

Instructions

This lab consists of a collection of sequential exercises from Chapter 7 in the course textbook (OpenIntro Statistics). Even-numbered exercises have been assigned, along with corresponding odd-numbered exercises for reference. When answering the assigned even-numbered exercises, it is recommended that the student consider using the odd-numbered exercises (as needed) for reference, since their respective answers are given in Appendix B (pp. 405 - 426). Further, while the textbook provides software-generated output for each exercise, the student shall be given the data set and asked to replicate the output, and for some [additional] questions generate new output using RStudio. Example R commands are provided for performing the same task on the aforementioned odd-numbered exercise(s). The student will need to make the appropriate modifications to the **highlighted** inputs in the example R commands in order to adopt them to the even-numbered exercises.

Unless otherwise indicated, the student shall answer all parts, e.g. (a), (b), (c), ... for each assigned exercise. *ALL RSTUDIO GENERATED OUTPUT SHOULD BE INCLUDED.*

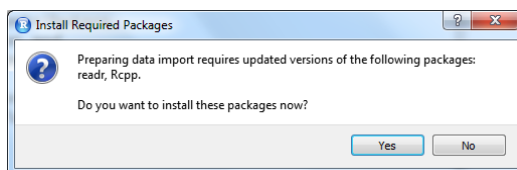
It is recommended that the student read through the assigned exercises and at least review Chapter 7 on pp. 331 - 356. NOTE: sections 7.1 to 7.4 are [technically] review content from the prerequisite course (MAT 120).

1. Assigned Exercise(s): 7.6 and 7.38 on pages 358 and 369

Reference Exercise(s): 7.37 on page 368

It is recommended that the student first load the **MARRIED** data set into RStudio using these step-by-step instructions.

- Download either the **MARRIED.TXT** and/or **MARRIED.XLSX** file to your computer.
- Open RStudio. In the *ENVIRONMENT* window (located in the upper right), click on the *Import Dataset* button and select *From CSV*, if you downloaded the TXT file, or *From Excel*, if you downloaded the XLSX file. Note: The *Import Dataset* option is also found under the *File* menu at the top of RStudio. If you are doing this for the first time, you may be met with the following dialog window; Click **Yes**.



- Click on the *Browse* button (upper right) and locate the downloaded **MARRIED.TXT** or **MARRIED.XLSX** file. If loading the TXT file, under the *Import Options*: area (box in middle of *Import Text Data* window), click on the drop-down menu labeled *Delimiter*: and select **Tab**. Finally, click on the *Import* button (lower right) to load the file.

In Exercise 7.6, prior to answering parts (a) through (d), please complete the following:

- The **MARRIED** data set contains four columns: `age_husband`, `height_husband`, `age_wife`, and `height_wife`, which correspond to the age (in years) and height (in inches; converted from mm) for the husband and wife of 170 randomly selected, married British couples.

The command below will generate the scatterplot of the wife's height against husband's height. Make the appropriate edits and generate the other scatterplot: the wife's age plotted against her husband's age, and capture/submit both plots. Use these plots (or those found in the textbook, which should be identical) to answer questions (a) and (b).

```
plot(
  x=MARRIED$height_husband,
  y=MARRIED$height_wife,
  xlab="Husband's height (in inches)",
  ylab="Wife's height (in inches)",
  pch=16
)
```

- The command below computes the correlation coefficient between the heights (in inches) of husbands and wives. Make the appropriate edits to compute the correlation coefficient between the ages (in years) of husbands and wives, and capture both values. Use these values to answer question (c).

```
cor(MARRIED$height_husband, MARRIED$height_wife)
```

- Make the appropriate edit(s) to the **cor** command above to compute the correlation coefficient using the heights measured in centimeters - you'll just need to multiply each column but the conversion factor 2.54 and re-run the command. In other words, add ***2.54** to the end of each column name in the above **cor** command. Use this result to answer question (d).

In Exercise 7.38, prior to answering parts (a) through (d), please do the following:

- Generate the table containing the summary output of the least squares fit. As an example for your reference, to generate the table found in exercise 7.37, the command (in blue), with corresponding output (in black) would be:

```
summary(fm <- lm(height_wife~height_husband,data=MARRIED))
```

Call:

```
lm(formula = height_wife ~ height_husband, data = MARRIED)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-7.0117	-1.4162	-0.1622	1.6592	6.9594

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	43.57555	4.68423	9.303	< 2e-16 ***
height_husband	0.28630	0.06862	4.172	4.82e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.334 on 168 degrees of freedom

Multiple R-squared: 0.09389, Adjusted R-squared: 0.0885

F-statistic: 17.41 on 1 and 168 DF, p-value: 4.825e-05

To display more precise coefficient estimates, enter the following command (in blue):

```
fm$coefficients
```

```
(Intercept) height_husband
43.5755466    0.2863048
```

- **Guide for Answering Part (a):** Refer to 7.4.2 pp. 353 - 354 in the textbook. While the wording of this problem may be a bit confusing, simply put, we are being asked to test the following hypotheses: $H_0: \beta_1 = 1$ vs. $H_A: \beta_1 \neq 1$. Typically, the hypothesized value is zero, rather than one, since we are looking to see if there is sufficient evidence that a linear relationship exists between the explanatory: age_husband and response: age_wife variables, i.e. $\beta_1 \neq 0$. The test statistic and p-value that were generated in the table - under the t value and $\text{Pr}(>|t|)$ columns, respectively - for the coefficient of age_husband assumes such. Therefore, we will actually need to [re]compute the test statistic and p-value using 1 for the hypothesized (null) value of the coefficient, rather than 0. I have provided the calculations for each - in exercise 7.37 - when $\beta_1 = 0$ below, you will need to update the highlighted values and recompute using $\beta_1 = 1$ and the respective coefficients for exercise 7.38. NOTE: the T score (test statistic) will be negative, so be sure to take its absolute value when computing the p-value.

Coefficients:						} Point estimate and standard error for slope coefficient from summary output of the least squares fit.
	Estimate	Std. Error	t value	$\text{Pr}(> t)$		
(Intercept)	43.57555	4.68423	9.303	< 2e-16	***	
height_husband	0.28630	0.06862	4.172	4.82e-05	***	

$$t = \frac{\text{estimate} - \text{null value}}{SE} = \frac{0.28630 - 0}{0.06862} \approx 4.172253$$

$$p.\text{value} = P(|T| > |t|) = P(|T| > 4.172253) \approx 4.827089 \times 10^{-5}$$

R Command: `pt(q=4.172253, df=168, lower.tail=F) * 2`

Aside, the following R commands will provide a more precise calculation of the test statistic and p-value by directly accessing the coefficients stored in the fm object. You would just need to update null.value to 1.

```

null.value <- 0
b1 <- summary(fm)$coefficients[2,1] # or coef(summary(fm))[2,1]
SE <- summary(fm)$coefficients[2,2] # or coef(summary(fm))[2,2]
t <- (b1-null.value)/SE; t
pt(q=abs(t),df=fm$df.residual,lower.tail=F)*2

```

- **Guide for Answering Part (b):** Refer to 7.2.3 on pp. 342 - 344 in the textbook. Using the coefficients listed under the Estimate column in the summary output of the least squares fit for exercise 7.37, which is reproduced below, we have

Coefficients:					
	Estimate	Std. Error	t value	$\text{Pr}(> t)$	
(Intercept)	43.57555	4.68423	9.303	< 2e-16	***
height_husband	0.28630	0.06862	4.172	4.82e-05	***

$$\widehat{\text{height_wife}} = 43.57555 + 0.28630 \text{ height_husband}$$

- **Guide for Answering Part (c):** Refer to 7.2.4 on p. 345 in the textbook.
- **Guide for Answering Part (d):** Refer to 7.2.6 on pp. 346 - 347 in the textbook.

- **Guide for Answering Part (e):** Given the discussion in the textbook on p. 335 (Section 7.1.2) and p. 346 (Example 7.20), the question is only expecting a point estimate. However, I would like you to utilize the `predict.lm` command in RStudio to provide an interval estimate, along with an appropriate interpretation. Below is the command (in blue) to generate the point estimate and prediction interval for 7.37(e), along with its corresponding console output (in black). Edit as necessary to answer 7.38(e).

```
predict.lm(
  object=fm,
  newdata=list(height_husband=69),
  interval="prediction",
  level=0.95
)
```

```
1      fit      lwr      upr
1 63.33058 58.70715 67.95401
```

Point estimate is listed under `fit`.
Limits for prediction interval are listed under `lwr` and `upr`.

Interpretation:

We are 95% confident the wife's height for a randomly selected husband in Britain that is 69 inches tall will be between 58.71 and 67.95 inches.

- **Guide for Answering Part (f):** See 7.2.5 on pp. 345 - 346 of the textbook.