STAT5002 Lab12 Question Sheet

Introduction to Statistics

STAT5002

1 Simple linear regression

In this workshop, we will first re-analysis the simple linear regression model we built in Week 4, where we considered simulated climate data based on real data from the Bureau of Meteorology at Canterbury Racecourse AWS {station 066194} collected in 2023. The simulated data contains several different daily measurements throughout Autumn (March-May).

You need the data file `AutumnCleaned.csv` to load the variables needed for this question. We will use the following variables.

- `X9am.temperature` (**daily temperature measured at 9 am**)
- `Minimum.temperature` (**minimum daily temperature**)

The temperature data are measured in Celsius. Please beware that the variable names are **case sensitive**.

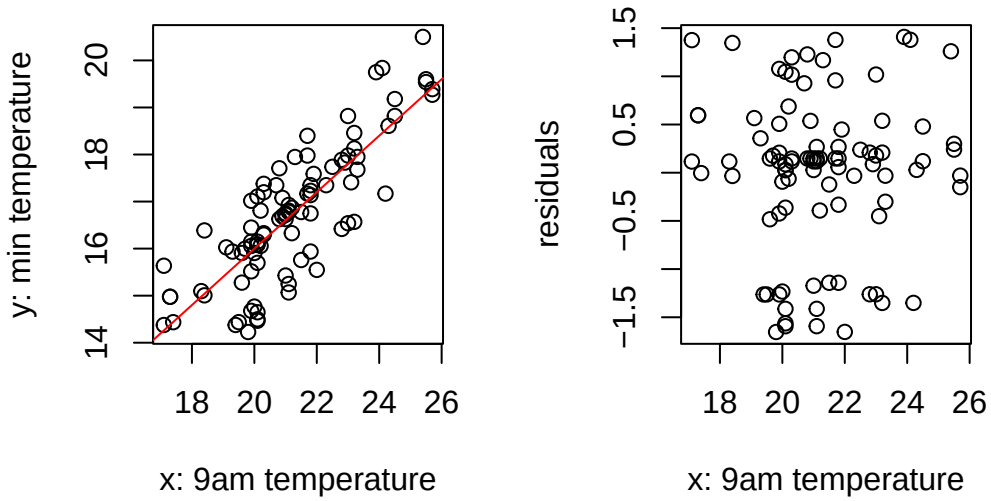
```
data = read.csv("data/AutumnCleaned.csv", header = T)
### the following displays the dimension of the data
dim(data)
```

```
[1] 92 16
```

```
T9am = data$X9am.temperature
Tmin = data$Minimum.temperature
```

We can use the function `lm()` to fit a linear regression model for predicting minimum daily temperature (Y) using daily temperature measured at 9 am (x). The following code shows the scatter plot of the data, the fitted linear regression line, and the residual plot.

```r
###
par(mfrow = c(1, 2))
plot(T9am, Tmin, xlab = "x: 9am temperature", ylab = "y: min temperature")
lm1 = lm(Tmin ~ T9am)
abline(lm1, col = "red")
plot(T9am, lm1$residuals, xlab = "x: 9am temperature", ylab = "residuals")
```



We want to test if the daily temperature measured at 9 am (x) has a **positive** linear association with the minimum daily temperature (Y). In the following, we want to first carry out these steps by "hand", and then compare our result with R.

1.1 Hypotheses

- What are the null and alternative hypotheses.

1.2 Checking assumptions

- What are the assumptions we need to check here?
- Use appropriate graphical summaries to check them?

1.3 Test statistic

Recall that the T-statistic takes the form of

$$T = \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} = \frac{\hat{b}_1}{\widehat{SE}(\hat{b}_1)} \sim t_{n-(*)}$$

where

$$\widehat{SE}(\hat{b}_1) = \sqrt{\frac{1}{n-(*)} \frac{\text{sum of squared residual}}{\text{sum of squared deviations in } x_1}}$$

- What is the value of $*$? In other words, what is the degrees of freedom of Student's t-distribution here?
- What is the value of b_1?
- Calculate the observed value of the test statistic.

1.4 P-value and conclusion.

- Calculate the P-value.
- What is your conclusion?

1.5 Validate your result using `summary()`

- Use the function `summary()` to validate your results.

1.6 Calculate a 95% confidence interval for the true slope.

2 Multiple linear regression

The effect of body height (`H`, in inches) and body weight (`W`, in pounds) on catheter length (`L`, in centimetres) was studied in a sample of $n = 12$ children with congenital heart disease. Because selecting an appropriate catheter length is critical for safe and effective cardiac procedures in pediatric patients, a multiple linear regression model was used to examine whether these physical measurements (`H` and `W`) can help predict the required catheter length (`L`). A data set is given as follows.

```
H = c(42.8, 63.5, 37.5, 39.5, 45.5, 38.5, 43, 22.5, 37, 23.5, 33, 58)
W = c(40, 93.5, 35.5, 30, 52, 17, 38.5, 8.5, 33, 9.5, 21, 79)
L = c(37, 49.5, 34.5, 36, 43, 28, 37, 20, 33.5, 30.5, 38.5, 47)
dat = data.frame(H, W, L)
dim(dat)
```

```
[1] 12  3
```

```
str(dat)
```

```
'data.frame':   12 obs. of  3 variables:
 $ H: num  42.8 63.5 37.5 39.5 45.5 38.5 43 22.5 37 23.5 ...
 $ W: num  40 93.5 35.5 30 52 17 38.5 8.5 33 9.5 ...
 $ L: num  37 49.5 34.5 36 43 28 37 20 33.5 30.5 ...
```

2.1 Plot the pairwise scatter plots and calculate the pairwise correlation coefficients.

- Check the scatter plots and the correlation coefficients. What are your observations regarding the associations among the three variables?

2.2 Using R, fit a multiple linear regression model.

- Write down the multiple linear regression model. How can you interpret the fitted model?
- Based on the summary of the multiple linear regression model, are the independent variables `H` and `W` significant for predicting the required catheter length (`L`), after adjusting for each other? You can use a 5% level of significance.

2.3 Model comparison

The height variable is correlated with the weight (which makes sense). If using the weight variable `W` alone can make a good prediction on the the catheter length `L`, we may not need to use the multiple linear regression model. Let's examine this.

- Build a simple linear regression model predicting the catheter length `L` using weight `W`.
- Write down the fitted simple linear regression model and interpret the model.
- What numerical summary can be used to compare the performance between the multiple linear regression model and the simple linear regression model?
- Calculate the numerical summary for both models (reading from `summary()`), what is your conclusion?