STAT5002 Lab12 Solution 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**.

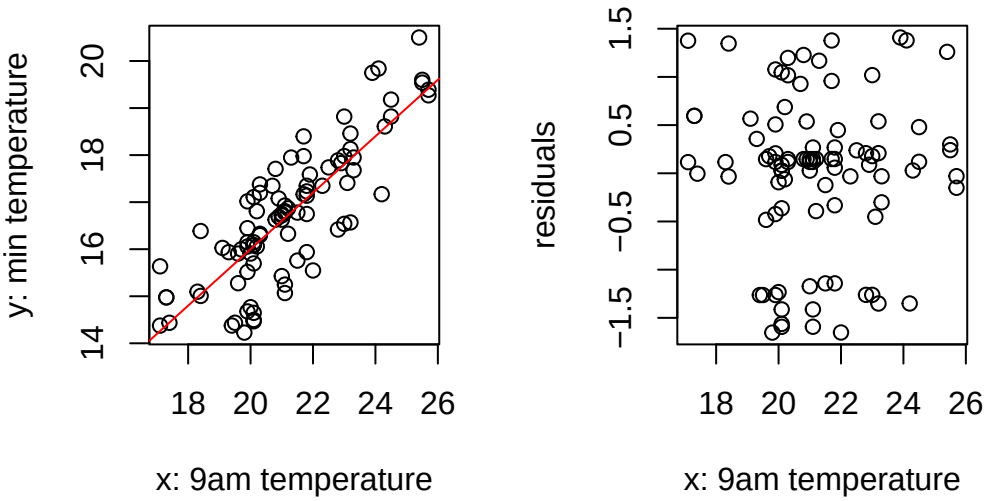
```
data = read.csv("data/AutumnCleaned.csv", header = T)
### the following displays the dimension of the data
dim(data)
```

```
[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.

Solution:

$H_0 : b_1 = 0$ – There is no linear association between 9 am temperature and the daily minimum temperature.

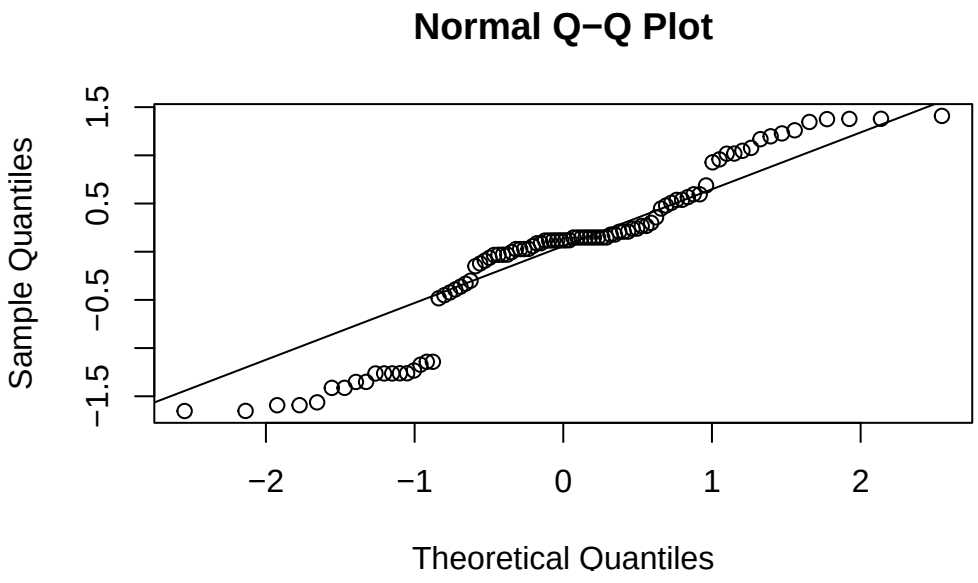
$H_1 : b_1 > 0$ – There is a positive linear association between 9 am temperature and the daily minimum temperature.

1.2 Checking assumptions

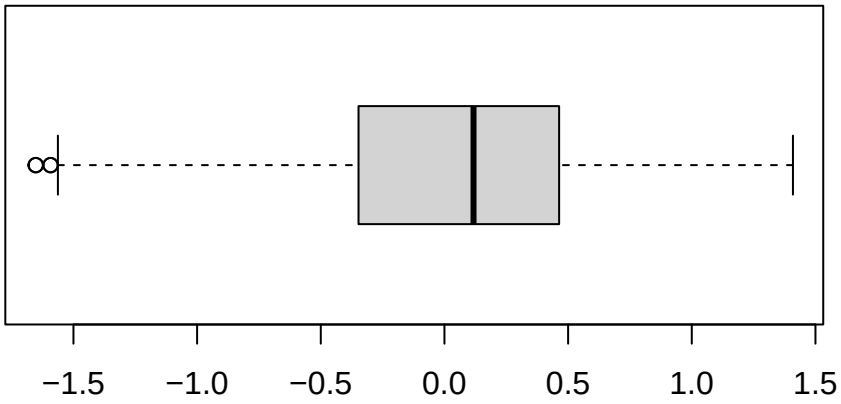
- What are the assumptions we need to check here?
- Use appropriate graphical summaries to check them?

Solution:

```
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```




```
boxplot(lm1$residuals, horizontal = T)
```



- Linearity: The relationship between the independent x and the dependent variable y is linear. The scatter plot and the residual plot appear to show a linear relationship between x and y in this case.
- Independence: It depends on the setup of the experiment. In this example, we don't have sufficient information to justify the independence.

- Homoscedasticity (Constant variance): The variance of the residuals appears to be constant across all levels of x.
- Normality of residuals: there are some quantile points in the QQ plot that do not follow the QQ line, especially in the lower tail. It suggests that the normality may not hold. The boxplot also suggests outliers in the lower tail.

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 estimated standard error and the observed value of the test statistic "by hand" (without using `summary()`).

Solution: We have two parameters (slope and intercept), so $* = 2$. The degrees of freedom is $n - 2$ in this case.

The value of b_1 is 0 (under H_0).

The following code calculates the estimated standard error and the observed value of the test statistic.

```
n = length(T9am)
SSE = sum(lm1$residual^2)
deviations = T9am - mean(T9am)
SSD = sum(deviations^2)
est.SE = sqrt(SSE/SSD/(n - 2))
est.SE
```

```
[1] 0.04350773
```

```
stat = lm1$coefficient[2]/est.SE
stat
```

```
    T9am
13.79065
```

1.4 P-value and conclusion.

- Calculate the P-value.
- What is your conclusion at the 1% level of significance?

Solution: The following code calculates the P-value. Note that it's a one-sided test and only large values of test statistic (corresponding to large slopes) argue against H_0.

Solution: The P-value is close to zero, so we reject H_0 (no linear association) at the 1% level of significance. This indirectly suggests that the temperature at 9 am is linearly associated with the daily minimum temperature.

```
pt(stat, df = n - 2, lower.tail = F)
```

```
        T9am
3.240475e-24
```

1.5 Validate your result using `summary()`

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

Solution: Your estimated standard error and observed value of T-statistic should agree with the summary (along the row `T9am`). `summary()` gives the P-value of the two-sided alternative, the one-sided P-value is just half of that.\

Note that the value is too small to be displayed, you can use `summary(lm1)$coefficients[,4]` to find it, which exactly doubles the one-sided P-value obtained by hand.

```
summary(lm1)
```

```
Call:
lm(formula = Tmin ~ T9am)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6530 -0.3397  0.1176  0.4559  1.4090
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.00000    0.92721   4.314  4.1e-05 ***
T9am         0.60000    0.04351  13.791  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8198 on 90 degrees of freedom
Multiple R-squared:  0.6788,    Adjusted R-squared:  0.6752
F-statistic: 190.2 on 1 and 90 DF,  p-value: < 2.2e-16
```

```r
p.values = summary(lm1)$coefficients[, 4]  # extract the P-value
p.values
```

```
(Intercept)        T9am
4.10069e-05 6.48095e-24
```

```r
p.values[2]  # P-value of T9am
```

```
       T9am
6.48095e-24
```

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

Solution: Recall the formula

$$P\left(\hat{b}_1 - u \times \widehat{SE}(\hat{b}_1) \leq b_1 \leq \hat{b}_1 + u \times \widehat{SE}(\hat{b}_1)\right) = 0.95$$

where u is the upper 2.5% percentage point under Student's t-distribution with $n - (p + 1)$ degrees of freedom. The following code calculates the 95% confidence interval for the give data.

```r
u = qt(0.975, df = n - 2)
est.SE * c(-1, 1) * u + lm1$coefficients[2]
```

```
[1] 0.5135643 0.6864357
```

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)
```

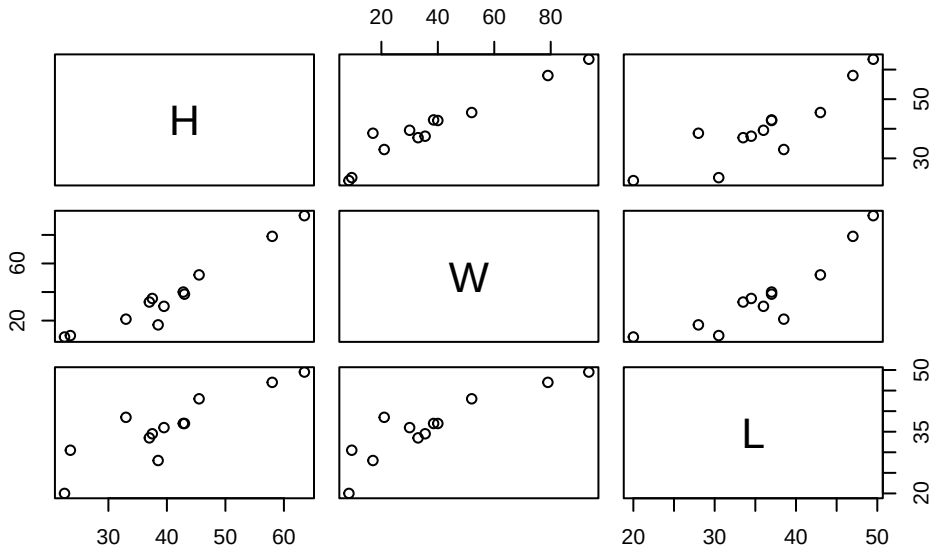
```
'data.frame':   12 obs. of  3 variables:
 $ H: num  42.8 63.5 37.5 39.5 45.5 38.5 43 22.5 37 23.5 ...
 $ W: num  40 93.5 35.5 30 52 17 38.5 8.5 33 9.5 ...
 $ L: num  37 49.5 34.5 36 43 28 37 20 33.5 30.5 ...
```

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?
- Do we need to be concerned about the effect of multicollinearity in this dataset?

Solution:

```
pairs(dat)  # Pairwise scatter plots
```




```
cor(dat)  # Pairwise Pearson correlation matrix
```

```
          H         W         L
H 1.0000000 0.9610936 0.8811691
W 0.9610936 1.0000000 0.8938226
L 0.8811691 0.8938226 1.0000000
```

- Both height and weight are strongly positively correlated with catheter length.
- Height and weight are also strongly correlated with each other ($r = 0.96$), which suggests multicollinearity.

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.

Solution:

```
model_full <- lm(L ~ H + W, data = dat)
summary(model_full)
```

```
Call:
```

```
lm(formula = L ~ H + W, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-7.048 -1.258 -0.259  1.899  7.004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.0084     8.7512   2.401   0.0399 *
H             0.1964     0.3606   0.545   0.5993
W             0.1908     0.1652   1.155   0.2777
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.943 on 9 degrees of freedom
Multiple R-squared:  0.8053,    Adjusted R-squared:  0.7621
F-statistic: 18.62 on 2 and 9 DF,  p-value: 0.0006336
```

We have the linear regression model

$$\hat{L} = 21.0084 + 0.1964 \cdot H + 0.1908 \cdot W$$

- Holding weight constant, an increase of 1 inch in height is associated with an estimated increase of 0.1964 cm in catheter length in average.
- Holding height constant, an increase of 1 pound in weight is associated with an increase of 0.1908 cm in catheter length in average.

After adjusting for each other, neither the height nor the weight is statistically significant at a 5% level of significance, as their P-values are smaller than 0.05. This is caused by the high correlation between height `H` and weight `W`.

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?

Solution:

```
model_weight <- lm(L ~ W, data = dat)
summary(model_weight)
```

```
Call:
lm(formula = L ~ W, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-7.994 -1.481 -0.135  2.091  7.040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.63746    2.00421  12.792 1.60e-07 ***
W            0.27727    0.04399   6.303 8.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.802 on 10 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7788
F-statistic: 39.73 on 1 and 10 DF,  p-value: 8.871e-05
```

$$\hat{L} = 25.63746 + 0.27727 \cdot W$$

- An increase of 1 pound in weight is associated with an increase of 0.27727 cm in catheter length in average. This is quite different compared to the previous multiple linear regression model.
- We can use the coefficient of determination (r^2) to compare the performance between the multiple linear regression model and the simple linear regression model.
- The coefficient of determination for the multiple linear regression model is 0.8053, whereas that of the simple linear regression model is 0.7989. By including the height variable only leads to a marginal improvement in the prediction performance.