Quiz 2

Ryan Richardson

19/11/2020

R Markdown

```
d <- read.dta('wcgs.dta')
summary(d)
```

```
##       age            arcus            behpat          bmi        chd69
##  Min.   :39.00   Min.   :0.0000   A1: 264   Min.   :11.19   No :2897
##  1st Qu.:42.00   1st Qu.:0.0000   A2:1325   1st Qu.:22.96   Yes: 257
##  Median :45.00   Median :0.0000   B3:1216   Median :24.39
##  Mean   :46.28   Mean   :0.2985   B4: 349   Mean   :24.52
##  3rd Qu.:50.00   3rd Qu.:1.0000             3rd Qu.:25.84
##  Max.   :59.00   Max.   :1.0000             Max.   :38.95
##                  NA's   :2
##       chol            dbp            dibpat          height            id
##  Min.   :103.0   Min.   : 58.00   Type B:1565   Min.   :60.00   Min.   : 2001
##  1st Qu.:197.2   1st Qu.: 76.00   Type A:1589   1st Qu.:68.00   1st Qu.: 3741
##  Median :223.0   Median : 80.00                 Median :70.00   Median :11406
##  Mean   :226.4   Mean   : 82.02                 Mean   :69.78   Mean   :10478
##  3rd Qu.:253.0   3rd Qu.: 86.00                 3rd Qu.:72.00   3rd Qu.:13115
##  Max.   :645.0   Max.   :150.00                 Max.   :78.00   Max.   :22101
##  NA's   :12
##      lnsbp           lnwght          ncigs           sbp        smoke
##  Min.   :4.585   Min.   :4.357   Min.   : 0.0   Min.   : 98.0   No :1652
##  1st Qu.:4.787   1st Qu.:5.043   1st Qu.: 0.0   1st Qu.:120.0   Yes:1502
##  Median :4.836   Median :5.136   Median : 0.0   Median :126.0
##  Mean   :4.850   Mean   :5.128   Mean   :11.6   Mean   :128.6
##  3rd Qu.:4.913   3rd Qu.:5.204   3rd Qu.:20.0   3rd Qu.:136.0
##  Max.   :5.438   Max.   :5.768   Max.   :99.0   Max.   :230.0
##
##        t1               time169        typchd69           uni
##  Min.   :-47.43147   Min.   :  18   Min.   :0.0000   Min.   :0.0007097
##  1st Qu.: -1.00337   1st Qu.:2842   1st Qu.:0.0000   1st Qu.:0.2573755
##  Median :  0.00748   Median :2942   Median :0.0000   Median :0.5157779
##  Mean   : -0.03336   Mean   :2684   Mean   :0.1363   Mean   :0.5052159
##  3rd Qu.:  0.97575   3rd Qu.:3037   3rd Qu.:0.0000   3rd Qu.:0.7559902
##  Max.   : 47.01623   Max.   :3430   Max.   :3.0000   Max.   :0.9994496
##  NA's   :39
##      weight       wghtcat        agec
##  Min.   : 78   < 140  : 232   35-40: 543
##  1st Qu.:155   140-170:1538   41-45:1091
##  Median :170   170-200:1171   46-50: 750
##  Mean   :170   > 200  : 213   51-55: 528
##  3rd Qu.:182                  56-60: 242
##  Max.   :320
##
```

Cleanup

Remove chd69, behpat, lnsbp, lnwght, smoke, t1, uni, wghtcat, agec, as they are captured in other variables.

Id is unrelated and can be removed.

Arcus is recoded as a factor variable, but because it is a 1,0 it is not strictly necessary.

We then remove NA variables

```
d$arcus = as.factor(d$arcus)
d$typchd69 = as.factor(d$typchd69)

dTrim = d[, -c(3,5,10,11,12,15,16,19,21,22)]
dTrim = na.omit(dTrim)
summary(dTrim)
```

```
##       age        arcus         bmi             chol            dbp
##  Min.   :39.00   0:2202   Min.   :11.19   Min.   :103.0   Min.   : 58.00
##  1st Qu.:42.00   1: 938   1st Qu.:22.96   1st Qu.:197.0   1st Qu.: 76.00
##  Median :45.00            Median :24.39   Median :223.0   Median : 80.00
##  Mean   :46.27            Mean   :24.52   Mean   :226.3   Mean   : 81.97
##  3rd Qu.:50.00            3rd Qu.:25.84   3rd Qu.:253.0   3rd Qu.: 86.00
##  Max.   :59.00            Max.   :38.95   Max.   :645.0   Max.   :136.00
##     dibpat         height          ncigs            sbp          time169
##  Type B:1557   Min.   :60.00   Min.   : 0.00   Min.   : 98.0   Min.   :  18
##  Type A:1583   1st Qu.:68.00   1st Qu.: 0.00   1st Qu.:120.0   1st Qu.:2843
##                Median :70.00   Median : 0.00   Median :126.0   Median :2942
##                Mean   :69.78   Mean   :11.58   Mean   :128.6   Mean   :2684
##                3rd Qu.:72.00   3rd Qu.:20.00   3rd Qu.:136.0   3rd Qu.:3036
##                Max.   :78.00   Max.   :99.00   Max.   :230.0   Max.   :3430
##  typchd69     weight
##  0:2885   Min.   : 78.0
##  1: 134   1st Qu.:155.0
##  2:  70   Median :170.0
##  3:  51   Mean   :169.9
##           3rd Qu.:182.0
##           Max.   :320.0
```

Build Model Using StepAIC

```
##
## Call:
## glm(formula = step$formula, family = "binomial", data = step$model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9827  -1.1363   0.6746   1.1712   1.4905
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.935e+00  1.130e+00  -3.483 0.000497 ***
## age          2.781e-02  6.762e-03   4.112 3.92e-05 ***
## height       3.079e-02  1.449e-02   2.125 0.033584 *
## ncigs        1.128e-02  2.546e-03   4.430 9.43e-06 ***
## sbp          6.580e-03  2.493e-03   2.639 0.008322 **
## time169     -1.848e-04  6.231e-05  -2.966 0.003014 **
## typchd691    4.928e-01  2.053e-01   2.400 0.016384 *
## typchd692    4.068e-01  2.684e-01   1.516 0.129563
## typchd693    6.596e-01  3.280e-01   2.011 0.044312 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4352.7  on 3139  degrees of freedom
## Residual deviance: 4248.8  on 3131  degrees of freedom
## AIC: 4266.8
##
## Number of Fisher Scoring iterations: 4
```

```
## (Intercept)         age      height       ncigs         sbp     time169
##   0.0195512   1.0281965   1.0312681   1.0113437   1.0066012   0.9998152
##   typchd691   typchd692   typchd693
##   1.6369177   1.5020597   1.9339906
```

Interpretation

The base odds of Type A Behaviour pattern to 1.9% when you are at age 0, height of 0, smoke 0 cigarettes a day, with a systolic bp of 0, and have had no CHD events

For every year above the base year, your odds increase by 100.8% of behaviour pattern a.

For every inch taller you are 103.1% as likely to have behaviour pattern a

For every additional cigarette you smoke a day you are 101.1% as likely to have behaviour pattern a

For every additional point in SBP you are 100.6% as likely to have behaviour pattern a

For the additional time unit since you experienced a CHD you are 163.69% more likely to have behaviour pattern a

If you had experienced CHD1 as likely to have behaviour pattern as

If you had experienced CHD2 as likely to have behaviour pattern as

If you had experienced CHD3 you are 193.39% as likely to have behaviour pattern as

```r
exp(bestModel$coefficients)
```

```
## (Intercept)         age      height       ncigs         sbp     time169
##   0.0195512   1.0281965   1.0312681   1.0113437   1.0066012   0.9998152
##   typchd691   typchd692   typchd693
##   1.6369177   1.5020597   1.9339906
```

The AUC for our best model is 0.597, with a threshold predictive value of 0.06. So anyone below that value was given a 'no', and anyone above that value was given a yes. This corresponds to an 1.06 odds of behaviour pattern A. Despite the low AUC, the threshold value seems decent, but the model can be improved.

This seems like the model may be over committing a small amount the 'yes' values in order to get a decent AUC, but it's hard to tell. Overall, the results aren't terrible but aren't great either, and the model can clearly be improved.

```
predictions = predict(bestModel)
rCurve = roc(dTrim$dibpat~predictions)
```

```
## Setting levels: control = Type B, case = Type A
```

```
## Setting direction: controls < cases
```

```
rCurve
```

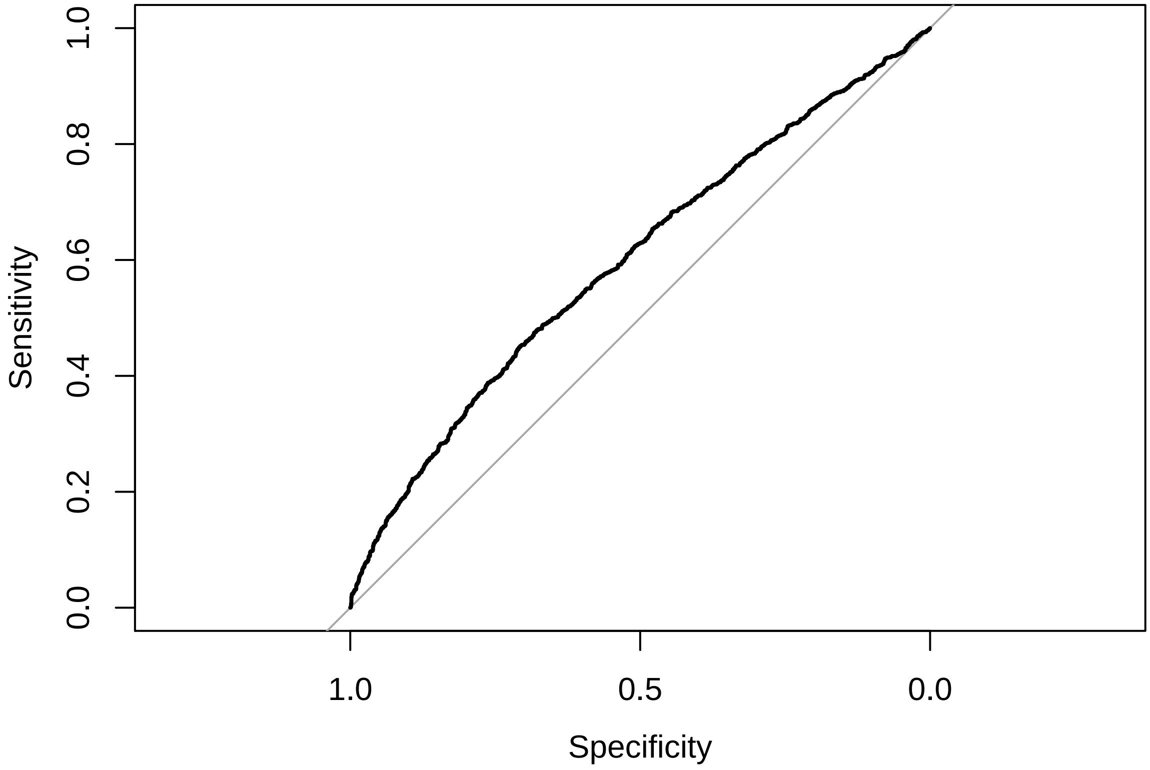
```
##
## Call:
## roc.formula(formula = dTrim$dibpat ~ predictions)
##
## Data: predictions in 1557 controls (dTrim$dibpat Type B) < 1583 cases (dTrim$dibpat Type A).
## Area under the curve: 0.597
```

```
coords(rCurve, "best", ret ="threshold")
```

```
## Warning in coords.roc(rCurve, "best", ret = "threshold"): The 'transpose'
## argument to FALSE by default since pROC 1.16. Set transpose = TRUE explicitly
## to revert to the previous behavior, or transpose = TRUE to silence this warning.
## Type help(coords_transpose) for additional information.
```

```
##    threshold
## 1 0.06246343
```

```
plot(rCurve)
```



Build Model Using StepAIC

AUC increased to 0.6162 and our threshold is now -0.133 for the model with 2nd order interactions. This corresponds to 87.5% odds of behaviour pattern A, which still seems to have swung the threshold away from over attributing to A, and instead now under attributing A, which leads me to believe that we are missing some additional information in this model and may have some better predictors.

```r
logit2 <- glm(dibpat ~.^2, data = dTrim, family = "binomial")
step2<- stepAIC(logit2, trace = FALSE)
bestModel2 = glm(step2$formula, data = step2$model, family="binomial")
summary(bestModel2)
```

```
##
## Call:
## glm(formula = step2$formula, family = "binomial", data = step2$model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8777  -1.1191   0.3826   1.1724   1.5696
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.794e+02  6.781e+01  -2.646  0.00814 **
## age               1.380e-01  4.900e-02   2.816  0.00487 **
## arcus1            2.204e+00  7.239e-01   3.045  0.00233 **
## bmi               4.729e+00  1.567e+00   3.018  0.00255 **
## chol              4.090e-01  2.238e-01   1.828  0.06761 .
## dbp              -8.173e-02  4.889e-02  -1.672  0.09458 .
## height            2.527e+00  9.747e-01   2.593  0.00953 **
## ncigs            -2.942e-02  2.354e-02  -1.250  0.21137
## sbp               6.887e-02  3.165e-02   2.176  0.02955 *
## time169           3.048e-02  1.601e-02   1.904  0.05695 .
## typchd691         7.115e+01  6.738e+01   1.056  0.29096
## typchd692        -5.302e+02  2.110e+02  -2.513  0.01197 *
## typchd693        -1.156e+01  7.980e+01  -0.145  0.88479
## weight           -2.189e-01  2.266e-01  -0.966  0.33409
## age:chol         -2.480e-04  1.583e-04  -1.567  0.11715
## age:time169      -2.007e-05  1.095e-05  -1.833  0.06685 .
## arcus1:sbp       -1.638e-02  5.585e-03  -2.933  0.00336 **
## bmi:chol         -7.207e-03  4.489e-03  -1.605  0.10840
## bmi:height       -4.840e-02  2.188e-02  -2.212  0.02694 *
## bmi:time169      -5.362e-04  3.206e-04  -1.672  0.09444 .
## bmi:typchd691    -1.384e+00  1.387e+00  -0.998  0.31824
## bmi:typchd692     1.105e+01  4.328e+00   2.553  0.01067 *
## bmi:typchd693     4.710e-01  1.596e+00   0.295  0.76793
## chol:height      -5.884e-03  3.211e-03  -1.833  0.06686 .
## chol:weight       1.121e-03  6.501e-04   1.725  0.08458 .
## dbp:ncigs         5.035e-04  2.872e-04   1.753  0.07959 .
## dbp:weight        4.594e-04  2.823e-04   1.627  0.10364
## height:time169   -4.390e-04  2.289e-04  -1.917  0.05520 .
## height:typchd691 -1.025e+00  9.605e-01  -1.067  0.28584
## height:typchd692  7.498e+00  2.992e+00   2.506  0.01221 *
## height:typchd693  8.360e-02  1.139e+00   0.073  0.94149
## sbp:weight       -3.462e-04  1.835e-04  -1.887  0.05919 .
## time169:weight    8.239e-05  4.601e-05   1.791  0.07334 .
## typchd691:weight  2.043e-01  1.982e-01   1.031  0.30255
## typchd692:weight -1.551e+00  6.112e-01  -2.538  0.01114 *
## typchd693:weight -3.180e-02  2.277e-01  -0.140  0.88891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4352.7  on 3139  degrees of freedom
## Residual deviance: 4188.2  on 3104  degrees of freedom
## AIC: 4260.2
##
## Number of Fisher Scoring iterations: 6
```

```
predictions2 = predict(bestModel2)
rCurve2 = roc(dTrim$dibpat~predictions2)
```

```
## Setting levels: control = Type B, case = Type A
```

```
## Setting direction: controls < cases
```

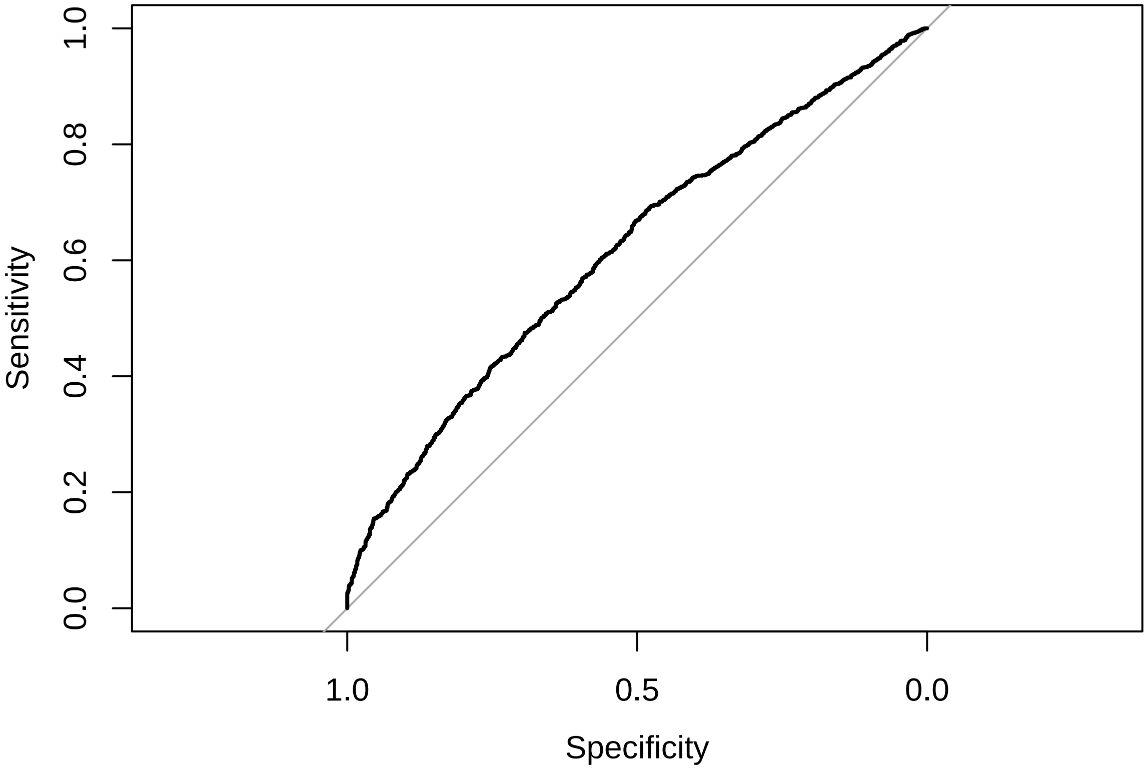
```
rCurve2
```

```
##
## Call:
## roc.formula(formula = dTrim$dibpat ~ predictions2)
##
## Data: predictions2 in 1557 controls (dTrim$dibpat Type B) < 1583 cases (dTrim$dibpat Type A).
## Area under the curve: 0.6162
```

```
coords(rCurve2, "best", ret ="threshold", transpose = TRUE)
```

```
##  threshold
## -0.1330499
```

```
plot(rCurve2)
```



Model w/ 3rd order Effects

Unfortunately, this was taking too long to run, so I left the code in but commented it out for the sake of time. I suspect that the AUC would increase with the additional interaction effects and that the threshold may also increase back towards a value closer to 100% based on the 2nd order interaction model.

```
#logit3 <- glm(dibpat~.^3, data = dTrim, family = "binomial")
#step3<- stepAIC(logit3, trace = FALSE)
#bestModel3 = glm(step3$formula, data = step3$model, family="binomial")
#summary(bestModel3)
#predictions3 = predict(bestModel3)
#rCurve3= roc(dTrim$dibpat~predictions3)
#rCurve3
#coords(rCurve3, "best", ret ="threshold")
#plot(rCurve3)
```

Question 2

Base model appears to be over dispersed based on on the spread of the Null deviance and residual deviance.

We can interpret the model as follows:

The base number o damage incidents each ship will experience is 4.92 (exp(1.5)), which is a Type A, built from 1960-1964, and operated from 1960-1974.

Type B ships experience 533% more damage events

Type C ships experience 25% as many damage events

Type D ships experience 42% as many damage events

Type E ships experiecne 82.7% as many damage events

Ships built from 1965-1969 experience 162% more damage events

Ships built from 1970-1974 experience 130% more damage events

Ships built from 1975-1979 experience 70.7% as many events

Ships built Operated between 1975-1979 experience 126.9% more damage events

```
ships <- read.table("https://data.princeton.edu/wws509/datasets/ships.dat", header=TRUE)

ships$type = as.factor(ships$type)
ships$construction = as.factor(ships$construction)
ships$operation = as.factor(ships$operation)
summary(ships)
```

```
##  type   construction   operation      months          damage
##  A:7   1960-64: 9    1960-74:15   Min.   :   45   Min.   : 0.00
##  B:7   1965-69:10    1975-79:19   1st Qu.:  371   1st Qu.: 1.00
##  C:7   1970-74:10                 Median : 1095   Median : 4.00
##  D:7   1975-79: 5                 Mean   : 4811   Mean   :10.47
##  E:6                              3rd Qu.: 2223   3rd Qu.:11.75
##                                   Max.   :44882   Max.   :58.00
```

```
shipModel = glm( damage ~ type + construction + operation, offset(log(months)),data=ships, family="pois
summary(shipModel)
```

```
##
## Call:
## glm(formula = damage ~ type + construction + operation, family = "poisson",
##     data = ships, weights = offset(log(months)))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.1055   -5.1049   -1.1605    0.3514    8.5839
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.59396    0.06907  23.077  < 2e-16 ***
## typeB                1.67400    0.05919  28.281  < 2e-16 ***
## typeC               -1.38730    0.12450 -11.143  < 2e-16 ***
## typeD               -0.85921    0.10680  -8.045 8.63e-16 ***
## typeE               -0.18922    0.08712  -2.172   0.0299 *
## construction1965-69  0.48280    0.04702  10.269  < 2e-16 ***
## construction1970-74  0.26685    0.04954   5.386 7.19e-08 ***
## construction1975-79 -0.34552    0.07233  -4.777 1.78e-06 ***
## operation1975-79     0.23855    0.03698   6.451 1.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4709.57  on 33  degrees of freedom
## Residual deviance:  964.55  on 25  degrees of freedom
## AIC: 1770.7
```

```
##
## Number of Fisher Scoring iterations: 6
```

```
exp(shipModel$coefficients)
```

```
##         (Intercept)               typeB               typeC               typeD
##           4.9232137           5.3334517           0.2497484           0.4234968
##               typeE construction1965-69 construction1970-74 construction1975-79
##           0.8276080           1.6206080           1.3058464           0.7078523
##    operation1975-79
##           1.2694020
```

Dispersion

Based on the dispersion test we reject the null hypothesis that dispersion is equal to 1. As a result, we can assume that the model is overdispersed and that we need to correct for this overdispersion in order to get an accurate estimate of the number of damage incidents over the life of the ship.

```
dispersiontest(shipModel)
```

```
##
##  Overdispersion test
##
## data:  shipModel
## z = 3.9678, p-value = 3.627e-05
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   3.742422
```

Quasi Poisson

With the quasi-Poisson model we see a very different model. Most of the significant variables from our initial model are no longer significant. We see that now ship types are still highly correlated to the number of damage incidents, which operation year is not significant. Construction year is only significant for ships made very early, which could be explained by the safety requirements on board when the ships were initially constructed. Overall, this model appears to still be a decent fit, mainly because it seems to be correcting for the dispersion of the data.

```
shipQuasi = glm( damage ~ type + construction + operation, offset(log(months)),data=ships, family="quasi
summary(shipQuasi)
```

```
##
## Call:
## glm(formula = damage ~ type + construction + operation, family = "quasipoisson",
##     data = ships, weights = offset(log(months)))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.1055   -5.1049   -1.1605    0.3514    8.5839
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.5940     0.4150   3.840 0.000745 ***
## typeB                 1.6740     0.3557   4.707 7.97e-05 ***
## typeC                -1.3873     0.7481  -1.854 0.075510 .
## typeD                -0.8592     0.6418  -1.339 0.192673
## typeE                -0.1892     0.5235  -0.361 0.720790
## construction1965-69   0.4828     0.2825   1.709 0.099853 .
## construction1970-74   0.2669     0.2977   0.896 0.378603
## construction1975-79  -0.3455     0.4346  -0.795 0.434128
## operation1975-79      0.2385     0.2222   1.074 0.293299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 36.107)
##
##     Null deviance: 4709.57  on 33  degrees of freedom
## Residual deviance:  964.55  on 25  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```