Df Sum of Sq    RSS    AIC
+ rses  1    95.365 2775.5 305.81
<none>              2870.8 306.65
+ bmi   1     6.720 2864.1 308.45
+ ses   1     1.016 2869.8 308.62

Step:  AIC=305.81
hiq ~ bdi + wtpercep + shpercep + socdesir + rses

       Df Sum of Sq    RSS    AIC
<none>              2775.5 305.81
+ bmi   1    4.0688 2771.4 307.69
+ ses   1    2.0244 2773.5 307.75

> foward_mod

Error in eval(expr, envir, enclos):  object ’foward_mod’ not found

> backward_mod <- stepAIC(lm(hiq~wtpercep + shpercep + rses + bdi +
+                          bmi + ses + socdesir,data=geller),direction="back

Start:  AIC=309.62
hiq ~ wtpercep + shpercep + rses + bdi + bmi + ses + socdesir

           Df Sum of Sq    RSS    AIC
- ses       1     2.428 2771.4 307.69
- bmi       1     4.473 2773.5 307.75
<none>                  2769.0 309.62
- bdi       1    85.678 2854.7 310.18
- rses      1    93.734 2862.7 310.41
- wtpercep  1   120.777 2889.8 311.20
- socdesir  1   217.403 2986.4 313.96
- shpercep  1   286.208 3055.2 315.88

Step:  AIC=307.69
hiq ~ wtpercep + shpercep + rses + bdi + bmi + socdesir

           Df Sum of Sq    RSS    AIC
- bmi       1     4.069 2775.5 305.81
<none>                  2771.4 307.69
- rses      1    92.714 2864.1 308.45
- bdi       1    93.085 2864.5 308.46
- wtpercep  1   118.605 2890.0 309.21
- socdesir  1   228.864 3000.3 312.35
- shpercep  1   299.504 3070.9 314.31

Step:  AIC=305.81
hiq ~ wtpercep + shpercep + rses + bdi + socdesir

           Df Sum of Sq    RSS    AIC
<none>                  2775.5 305.81
- rses      1    95.365 2870.8 306.65
- bdi       1    99.371 2874.9 306.77
- wtpercep  1   145.967 2921.4 308.12
- socdesir  1   237.390 3012.9 310.71
- shpercep  1   298.124 3073.6 312.38

> backward_mod


Call:
lm(formula = hiq ~ wtpercep + shpercep + rses + bdi + socdesir, 
    data = geller)

Coefficients:
(Intercept)     wtpercep     shpercep         rses          bdi  
    43.3005      -2.4395      -2.1335      -0.2160       0.2706  
   socdesir  
    -0.9600  

> stepwise_mod <- stepAIC(mod0,scope=~wtpercep + shpercep + rses + bdi +
+                           bmi + ses + socdesir,direction="both")

Start:  AIC=371.98
hiq ~ 1
```

```
           Df Sum of Sq    RSS    AIC
+ bdi       1   2959.05 3914.2 326.69
+ rses      1   2550.62 4322.6 335.03
+ wtpercep  1   2192.00 4681.2 341.72
+ shpercep  1   2113.53 4759.7 343.12
+ socdesir  1    873.06 6000.2 362.57
+ bmi       1    329.04 6544.2 369.86
+ ses       1    219.72 6653.5 371.25
<none>                  6873.2 371.98

Step:  AIC=326.69
hiq ~ bdi

           Df Sum of Sq    RSS    AIC
+ wtpercep  1    622.73 3291.4 314.13
+ shpercep  1    527.43 3386.8 316.53
+ bmi       1    183.80 3730.4 324.65
+ socdesir  1    150.92 3763.3 325.39
+ rses      1    143.89 3770.3 325.54
<none>                  3914.2 326.69
+ ses       1      1.50 3912.7 328.66
- bdi       1   2959.05 6873.2 371.98

Step:  AIC=314.13
hiq ~ bdi + wtpercep

           Df Sum of Sq    RSS    AIC
+ shpercep  1    178.34 3113.1 311.45
+ rses      1    117.36 3174.1 313.08
+ socdesir  1     98.93 3192.5 313.57
<none>                  3291.4 314.13
+ ses       1     19.47 3272.0 315.64
+ bmi       1     11.48 3280.0 315.84
- wtpercep  1    622.73 3914.2 326.69
- bdi       1   1389.79 4681.2 341.72

Step:  AIC=311.45
hiq ~ bdi + wtpercep + shpercep

           Df Sum of Sq    RSS    AIC
+ socdesir  1    242.25 2870.8 306.65
+ rses      1    100.23 3012.9 310.71
<none>                  3113.1 311.45
+ bmi       1     17.21 3095.9 312.99
+ ses       1      9.87 3103.2 313.19
- shpercep  1    178.34 3291.4 314.13
- wtpercep  1    273.65 3386.8 316.53
- bdi       1    998.14 4111.2 332.82

Step:  AIC=306.65
hiq ~ bdi + wtpercep + shpercep + socdesir

           Df Sum of Sq    RSS    AIC
+ rses      1     95.36 2775.5 305.81
<none>                  2870.8 306.65
+ bmi       1      6.72 2864.1 308.45
+ ses       1      1.02 2869.8 308.62
- wtpercep  1    148.41 3019.3 308.88
- socdesir  1    242.25 3113.1 311.45
- shpercep  1    321.66 3192.5 313.57
- bdi       1    570.06 3440.9 319.86

Step:  AIC=305.81
hiq ~ bdi + wtpercep + shpercep + socdesir + rses

           Df Sum of Sq    RSS    AIC
<none>                  2775.5 305.81
- rses      1    95.365 2870.8 306.65
- bdi       1    99.371 2874.9 306.77
+ bmi       1     4.069 2771.4 307.69
+ ses       1     2.024 2773.5 307.75
- wtpercep  1   145.967 2921.4 308.12
- socdesir  1   237.390 3012.9 310.71
- shpercep  1   298.124 3073.6 312.38

> stepwise_mod


Call:
lm(formula = hiq ~ bdi + wtpercep + shpercep + socdesir + rses,
    data = geller)

Coefficients:
(Intercept)          bdi     wtpercep     shpercep     socdesir
    43.3005       0.2706      -2.4395      -2.1335      -0.9600
       rses
    -0.2160

> ### note that the stepAIC finds the same model in each case
> ### these results differ quite a bit from the procedures
> ### implemented in SPSS, which are based on F tests
> ### in the answer below I'm discussing SPSS results
```

Answer: With stepwise and forward, we end up with the following model:

$$\texttt{HIQ}_i = 17.96 + .50 \times \texttt{BDI}_i + .068 \times \texttt{SAWBS}_i - 3.24 \times \texttt{WtPercep}_i + e_i$$

with an $R^2 = .615$. With backward, we get the following model

$$\texttt{HIQ}_i = 39.90 + .069 \times \texttt{SAWBS}_i - 2.44 \times \texttt{ShPercep}_i - 0.30 \times \texttt{RSES}_i - 1.19 \times \texttt{SocDesir}_i + e_i$$

with $R^2 = 0.656$. The reason that these methods end with different models is due to the correlation between predictors, making some predictors redundant. But this redundancy in a practical way depends on the other predictors in the model and it's hard to predict what will happen. In any case, you should be aware that these methods can arrive at different models. If I had to choose, I would probably go with the simpler model; the difference in R^2 between the models is not that large to justify the additional parameter. On the other hand, the more complex model may fit better with some theories of eating disorders.