DIG Secondary Analysis

Costa Stavrianidis

4/20/2022

Reading in the data/cleaning

```
DIG <- read.csv("dig.csv", header = TRUE)

# Removing patients with missing data for variables
DIG <- DIG %>% drop_na(c(KLEVEL, WHF, TRTMT, AGE, BMI, NSYM, SEX, RACE, CHESTX,
                         EJF_PER, EJFMETH, CREAT, CHFDUR, HEARTRTE, SYSBP,
                         FUNCTCLS, CHFETIOL, PREVMI, ANGINA, DIABETES, HYPERTEN,
                         DIGUSE, DIURETK, DIURET, KSUPP, ACEINHIB, NITRATES,
                         HYDRAL, VASOD, DIGDOSE))

# Removing patients with anomalous KLEVEL
DIG$KLEVEL <- ifelse(DIG$KLEVEL > 10 | DIG$KLEVEL < 2.5, NA, DIG$KLEVEL)
DIG <- DIG %>% drop_na(KLEVEL)

# Converting binary/categorical variables to factors
columns <- c("TRTMT", "NSYM", "SEX", "RACE", "EJFMETH",
             "FUNCTCLS", "CHFETIOL", "PREVMI", "ANGINA", "DIABETES",
             "HYPERTEN", "DIGUSE", "DIURETK", "DIURET", "KSUPP",
             "ACEINHIB", "NITRATES", "HYDRAL", "VASOD")
DIG$WHF <- factor(DIG$WHF, levels = c(0,1),
                  labels = c("WHF No Event", "WHF Event"))
DIG <- DIG %>% mutate_at(columns, factor)
```

Table 1

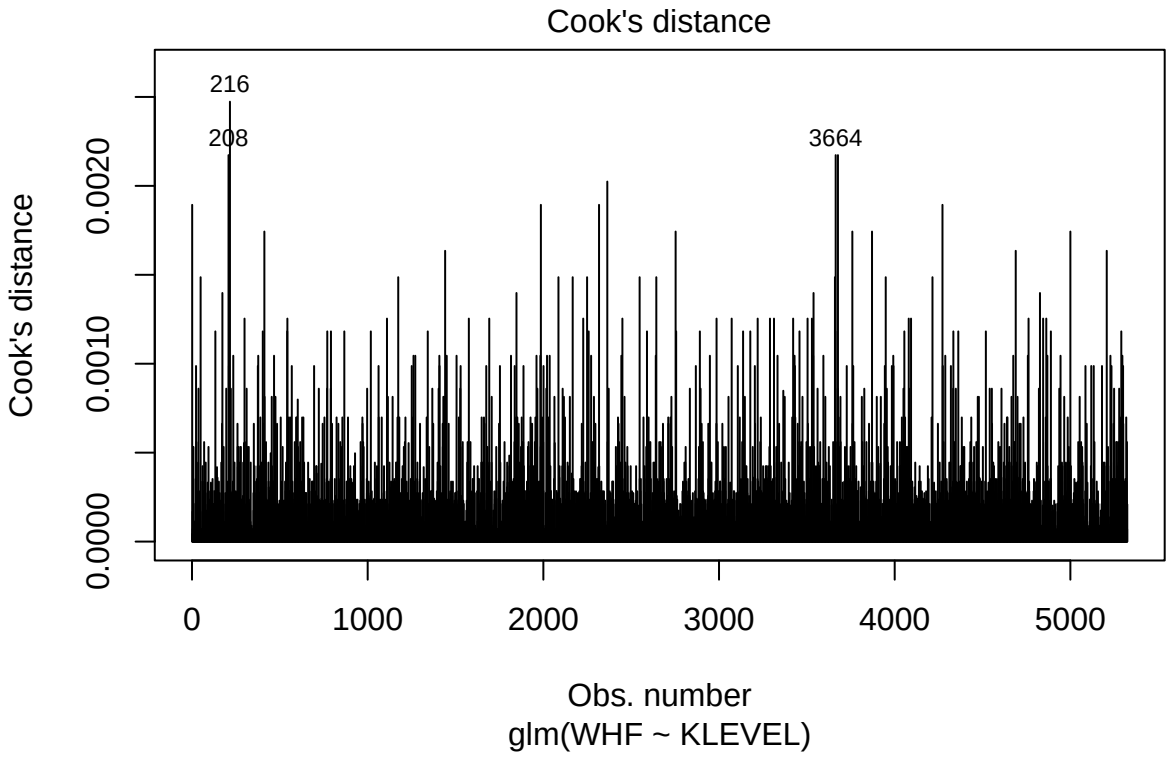
```
table1(~ KLEVEL + TRTMT + AGE + BMI + NSYM + SEX + RACE | WHF, data = DIG)
```

```
## Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

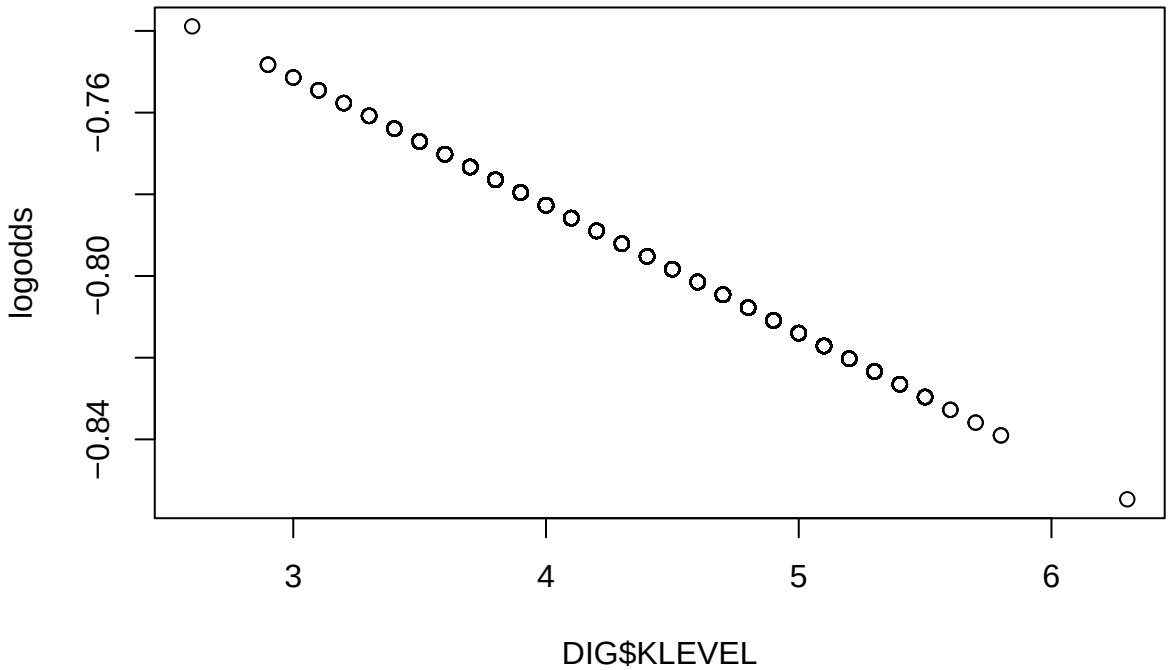
	WHF No Event	WHF Event	Overall
	(N=3666)	(N=1658)	(N=5324)
KLEVEL			
Mean (SD)	4.35 (0.428)	4.34 (0.439)	4.35 (0.432)
Median [Min, Max]	4.30 [2.60, 6.30]	4.30 [2.90, 5.60]	4.30 [2.60, 6.30]
TRTMT			
0	1713 (46.7%)	935 (56.4%)	2648 (49.7%)
1	1953 (53.3%)	723 (43.6%)	2676 (50.3%)
AGE			
Mean (SD)	63.4 (10.9)	63.5 (10.7)	63.4 (10.9)
Median [Min, Max]	65.0 [21.0, 91.0]	64.0 [22.0, 92.0]	64.5 [21.0, 92.0]
BMI			
Mean (SD)	27.1 (5.12)	27.2 (5.31)	27.1 (5.18)
Median [Min, Max]	26.5 [14.4, 58.3]	26.5 [15.1, 62.7]	26.5 [14.4, 62.7]
NSYM			
0	38 (1.0%)	15 (0.9%)	53 (1.0%)
1	82 (2.2%)	34 (2.1%)	116 (2.2%)
2	250 (6.8%)	115 (6.9%)	365 (6.9%)
3	338 (9.2%)	144 (8.7%)	482 (9.1%)
4	2958 (80.7%)	1350 (81.4%)	4308 (80.9%)
SEX			
1	2873 (78.4%)	1272 (76.7%)	4145 (77.9%)
2	793 (21.6%)	386 (23.3%)	1179 (22.1%)
RACE			
1	3248 (88.6%)	1382 (83.4%)	4630 (87.0%)
2	418 (11.4%)	276 (16.6%)	694 (13.0%)

Primary Hypothesis Assumptions

```
# Fitting model
DIG_model1 <- glm(WHF ~ KLEVEL, data = DIG, family = "binomial")

# Checking for outliers
plot(DIG_model1, which = 4, id.n = 3)
```



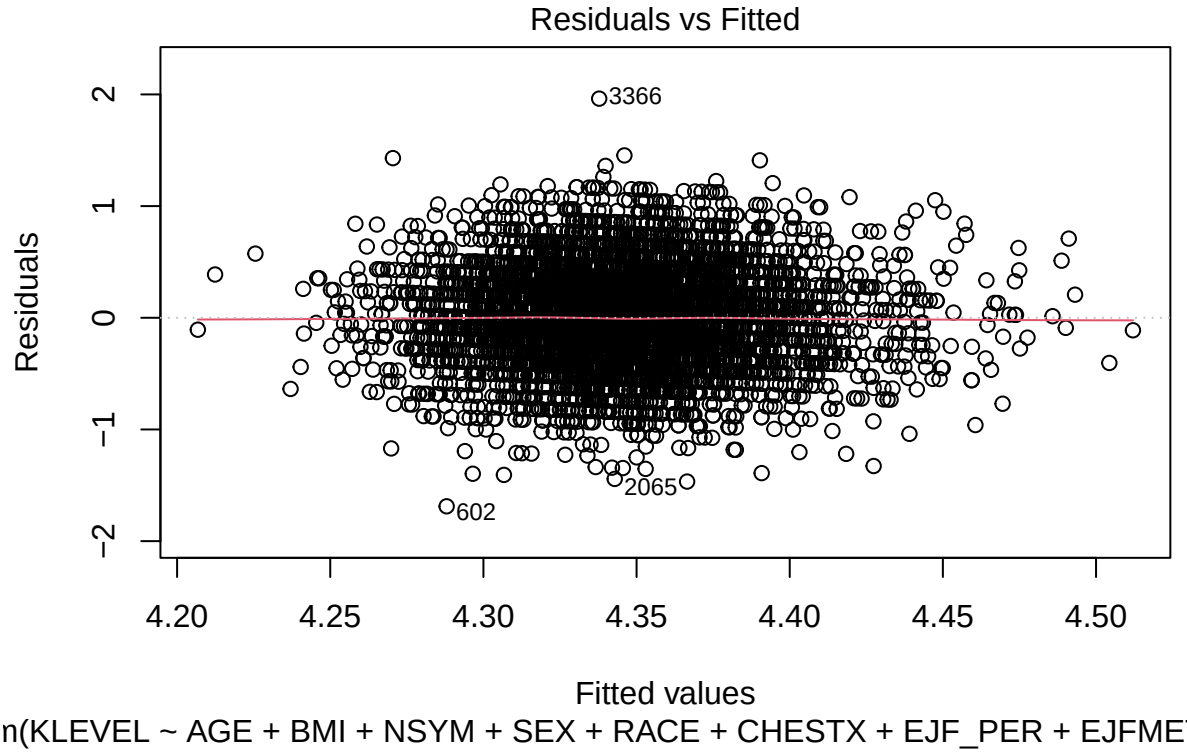

```
# Checking for linear relationship between KLEVEL and logodds of WHF
logodds <- DIG_model1$linear.predictors
plot(logodds ~ DIG$KLEVEL)
```



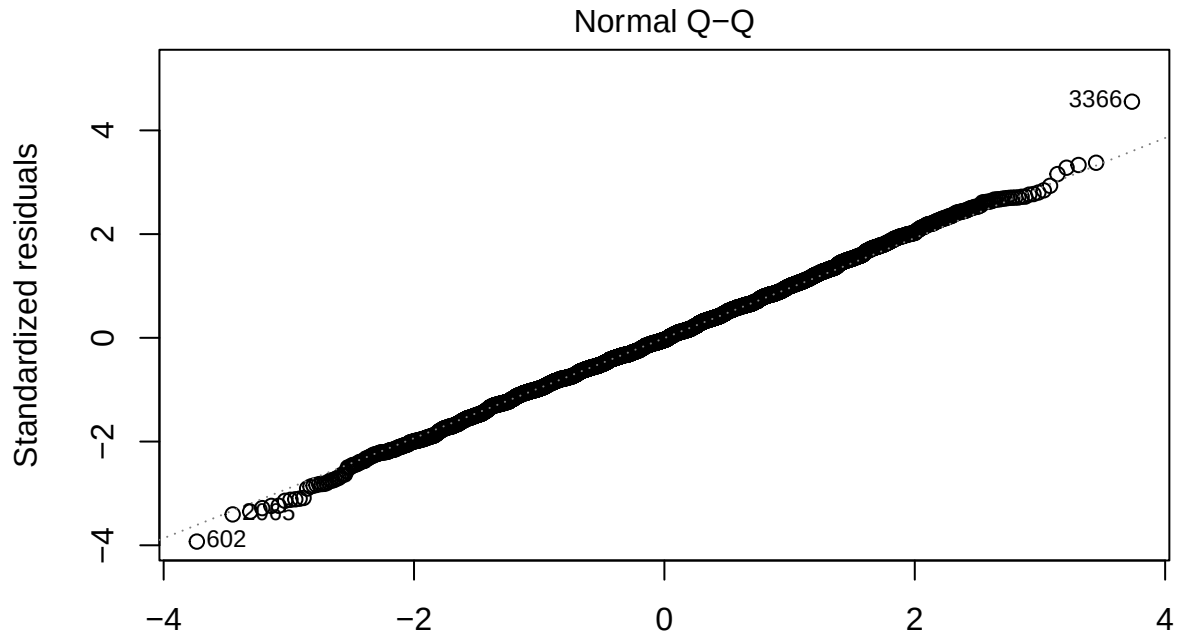
The Cook's distance plot does not appear to have any clear outliers or influential observations. There appears to be a linear relationship between the predictor, KLEVEL, and the log-odds of the outcome WHF. Assumptions seem to be meeting, thus we will continue with the model for the results.

Secondary Hypothesis (1.2.2) Assumptions

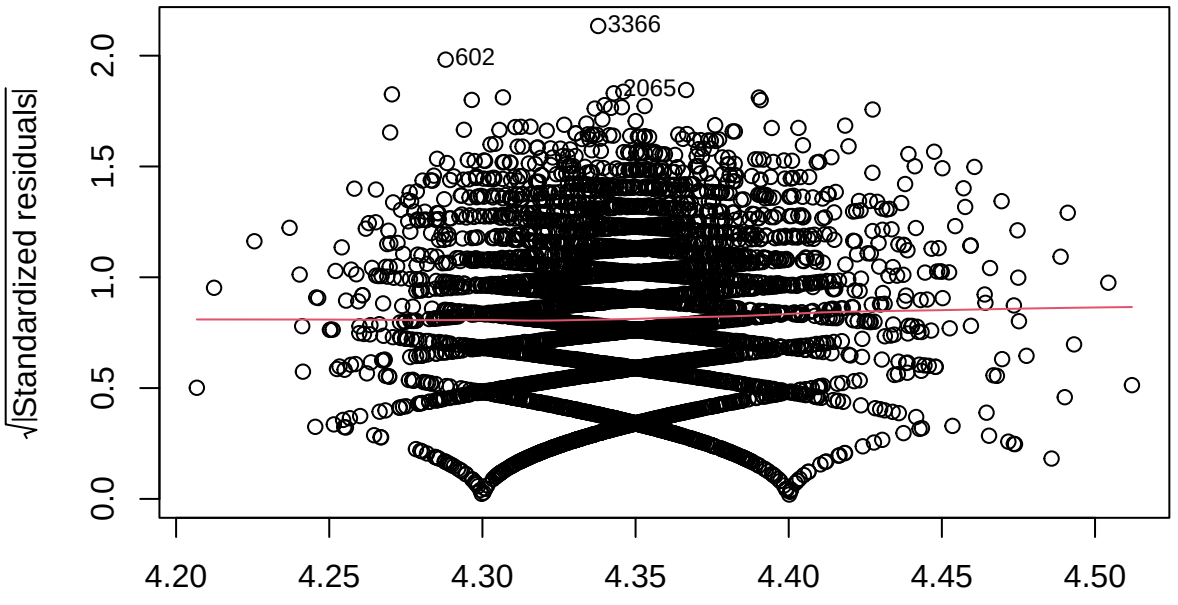
```
# Fitting model
DIG_model2 <- lm(KLEVEL ~ AGE + BMI + NSYM + SEX + RACE + CHESTX + EJF_PER +
                   EJFMETH + CREAT + CHFDUR + HEARTRTE + SYSBP + FUNCTCLS +
                   CHFETIOL + PREVMI + ANGINA + DIABETES + HYPERTEN + DIGUSE +
                   DIURETK + DIURET + KSUPP + ACEINHIB + NITRATES + HYDRAL +
                   VASOD + DIGDOSE, data = DIG)

# Checking assumptions
plot(DIG_model2)
```



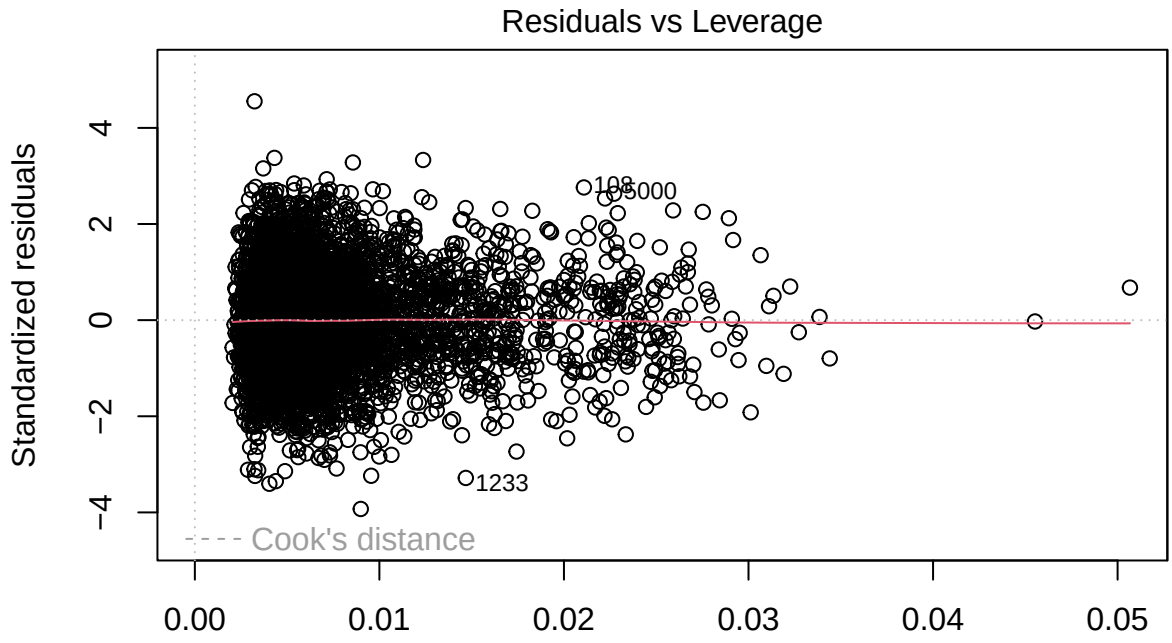
n(KLEVEL ~ AGE + BMI + NSYM + SEX + RACE + CHESTX + EJF_PER + EJFMETH +



n(KLEVEL ~ AGE + BMI + NSYM + SEX + RACE + CHESTX + EJF_PER + EJFMETH +



n(KLEVEL ~ AGE + BMI + NSYM + SEX + RACE + CHESTX + EJF_PER + EJFMETH +



n(KLEVEL ~ AGE + BMI + NSYM + SEX + RACE + CHESTX + EJF_PER + EJFMETH +

```
# Checking multicollinearity
car::vif(DIG_model2)
```

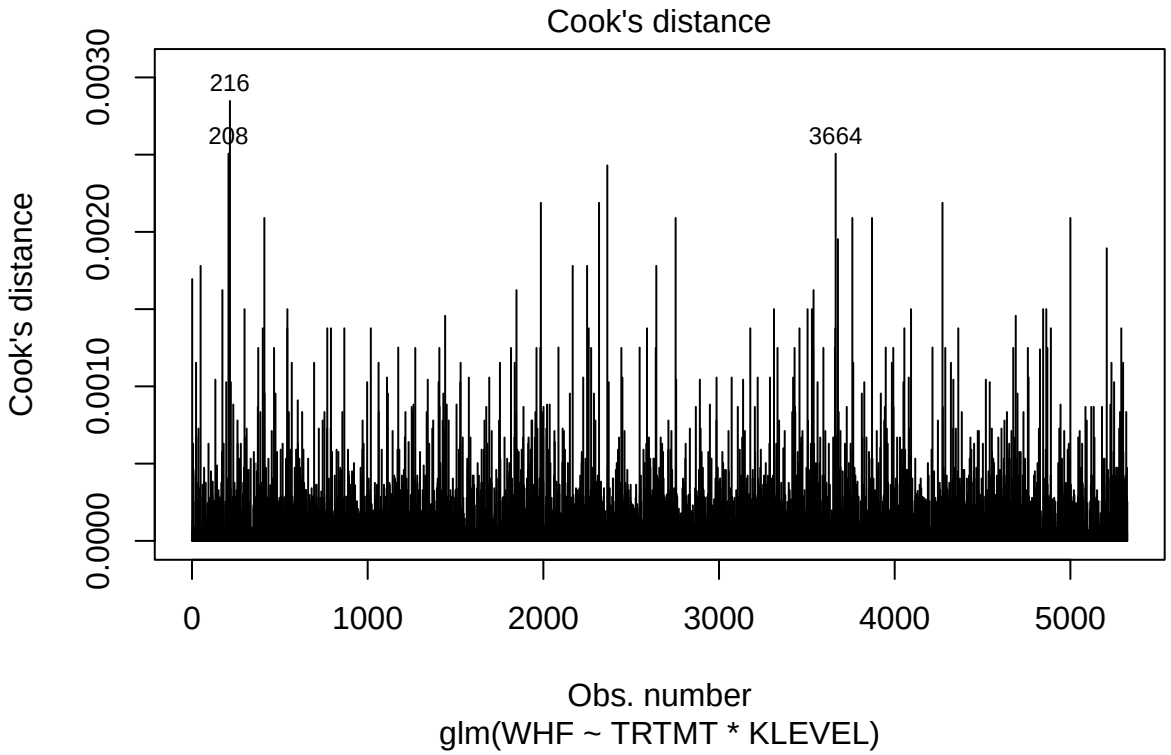
```
##                GVIF Df GVIF^(1/(2*Df))
## AGE        1.008713  1        1.004347
## BMI        1.007253  1        1.003620
## NSYM       1.019252  4        1.002387
## SEX        1.108235  1        1.052728
## RACE       1.141827  1        1.068563
## CHESTX     1.172198  1        1.082681
## EJF_PER    1.116472  1        1.056633
## EJFMETH    1.013916  2        1.003461
## CREAT      1.009122  1        1.004550
## CHFDUR     1.008593  1        1.004287
## HEARTRTE   1.008173  1        1.004078
## SYSBP      1.005666  1        1.002829
## FUNCTCLS   1.109804  3        1.017516
## CHFETIOL   1.185445  5        1.017157
## PREVMI     1.006866  1        1.003427
## ANGINA     1.005961  1        1.002976
## DIABETES   1.007308  1        1.003647
## HYPERTEN   1.004879  1        1.002437
## DIGUSE     1.044035  1        1.021780
## DIURETK    1.043894  1        1.021711
## DIURET     1.188731  1        1.090289
## KSUPP      1.109899  1        1.053518
## ACEINHIB   1.008524  1        1.004253
## NITRATES   1.008900  1        1.004440
## HYDRAL     1.020566  1        1.010231
## VASOD      1.003501  1        1.001749
```

```
## DIGDOSE  1.007217  1         1.003602
```

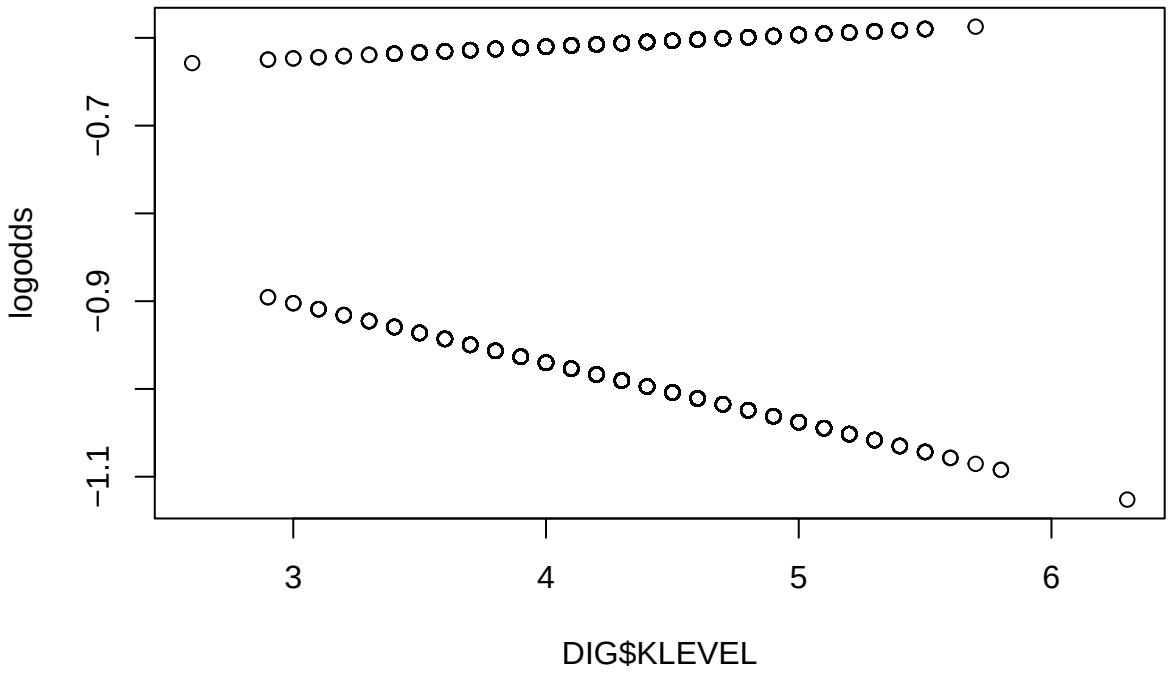
Residuals vs. Fitted plot seems to have a horizontal red line around 0, and there does not seem to be a fitted pattern to the plot. It appears we are meeting the linearity assumptions for our model. The Normal Q-Q plot appears to follow a straight line (with a few outliers), thus the residuals appear to be normally distributed. In the Scale-Location plot, we can see the red line is horizontal. The points appear to be somewhat evenly distributed around the red line, however there appear to be some patterns that they follow below the line. We will continue fitting the model, however this potential violation of homoscedasticity will be kept in mind for the discussion portion of the analysis. In the Residuals vs. Leverage plot, there aren't any observations that are particularly influential for the regression, so we will continue fitting the model. Calculating the variance inflation factor for each predictor, it does not seem that any coefficients of predictors in particular are having their variance inflated due to multicollinearity.

Secondary Hypothesis (1.2.3) Assumptions

```r
# Fitting model
DIG_model3 <- glm(WHF ~ TRTMT * KLEVEL, data = DIG, family = "binomial")

# Checking for outliers
plot(DIG_model3, which = 4, id.n = 3)
```




```r
# Checking for linear relationship between predictors and logodds of WHF
logodds <- DIG_model3$linear.predictors
plot(logodds ~ DIG$KLEVEL)
```



The Cook's distance plot does not appear to have any clear outliers or influential observations. There appears to be a linear relationship between KLEVEL and the log-odds of the outcome WHF for both treatment groups (evidenced by the two separate lines in the plot). Assumptions seem to be meeting, thus we will continue with the model for the results.

Primary Hypothesis Results

```
summary(DIG_model1)
```

```
##
## Call:
## glm(formula = WHF ~ KLEVEL, family = "binomial", data = DIG)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8837  -0.8655  -0.8610   1.5226   1.5452
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.65755    0.29947  -2.196   0.0281 *
## KLEVEL      -0.03129    0.06861  -0.456   0.6484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6604.2  on 5323  degrees of freedom
## Residual deviance: 6604.0  on 5322  degrees of freedom
## AIC: 6608
##
## Number of Fisher Scoring iterations: 4
```

```
cbind("OR" = exp(coef(DIG_model1)), exp(confint(DIG_model1)))
```

```
## Waiting for profiling to be done...
```

```
##                     OR     2.5 %    97.5 %
## (Intercept) 0.5181193 0.2879317 0.9315405
## KLEVEL      0.9691964 0.8471927 1.1086632
```

KLEVEL predictor has a coefficient of -0.03129 with a p-value of 0.6484 for the Wald test. Final model is: logit(p) = -0.66 - 0.03*KLEVEL

Secondary Hypothesis (1.2.2) Results

```
outcome <- DIG$KLEVEL
predictors <- data.matrix(DIG[, c("AGE", "BMI", "NSYM", "SEX", "RACE",
                                  "CHESTX", "EJF_PER", "EJFMETH", "CREAT",
                                  "CHFDUR", "HEARTRTE", "SYSBP", "FUNCTCLS",
                                  "CHFETIOL", "PREVMI", "ANGINA", "DIABETES",
                                  "HYPERTEN", "DIGUSE", "DIURETK", "DIURET",
                                  "KSUPP", "ACEINHIB", "NITRATES", "HYDRAL",
                                  "VASOD", "DIGDOSE")])

# Find optimal lambda tuning parameter for LASSO regression with
# k-fold cross-validation
cv_model <- cv.glmnet(predictors, outcome, alpha = 1)
lambda_model <- cv_model$lambda.min

# Run LASSO regression and summarize
DIG_model_lasso <- glmnet(predictors, outcome, alpha = 1, lambda = lambda_model)
coef(DIG_model_lasso)
```

```
## 28 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept) 4.345868e+00
## AGE         .
## BMI         .
## NSYM        .
## SEX         .
## RACE        .
## CHESTX      .
## EJF_PER     .
## EJFMETH     .
## CREAT       1.224074e-17
## CHFDUR      .
## HEARTRTE    .
## SYSBP       .
## FUNCTCLS    .
## CHFETIOL    .
## PREVMI      .
## ANGINA      .
## DIABETES    .
## HYPERTEN    .
## DIGUSE      .
## DIURETK     .
```

```
## DIURET      .
## KSUPP       .
## ACEINHIB    .
## NITRATES    .
## HYDRAL      .
## VASOD       .
## DIGDOSE     .
```

Only variable kept in the model is CREAT, with a coefficient of 1.22e-17. Final model is: KLEVEL = 4.35 + 1.22e-17*CREAT

Secondary Hypothesis (1.2.3) Results

```
summary(DIG_model3)
```

```
##
## Call:
## glm(formula = WHF ~ TRTMT * KLEVEL, family = "binomial", data = DIG)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9402  -0.9316  -0.7948   1.4420   1.6561
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.66355    0.41933  -1.582    0.114
## TRTMT1        -0.03534    0.60246  -0.059    0.953
## KLEVEL         0.01338    0.09614   0.139    0.889
## TRTMT1:KLEVEL -0.08119    0.13803  -0.588    0.556
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6604.2  on 5323  degrees of freedom
## Residual deviance: 6561.0  on 5320  degrees of freedom
## AIC: 6569
##
## Number of Fisher Scoring iterations: 4
```

```
cbind("OR" = exp(coef(DIG_model3)), exp(confint(DIG_model3)))
```

```
## Waiting for profiling to be done...
```

```
##                       OR     2.5 %   97.5 %
## (Intercept)   0.5150181 0.2261667 1.171000
## TRTMT1        0.9652724 0.2962432 3.143930
## KLEVEL        1.0134741 0.8393295 1.223654
## TRTMT1:KLEVEL 0.9220223 0.7034053 1.208473
```

Final models are: placebo: logit(p) = -0.664 + 0.013*KLEVEL

treatment: logit(p) = (-0.664 - 0.035) + (0.013 - 0.081)*KLEVEL = -0.699 - 0.068KLEVEL

The coefficient of the interaction term is -0.08119 with a p-value of 0.556 from the Wald test.

Odds of being hospitalized due to worsening heart failure increases by 1% for every 1 unit increase in serum potassium level for patients that received the placebo. The odds of being hospitalized due to worsening heart

failure decreases by a factor of 1.01 * 0.92 = 0.93 (7%) for every 1 unit increase in serum potassium level for patients that received the Digoxin treatment.