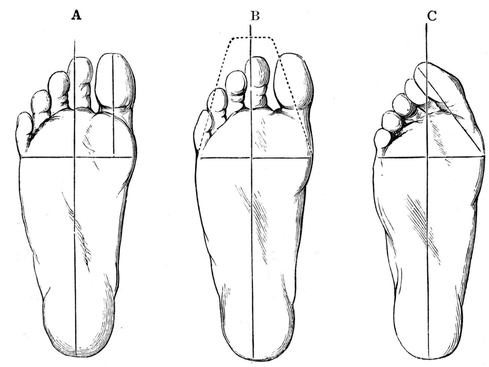
MATH1324 - Class Worksheet

Week 11 - Linear Regression

Working in small groups or pairs, complete the following exercises.

Datasets: `Hand_length.csv` (see Data Repository)

Hand and Foot Length



Activity Description

Does hand length predict a person's foot length? Export the class `Hand_length.csv` dataset collected in class (see Data Repository), screen and clean, and import into RStudio. Complete the following exercises:

1. Visualise the relationship between hand and foot length. Does there appear to be a linear relationship?

2. Fit a linear regression model to the data and assess the fit. Is the model statistically significant?

3. Test the main assumptions for linear regression. Are the assumptions safe?

4. Interpret and test the statistical significance of the regression intercept and slope.

5. Use the estimated linear regression model and your hand length to predict your foot size. Compare your predicted foot size to your actual foot size. Is it a close prediction?

6. Draw an overall conclusion in the context of the example.

Optional

7. Does the relationship between hand and foot length depend on gender?

R Hints

Use this example R code if you get stuck:

Hand and Foot Length

```
> library(mosaic)

> xyplot(foot_length ~ hand_length, data = Hand_length,
     ylab ="Foot length (mm)",xlab="Hand length (mm)")

> model1 <- lm(foot_length ~ hand_length, data = Hand_length)

> msummary(model1)

> confint(model1)

> mplot(model1, 1)

> qqPlot(model1$residuals, dist="norm")
```

Solutions and Answers

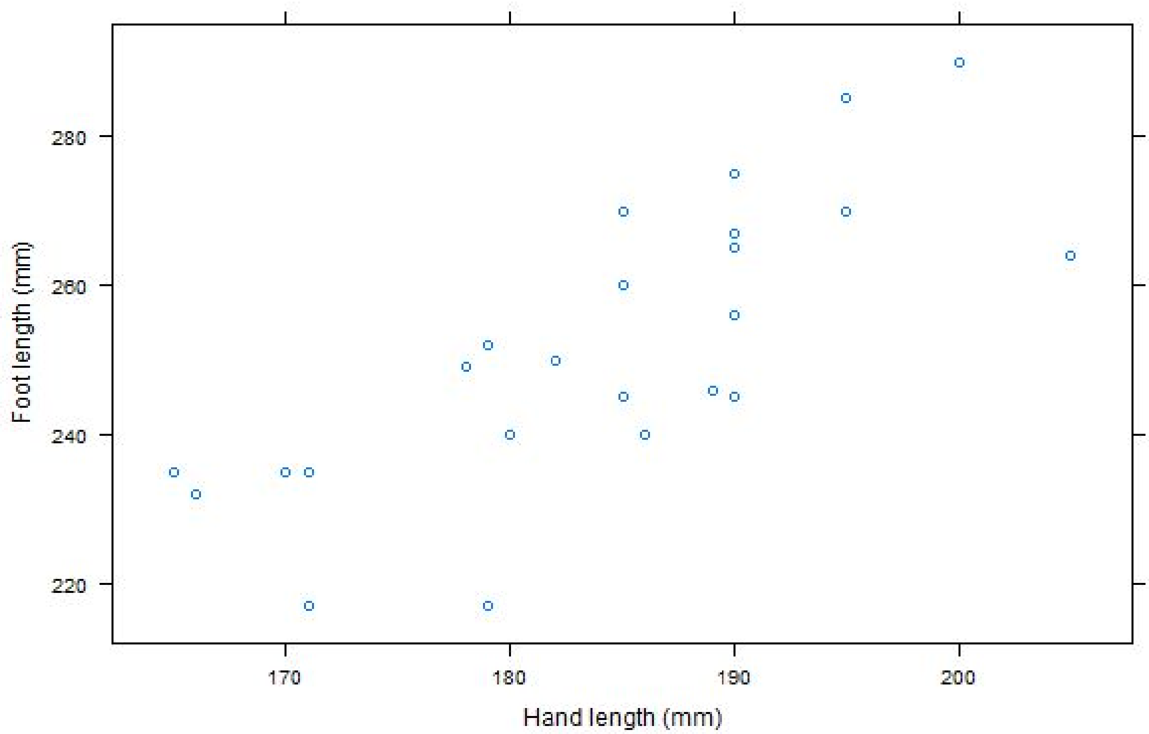
Hand and Foot Length

1. Visualise the relationship between hand and foot length. Does there appear to be a linear relationship?

First, filter out the two cases (18 and 25) reported in cms. These cases distort the relationship because the other measures, reported in mm, are more precise.

```
> Hand_length_sub <- subset(Hand_length, subset = hand_length > 100)
```

```
> xyplot(foot_length ~ hand_length, data = Hand_length_sub,
        ylab ="Foot length (mm)",xlab="Hand length (mm)")
```



The relationship appears linear and positive.

2. Fit a linear regression model to the data and assess the fit. Is the model statistically significant?

```
> model1 <- lm(foot_length ~ hand_length, data = Hand_length_sub)
> msummary(model1)
            Estimate Std. Error t value Pr(>|t|)
```

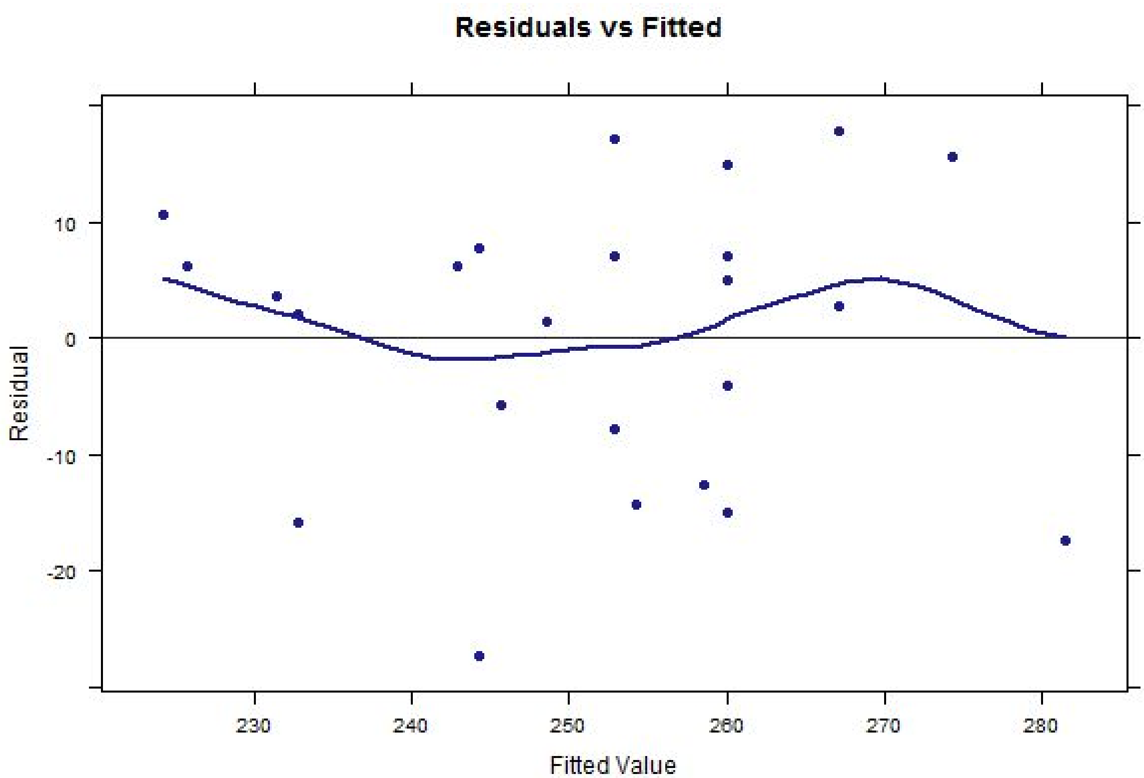
```
(Intercept) -11.5551    45.6234  -0.253    0.802
hand_length   1.4293     0.2478   5.768  7.1e-06 ***

Residual standard error: 12.37 on 23 degrees of freedom
Multiple R-squared:  0.5912,  Adjusted R-squared:  0.5735
F-statistic: 33.27 on 1 and 23 DF,  p-value: 7.098e-06
```

The linear model was statistically significant, $F(1,25) = 33.27$, $p < .001$. Hand length explained 59.12% of the variability in foot length.

3. Test the main assumptions for linear regression. Are the assumptions safe?

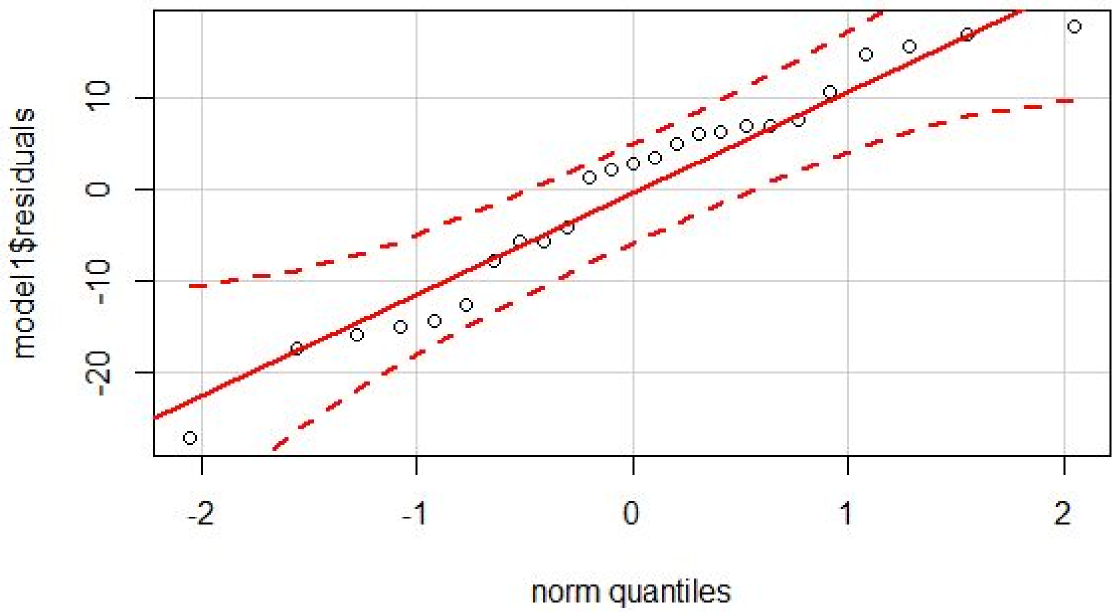
Independence was assumed as each foot and hand measurement came from different people. The scatter plot suggested a linear relationship. Other non-linear relationships were ruled out.

```
> mplot(model1, 1)
```



Homoscedasticity looked OK. The variance in residuals appeared constant across predicted values.

```
> qqPlot(model1$residuals, dist="norm")
```



Residuals appeared to be approximately normally distributed.

4. Interpret and test the statistical significance of the regression intercept and slope.

```
> msummary(model1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.5551    45.6234  -0.253    0.802
hand_length   1.4293     0.2478   5.768  7.1e-06 ***

Residual standard error: 12.37 on 23 degrees of freedom
Multiple R-squared:  0.5912,  Adjusted R-squared:  0.5735
F-statistic: 33.27 on 1 and 23 DF,  p-value: 7.098e-06

> confint(model1)
                   2.5 %    97.5 %
(Intercept) -105.9343099 82.824169
hand_length    0.9166377  1.941881
```

The estimated average foot length when hand length = 0 was -11.56 mm.

The intercept of the regression was not statistically significant, a = -11.55, p = .802, 95% *CI* (-105,93, 82.82).

For every one unit increase in hand length (mm), foot length was estimated to increase on average by 1.43 mm.

The slope of the regression for hand length was statistically significant, $b = 1.43$, $p < .001$, 95% *CI* (0.92, 1.94).

5. Use the estimated linear regression model and your hand length to predict your foot size. Compare your predicted foot size to your actual foot size. Is it a close prediction?

Say your hand the foot measurement were 179 and 217 mm respectively.

Using the estimated regression equation:

$$\text{Predicted Foot length} = -11.55 + 1.43*179$$

$$\text{Predicted Foot length} = -11.55 + 1.43*179 = -11.55 + 255.97 = 244.42$$

This prediction has an error or residual of 217 - 244.42 = - 27.42 mm. The regression model did not predict this value well.

6. Draw an overall conclusion in the context of the example.

Overall, there was a statistically significant positive linear relationship between hand length and foot length. A person's hand length was estimated to explain up to 59% of the variability in foot length.

Optional

7. Does the relationship between hand and foot length depend on gender?

```
> Hand_length_male <- subset(Hand_length_sub, subset = Gender == "Male")
> xyplot(foot_length ~ hand_length, data = Hand_length_male,
+        ylab ="Foot length (mm)",xlab="Hand length (mm)",)
> model2 <- lm(foot_length ~ hand_length, data = Hand_length_male)
> msummary(model2)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.4180    56.6996   0.307 0.762917
hand_length   1.2904     0.3023   4.269 0.000673 ***

Residual standard error: 11.16 on 15 degrees of freedom
Multiple R-squared:  0.5485,  Adjusted R-squared:  0.5184
F-statistic: 18.22 on 1 and 15 DF,  p-value: 0.0006729
```

```
>
> Hand_length_female <- subset(Hand_length_sub, subset = Gender == "Female")
> xyplot(foot_length ~ hand_length, data = Hand_length_female,
+        ylab ="Foot length (mm)",xlab="Hand length (mm)",)
> model3 <- lm(foot_length ~ hand_length, data = Hand_length_female)
> msummary(model3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.5388    98.4730   1.194    0.278
hand_length   0.6617     0.5578   1.186    0.280

Residual standard error: 12.12 on 6 degrees of freedom
Multiple R-squared:   0.19,   Adjusted R-squared:  0.055
F-statistic: 1.407 on 1 and 6 DF,  p-value: 0.2803
```

The relationship was stronger in males. The relationship was not statistically significant for females due to insufficient data.