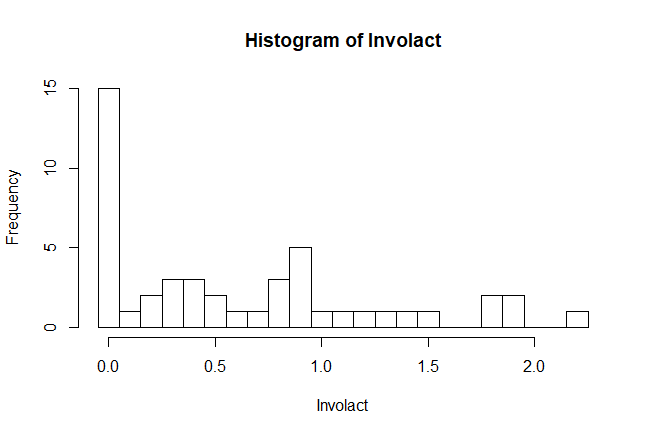
**STAT34700-HW2**

**Problem 3**

**(a)**

In the *chredlin* data, the value of *involact* distributes on the range (0,2.25). If we want to illustrate the number of zero values, we should plot a histogram of the distribution of *involact* with width 0.1 and range (-0.05, 2,25).



There are total 47 samples and 15 of them have zero *involact.* Hence, 15/47 of responses is zero.

**(b)**

Fit a Gaussian linear model with *involact* as the response with the other five variables: *race, fire, theft, age, log(income)* as predictors.

> fit\_3b = lm(involact ~ race + fire + theft + age + log(income), data=chredlin)

**The Gaussian linear model formula:**

**Output:**

> sumary(fit\_3b)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.1855396 1.1002549 -1.0775 0.2875500

race 0.0095022 0.0024896 3.8168 0.0004485

fire 0.0398560 0.0087661 4.5466 4.758e-05

theft -0.0102945 0.0028179 -3.6533 0.0007276

age 0.0083356 0.0027440 3.0377 0.0041345

log(income) 0.3457615 0.4001234 0.8641 0.3925401

n = 47, p = 6, Residual SE = 0.33453, R-Squared = 0.75

From the summary of this Gaussian linear model fit, the corresponding p-values of coefficients

for predictors *race, fire, theft* and *age* are smaller than the significance level . As a

result, we can reject the null hypothesis that at significance

level .The corresponding p-value of coefficient for predictor *log(income)* is greater thanthe significance level . As a result, we cannot reject the null hypothesis that .

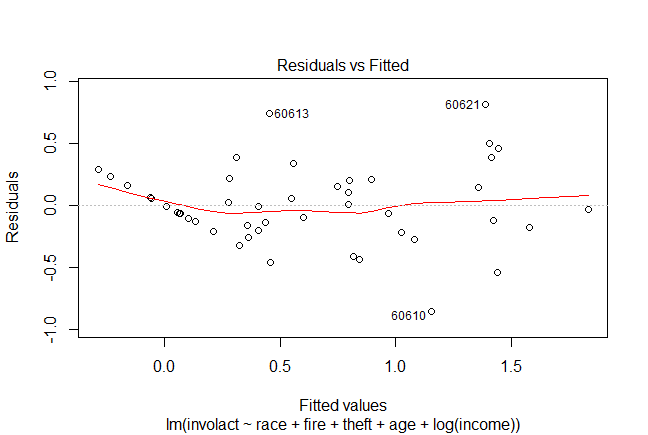
**Conclusion:**

Therefore, we can conclude that the predictors *race, fire, theft* and *age* are significant to the

response and the predictor *log(income)* is not significant to the response in this model.

**(c)**

Plot the residuals against the fitted values:



**How are the zero response values manifested on the plot?**

From the plot we can see that the residuals of zero response values are linearly distributed on the residual plot with a slope = -1. When the observed response is zero, the residual is ( is the fitted value). As a result, the residuals of zero response values and the fitted values have a linear relationship with negative correlation.

**What impact do these cases have on the interpretation of the plot?**

The residuals of zero response values are linearly distributed on the residual plot. When we make interpretation of the residual plot, we can say that there is a linear trend in it, which violates the assumption of Gaussian linear model. If the fit does not violate the constant variance assumption, with these zero response values, we may get a wrong diagnostic. As a result, the zero response values may affect the interpretation of residual plot.

**(d)**

Create a binary response variable *binary* in the data. If the value of *involact* is 0, then the value of *binary* is 0. If the values of *involact* is not 0, then the value of *binary* is 1.

**Fit a binary GLM:**

We fit a binary GLM with *binary* as the response, *race, fire, theft, age* and *log(income)* as five numerical predictors. In this generalized linear model, we use logit link function and binomial distribution family.

> fit\_3d = glm(binary ~ race + fire + theft + age + log(income), family=binomial, data=chredlin)

Warning messages:

1: glm.fit: algorithm did not converge

2: glm.fit: fitted probabilities numerically 0 or 1 occurred

**The GLM formula:**

**Output:**

> sumary(fit\_3d)

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.2262e+04 1.0590e+06 0.0116 0.9908

race 3.0422e+01 6.2072e+03 0.0049 0.9961

fire -8.0740e+01 1.9314e+04 -0.0042 0.9967

theft 1.6557e+01 5.2572e+03 0.0031 0.9975

age 2.1108e+00 4.9427e+03 0.0004 0.9997

log(income) -5.1513e+03 3.8595e+05 -0.0133 0.9894

n = 47 p = 6

Deviance = 0.00000 Null Deviance = 58.86527 (Difference = 58.86527)

**What problem occurred during this fit?**

From the summary of this GLM fit, we find that z-values corresponding to all coefficients are extremely small and p-values are all very large compared to the significance level .

The main reason of this problem is that the data is sparse. In the binary responses of GLM, if there are unbalanced 0 and 1, we can say that the data here is sparse. With sparse data, the standard errors can be overestimated and so the z-values too small and the significance of an effect could be missed. That’s why this problem occurs in the fitted model.

In addition, when we fit the GLM, there are two warnings: algorithm did not converge; fitted probabilities numerically 0 or 1 occurred. The main reason of these warnings is that the model has a perfect fit to the data here. The residual deviance is zero indicating a perfect fit but none of the predictors are significant due to the high standard errors. If we plot the data in any two dimensions, we can see that the two groups with different response values are linearly separable so that a perfect fit is possible. As a result, there will be several perfect fitted lines then the algorithm did not converge. If we can fi the model perfectly, all the fitted value will be 0 or 1.

**(e)**

**Fit a binary GLM:**

We fit a binary GLM with *binary* as the response, *race* and *age* as two numerical predictors. In this generalized linear model, we use logit link function and binomial distribution family.

> fit\_3e = glm(binary ~ race + age, family=binomial, data=chredlin)

**The GLM formula:**

**Output:**

> sumary(fit\_3e)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -13.09746 7.44557 -1.7591 0.07856

race 0.32539 0.16601 1.9600 0.04999

age 0.14675 0.08794 1.6688 0.09517

n = 47 p = 3

Deviance = 9.22863 Null Deviance = 58.86527 (Difference = 49.63665)

**Interpret the z-statistics:**

The z-value corresponding to the coefficient of *race* is 1.9600 and the corresponding p-value is 0.04999, which is smaller than the significance level . Then we can reject the null hypothesis that and conclude that the predictor *race* is significant to the response in this model.

The z-value corresponding to the coefficient of *age* is 1.9600 and the corresponding p-value is 0.09517, which is greater than the significance level . Then we cannot reject the null hypothesis that and conclude that the predictor *age* is not significant to the response in this model.

**Test the significance of the two predictors using the difference-in-deviances test:**

**1) To test the significance of the *race***, we can use the deviance to compare two models: model 1 with *race* and model 2 without *race* (only with *age*). The model 1 here is just the full model above.

The likelihood test statistic is Deviance of model 2 – Deviance of model 1, which asymptotically Chi-squared distributed with df = 1 (assuming that the model 2 is correct and the distributional assumptions hold).

**Fit model 2 (only with predictor *age*):**

**Output:**

> sumary(fit\_3ea)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.645516 1.177006 -2.2477 0.024598

age 0.059040 0.019585 3.0145 0.002574

n = 47 p = 2

Deviance = 45.40762 Null Deviance = 58.86527 (Difference = 13.45765)

**Deviance Test Output:**

> pchisq(45.408-9.2286,1,lower=FALSE)

[1] 1.799629e-09

**Conclusion:**

The null hypothesis is that the model without *race* is correct. Because the p-value is smaller than significance level , then we can reject the null hypothesis and conclude that *race* is a significant predictor to the response.

**2) To test the significance of the *age***, we can use the deviance to compare two models: model 1 with *age* and model 3 without *age* (only with *race*). The model 1 here is just the full model above.

The likelihood test statistic is Deviance of model 3 – Deviance of model 1, which asymptotically Chi-squared distributed with df = 1 (assuming that the model 2 is correct and the distributional assumptions hold).

**Fit model 3 (only with predictor *race*):**

**Output:**

> sumary(fit\_3er)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.57547 0.92109 -2.7961 0.005172

race 0.27755 0.10910 2.5440 0.010959

n = 47 p = 2

Deviance = 18.26929 Null Deviance = 58.86527 (Difference = 40.59598)

**Deviance Test Output:**

> pchisq(18.296-9.2286,1,lower=FALSE)

[1] 0.002602068

**Conclusion:**

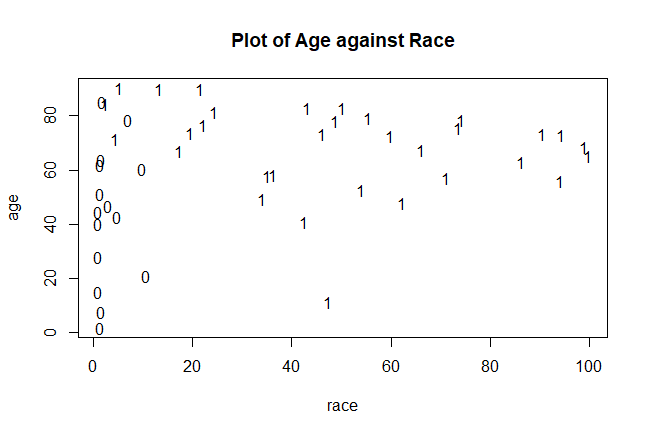
The null hypothesis is that the model without *age* is correct. Because the p-value is smaller than significance level , then we can reject the null hypothesis and conclude that *age* is a significant predictor to the response.

**Which test for the significance of the predictors should be preferred?**

The difference in deviances test should be preferred here. From part (d), we know that the data is sparse and the standard errors can be overestimated and so the z-values too small and the significance of an effect could be missed. As a result, we should use difference-in-deviance test to test the significance of the predictors.

**(f)**

Make plot of race against age which also distinguishes the two levels of the response variable:



**Interpret the plot and connect it to the previous model output:**

In the plot, we can see that for points have 0 as response value, their values of *race* are all less than 20. For points have 1 as response value, their values of *race* are all most greater than 10. As a result, we can distinguish the two levels of the response variable by different *race* values. It means that the predictor *race* is significant to the response in this model, which coincides with our answer in part (e).

However, for points have 0 as response value, their values of *age* are distributed from 0 to 100. For points have 1 as response value, their values of *age* are also distributed from 0 to 100. As a result, we cannot distinguish the two levels of the response variable by different *age* values. It means that the predictor *age* is not significant to the response in this model, which coincides with our answer in part (e).

In addition, the data points with 0 and 1 as response values are not linearly separable in the plot so there will be no perfect fit for this model with only *race* and *age.* Hence, when we fit the model, the algorithm did not converge and the fitted values are not only 0 and 1, which coincides with the answer in part (d).

**(g)**

**Fit a binary GLM with probit link function:**

We fit a binary GLM with *binary* as the response and the two variables: *race* and *age* as predictors. In this generalized linear model, we use probit link function and binomial distribution family.

> fit\_3g = glm(binary ~ race + age, family=binomial(link=probit), data=chredlin))

**The GLM formula:**

**Output:**

> sumary(fit\_3g)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -7.559838 4.168736 -1.8135 0.06976

race 0.186545 0.089134 2.0929 0.03636

age 0.085034 0.049390 1.7217 0.08513

n = 47 p = 3

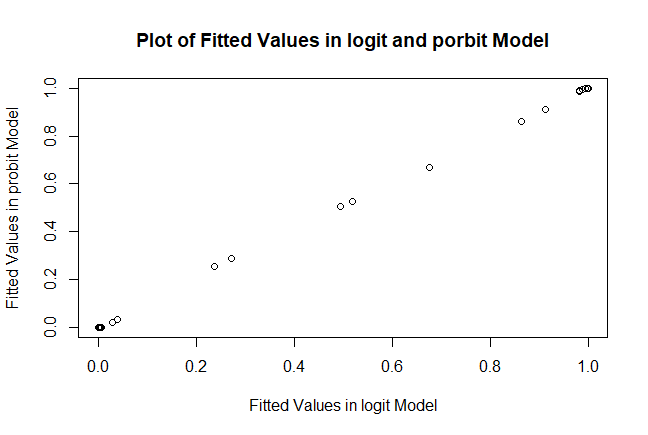
Deviance = 8.97862 Null Deviance = 58.86527 (Difference = 49.88665)

Comparing the model output between the logit and probit models:

Similarity: the z-values and p-values corresponding to different terms are similar. If we use the interpretation of z-value to test the significance of predictors, the result is the same in these two models. The deviance value is also similar.

Difference: the estimated coefficients and their standard errors in logit and probit models are different.

**Plot the predicted values on the probability scale against each other:**



From the plot we can see that there is a linear relationship between the fitted values in these two models. All the fitted values conditional on same predictor values in the two models are same. As a result, the probit and logit model will give us same fitted values of response.

**(h)**

The binary response model with probit link function is the most comparable to the Gaussian linear model.

In the GLM with probit link function:

The response

So .

,

Since the error in the Gaussian linear model is normally distributed, the response Y in the Gaussian linear model also follows a normal distribution. Hence, the binary response model with probit link function is the most comparable to the Gaussian linear model.

**Contrast the drawbacks between the Gaussian and binary response models:**

The drawback of binary response models is the loss of information. Because when we transfer the response into binary response, we lose the specific response values.

The drawback of Gaussian model is that using this model, we will focus too much on the outliers, especially for the zero values in this data.

**R Code**

# Problem 3 --------------------------------------------------------------

install.packages("faraway")

library(faraway)

#a

data(chredlin)

help(chredlin)

hist(chredlin$involact, breaks=seq(-0.05,2.25,by=0.1) ,xlab="Involact", main="Histogram of Involact")

#b

fit\_3b = lm(involact ~ race + fire + theft + age + log(income), data=chredlin)

sumary(fit\_3b)

#c

residuals(fit\_3b)

plot(fit\_3b)

#d

#transfer the response into binary response

chredlin$binary=rep(1,47)

for (i in 1:47){

if (chredlin[i,5]==0){

chredlin[i,8]=0

}

}

#fit the binary model

fit\_3d = glm(binary ~ race + fire + theft + age + log(income), family=binomial, data=chredlin)

sumary(fit\_3d)

#e

#fit the model with race and age

fit\_3e = glm(binary ~ race + age, family=binomial, data=chredlin)

summary(fit\_3f)

#fit the model with race

fit\_3er = glm(binary ~ race, family=binomial, data=chredlin)

sumary(fit\_3fr)

#fit the model with age

fit\_3ea = glm(binary ~ age, family=binomial, data=chredlin)

sumary(fit\_3fa)

#teat age

pchisq(18.296-9.2286,1,lower=FALSE)

#test race

pchisq(45.408-9.2286,1,lower=FALSE)

#f

plot(chredlin$race, chredlin$age, pch="", main="Plot of Age against Race", xlab="race", ylab="age")

text(chredlin$race, chredlin$age, chredlin$binary, cex=1.0)

#g

#fit the probit model

fit\_3g = glm(binary ~ race + fire + theft + age + log(income), family=binomial(link=probit), data=chredlin)

sumary(fit\_3g)

#plot the fitted values in probit and logit models

plot(fitted.values(fit\_3e),fitted.values(fit\_3g), main="Plot of Fitted Values in logit and porbit Model", xlab="Fitted Values in logit Model", ylab="Fitted Values in probit Model")