lab 2

Mason Wong

15th March 2022

```
library(tidyverse)
library(dplyr)
# library(broom)
# library(patchwork)
knitr::opts_chunk$set(echo = TRUE, tidy = FALSE, fig.align = "center", fig.pos = "H")
```

Question 1

```
olympic = read.table("~/Desktop/R-programming/3002-labs/lab-2/olympic.txt", sep = "\t", header = TRUE)
head(olympic, 6)
```

```
##   HighJump DiscusThrow LongJump Year
## 1    71.25     1147.50  249.750   -4
## 2    74.80     1418.90  282.875    0
## 3    71.00     1546.50  289.000    4
## 4    75.00     1610.00  294.500    8
## 5    76.00     1780.00  299.250   12
## 6    76.25     1759.25  281.500   20
```

```
summary(olympic)
```

```
##     HighJump      DiscusThrow      LongJump          Year
##  Min.   :71.00   Min.   :1148   Min.   :249.8   Min.   :-4.00
##  1st Qu.:76.19   1st Qu.:1775   1st Qu.:296.2   1st Qu.:22.00
##  Median :78.97   Median :2033   Median :308.2   Median :52.00
##  Mean   :80.92   Mean   :2053   Mean   :311.3   Mean   :47.65
##  3rd Qu.:86.25   3rd Qu.:2435   3rd Qu.:331.6   3rd Qu.:74.00
##  Max.   :92.75   Max.   :2657   Max.   :350.5   Max.   :96.00
##  NA's   :3       NA's   :3
```

```
tail(olympic, 6)
```

```
##    HighJump DiscusThrow LongJump Year
## 18    88.50      2657.4   328.50   76
## 19    92.75      2624.0   336.25   80
## 20    92.50      2622.0   336.25   84
## 21       NA          NA   343.25   88
## 22       NA          NA   341.50   92
## 23       NA          NA   334.75   96
```

The possible unusual features about the `olympic` dataset are: - The distances are in inches. We should aim to convert them into meters - The years are relative to 1900's - There are 3 NA observations for the last 3 observations.

We change the inches to meters

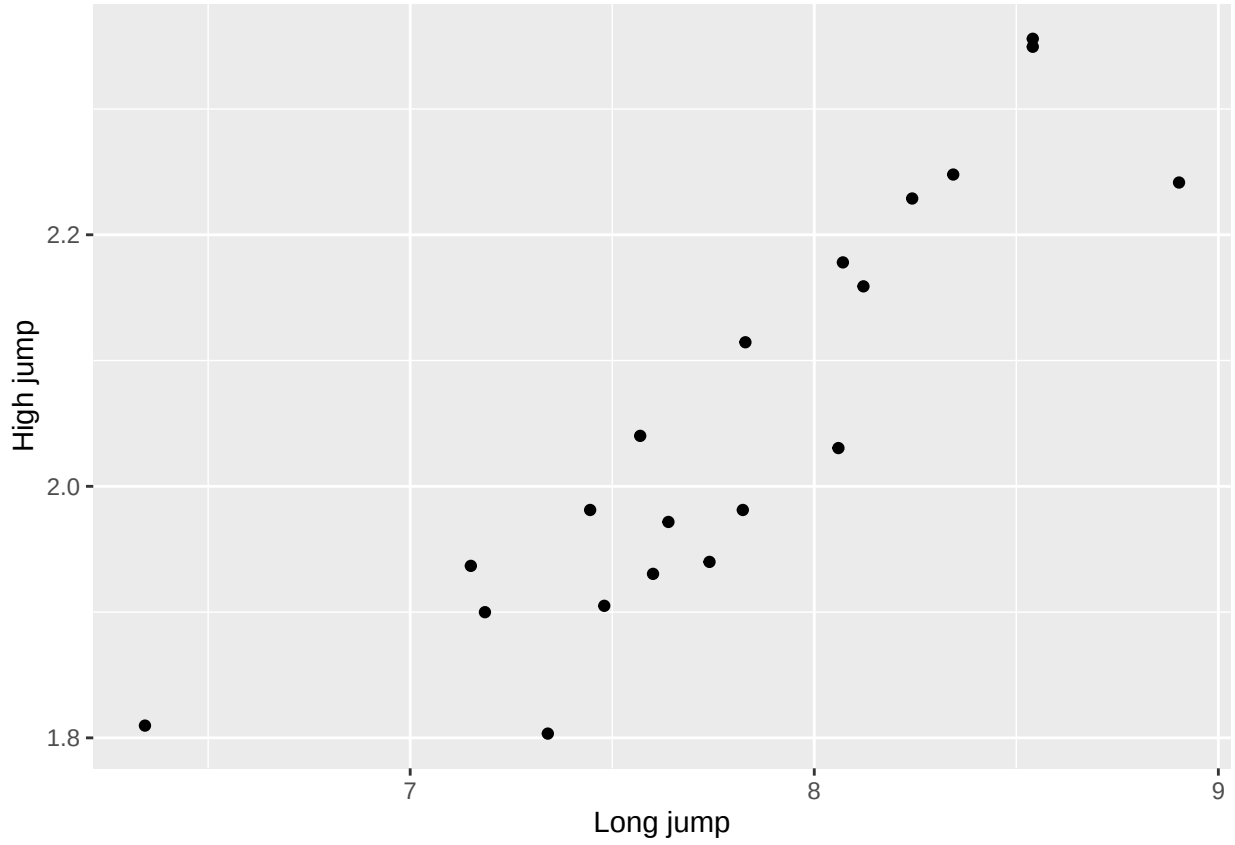
```
convert_inch_to_m = function(x) {
  return(x/39.3701)
}
olympicMetric = olympic %>%
  mutate(across(c(HighJump, DiscusThrow, LongJump), convert_inch_to_m)) %>%
  mutate(Year = Year + 1900)
```

Question 2

(a)

```
olympicMetric %>%
  ggplot(aes(x = LongJump, y = HighJump)) +
  geom_point() +
  labs(x = "Long jump", y = "High jump")
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



We see that there appears to be a linear trend.

(b) we fit a simple linear regression model.

```
olympicLm = lm(HighJump ~ LongJump, data = olympicMetric)
```

(c) we find a least squares estimate for the parameters $(\beta_1, \beta_2, \sigma^2)$ using a `summary` output from `olympicLm`

```
data_summary = summary(olympicLm)
data_summary
```

```
##
## Call:
## lm(formula = HighJump ~ LongJump, data = olympicMetric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13574 -0.07615  0.01865  0.05390  0.12339
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07790    0.25016   0.311    0.759
## LongJump     0.25355    0.03199   7.925  2.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08177 on 18 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.7773, Adjusted R-squared:  0.7649
## F-statistic: 62.81 on 1 and 18 DF,  p-value: 2.8e-07
```

From the R output we see that

- $\hat{\beta}_0 = 0.0779048$
- $\hat{\beta}_1 = 0.2535541$
- $\hat{\sigma} = 0.0817736$

We construct a 95% confidence interval for $\hat{\beta}_1$. It is given by:

```
t_val = qt(p = 1 - (0.05)/2, df = data_summary$df[2], lower.tail = TRUE)
```

$$0.2535541 \pm 0.0672135$$

Which is equal to:

$$0.1863407, \quad 0.3207676$$

We also manually verify the result with:

```
confint(olympicLm)
```

```
##                  2.5 %    97.5 %
## (Intercept) -0.4476659 0.6034756
## LongJump     0.1863407 0.3207676
```

(e) We now use the anova function to produce an ANOVA table for testing

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

```
anova(olympicLm)
```

```
## Analysis of Variance Table
##
## Response: HighJump
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## LongJump   1 0.42003 0.42003  62.813 2.8e-07 ***
## Residuals 18 0.12036 0.00669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The t values given in the summary table is 7.925 and the f value in the summary table is given by 62.813. We see that:

$$(7.925)^2 = 62.813$$

Ignoring rounding errors. Hence, we have that the f statistics is the square of the t statistic.

(f) We test the hypothesis of

$$H_0 : \beta_1 = 0.25 \qquad vs \qquad H_1 : \beta_1 > 0.25$$

1. Under the Null hypothesis we have our test statistic to be

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{18}$$

Now since we have that $\hat{\beta}_1 = 0.2535541$ and $SE(\hat{\beta}_1) = 0.0319924$ we have that the observed statistic is:

$$t_0 = 0.1110935$$

2. Hence our p value is given by:

$$\text{p value} = P(T > t_0) = 0.4563858$$

Since the p value is greater than or equal to 0.05 we fail to reject the null hypothesis