Intro to R Programming: Lab 3

All data files for this lab need to be downloaded from the module's Moodle page first.

Task 1

Read the data files `health.txt` and `cia.csv` into R. Make sure missing values are read in correctly.

Task 2

Use the function `write.table` to save the data frame `health` you created in the first task as a comma separated file usign column names, but no row names. Use * for missing values. Onve you have created the file, open it in an editor or Excel and check its content.

Task 3

In this task you will learn about the function `cut`, which can be used to discretise a numerical variable.

For instance, consider a variable `speed` created using

```
speed <- c(34, 49, 52, 24, 60, 74, 55)
```

Suppose we want to turn this continuous variable into a discrete variable with only three categories. We can do so using the function `cut`:

```
speed.discretised <- cut(speed, breaks=3)
speed.discretised
```

```
## [1] (23.9,40.7] (40.7,57.3] (40.7,57.3] (23.9,40.7] (57.3,74]   (57.3,74]
## [7] (40.7,57.3]
## Levels: (23.9,40.7] (40.7,57.3] (57.3,74]
```

If we give no additional arguments to `cut`, it tries to determine the cutoff points automatically. You can also set these and the corresponding labels manually:

```
speed.discretised <- cut(speed, breaks=c(0,40,60,100),
                         labels=c("slow", "medium", "fast"))
speed.discretised
```

```
## [1] slow   medium medium slow   medium fast   medium
## Levels: slow medium fast
```

Speeds > 0 and ≤ 40 are classed as `slow`, speeds > 40 and ≤ 60 are classed as `medium`, etc.

Use what you have learned in the example to add a new column called `ExpectancyGroup` that takes the values `low` (`LifeExpectancy`≤ 40), `medium` ($40 <$`LifeExpectancy`≤ 70), and `high` (`LifeExpectancy`> 70) to the data frame `health` from the previous task.

Task 4

The file `maternity.csv` contains data (obtained from `data.gov.uk`) on the number of pregnant women who were still smoking at delivery for each Health Authority in England in the first quarter 2011. The data also contains the number of mothers breastfeeding when their baby is 6–8 weeks old. The file contains the following variables:

`HealthAuthority`	Name of the Health Authority (Primary Care Trust)
`Region`	Name of the Region (Strategic Health Authority)
`Deprivation`	Average deprivation score
`Maternities`	Number of maternities
`Smoking`	– with the mother smoking at delivery
`SmokingUnknown`	– for which the smoking status of the mother could not be determined
`Breastfeeding`	– with the mother breastfeeding at 6 weeks
`BreastfeedingUnknown`	– for which the breastfeeding status of the mother could not be determined

Use R to determine ...

1. for each Health Authority the proportion of smoking pregnant women and the proportion of breastfeeding mothers;
2. the name of the Health Authority which both the smallest and the largest proportion of smoking pregnant women / breastfeeding mothers;
3. the percentage of smoking pregnant women / breastfeeding mothers in the North West and in London;
4. the percentage of smoking pregnant women / breastfeeding mothers for Health Authorities with an average deprivation score of at most 10 and at least 40; and
5. the percentage of breastfeeding mothers for Health Authorities with more than 25% (and less than 15%) smoking pregnant women.

Task 5

In this task you we will implement simple linear regression, both using the "sum formulae" and using matrix algebra. Load in the alligator.csv file from moodle.

`Length`	Snout vent length in inches (distance between back of head and end of nose)
`Weight`	Weight of the animal in pounds

1. Create two vectors `x` and `y`, such that `x` contains the logarithm of the snout vent length (covariate) and `y` contains the logarithm of the weight (response).
2. In the linear regression model $E(y_i) = \beta_0 + \beta_1 x_i$ with one covariate $\mathbf{x} = (x_1, \ldots, x_n)$ and response $\mathbf{y} = (y_1, \ldots, y_n)$ the least-squares estimate of the regression coefficient $\boldsymbol{\beta} = (\beta_0, \beta_1)$ can be found using the formulae

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}.$$

2.1. Compute the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the above formulae.

2.2. Create a vector `y.hat` that contains the predictions $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

2.3. The estimate of the variance of the residuals is $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$. Use R to compute this estimate.

2.4. Use R to compute the coefficient of determination: $R^2 = 1 - \dfrac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

3. In this part you will compute the least-squares estimate using matrix algebra:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

by solving the system of equations

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

3.1. Create the design matrix

$$\mathbf{X} = (\mathbf{1}\ \mathbf{x}) = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

3.2. Create a matrix `XtX` which holds $\mathbf{X}'\mathbf{X}$ and a vector `Xty` which holds $\mathbf{X}'\mathbf{y}$.

3.3. You can now solve the above system of equations by solving $\texttt{XtX}\hat{\boldsymbol{\beta}} = \texttt{Xty}$.

3.4 Compute the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and store them in a vector `y.hat`.

Parts (b) and (c) should give you the same regression coefficients and fitted values.