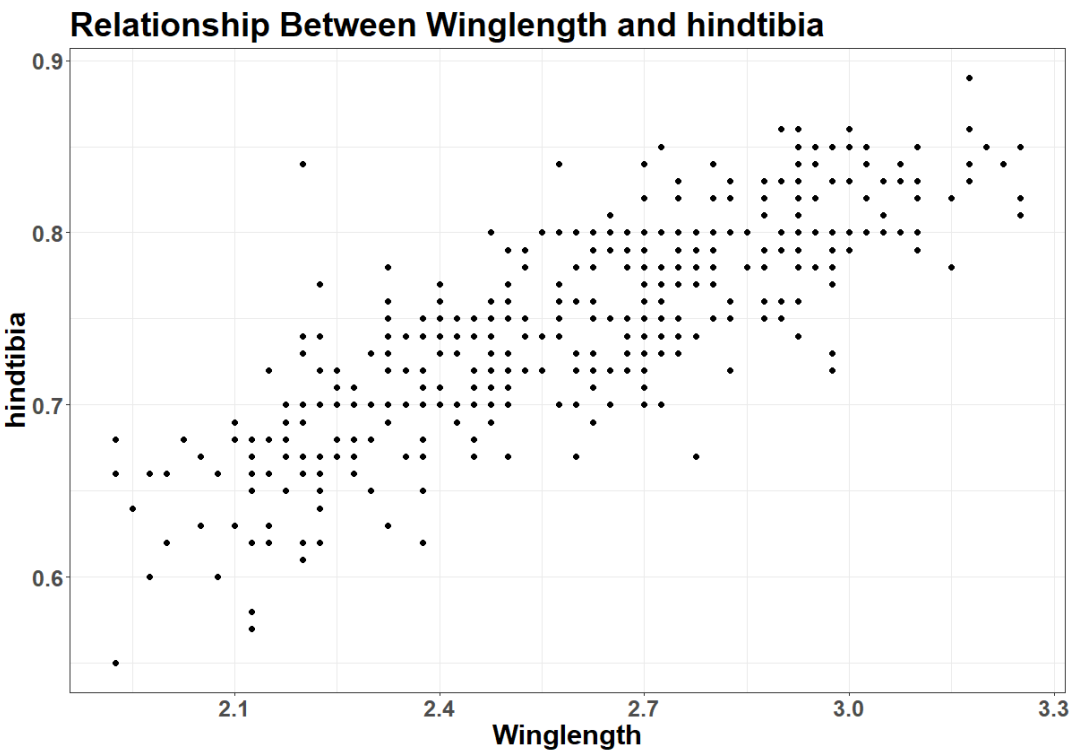
Problem 1

a. **Regression between Winglength and hindtibia**

Code for the Problem

```
library(ggplot2)
library(tidyverse)
library(ggtext)
library(ggpubr)
library(ggplot2)

dat <- read.csv('Drosophila.csv')

dat %>%
  head(3)

dat %>%
  summary()


## Regression

ggscatter(dat, x = 'winglength', y = 'hindtibia',
          xlab = "Winglength", ylab = "hindtibia")+
  theme_bw()+
  theme(text=element_text(family="Times", face="bold", size=30))+
  ggtitle("Relationship Between Winglength and hindtibia")
```

Output



Explanation

The scatterplot above shows the relationship between winglength and hintibia of the flies. The data frame was given the label dat. From the scatter plot we can observe a positive relationship between the two variables where an increase in hintibia in flies results to proportionate increase in the flies' winglength.

The library(ggpubr) allowed the code to run the scatter plot while the ggplot2 library allowed the use of themes in the code above.

b) **What does the conditional random variable hindtibia | winglength = 3.15 represent in this context? Please use words.**

Conditional random variable is normally used in showing probability of occurrence between two variables. In this case, 3.15 means that the value of the hintibia will be dependent or given when the wing length is at 3.15mm. This where the two variables will meet at the line of best fit.

(c)_ [2 pts] Fit a simple linear regression model that uses the wing length to predict the hind tibia length in R and produce the model summary. (Reminder: show R code and output.)

Equation

Hint tibia = b0 + b1 * winglenth

Where b0 is the intercept and b1 the beta coefficient.

We us in r the function lm() to get the beta coefficient of the linear model described in the equation above.

Code

```
## Statistical Mode

dat_model <- lm(hindtibia ~ winglength, data = dat)

dat_model

summary(dat_model)
```

Output 1

```
> dat_model <- lm(hindtibia ~ winglength, data = dat)
> dat_model

Call:
lm(formula = hindtibia ~ winglength, data = dat)

Coefficients:
(Intercept)   winglength
     0.3168       0.1665
```

Output 2

```
> summary(dat_model)

Call:
lm(formula = hindtibia ~ winglength, data = dat)

Residuals:
      Min        1Q    Median
-0.108960 -0.023203  0.000263
       3Q       Max
 0.025163  0.156797

Coefficients:
            Estimate Std. Error
(Intercept)  0.31683    0.01473
winglength   0.16653    0.00564
            t value Pr(>|t|)
(Intercept)   21.51   <2e-16
winglength    29.52   <2e-16

(Intercept) ***
winglength  ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
  0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03475 on 427 degrees of freedom
Multiple R-squared:  0.6712,    Adjusted R-squared:  0.6704
F-statistic: 871.7 on 1 and 427 DF,  p-value: < 2.2e-16

>
```

Explanation

From the above results we state that a value in flies winglength can be expected to provide a hindtibia value by using the equation below where winglenth is any value given to winglenth.

0.32 + 0.0475*winglenth.

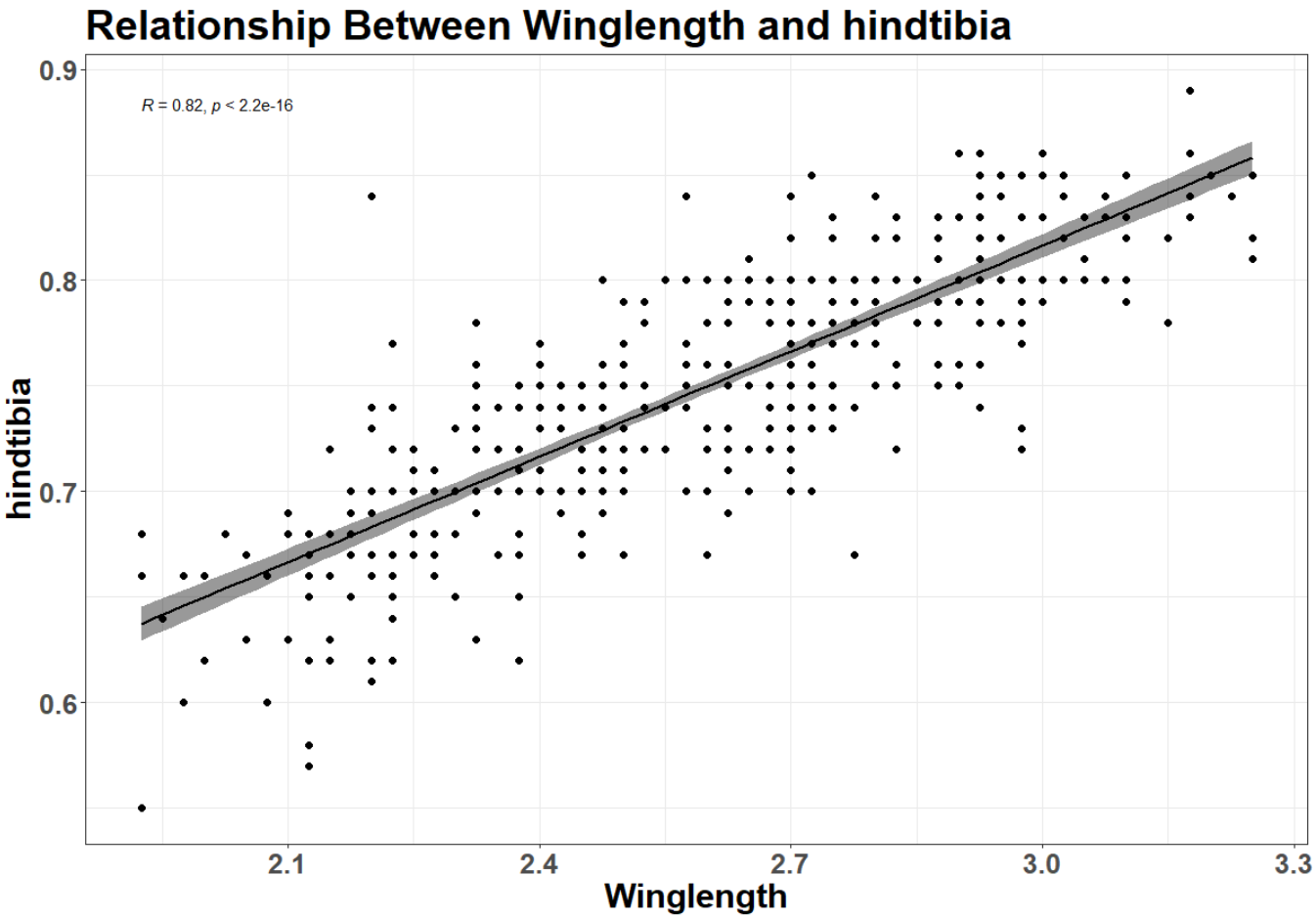
The r function here used the library package tidyverse ad function lm() in conducting the statistical model test.

d) Scatter Plot with fitted regression model

i. Code for the Problem

```
library(ggplot2)
library(tidyverse)
library(ggtext)
library(ggpubr)
library(ggplot2)

dat <- read.csv('Drosophila.csv')

dat %>%
  head(3)

dat %>%
  summary()


## Regression

ggscatter(dat, x = 'winglength', y = 'hindtibia',
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Winglength", ylab = "hindtibia")+
  theme_bw()+
  theme(text=element_text(family="Times", face="bold", size=23))+
  ggtitle("Relationship Between Winglength and hindtibia")
```

ii. Output to the Code



iii. Explanation

The 'reg.line' in the code refers to the best line fit or regression line, conf.coef refers to the Pearson correlation coefficient value. The conf.int is the confidence interval whereas xlab is where the axes are labeled.

The Pearson vale R= 0.82 indicates a strong positive correlation between the two variables where as one increases the other variable increases proportionally. This is also observed as most values in the diagram are gathered along the line of best fit. The p value (p< 2.2e-16) indicates that the relationship is statistically significant meaning the null hypothesis which states that there is no relationship between the two variables can be rejected.

e) Prediction of results

```
> dat_model <- lm(hindtibia ~ winglength, data = dat)
> dat_model

Call:
lm(formula = hindtibia ~ winglength, data = dat)

Coefficients:
(Intercept)   winglength
     0.3168       0.1665
```

Explanation

To predict a value using fitted regression model we normally use the intercept value (b0) and the slope coefficient intercept value (b1) as in the equation below:

Hint tibia = b0 + b1 * winglenth

The intercept and slope coefficients were already calculated when we calculating the summary for the model test.

Intercept (b0) = 0.3168

Slope coefficient (b1) = 0.1665

Hence to calculate for hinttibia when the wing length is 3.15mm we use the formula as stiputated below:

Hinttibia = 0.3168 + 0.1665*3.15 = 0.8413

f) Find the fruit flies with wing length equal to 3.15 mm

Explanation

There are only two flies with wing length equal to 3.15 mm, with tibia length of 0.78 mm and 0.82 mm respectively. Since we were allowed to use Microsoft excel, I filtered the dataset to find the given fruit flies.

g) Changing the unit of Winglength from mm to cm

Explanation

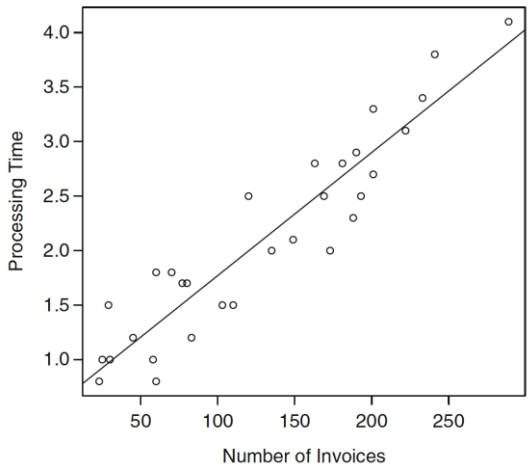
When the units are changed to cm the slope will still remain the same as the change does not affect the correlation between the two variables. This is usually because the measurements used in correlation coefficient in the study standardizes the variables.

Problem 2

The manager of the purchasing department of a large company would like to develop a regression model to predict the amount of time it takes to process a given number of invoices. Data are collected on the number of invoices processed and the total time taken (in hours). Please utilize the graph and R output below to complete the following tasks.

Note: There is no data file for Problem 2.

Scatterplot with the fitted line:



Summary of the model fitted using the Least Square Estimation method:

```
Call:
lm(formula = Time ~ Invoices)

Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept) 0.6417099     0.1222707     5.248    1.41e-05 ***
Invoices    0.0112916     0.0008184    13.797    5.17e-14 ***
---
Residual standard error: 0.3298 on 28 degrees of freedom
Multiple R-Squared: 0.8718, Adjusted R-squared: 0.8672
F-statistic: 190.4 on 1 and 28 DF, p-value: 5.175e-14
```

(a)_[1 pt] What is the response variable? There is no need to explain. *Hint: look at the model summary.*

Time

(b)_[1 pt] What is the sample size? There is no need to explain.

28

(c)_[2 pts] Compute the RSS of the fitted model. Please show your work.

The first step involves fitting the model. In our case we will use 50 invoices to predict time as in the equation below.

Predicted Time = 0.6417 + 0.01129*50 =1.2 hours

After fitting the given model, we then calculate the RSS using the formula below.

RSS = ∑_(i=1)^n (y_i - ŷ_i)^2

Where n is the sample,

Y_1 is the observed time

ŷ_i is the predicted time

Y_1 is 1.5

ŷ_i is 1.2

n is 28

RSS = ∑_(i=1)^n (y_i - ŷ_i)^2 = ∑_(i=1)^28 (1.5 - 1.2)^2 = ∑_(i=1)^28 (0.3)^2 = 28 * 0.3^2 = 28 * 0.09 = 2.52

So, the residual sum of squares for the given sample is 2.52. Which mean the model has a good fit for prediction.

(d)_[2 pts] Compute the SXX value (i.e., $\sum_{i=1}^{n} \left(x_i - \underline{x}\right)^2$). Please show your work. Hint: The variance of $\hat{\beta}_1$ is $\frac{\sigma^2}{SXX}$. Therefore the estimated standard error of $\hat{\beta}_1$ is $\frac{\hat{\sigma}}{\sqrt{SXX}}$. *Hint: your final answer should be above 100,000.*

The estimated standard error of the slope coefficient β^_1 is given as 0.0008184. We can use this to estimate the variance of β^_1, which is σ^2/SXX. Thus, SXX can be calculated as:

SXX = σ^2 / (estimated standard error of β^_1)^2 = 3.05 / (0.0008184^2) = 151568

(e)_[2 pts] The intercept (β_0) represents the start-up time for processing invoices. Use the hypothesis test to evaluate if we need any start-up time ($H_0: \beta_0 = 0$ vs. $H_A: \beta_0 > 0$). Provide the test statistic value and the p value. Based on the test, does the dataset provide strong evidence that the start-up time is greater than 0? You may use 0.05 as the significance level. Please show your work.

Hypothesis test will be as follows:

H_0: β_0 = 0

H_A: β_0 > 0

The null hypothesis (H_0) states that the start-up time is 0, and the alternative hypothesis (H_A) states that the start-up time is greater than 0.

With a significance level of 0.05, if the p-value is less than 0.05, we reject the null hypothesis and conclude that there is strong evidence that the start-up time is greater than 0.

In this case, the p-value (1.41e-05) is much less than 0.05, so we reject the null hypothesis and conclude that there is strong evidence that the

start-up time is greater than 0. The dataset provides strong evidence that a start-up time is required.

(f)_[2 pts] A best practice benchmark for invoice processing is 0.01 hour (or 0.6 min) for each additional invoice. Conduct a hypothesis test to determine if the mean processing time for an additional invoice at this company is 0.01 hour. Please test the null hypothesis against a two-sided alternative hypothesis. Provide the hypotheses, the test statistic value, and the p value. Based on the test, does the dataset provide strong evidence that the mean processing time of an invoice at this company is different from 0.01 hours? You may use 0.05 as the significance level. Please show your work.

To perform the hypothesis test, we used a t-test. The null hypothesis is the mean processing time of an invoice at this company which is equal to 0.01 hours (i.e. $H_0: \beta_1 = 0.01$).

The alternative hypothesis is that the mean processing time is different from 0.01 hours (i.e. $H_A: \beta_1 \neq 0.01$).

Using the summary model provided, the estimate of the mean processing time of an invoice is 0.0112916 hours. The standard error of the estimate is 0.0008184. The t-statistic is given by:

$t = (0.0112916 - 0.01) / 0.0008184 = 1.566$

The degrees of freedom for this test is 28. Using a two-sided t-test with a significance level of 0.05, we find a critical value for a t-distribution with 28 degrees of freedom as ±1.833.

Since the t-statistic of 1.566 is not greater than the critical value of 1.833, we fail to reject the null hypothesis. We do not have strong evidence to conclude that the mean processing time of an invoice at this company is different from 0.01 hours.

Using a t-distribution table we get a p-value for a t-statistic of 1.566 and 28 degrees of freedom to be approximately 0.13. Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis.