506CourseWork

Table of contents

For the love of god don't forget

Do not display too much raw R output (e.g. don't display the full output of 'summary(model)'), but edit this down to the essentials. Ensure to include justification for each step of your analyses, providing comments alongside your R code to explain what you are doing and add appropriate titles and labelled axes to your plots.

Question 1

We have the model:

$Yi \sim N(\frac{\theta 1 xi}{\theta 2 + xi}, \sigma^2)$

Question 1 a)

Due to the visible non-linearity of the model, we would be required to significantly transform our data to get a linear model that would have an acceptable fit of the data. We can also see that the response data seems to be only positive while a normal distribution goes from]-∞,∞[. Such arbitrary transformation increases the complexity of the model, making it less interpretable and not respect the nature of the data.

Linear regression models are based on the assumption that the relationship between the independent and dependent variables is linear. If the relationship between the variables is non-linear, a linear regression model may not be appropriate to use. In such cases, transforming the data to make the relationship linear may not result in an accurate representation of the true relationship, and can lead to overfitting or underfitting. Additionally, transforming the data can result in a loss of interpretability of the results, as it can be difficult to understand the meaning of the transformed variables.

Another issue with using a linear regression model for non-linear data is that the residuals, which represent the difference between the observed and predicted values, may not be normally distributed, which is another assumption of linear regression models. This can lead to biased or incorrect results.

In conclusion, when the data is non-linear, a linear regression model may not be the best choice for modelling the relationship between the variables, and alternative methods need to be considered.

Question 1 b)

The Yi are independent so the likelihood is a product of the individual pdfs.

TODO GET BETTER WORDING FOR THE REPLACE PART

Likelihood of a normal distribution where $L(yi|\mu, \sigma^2) =$

$= \Pi_{i=1}^{n} fX(yi|\mu, \sigma^2)$

$= \Pi_{i=1}^{n}((2\pi\sigma^2)^{-\frac{1}{2}} * exp(-\frac{1}{2} * \frac{(yi-\mu)^2}{\sigma^2})) =$

$= (2\pi\sigma^2)^{-\frac{n}{2}} * exp(-\frac{1}{2\sigma^2} * \sum_{i=1}^{n}(yi-\mu)^2)$

Replacing the μ with the respective θs and n, we have the likelihood as:

$L(\beta 0, \beta 1, \sigma^2; x, y) =$

$= \prod_{i=1}^{n} p(\beta 0, \beta 1, \sigma^2; x, y) =$

$= (2\pi\sigma^2)^{-\frac{100}{2}} * exp(-\frac{1}{2\sigma^2} * \sum_{i=1}^{100}(yi - \frac{\theta 1 xi}{\theta 2 + xi})^2)$

The log-likelihood of a normal distribution is:

$l(yi|\mu, \sigma^2) =$

$= ln(L(yi|\mu, \sigma^2)) =$

$= ln((2\pi\sigma^2)^{-\frac{n}{2}} * exp(-\frac{1}{2\sigma^2} * \sum_{i=1}^{n}(yi - \mu)^2)) =$

$= ln((2\pi\sigma^2)^{-\frac{n}{2}}) + ln(exp(-\frac{1}{2\sigma^2} * \sum_{i=1}^{n}(yi - \mu)^2)) =$

$= -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{n}(yi - \mu)^2) =$

$= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{n}(yi - \mu)^2)$

Once again replacing the μ with the respective θs and n, we have the log-likelihood as:

$l(\beta 0, \beta 1, \sigma^2; x, y) =$

$= -\frac{100}{2}ln(2\pi) - \frac{100}{2}ln(\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{100}(yi - \frac{\theta 1 xi}{\theta 2 + xi})^2) =$

$= -50ln(2\pi) - 50ln(\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{100}(yi - \frac{\theta 1 xi}{\theta 2 + xi})^2)$

Question 1 c)

```
n = nrow(nlmodel)


### Create a function to evaluate minus the log-likelihood
myLike = function(variables) {

    theta1 = variables[1]  #theta1
    theta2 = variables[2]  #theta2
    sigma = variables[3]  #sigma

    mu = ((theta1 * nlmodel$x)/(theta2 + nlmodel$x))

    # Log-likelihood
    result = (-(n/2) * log(2 * pi)) - ((n/2) * log(sigma^2)) - (1/(2 * (sigma^2))) *
        (sum((nlmodel$y - mu)^2))

    # Returning negative log-likelihood
    return(-result)
}
```

Question 1 d)

From the graph data we can clearly see that the deviantion is approximatly 15 from the value scatter which is easier to destinguish and measure between the $[0.5, 1]$ interval. We can observethat when x -> 1 y is approximately 210 and as x -> 0 y is approximately to 50

To determine the thetas we will first see see that when x approximates to zero we can observe that:

$\lim_{x->0} \frac{\theta 1 xi}{\theta 2 + xi} \implies \frac{1}{\theta 2}$

As such $ Y$ ~ $N(\frac{\theta 1 xi}{\theta 2 + xi})$ becomes $ 50$ ~ $\frac{1}{\theta 2}$

Solving it for $\theta 2$ we get $\theta 2 = 1/50 \iff \theta 2 = 0.02$

Now that we have theta2 we can use the approximation of x to 1 to determine the value of $\theta 1$

$\lim_{x->1} \frac{\theta 1 xi}{\theta 2 + xi} \implies \frac{\theta 1}{\theta 2 + x}$

As such $ Y$ ~ $N(\frac{\theta 1 xi}{\theta 2 + xi})$ becomes $215 \sim \frac{\theta 1 * 1}{\theta 2 + 1}$

Solving it for $\theta 1$ we get $\theta 1 = 215/1.02 \iff \theta 1 \approx 210.7$

```
# Estimating the MLE
out <- nlm(myLike,
  p = c(210.7,0.02,15), #plugging in the starting values
  hessian = T,
  iterlim = 10000,
  steptol = 1e-10)

# Reporting estimates
variableEstimates = out$estimate
out$estimate
```

```
[1] 214.65008415   0.06353447  13.61564428
```

Question 1 e)

```
# Invert the negated Hessian to obtain the Observed Information Matrix
OIM <- solve(out$hessian)

# The diagonal entries are the variances of beta0 and beta1
# respectively so # obtain them
```

```r
VarianceBeta <- diag(OIM)

# and then square root them to obtain standard errors
stand_error <- sqrt(VarianceBeta)

# reporting standard errors
stand_error
```

```
[1] 2.674798031 0.005140379 0.963024267
```

The formula to calculate a 99% confidence interval is: $\pm$ 2.576 * SE( )

```r
# Estimating CIs
CIs <- cbind(variableEstimates - 2.576 * stand_error, variableEstimates +
    2.576 * stand_error)

# Reporting the CIs
CIs
```

```
             [,1]         [,2]
[1,] 207.75980442 221.54036388
[2,]   0.05029285   0.07677609
[3,]  11.13489377  16.09639479
```

Question 1 f)

H0 : $\beta_2 = 0{,}08$ vs. H1 : $\beta_2 \neq 0{,}08$

```r
## Hypothesis thesis without using confidence interval

z_stat <- (variableEstimates[2] - 0.08)/stand_error[2]

# Print the test values
z_stat  ## significance tests
```

```
[1] -3.203174
```

So now we need to decide if this value of the z-statistic is extreme at the 10% significance level

```r
### Note that equivalently we can look at the 95% quantile of N(0,1)
qnorm(0.95, 0, 1)
```

```
[1] 1.644854
```

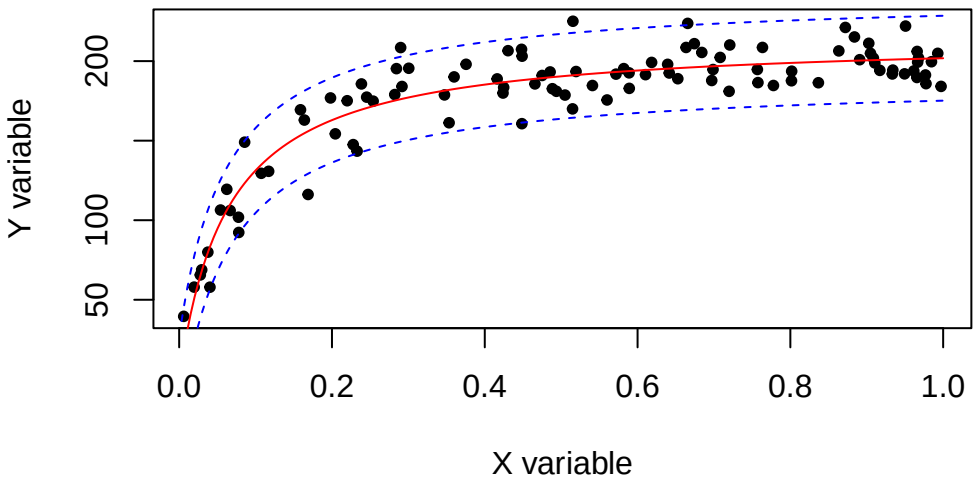
```r
qnorm(0.05, 0, 1)
```

```
[1] -1.644854
```

As we can see the value from the z-statistician test is considerably lower than -1,645, meaning that it is an extreme value and therefore rejecting the null hypothesis that $\theta 2$ is 0,08.

Question 1 g)

```r
# Plotting initial starting guess
xx <- seq(0, 1, len = 200)

# Estimating mean relationship (mean mu = )
mu <- (out$estimate[1] * xx)/(out$estimate[2] + xx)

# Getting standard Deviation
standardDeviation <- out$estimate[3]


# Getting 95% interval from the quantiles of a Normal distribution
plot(nlmodel$x, nlmodel$y, pch = 20, xlab = "X variable", ylab = "Y variable")

lines(xx, qnorm(0.025, mean = mu, sd = standardDeviation), col = "blue",
    lty = "dashed")
lines(xx, qnorm(0.975, mean = mu, sd = standardDeviation), col = "blue",
    lty = "dashed")

lines(xx, qnorm(0.5, mean = mu, sd = standardDeviation), col = "red")
```



From the estimations produced through our model we can see that the current model with a 95% prediction fits the data quite nicely, having only 4 of the 100 observations shortly out of the 95% prediction interval.

However it should be noted that even though the model has a good performance, a normal distribution has the assumption that data can take any value in the real line however this data is bounded between the [0,1] interval.

Considering that the variance increases trough the model it further indicates that a normal distribution should be switched for another distribution that better respects the nature of our data.

Question 2

Model 1:

$Yi \sim Pois(\lambda i)$

$log(\lambda i) = \beta 0 + \beta 1 xi$

Model 2:

$Yi \sim N(\mu i, \sigma^2)$

$log(\mu i) = \gamma 0 + \gamma 1 xi$

Question 2 a)

As we can from the graph and what we can determine from the nature of the data represented in such graph the recorded number of AIDS cases is a count variable and the counts are non-negative integers.

The first model, a Poisson distribution, would be a more appropriate choice. The Poisson distribution is a discrete distribution that models count data which respects the nature of the data

The second model, a Normal distribution, would not be the best fit since its range is from]-∞,∞[and expects continuous values, not respecting the nature of the data.

The log-link function in both models ensures that the predicted values are always positive. which makes sense for poisson and not for normal //TODO redo this pls

Question 2 b)

The Yi are independent so the likelihood is a product of the individual pdfs.

$L(\theta 1, theta2, \sigma^2; y, x)$

```r
# Fitting in R

model2 = glm(cases ~ date, data = aids, family = poisson(link = "log"))

# Summarise the model
summary(model2)
```

```
Call:
glm(formula = cases ~ date, family = poisson(link = "log"), data = aids)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-7.768  -4.042  -0.335   3.048   7.281

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -16.875879   0.350353  -48.17   <2e-16 ***
date          0.247169   0.003856   64.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

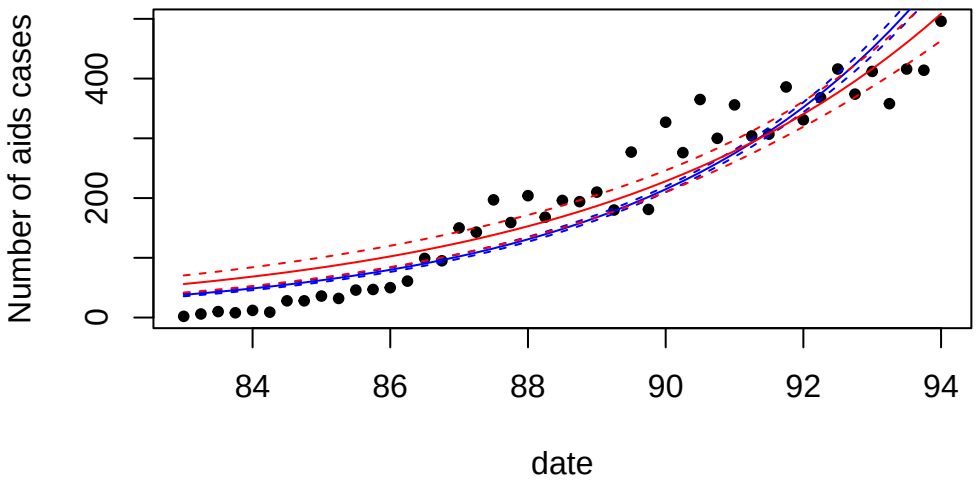
```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5738.16  on 44  degrees of freedom
Residual deviance:  854.02  on 43  degrees of freedom
AIC: 1153.9

Number of Fisher Scoring iterations: 5
```

```r
# Fitting model 1
model1 = glm(cases ~ date, data = aids, family = gaussian(link = "log"))

# Summarise the model
summary(model1)
```

```
Call:
glm(formula = cases ~ date, family = gaussian(link = "log"),
    data = aids)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-79.54  -50.35  -12.50   24.94  112.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.6047     1.2663  -9.954 9.94e-13 ***
date          0.2004     0.0138  14.523  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 2448.904)

    Null deviance: 974004  on 44  degrees of freedom
Residual deviance: 105306  on 43  degrees of freedom
AIC: 482.81

Number of Fisher Scoring iterations: 7
```

```r
## We can use this to obtain 95% confidence intervals on our estimated
## relationship
xx <- seq(83, 94, len = 45)
```

```r
## predict Poisson
predPoisson = predict(model2, newdata = aids, type = "response", se.fit = T)


## Mean and Confidence Intervals estimation
muPoisson = predPoisson$fit
muPoisson_upper = (predPoisson$fit + qnorm(1 - 1.96/2) * predPoisson$se.fit)  ##exp(p$fit+
muPoisson_lower = (predPoisson$fit - qnorm(1 - 1.96/2) * predPoisson$se.fit)  ## this is t


## predict normal
predNormal = predict(model1, newdata = aids, type = "response", se.fit = T)

## Mean and Confidence Intervals estimation
muNormal = predNormal$fit
muNormal_upper = predNormal$fit + qnorm(1 - 1.96/2) * predNormal$se.fit  ## p$fit+qnorm(1-
muNormal_lower = predNormal$fit - qnorm(1 - 1.96/2) * predNormal$se.fit  ## p$fit-qnorm(1-

## plot the data

plot(aids$date, aids$cases, pch = 20, xlab = "date", ylab = "Number of aids cases")

lines(xx, muPoisson, col = "blue")
lines(xx, muPoisson_upper, col = "blue", lty = "dashed")
lines(xx, muPoisson_lower, col = "blue", lty = "dashed")

lines(xx, muNormal, col = "red")
lines(xx, muNormal_upper, col = "red", lty = "dashed")
lines(xx, muNormal_lower, col = "red", lty = "dashed")
```



As we can see from the plot alone, the linear model fits the data better than the poisson model. The poisson model is clearly not a good model since the big majority of the model is either overestimating or underestimating the data, which is a clear indicator that this model is inadequate. The Normal distribution, seems to fit somewhat the data due to the data although non-linear not deviating too much from a line but it is still clearly that it is not an adequate model since there is too many points that are under or overestimated, making it also a not very adequate model.

Question 2 c)

```
# Model comparison aka IAC time

# the formula to for the AIC: -2l + 2p

AIC(model1)
```

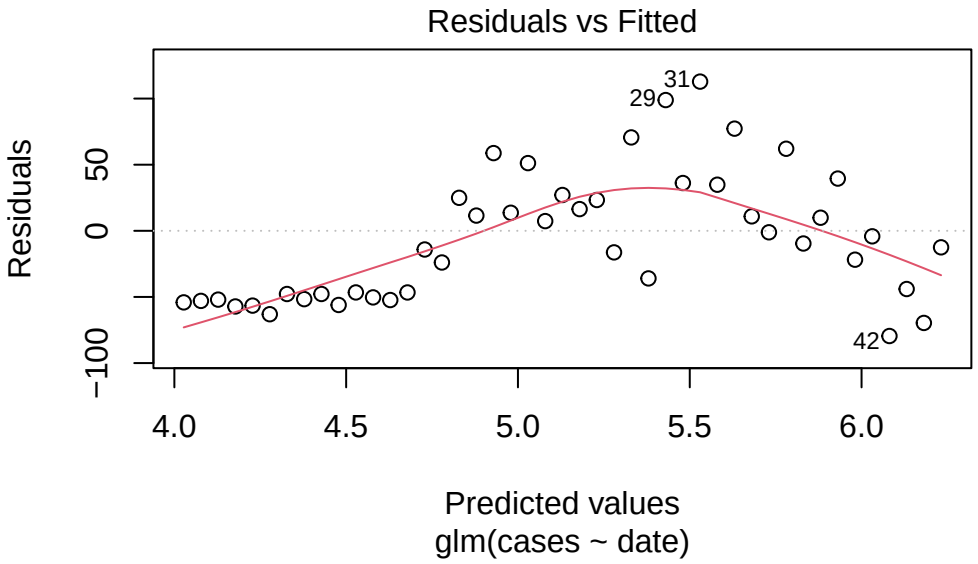
```
[1] 482.8128
```

```
AIC(model2)
```

```
[1] 1153.873
```

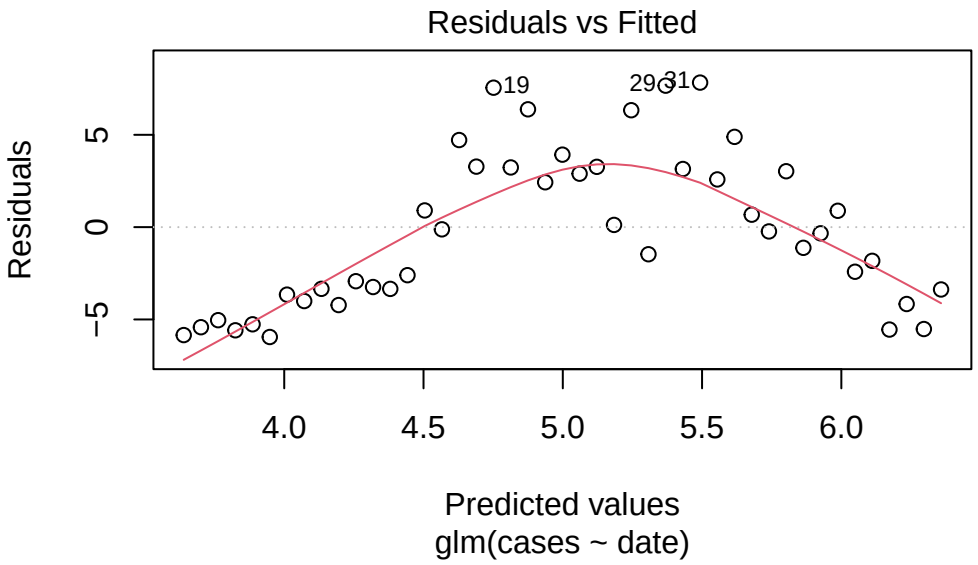
As we can see from the Akaike Information Criterion, model 1 as a much lower AIC, meaning model 1 has a much better fit than model 2 since when comparing AICs, the model with the lower values has the better fit ## Question 2 d)

```
plot(model1, 1)
```



As we can see from the Residual bs fitted model, there seems to be pattern in how the model overfits and underfits the data in a way tha would require more flexibility from a multi polynomial equation.

```
plot(model2, 1)
```



The Poisson model residuals seem to follow a quadratic function, // does this mean it required a quadratic term?