stat359_A4_wducharme

Wesley Ducharme

2024-11-18

Question 1

```
lungcancer_data <- read.csv(file="C:/Users/wesch/uvic/stat359 Data
↪  Analysis/data1/LungCancer.csv", header=TRUE)
attach(lungcancer_data)

names(lungcancer_data)
```

```
## [1] "Case"   "Smoker"
```

```
table(lungcancer_data)
```

```
##      Smoker
## Case   0   1
##    0  60 650
##    1  22 687
```

Turning the table into a matrix for furthur analysis

```
lungcancer_table <- as.matrix(table(lungcancer_data))
lungcancer_table
```

```
##      Smoker
## Case   0   1
##    0  60 650
##    1  22 687
```

```
# Table of expected values
expected_counts <- chisq.test(lungcancer_table,correct=FALSE)$expected

#Test statistic mock of what was done by hand to double check value
observed <- lungcancer_table
test_stat <- sum((observed - expected_counts)^2 / expected_counts)
test_stat
```

```
## [1] 18.63299
```

Test statistic calculated is 18.6 and the distribution is a chi-squared dist with 1 degree of freedom. Degrees of freedom was determined by ((NumberOfColumns - 1)*(NumberOfRows - 1)) = ((2-1)(2-1)) = 1

Computing p-value for test of association H0: Lungcancer is independent of smoking. Ha: They are not independent.

```
#Getting p-value
chisq.test(lungcancer_table,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  lungcancer_table
## X-squared = 18.633, df = 1, p-value = 1.585e-05
```

```
#Observed values
lungcancer_table
```

```
##     Smoker
## Case   0   1
##    0  60 650
##    1  22 687
```

```
#Expected values
chisq.test(lungcancer_table,correct=FALSE)$expected
```

```
##     Smoker
## Case        0        1
##    0 41.02889 668.9711
##    1 40.97111 668.0289
```

P-value is very small therefore we have very strong evidence against the null hypothesis of which we then reject. Conclude that there is an association between smoking and lung cancer.

When comparing the observed and expected values it can be seen that the observed values for being a non smoker with no lung cancer and being a smoker with lung cancer are high than the expected values. It can also be seen that the observed values for being a non-smoker with lungcancer and being a smoker without lung-cancer are lower than expected.

Question 2

Making a matrix of the observed data to compute the table of expected values.

```
observed2 <- matrix(c(7, 7, 7, 13,
                      27, 34, 12, 18,
                      55, 52, 11, 24),
                    nrow = 3, byrow = TRUE)
rownames(observed2) <- c("Moderate-advanced", "Minimal", "Not Present")
colnames(observed2) <- c("O", "A", "AB", "B")

# Totals
```

```
row_totals <- rowSums(observed2)
col_totals <- colSums(observed2)
grand_total <- sum(observed2)

# Compute expected counts
expected <- outer(row_totals, col_totals) / grand_total
dimnames(expected) <- dimnames(observed2)

# Display observed and expected counts
observed
```

```
##     Smoker
## Case   0   1
##    0  60 650
##    1  22 687
```

```
expected
```

```
##                          O        A        AB         B
## Moderate-advanced 11.33333 11.84270  3.820225  7.003745
## Minimal           30.33333 31.69663 10.224719 18.745318
## Not Present       47.33333 49.46067 15.955056 29.250936
```

After calculating the value by hand, the observed value of the test statistic is 17.91. The distribution is a chi-squared dist with 6 degree of freedom. Degrees of freedom was determined by ((NumberOfColumns - 1)*(NumberOfRows - 1)) = ((4-1)(3-1)) = 6

Getting the p-value for the test of association.

```
#Getting p-value from our calculated values
chi_sq <- 17.91
df <- 6

p_value <- pchisq(chi_sq, df, lower.tail = FALSE)
p_value
```

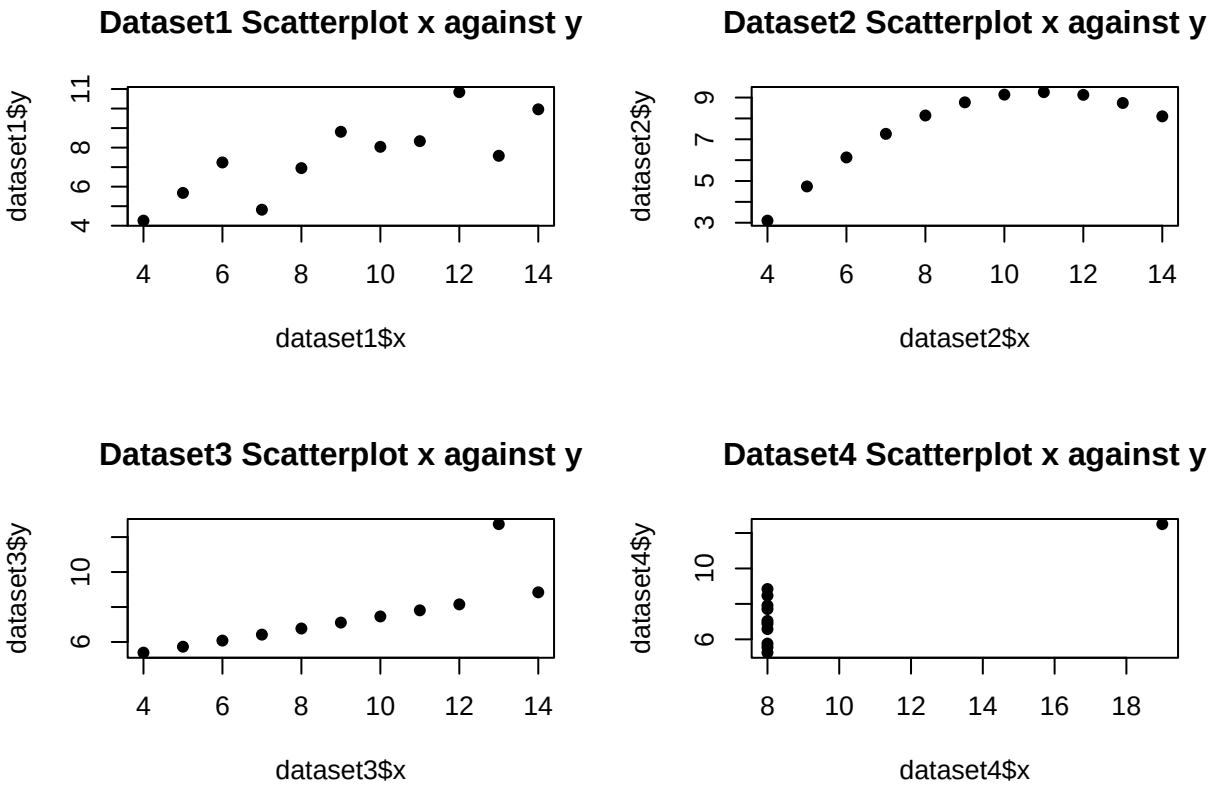
```
## [1] 0.00646109
```

P-value is 0.0064 so we have very strong evidence against the NULL hypothesis and so we reject it. Can then conclude that there is an association between disease and blood group in the ABO system.

When comparing the expected and observed counts it can be seen that most the observed data is approximately 3 to 5 counts different than the expected counts. Both in the greater or less than direction depending on variables. The only observation that seems to be very close to its expected value is the Minimal severity by B blood group observation. The observation with the largest difference between observed and expected values is the Not present severity by O blood group observation.

Question 3

```r
anscombe <- read.csv(file="C:/Users/wesch/uvic/stat359 Data Analysis/data1/anscombe.csv",
↪  header=TRUE)

#For part a)
# Separating the data
dataset1 <- data.frame(x = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
                       y = c(8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82,
                       ↪  5.68))
dataset2 <- data.frame(x = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
                       y = c(9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26,
                       ↪  4.74))
dataset3 <- data.frame(x = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
                       y = c(7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42,
                       ↪  5.73))
dataset4 <- data.frame(x = c(8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8),
                       y = c(6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91,
                       ↪  6.89))

par(mfrow=c(2,2))
plot(dataset1$x,dataset1$y,pch=16)
title("Dataset1 Scatterplot x against y")

plot(dataset2$x,dataset2$y,pch=16)
title("Dataset2 Scatterplot x against y")

plot(dataset3$x,dataset3$y,pch=16)
title("Dataset3 Scatterplot x against y")

plot(dataset4$x,dataset4$y,pch=16)
title("Dataset4 Scatterplot x against y")
```



In Dataset1 shows a roughly linear trend. So, a linear model would probably be appropriate.

In Dataset2 the relationship clearly follows a non-linear trend so a linear model is not appropriate.

In Dataset3 the data follows a very clear linear trend up to the outlier at (13, 12.74) and then seem to continue to be be linear afterwards. So a linear model should be appropriate.

In Dataset4 almost all data points have the same X value except for one outlier at X = 19. So, a linear model would not be suitable here.

For part b)

```
model1 <- lm(y ~ x, data = dataset1)
model2 <- lm(y ~ x, data = dataset2)
model3 <- lm(y ~ x, data = dataset3)
model4 <- lm(y ~ x, data = dataset4)

# Extracting R^2 values
r2_1 <- summary(model1)$r.squared
r2_2 <- summary(model2)$r.squared
r2_3 <- summary(model3)$r.squared
r2_4 <- summary(model4)$r.squared

r2_table <- data.frame(
  Dataset = c("Dataset1", "Dataset2", "Dataset3", "Dataset4"),
  R2 = c(r2_1, r2_2, r2_3, r2_4)
)
r2_table
```

```
##    Dataset        R2
```

```
## 1 Dataset1 0.6665425
## 2 Dataset2 0.6662420
## 3 Dataset3 0.6663240
## 4 Dataset4 0.6667073
```

It can be seen that the R-squared for each dataset is approximately 0.6665. Judging from the R-sqaured alone you would think that each model would be a good fit for the data but as we seen with our data plots above a linear model would not be suitable for datasets 2 and 4. These leads to the point that R-sqaured should not be the statistic that is used when determing if a model is a good fit for the data.

Question 4

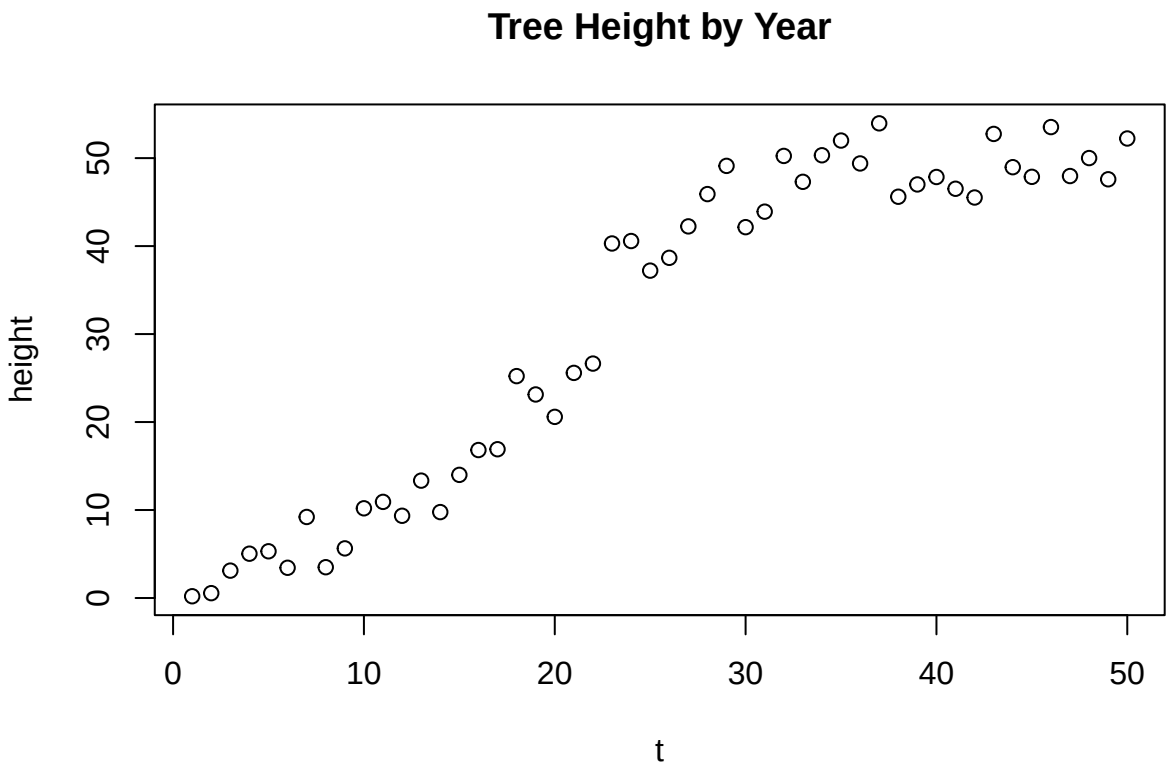
```
library(nls2)
```

```
## Warning: package 'nls2' was built under R version 4.3.3
```

```
## Loading required package: proto
```

```
## Warning: package 'proto' was built under R version 4.3.3
```

```
tree_growth <- read.table(file="C:/Users/wesch/uvic/stat359 Data
↪  Analysis/data1/growth.txt", header=TRUE)
attach(tree_growth)

plot(t,height)
title("Tree Height by Year")
```



Initial a-value and the b-values are computed below

```
#a value for all curve models
a <- max(height)
a
```

```
## [1] 53.95014
```

```
#So will say asymptote is at height = 54 so a = 54
a <- 54

#b value for curve (a)
b_a <- (52/min(height)) - 1
b_a
```

```
## [1] 259
```

```
#b value for curve (b)
b_b <- -log(min(height)/52)
b_b
```

```
## [1] 5.560682
```

```
#b value for curve (c). C does not get removed like in the other model equations when
↪  solving for b. assuming c for curve (c) is 1 to test if we get a convergence for
↪  these initial values.
b_c <- -log(1 - min(height) / a) / 1
b_c
```

```
## [1] 0.003710579
```

Initial c values is computed below

```
# c value for curve (a)
c_a <- (-1/mean(t))*log(((a/mean(height))-1)/b_a)
c_a
```

```
## [1] 0.2311196
```

```
# c value for curve (b)
c_b <- -(1/mean(t))*log((-log(mean(height)/a))/b_b)
c_b
```

```
## [1] 0.09152787
```

```
# c value for curve (c)
c_c <- 1
c_c
```

```
## [1] 1
```

Now fitting the models.

The Logistic model (a)

```
modela<-nls(height ~ a/(1+b*exp(-c_a*t)),start = list(a=a,b=b_a,c_a=c_a))
summary(modela)
```

```
##
## Formula: height ~ a/(1 + b * exp(-c_a * t))
##
## Parameters:
##     Estimate Std. Error t value Pr(>|t|)
## a    50.4208     0.8473  59.509  < 2e-16 ***
## b    47.1377    12.1605   3.876 0.000328 ***
## c_a   0.1993     0.0141  14.132  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.069 on 47 degrees of freedom
##
## Number of iterations to convergence: 8
## Achieved convergence tolerance: 2.382e-06
```

The Gompertz model (b)

```
modelb <- nls(height ~ a*exp(-b*exp(-c_b*t)), start = list(a=a, b=b_b, c_b=c_b))
summary(modelb)
```

```
##
## Formula: height ~ a * exp(-b * exp(-c_b * t))
##
## Parameters:
##     Estimate Std. Error t value Pr(>|t|)
## a   52.21791    1.33362  39.155  < 2e-16 ***
## b    7.57901    1.40071   5.411 2.07e-06 ***
## c_b  0.12491    0.01156  10.802 2.50e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.617 on 47 degrees of freedom
##
## Number of iterations to convergence: 9
## Achieved convergence tolerance: 9.801e-06
```

The Von Bertalanffy model

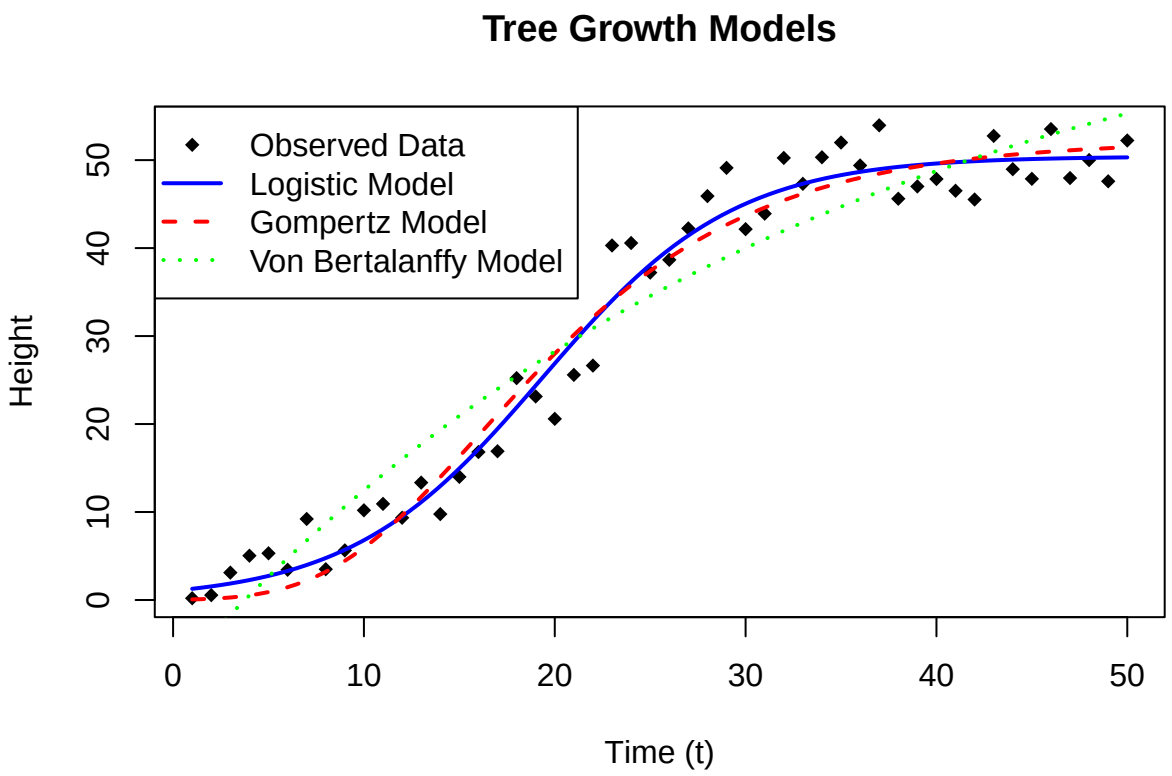
```
modelc <- nls(height ~ a-a*exp(-b*(t+c_c)), start = list(a=a, b=b_c, c_c=c_c))
summary(modelc)
```

```
##
## Formula: height ~ a - a * exp(-b * (t + c_c))
##
## Parameters:
##      Estimate Std. Error t value Pr(>|t|)
## a   74.412433   9.951159   7.478 1.55e-09 ***
## b    0.029390   0.006914   4.251    1e-04 ***
## c_c -3.760695   0.834255  -4.508 4.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.366 on 47 degrees of freedom
##
## Number of iterations to convergence: 15
## Achieved convergence tolerance: 6.298e-06
```

All models converge with our chosen and calculated initial starting values so they are good values.

Will make a figure of the estimated lines for each curve.

```r
plot(t, height, main = "Tree Growth Models", xlab = "Time (t)", ylab = "Height", pch =
↪  18, col = "black")

# Generating the estimated values for each model
t_seq <- seq(min(t), max(t), length.out = 100) # Sequence of time points for smooth
↪  curves

# Logistic model estimations
logistic_est <- predict(modela, newdata = data.frame(t = t_seq))

# Gompertz model estimations
gompertz_est <- predict(modelb, newdata = data.frame(t = t_seq))

# Von Bertalanffy model estimations
vonbertalanffy_est <- predict(modelc, newdata = data.frame(t = t_seq))

# Curves for the plot
# Logistic model
lines(t_seq, logistic_est, col = "blue", lwd = 2, lty = 1)
# Gompertz model
lines(t_seq, gompertz_est, col = "red", lwd = 2, lty = 2)
 # Von Bertalanffy model
lines(t_seq, vonbertalanffy_est, col = "green", lwd = 2, lty = 3)

# Legend
legend("topleft",
       legend = c("Observed Data", "Logistic Model", "Gompertz Model", "Von Bertalanffy
       ↪  Model"),
       col = c("black", "blue", "red", "green"),
       pch = c(18, NA, NA, NA),
       lty = c(NA, 1, 2, 3),
       lwd = c(NA, 2, 2, 2))
```



Computing the confidence interval for parameter a. Since the limit is describing the values for the expected maximum amount a tree is expected to grow for a time t.

```
# 95% confidence intervals for "a" in each model
# Logistic model (a)
a_a_CI<-c(50.4208-1.96*0.8473,50.4208+1.96*0.8473)
a_a_CI
```

```
## [1] 48.76009 52.08151
```

```
# Gompertz model (b)
a_b_CI<-c(52.21791-1.96*1.33362,52.21791+1.96*1.33362)
a_b_CI
```

```
## [1] 49.60401 54.83181
```

```
# Von Bertalanffy model (c)
a_c_CI<-c(74.412433-1.96*9.951159,74.412433+1.96*9.951159)
a_c_CI
```

```
## [1] 54.90816 93.91670
```

Based on a comparison of the models RSEs and observing the Tree Growth Models plot, the best model seems to be the Logistic model (a). This is because it has the lowest RSE and seems to fit the data the best based on the visual provided by the above plot.

Will Now calculate and plot an estamate of the derivative.

```r
#Finding the derivative
params <- coef(modela)
a <- params["a"]; b <- params["b"]; c <- params["c_a"]
derivative <- (a * b * c * exp(-c * t_seq)) / ((1 + b * exp(-c * t_seq))^2)

#The derivative plot
plot(t_seq, derivative, type = "l", col = "blue", lwd = 2,
     xlab = "Time (t)", ylab = "Rate of Growth (df(t)/dt)",
     main = paste("Growth Rate Over Time for", "Modela"))
abline(h = 0, lty = 2, col = "red")
```