MATH 5900 – ADVANCED DATA ANALYTICS

ASSIGNMENT 3

TAIWO JEGEDE – E00755963

APPLIED DATA SCIENCE

SOFTWARE USED: R

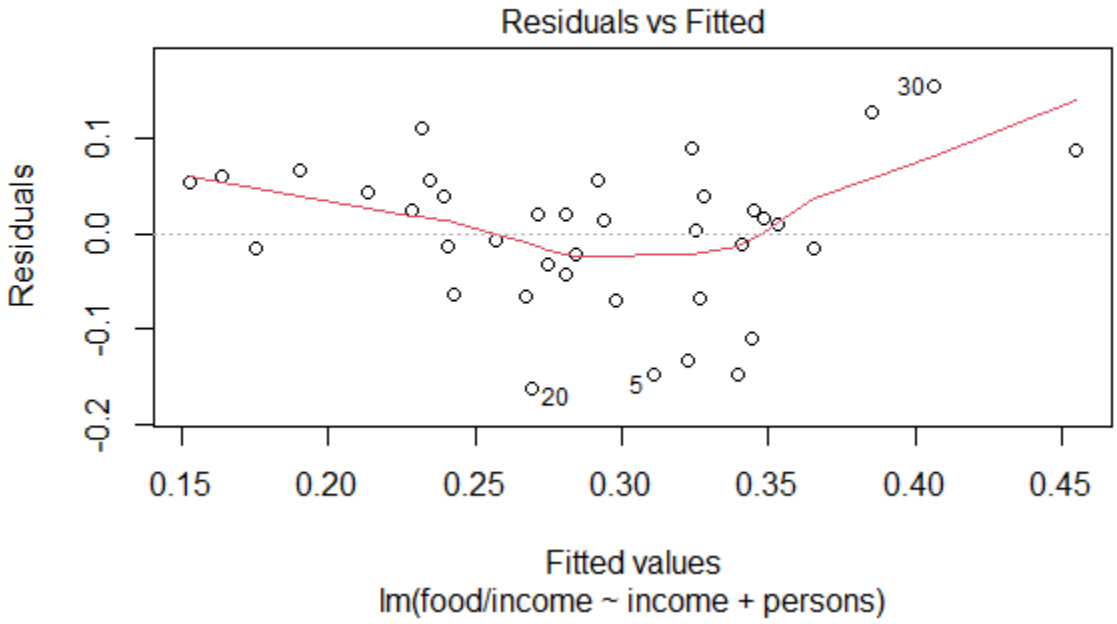
QUESTION 1

i. An appropriate linear model:

lm_model <- lm(food/income ~ income + persons, data = FoodExpenditure

ii. Assumptions and fitness of the model:

Linearity:



Here, we assume that the relationship between the predictors (income and persons) and the response variable (food/income) is linear. So the change in income and household size have a constant effect on the proportion of food expenditure to income. From our plot, we can see that the residuals are scattered around zero.

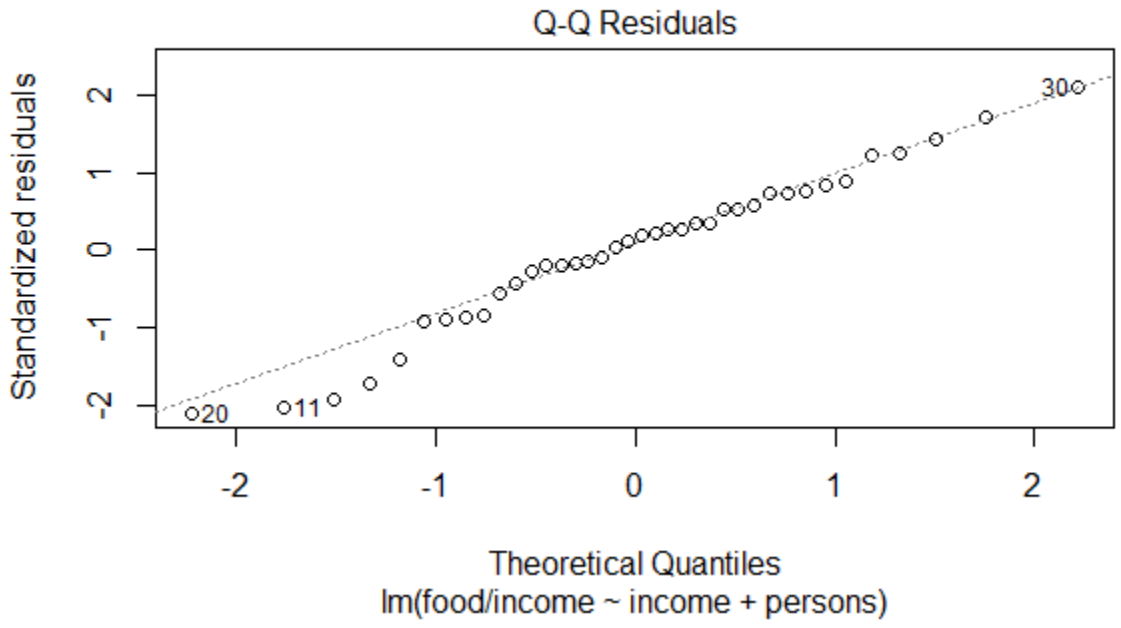
Independence: We assume that the responses are independent of each other

Non-constant variance:

```
> ncvTest(lm_model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.114319, Df = 1, p = 0.077607
>
```

I used the Non Constant variance test to check if the residual is contant accross all the levels of predictors. As we can see the Pvalue is greater than 0.05, which means that we do not reject the null hypothesis.

Normality:




```
        Shapiro-Wilk normality test

data:  residuals(lm_model)
W = 0.97101, p-value = 0.4197
```

```
> ad.test(lm_model$residuals)

        Anderson-Darling normality test

data:  lm_model$residuals
A = 0.40212, p-value = 0.3422
```

Using the Q-Q Plot, the Shapiro-Wilk test and the Anderson-Darling normality test, the normality of residuals implies that the errors follow a normal distribution.

iii. Beta regression model:

foodPer = food/income

beta_model <- betareg(foodPer ~ income + persons, link="logit")

```
Call:
betareg(formula = foodPer ~ income + persons, link = "logit")

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.7818 -0.4445  0.2024  0.6852  1.8755

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.622548   0.223854  -2.781 0.005418 **
income      -0.012299   0.003036  -4.052 5.09e-05 ***
persons      0.118462   0.035341   3.352 0.000802 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)    35.61       8.08   4.407 1.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 45.33 on 4 Df
Pseudo R-squared: 0.3878
Number of iterations: 28 (BFGS) + 4 (Fisher scoring)
```

iv. Because the response variable in this instance is a proportion between 0 and 1, which makes it inappropriate for logistic regression, logistic regression using a "events/trials" format is normally used for binary response variables, which is why the data cannot be modelled using this method.

v. Interpretation of coefficients:

The intercept coefficient (-0.622548) = the expected log odds of the proportion of food expenditure to income when both income and household size are zero.

The coefficient for income (-0.012299) = the change in the log odds of the proportion of food expenditure to income for a one-unit increase in income, holding the number of persons in the household constant.

The coefficient for persons (0.118462) = the change in the log odds of the proportion of food expenditure to income for a one-unit increase in the number of persons in the household, holding income constant.

Impact of household income and household size on the rate of food expenditures:

The coefficient for income is negative (-0.012299), indicating that as household income increases, the rate of food expenditures relative to income tends to decrease.

The coefficient for household size (persons) is positive (0.118462), This implies that larger households tend to allocate a higher proportion of their income to food expenditures.

QUESTION 2

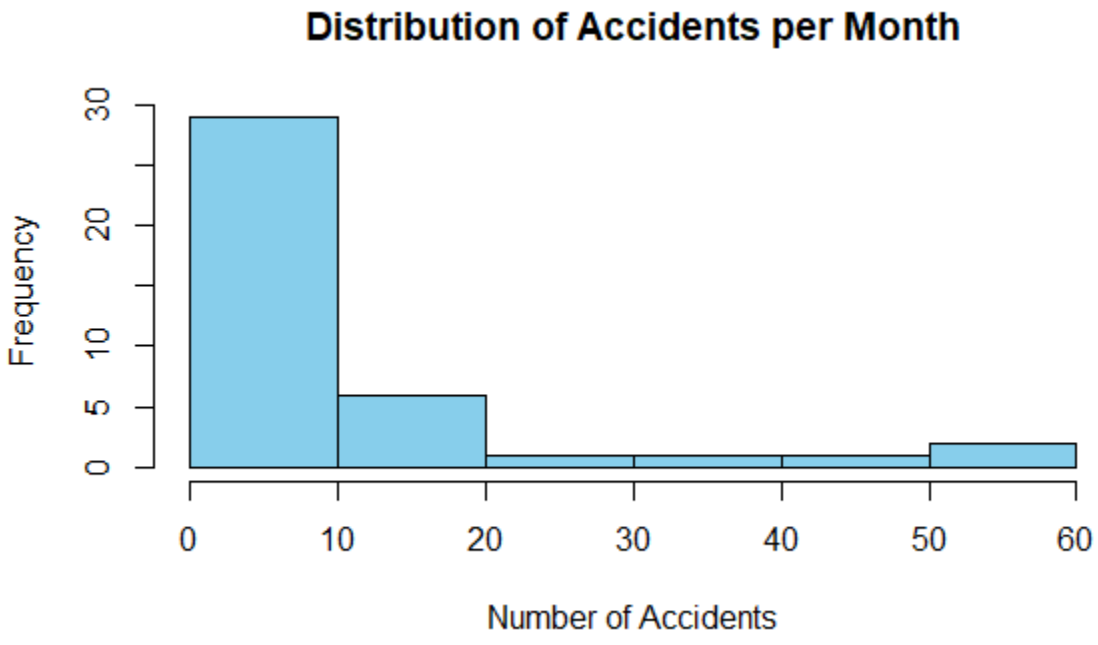
i. Providing descriptive statistics:

Using the summary function:

```
> summary(ShipData)
   accidents      operational      construction1      construction2      construction3
 Min.   : 0.0   Min.   :0.0   Min.   :0.00   Min.   :0.00   Min.   :0.00
 1st Qu.: 0.0   1st Qu.:0.0   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00
 Median : 2.0   Median :0.5   Median :0.00   Median :0.00   Median :0.00
 Mean   : 8.9   Mean   :0.5   Mean   :0.25   Mean   :0.25   Mean   :0.25
 3rd Qu.:11.0   3rd Qu.:1.0   3rd Qu.:0.25   3rd Qu.:0.25   3rd Qu.:0.25
 Max.   :58.0   Max.   :1.0   Max.   :1.00   Max.   :1.00   Max.   :1.00

    exposure       service_months
 Min.   : 3.807   Min.   :    0.0
 1st Qu.: 5.911   1st Qu.:  175.8
 Median : 6.999   Median :  782.0
 Mean   : 7.049   Mean   : 4088.6
 3rd Qu.: 7.707   3rd Qu.: 2078.5
 Max.   :10.712   Max.   :44882.0
 NA's   :6
```

Using an histogram:



Using the "skewness"

```
> skewness(ShipData$accidents)
[1] 2.021616
```

The skewness value of 2.021616 indicates that the distribution of accident counts per month is positively skewed, or right-skewed.

ii. The outcome variable, which is the monthly number of accidents, is discrete by nature and represents counts of events. But for count data, linear regression is inappropriate since it implies a continuous outcome variable. When applying a linear regression model to count data, it is possible to get projected values that are not positive integers or even negative, which is not relevant in this situation.

iii. Fitting a Poisson Count Regression model:

```
Call:
glm(formula = accidents ~ exposure + construction1 + construction2 +
    construction3, family = "poisson", data = ShipData)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.40041    0.40599 -10.839  < 2e-16 ***
exposure       0.82247    0.04539  18.121  < 2e-16 ***
construction1 -0.61076    0.22555  -2.708  0.00677 **
construction2  0.10250    0.20013   0.512  0.60851
construction3  0.28907    0.19294   1.498  0.13406
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 614.539  on 33  degrees of freedom
Residual deviance:  59.681  on 29  degrees of freedom
  (6 observations deleted due to missingness)
AIC: 167.55
```

iv. Assessing Overdispersion:

```
> residual_deviance <- poisson_model$deviance
> df <- poisson_model$df.residual
> overdispersion <- residual_deviance / df
> overdispersion
[1] 2.057954
```

An overdispersion value of 2.0 indicates that the observed variability in the number of accidents per month exceeds what would be expected under the Poisson distribution assumption, suggesting that the Poisson model may not adequately capture the data's variability, thus requiring consideration of alternative models like Negative Binomial regression.

v. Fitting a Negative Binomial Count Regression model

```
Call:
glm.nb(formula = accidents ~ exposure + construction1 + construction2 +
    construction3, data = ShipData, init.theta = 11.92122411,
    link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.64545    0.55211  -8.414   <2e-16 ***
exposure       0.85120    0.06343  13.420   <2e-16 ***
construction1 -0.67189    0.33494  -2.006   0.0449 *
construction2  0.10562    0.28753   0.367   0.7134
construction3  0.34072    0.27190   1.253   0.2102
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(11.9212) family taken to be 1)

    Null deviance: 328.320  on 33  degrees of freedom
Residual deviance:  40.355  on 29  degrees of freedom
  (6 observations deleted due to missingness)
AIC: 166.21

Number of Fisher Scoring iterations: 1


              Theta:  11.92
          Std. Err.:  9.92

 2 x log-likelihood:  -154.205
> |
```

From the result, when we divide the Residual deviance by the degree of freedom, we get 1.39. Also, the dispersion parameter = 11.92 and since this value is close to 1, it indicates that the negative binomial model adequately accounts for overdispersion in the data.

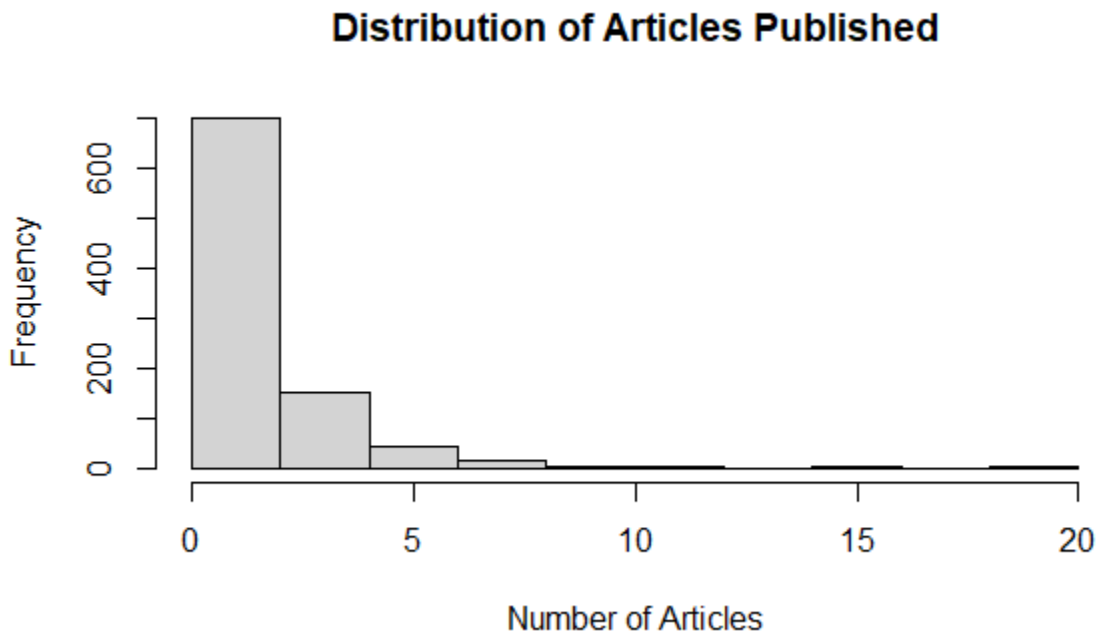
vi. Interpreting the coefficients and its impact on health-related accidents.

$\log(\mu) = -4.64545 + 0.85120 \times \text{exposure} - 0.67189 \times \text{construction1} + 0.10562 \times \text{construction2} + 0.34072 \times \text{construction3}$

From my observation, the exposure level and the construction era (specifically, construction era 1) significantly impact the expected number of health-related accidents per month on naval ships. Specifically, higher exposure levels are associated with increased accident counts, while ships from construction era 1 tend to have fewer accidents compared to ships from other eras. However, the construction eras 2 and 3 do not show a statistically significant impact on accident rates.

QUESTION 3

1. **Description of the data being analyzed**



From the histogram, we can observe more zeros (Excess Zeros). When we also use the "table" function, we get

```
> table(bioChemists$art)

  0   1   2   3   4   5   6   7   8   9  10  11  12  16  19
275 246 178  84  67  27  17  12   1   2   1   1   2   1   1
```

We notice that about 275 observations are Zeros.

We can also find the mean and variance:

```
> var(bioChemists$art)
[1] 3.709742
> mean(bioChemists$art)
[1] 1.692896
```

From the above result, we can see that the variance is greater than the mean.

Normally, when we notice these outputs, we would generally pursue Zero-Inflated or Hurdle models. I will be using the Hurdle model.

2. **Proposing an appropriate model**

Fit Poisson regression model

hurdlePoissonReg = hurdle(art ~ fem + mar + phd, data = bioChemists, dist="poisson")

Fit Negative Binomial regression model

hurdlePoissonReg = hurdle(art ~ fem + mar + phd, data = bioChemists, dist="negbin")

```
Call:
hurdle(formula = art ~ fem + mar + phd, data = bioChemists, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.9925 -0.8685 -0.2920  0.4149  9.0792

Count model coefficients (truncated negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.35357    0.20806   1.699  0.08926 .
femWomen    -0.27769    0.10004  -2.776  0.00551 **
marMarried   0.03519    0.10634   0.331  0.74071
phd          0.05984    0.04954   1.208  0.22713
Log(theta)   0.32998    0.22767   1.449  0.14723
Zero hurdle model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.33380    0.28494   1.171   0.2414
femWomen    -0.22610    0.15013  -1.506   0.1320
marMarried   0.11204    0.15791   0.709   0.4780
phd          0.17732    0.07409   2.393   0.0167 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 1.3909
Number of iterations in BFGS optimization: 12
Log-likelihood: -1599 on 9 Df
```

3. **The interpretations:**

Observing all the predictors, from if they have a phd or not, to whether they are married or not seem to either have an impact or not.

The phd variable, has a significant impact on both the likelihood of publishing articles and the expected publication volume for those who actually publish.

Being a female also seem to have an impact on the likelihood of publishing articles, with women having a lower expected count of articles published and lower odds of observing excess zeros but then it is not statistically significant.

Marital status does not have a significant impact on both the likelihood of publishing articles or the expected publication volume for those who do publish.

QUESTION 4

i. Beta regression model vs Logistic Regression model

Beta Regression: Here the coefficients represent the change in the log odds of the response variable for a one-unit change in the predictor variables, holding other predictors constant. The interpretation focuses on the effect of predictors on the log odds of the proportion or rate.

Logistic Regression: In logistic regression, the coefficients represent the change in the log odds of the event occurring for a one-unit change in the predictor variables, holding other predictors constant. The interpretation focuses on the effect of predictors on the log odds of the binary outcome.

When the interpretations of coefficients would be similar

The interpretations of coefficients in the beta regression model may resemble those in logistic regression if the response variable is changed to reflect a binary outcome and the logit link function is applied.

ii. Why count regression model

Because count regression models can be applied to time series data, unusual events, overdispersion, count or frequency variables, non-negative integers, and large number of levels

in categorical outcomes. When compared to conventional normal regression models, they offer more precise parameter estimates and better capture the features of count data.

iii. Overdispersion happens when the observed variance is greater than the variance expected by the fitted model. When there is overdispersion, it affect the count regression model by leading to underestimated standard errors, and potentially biased estimates.

Detecting Overdispersion: Using an overdispersion test, we may identify overdispersion. For example, if the p-value is less than 0.05, we can conclude that overdispersion is present, reject the null hypothesis, and state that the variance does not equal the mean.

What to do when we find meaningful overdispersion: When we find overdispersion, we should apply the Negative binomial regression model.

iv. Poisson Count Models vs. Hurdle Count Models

When we are using a Poisson count model, the response variable is assumed to follow a Poisson distribution, and the mean and variance are equal. But when we are using Hurdle count models, we use it when we have data with excess zero, we sometimes call it Advanced Counts - Excess Zeros.

So we can say that while Poisson count model focuses on modelling the rate of occurrences of an event, what hurdle count model does is to deal with data with excess zeros by separately modelling the probability of zero counts and the count distribution for non-zero counts