Poission Regression Model Project

2023-05-18

Description of dataset and problem

Project link: https://github.com/Arex-1/Math-449-Project

Ship Accidents dataset: https://search.r-project.org/CRAN/refmans/AER/html/ShipAccidents.html

For stating the problem and describe the data set:

Problem: Determine the factors contributing to the number of accidents a ship is involved in

Data: The analysis used the Ship Accidents dataset, as included in the AER package. The data originally came from Generalized Linear Models (1989) by McCullagh and Nelder.

Type: It gives the different types of ships that each one is. There are 5 different types of ships: A, B, C, D, E.

Construction: The year the ship was being built

Operation: The years that the ship was working

Service: how many months that the ship was working

Incidents: This explains how many incidents during that time period a certain ship was in.

Model interpretation and inferences:

- Ship type
  - Controlling for operation timeframe and service life, Type B ships experienced an estimated 75% more incidents than Type A ships on average. $\exp(\hat{\beta}_B) = \exp(0.557) = 1.747$, and $100 \times (1.747 - 1) \approx 75$.
  - Again controlling for operation timeframe and service life, ships of Type C experienced around 70% fewer incidents than Type A ships on average.
  - Type D ships experienced 57% fewer incidents than Type A ships.
  - Type E ships experienced around 11% fewer incidents than Type A ships. However, this coefficient was not statistically significant.
  - **The coefficients for ships of Type B-D were statistically significant, with p-values $<$ 0.05 for the null hypothesis that the true coefficient is not equal to zero. This indicates strong evidence that these types of ships did not experience the same number of incidents as Type A ships. The coefficient for Type E ships had a p-value of 0.628, meaning that there was no evidence that Type E ships experienced a different number of incidents than Type A ships.**
- Ship operation period
  - Ships in operation between 1975 and 1979 experienced an estimated 65% more incidents than ships in operation between 1960 and 1974.
  - **This coefficient was highly statistically significant, with p-value = 0.00030.**
- Ship service duration

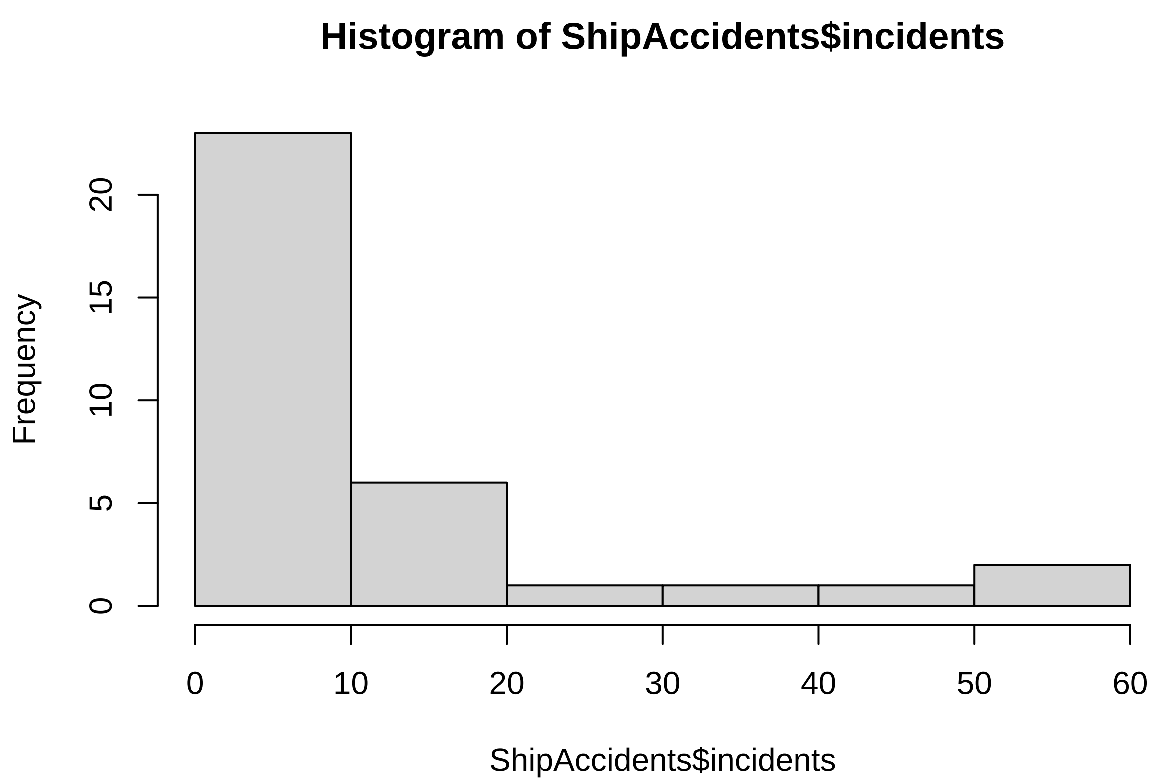
- A ship being in operation for an additional month increases the expected number of incidents by 0.0076%. A ship being in operation for an addition year increases the expected number of incidents by 0.09%, since $\exp(12 * \hat{\beta}_{\text{service}}) = 1.000906$
- **This coefficient was highly statistically significant, with a p-value near 0.**

```r
data('ShipAccidents', package='AER')
ShipAccidents = subset(ShipAccidents, service > 0)
```

```r
ShipAccidents
```

```
##    type construction operation service incidents
## 1     A      1960-64   1960-74     127         0
## 2     A      1960-64   1975-79      63         0
## 3     A      1965-69   1960-74    1095         3
## 4     A      1965-69   1975-79    1095         4
## 5     A      1970-74   1960-74    1512         6
## 6     A      1970-74   1975-79    3353        18
## 8     A      1975-79   1975-79    2244        11
## 9     B      1960-64   1960-74   44882        39
## 10    B      1960-64   1975-79   17176        29
## 11    B      1965-69   1960-74   28609        58
## 12    B      1965-69   1975-79   20370        53
## 13    B      1970-74   1960-74    7064        12
## 14    B      1970-74   1975-79   13099        44
## 16    B      1975-79   1975-79    7117        18
## 17    C      1960-64   1960-74    1179         1
## 18    C      1960-64   1975-79     552         1
## 19    C      1965-69   1960-74     781         0
## 20    C      1965-69   1975-79     676         1
## 21    C      1970-74   1960-74     783         6
## 22    C      1970-74   1975-79    1948         2
## 24    C      1975-79   1975-79     274         1
## 25    D      1960-64   1960-74     251         0
## 26    D      1960-64   1975-79     105         0
## 27    D      1965-69   1960-74     288         0
## 28    D      1965-69   1975-79     192         0
## 29    D      1970-74   1960-74     349         2
## 30    D      1970-74   1975-79    1208        11
## 32    D      1975-79   1975-79    2051         4
## 33    E      1975-79   1975-79      45         0
## 35    E      1965-69   1960-74     789         7
## 36    E      1965-69   1975-79     437         7
## 37    E      1970-74   1960-74    1157         5
## 38    E      1970-74   1975-79    2161        12
## 40    E      1975-79   1975-79     542         1
```

```r
hist(ShipAccidents$incidents)
```




```r
table(ShipAccidents$type, ShipAccidents$operation)
```

```
##
##     1960-74 1975-79
##   A       3       4
##   B       3       4
##   C       3       4
##   D       3       4
##   E       2       4
```

Full model (all predictors)

All variables in the full regression model (with log link function) are significant (not including `typeE`, but the rest of the types are significant, so we leave it in), but since the dataset is so small, it seems appropriate to not include `construction`.

```r
full_poisson = glm(incidents ~ ., data=ShipAccidents, family=poisson())
summary(full_poisson)
```

```
##
## Call:
```

```
## glm(formula = incidents ~ ., family = poisson(), data = ShipAccidents)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8497  -1.4945  -0.4671   0.7131   2.9049
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.657e-01  2.776e-01   0.957 0.338454
## typeB               7.030e-01  2.187e-01   3.214 0.001308 **
## typeC              -1.196e+00  3.275e-01  -3.653 0.000259 ***
## typeD              -8.320e-01  2.877e-01  -2.892 0.003831 **
## typeE              -2.492e-01  2.360e-01  -1.056 0.291027
## construction1965-69  1.048e+00  1.798e-01   5.832 5.47e-09 ***
## construction1970-74  1.446e+00  2.264e-01   6.387 1.70e-10 ***
## construction1975-79  6.788e-01  2.774e-01   2.447 0.014399 *
## operation1975-79    7.031e-01  1.358e-01   5.178 2.24e-07 ***
## service             6.421e-05  8.593e-06   7.473 7.86e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 614.54  on 33  degrees of freedom
## Residual deviance:  77.45  on 24  degrees of freedom
## AIC: 195.32
##
## Number of Fisher Scoring iterations: 5
```

Choosing a better model

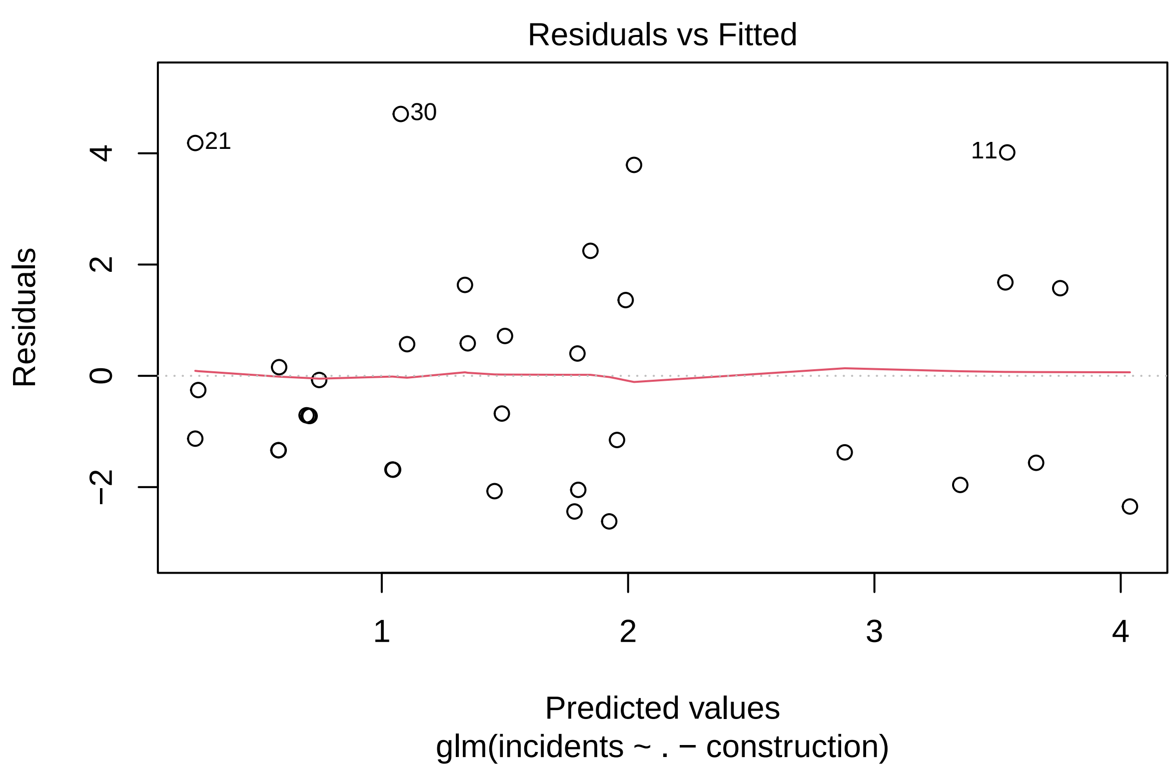
```
poisson_no_constr = glm(incidents ~ .-construction, data=ShipAccidents, family=poisson())
summary(poisson_no_constr)
```

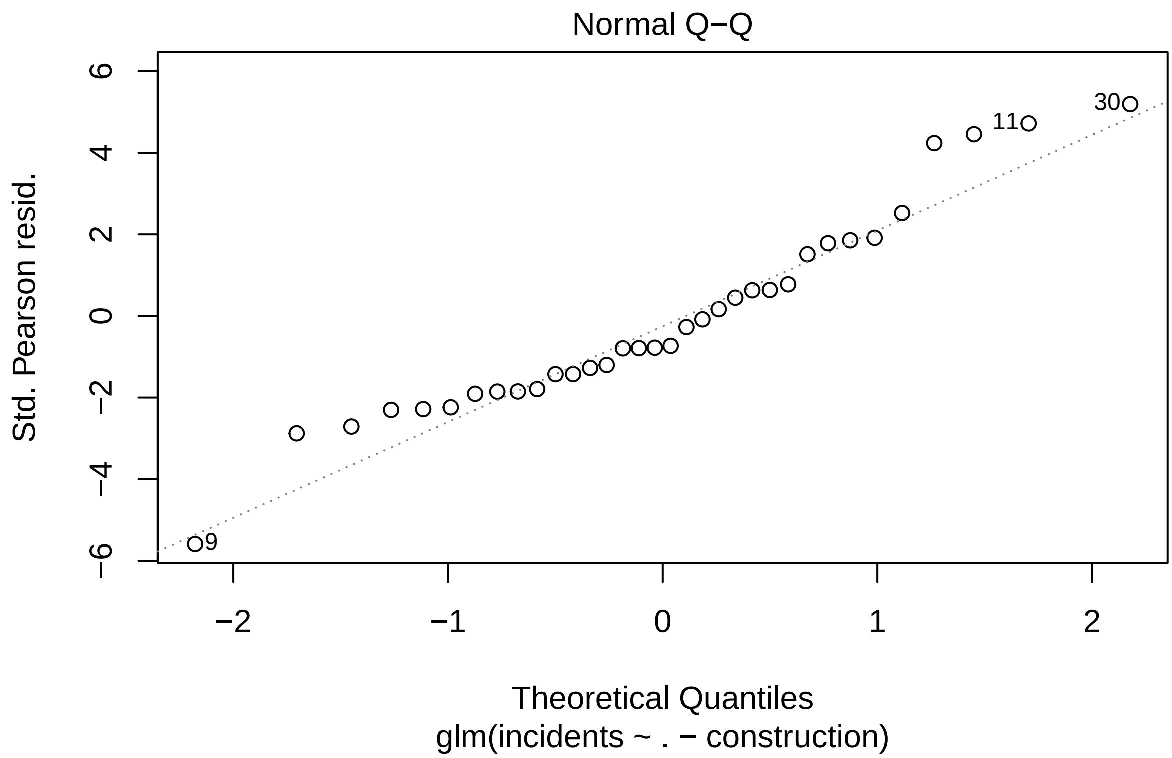
```
##
## Call:
## glm(formula = incidents ~ . - construction, family = poisson(),
##     data = ShipAccidents)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6998  -1.8910  -0.7528   1.1199   3.6531
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.454e+00  1.828e-01   7.955 1.80e-15 ***
## typeB            1.209e+00  2.037e-01   5.936 2.92e-09 ***
## typeC           -1.235e+00  3.274e-01  -3.773 0.000161 ***
## typeD           -8.814e-01  2.875e-01  -3.066 0.002171 **
## typeE           -1.405e-01  2.348e-01  -0.598 0.549685
## operation1975-79  4.675e-01  1.352e-01   3.458 0.000545 ***
## service          3.062e-05  5.816e-06   5.265 1.40e-07 ***
```

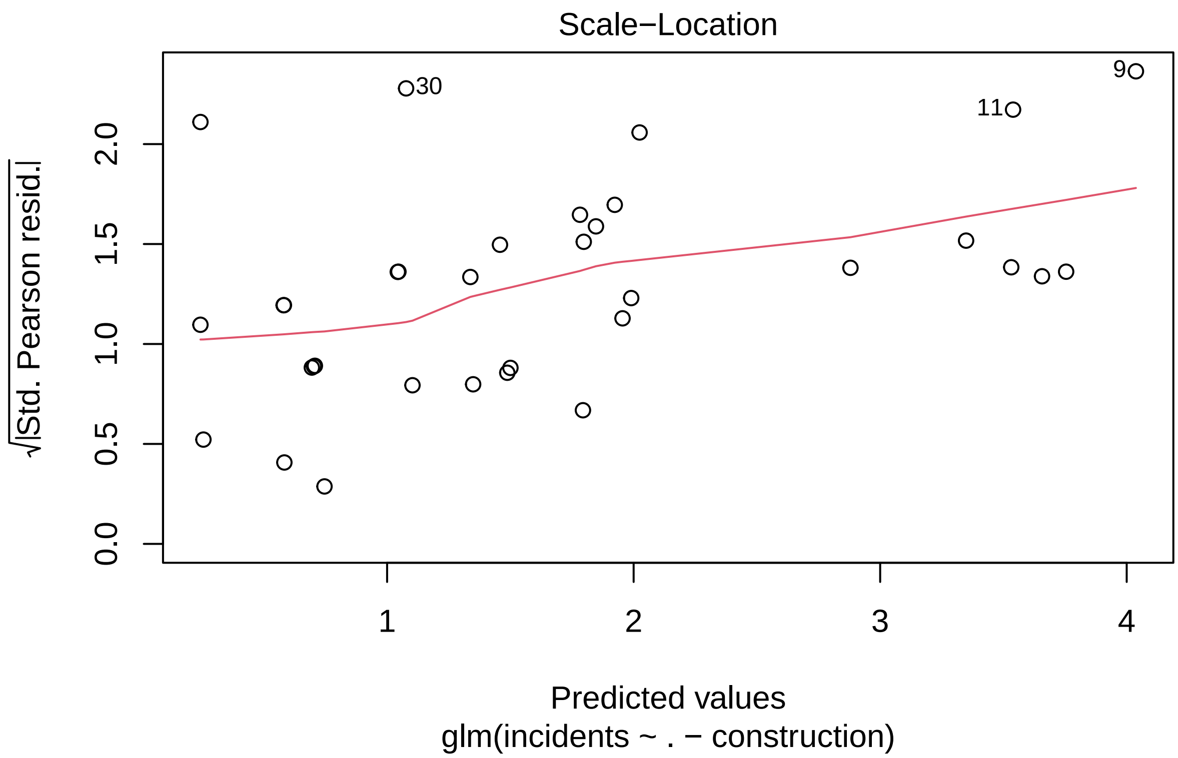
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 614.54  on 33  degrees of freedom
## Residual deviance: 140.82  on 27  degrees of freedom
## AIC: 252.69
##
## Number of Fisher Scoring iterations: 6
```

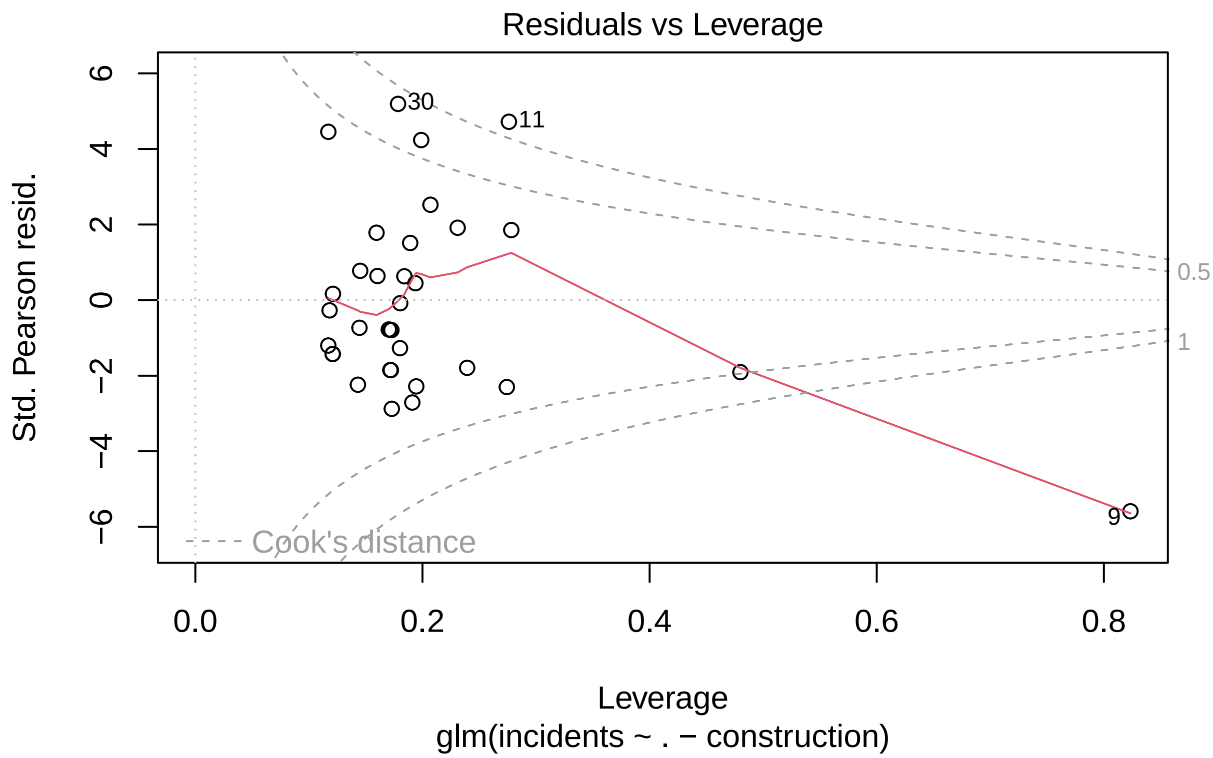
Residual plot looks good, but ship 9 has large leverage and high cook's distance, so it seems appropriate to exclude it.

```
plot(poisson_no_constr)
```




Normal Q−Q
Std. Pearson resid.
6
4
2
0
−2
−4
−6
30
11
9
−2
−1
0
1
2
Theoretical Quantiles
glm(incidents ~ . − construction)


Scale−Location
30
11
9
$\sqrt{|\text{Std. Pearson resid.}|}$
0.0
0.5
1.0
1.5
2.0
1
2
3
4
Predicted values
glm(incidents ~ . − construction)



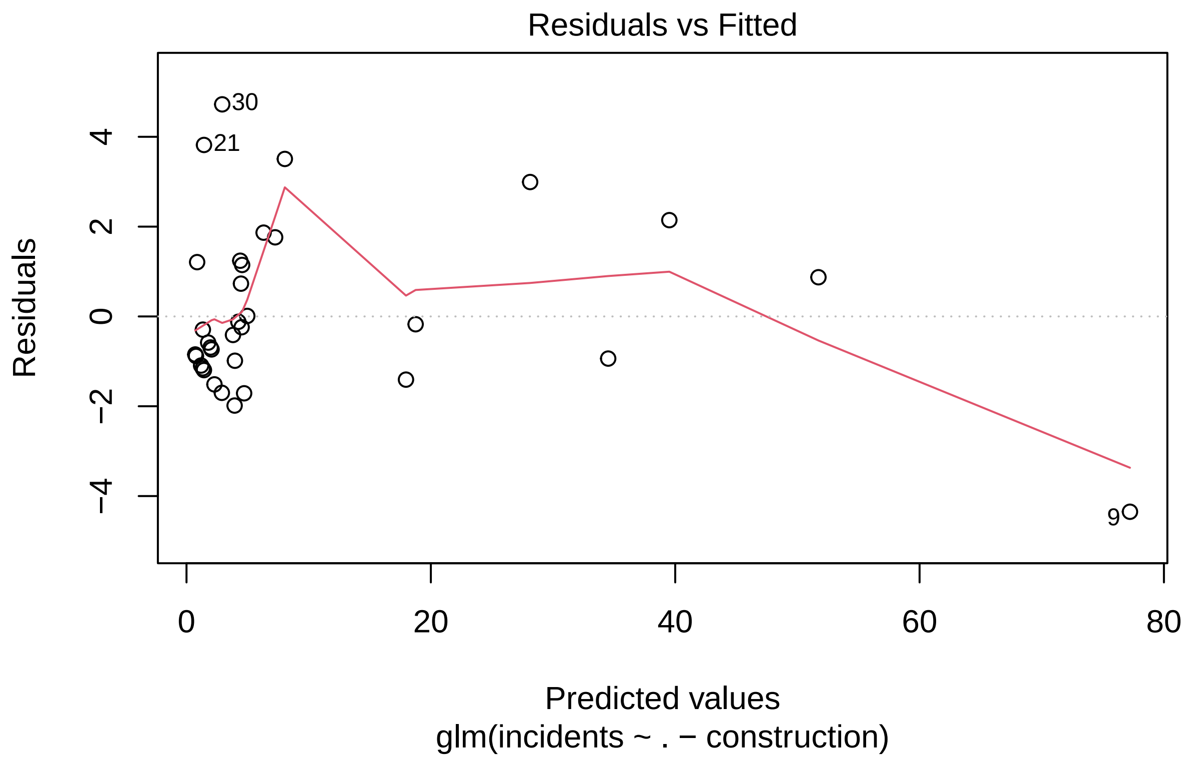

```r
# had to supply starting values to get code to run
poisson_ident_no_constr = glm(incidents ~ .-construction,
                              data=ShipAccidents, family=poisson(link='identity'),
                              start = coef(poisson_no_constr))
summary(poisson_ident_no_constr)
```

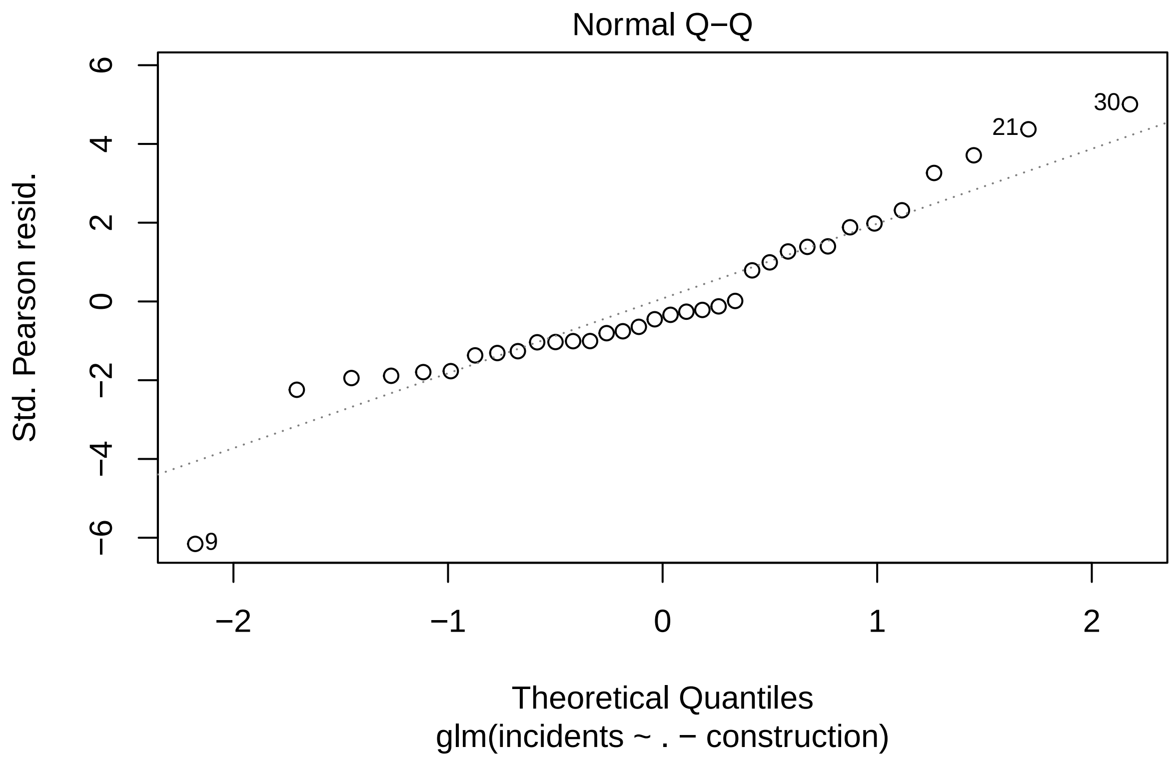
```
##
## Call:
## glm(formula = incidents ~ . - construction, family = poisson(link = "identity"),
##     data = ShipAccidents, start = coef(poisson_no_constr))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8124  -1.4350  -0.3684   1.0545   3.6051
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.0890008  0.8198472   2.548   0.0108 *
## typeB             4.8056877  3.4513794   1.392   0.1638
## typeC            -1.8844572  0.9048689  -2.083   0.0373 *
## typeD            -1.7645747  0.8650029  -2.040   0.0414 *
## typeE             1.0729280  1.1724683   0.915   0.3601
## operation1975-79  0.7067360  0.5635790   1.254   0.2098
## service           0.0015669  0.0001936   8.095 5.73e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

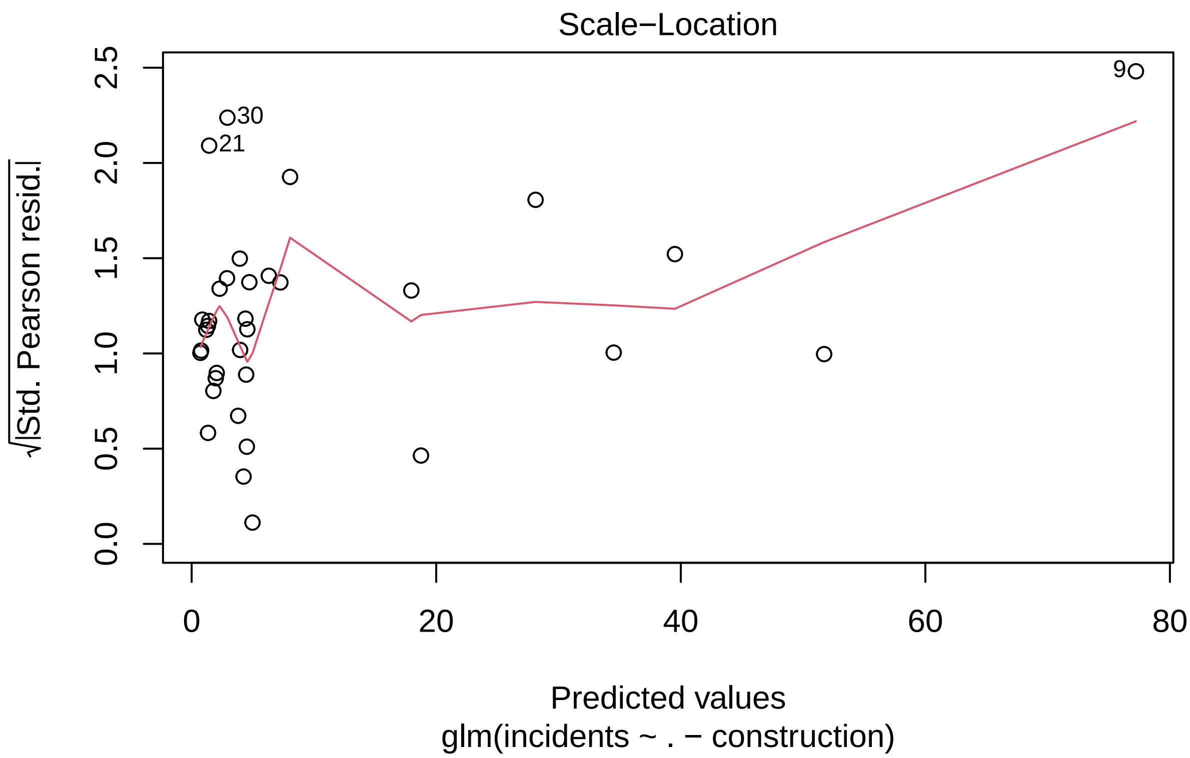
```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 614.54  on 33  degrees of freedom
## Residual deviance: 115.09  on 27  degrees of freedom
## AIC: 226.96
##
## Number of Fisher Scoring iterations: 13
```

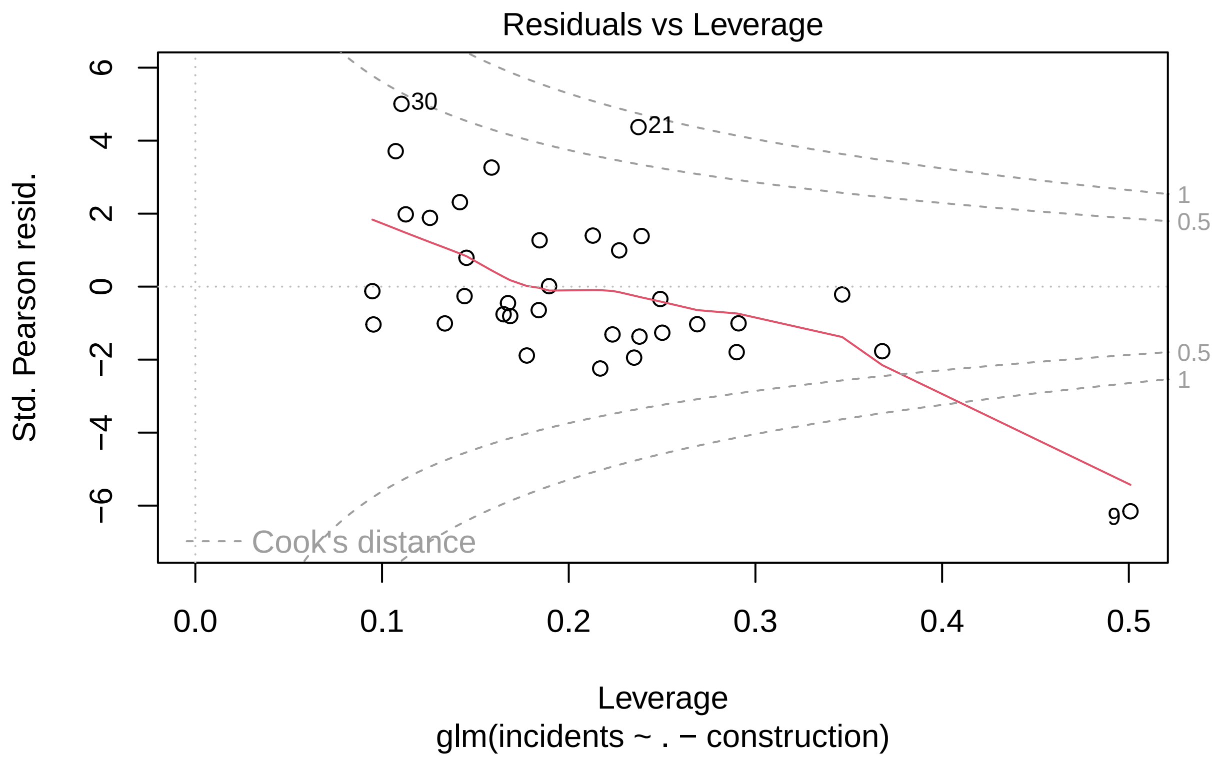
Ship 9 is an influential point in this model as well, so we will fit new models that exclude ship 9.

```
plot(poisson_ident_no_constr)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(incidents ~ . − construction)
30
21
9
−6
−4
−2
0
2
4
6
−2
−1
0
1
2


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
0
20
40
60
80
30
21
9
Predicted values
glm(incidents ~ . − construction)



Selecting the best model

The summary() function and tidy() function (from `broom`) both give information for performing inference on the coefficients for the regression model. Using the tidy() function lets us easily exponentiate the estimated coefficients for the model.

```
## recommended model (log link)
## take out ship 9
poisson_no_constr_no9 = glm(incidents ~ .-construction,
                            data=ShipAccidents[rownames(ShipAccidents)!=9,],
                            family=poisson())
summary(poisson_no_constr_no9)
```

```
##
## Call:
## glm(formula = incidents ~ . - construction, family = poisson(),
##     data = ShipAccidents[rownames(ShipAccidents) != 9, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5982  -1.8505  -0.2576   0.5790   3.5264
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)       1.365e+00  1.867e-01   7.311 2.66e-13 ***
## typeB             5.576e-01  2.512e-01   2.220 0.026442 *
## typeC            -1.207e+00  3.274e-01  -3.687 0.000227 ***
## typeD            -8.462e-01  2.876e-01  -2.943 0.003256 **
## typeE            -1.136e-01  2.348e-01  -0.484 0.628451
## operation1975-79  4.979e-01  1.376e-01   3.619 0.000295 ***
## service           7.550e-05  1.072e-05   7.043 1.88e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 566.61  on 32  degrees of freedom
## Residual deviance: 111.24  on 26  degrees of freedom
## AIC: 217.6
##
## Number of Fisher Scoring iterations: 6
```

We can define coefficients $\beta_B, \beta_C, \beta_D, \beta_E, \beta_{\text{75-79}}, \beta_{\text{service}}$ to be the "true" coefficients for the poisson regression. Since we are using a log link function, the model is estimated as

$$\begin{aligned}\widehat{\text{Accidents}} &= \hat{\beta}_0 + \hat{\beta}_B x_B + \hat{\beta}_C x_C + \hat{\beta}_D x_D + \hat{\beta}_E x_E + \hat{\beta}_{\text{75-79}} x_{\text{75-79}} + \hat{\beta}_{\text{service}} x_{\text{service}} \\ &= 1.36 + 0.558 x_B - 1.21 x_C - 0.846 x_D - 1.114 x_E + 0.498 x_{\text{75-79}} + 7.55 \times 10^{-5} x_{\text{service}}\end{aligned}$$

where $x_B, \ldots, x_E$ are indicator variables representing that a ship is of type B, C, D, or E (with type A as the reference level); $x_{\text{75-79}}$ is an indicator representing that the ship was in operation during 1975-1979 (with 1960-74 as the reference level); and x_{service} is a numerical variable representing the number of years that the ship was in service.

(See top of document for prediction and inference.)

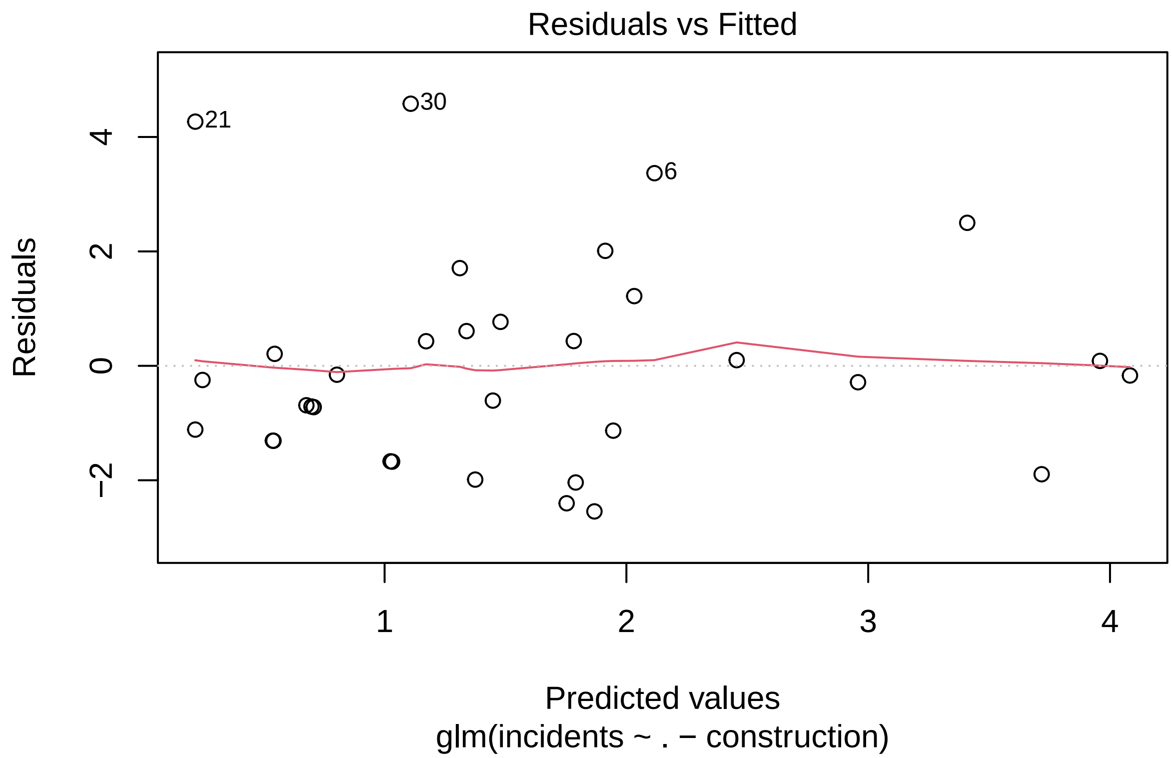
```
broom::tidy(poisson_no_constr_no9, exponentiate=TRUE)
```

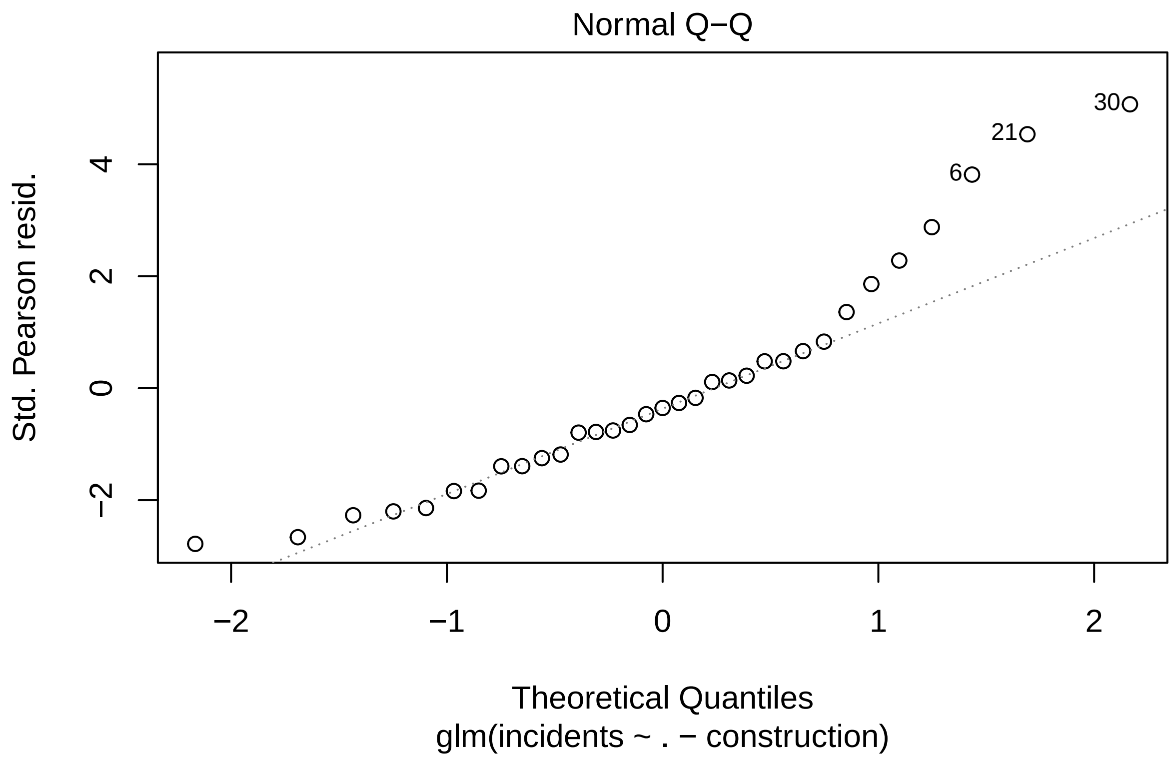
```
## # A tibble: 7 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         3.92  0.187         7.31  2.66e-13
## 2 typeB               1.75  0.251         2.22  2.64e- 2
## 3 typeC               0.299 0.327        -3.69  2.27e- 4
## 4 typeD               0.429 0.288        -2.94  3.26e- 3
## 5 typeE               0.893 0.235        -0.484 6.28e- 1
## 6 operation1975-79    1.65  0.138         3.62  2.95e- 4
## 7 service             1.00  0.0000107     7.04  1.88e-12
```

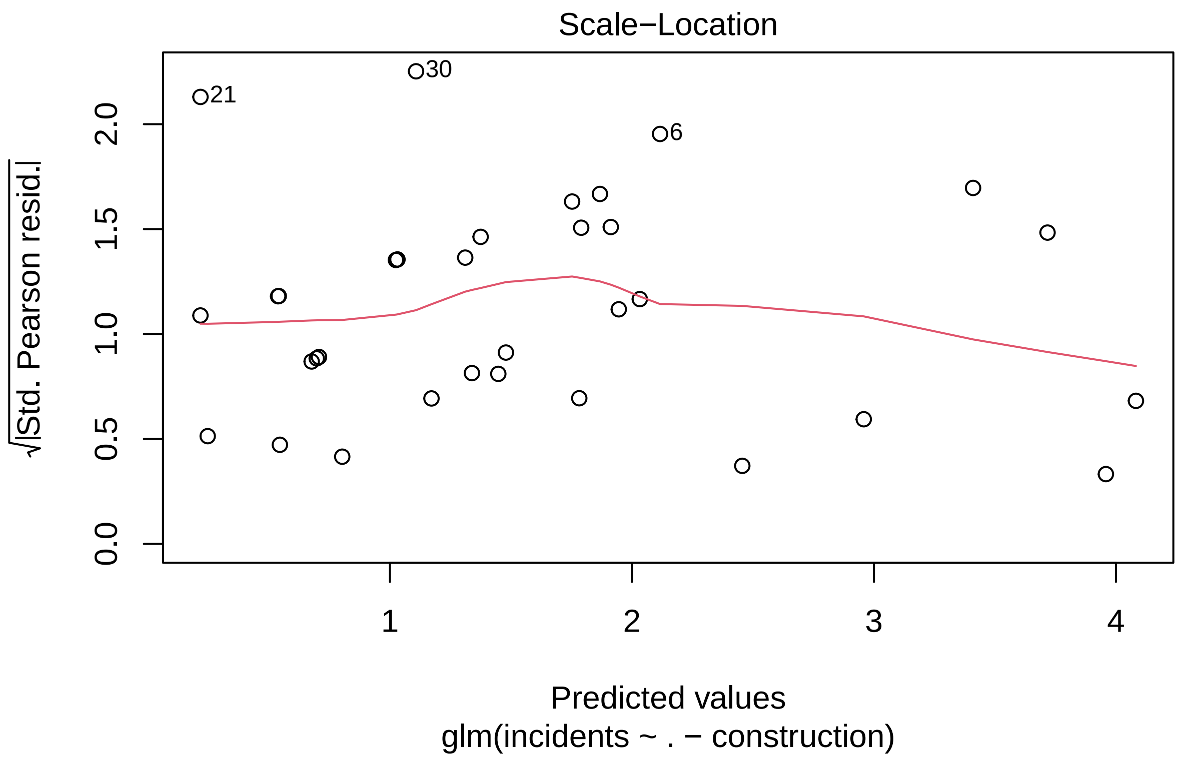
Diagnostic plots

The residual plot for the log link model still look good, and now the worst influential point has been removed, and the remaining influential points seem tolerable. Residual plot looks better for the log link than for the identity link. Some fitted values are much larger than others in the identity link model, and there is not a very good spread in the plot. However, the Q-Q plot for the identity link looks more even than that for the log link.

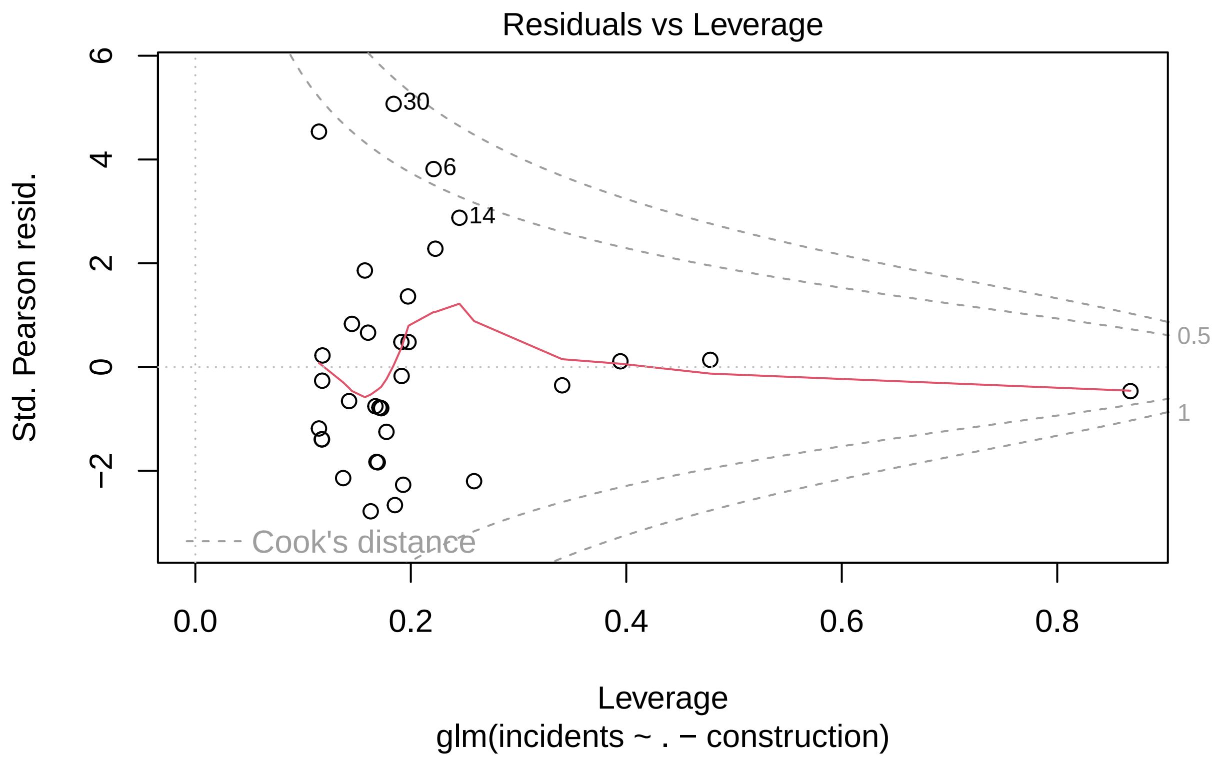
```
plot(poisson_no_constr_no9)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(incidents ~ . − construction)
30
21
6
−2
0
2
4
−2
−1
0
1
2



Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
21
30
6
1
2
3
4
Predicted values
glm(incidents ~ . − construction)

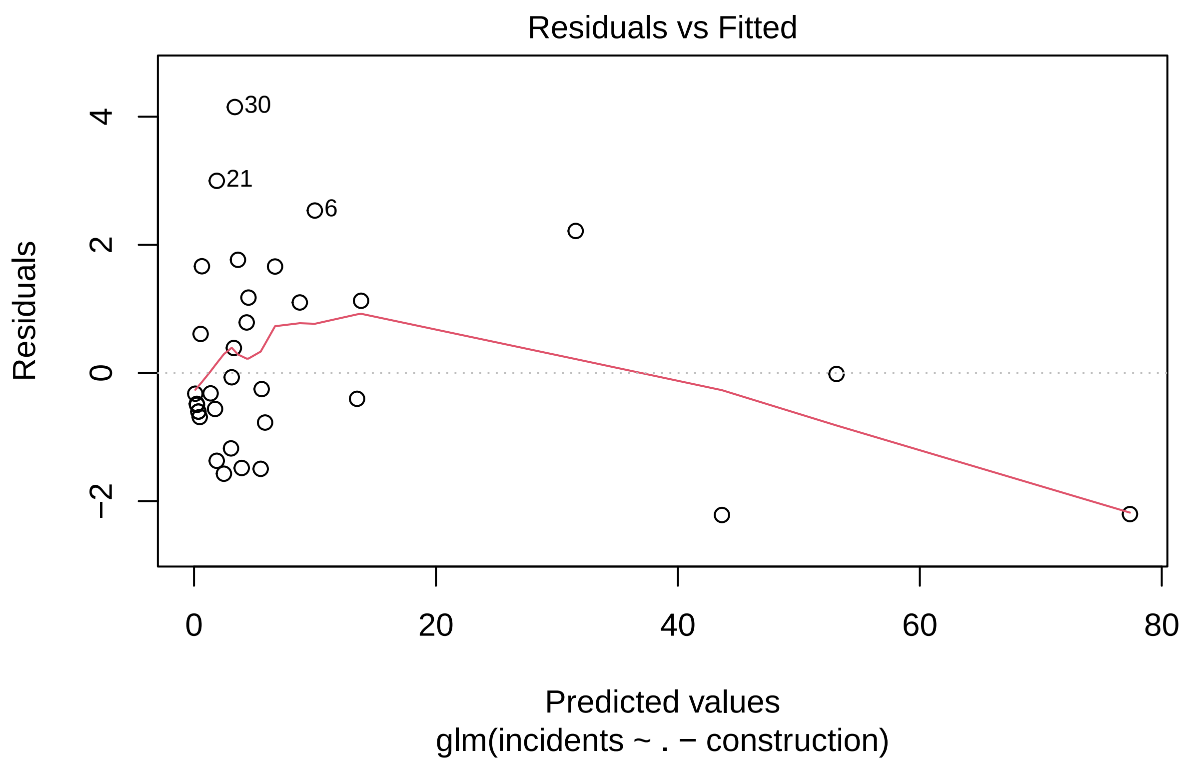


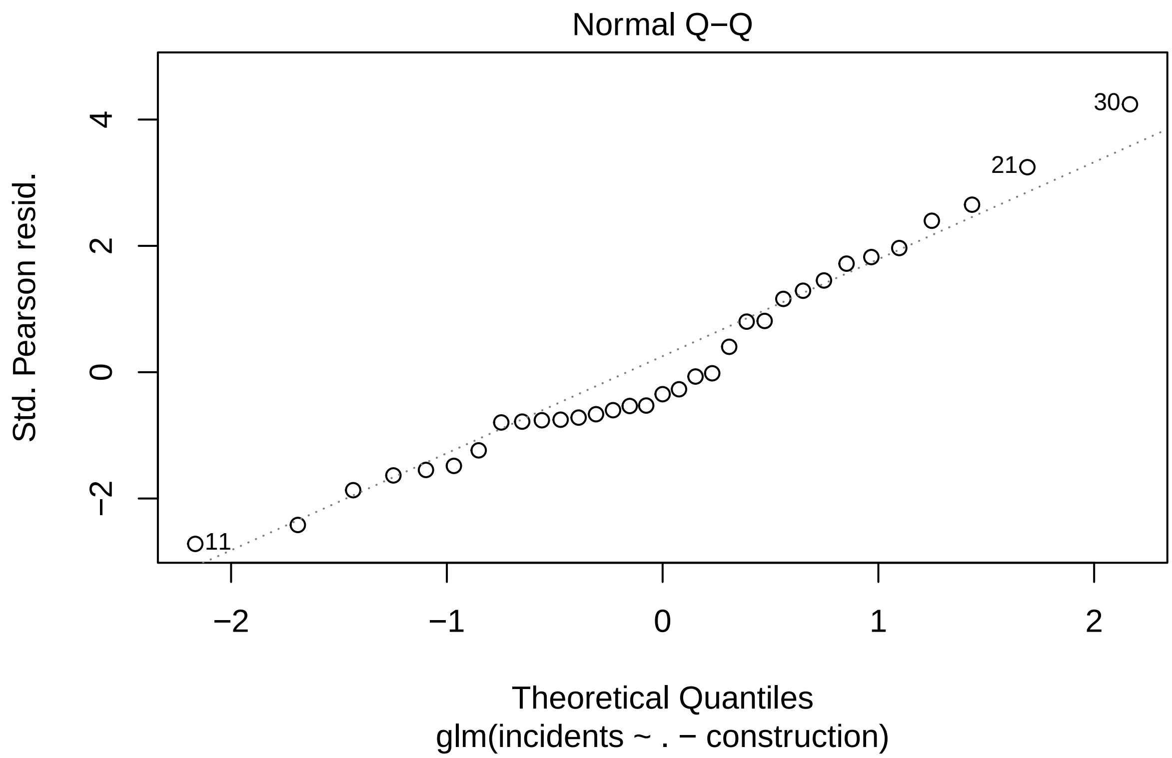

```r
# had to supply starting values to get code to run
poisson_ident_no_constr_no9 = glm(incidents ~ .-construction,
                                  data=ShipAccidents[rownames(ShipAccidents)!=9,],
                                  family=poisson(link='identity'),
                                  start = coef(poisson_no_constr),
                                  control = glm.control(maxit=50))
summary(poisson_ident_no_constr_no9)
```

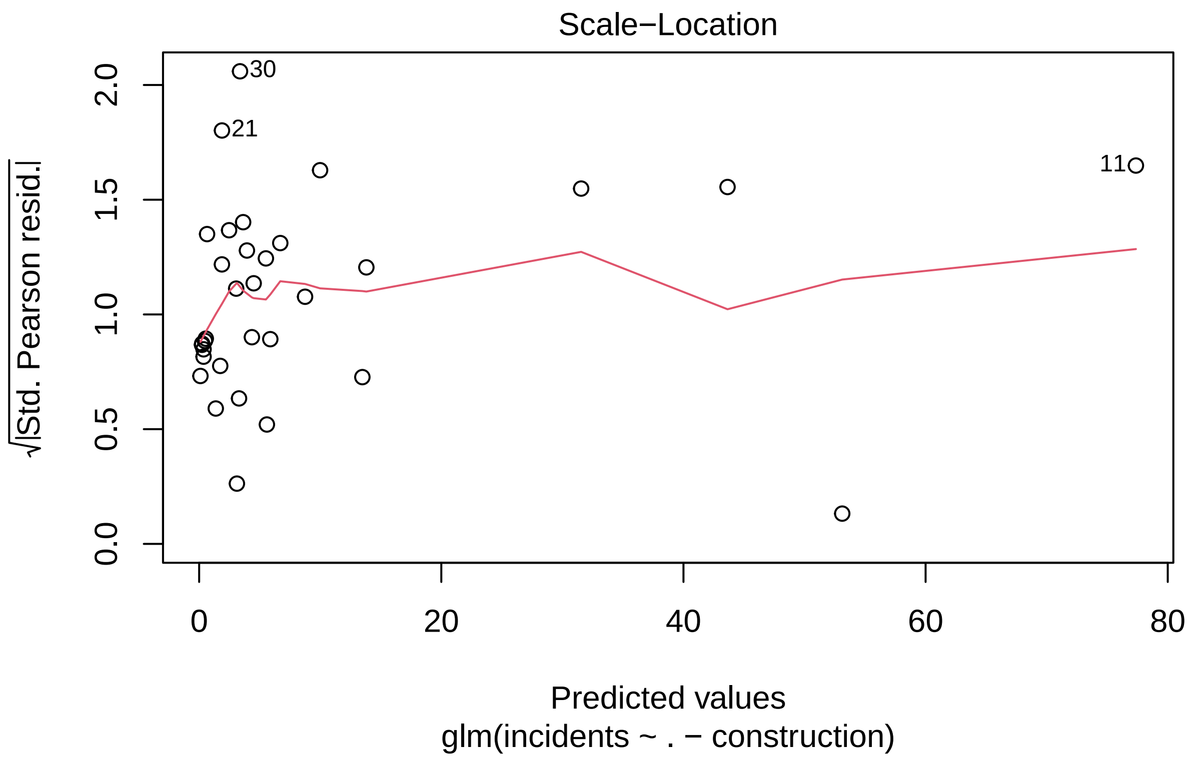
```
##
## Call:
## glm(formula = incidents ~ . - construction, family = poisson(link = "identity"),
##     data = ShipAccidents[rownames(ShipAccidents) != 9, ], start = coef(poisson_no_constr),
##     control = glm.control(maxit = 50))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3617  -0.8521  -0.3347   1.0762   3.2774
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.1289980  0.3831418  -0.337   0.7364
## typeB            -7.3390898  3.6319083  -2.021   0.0433 *
## typeC            -0.3091341  0.5772861  -0.535   0.5923
## typeD            -0.2522768  0.4012028  -0.629   0.5295
## typeE             2.2925637  0.9015085   2.543   0.0110 *
## operation1975-79  0.1742691  0.3560276   0.489   0.6245
```

```
## service           0.0029655  0.0002716  10.919   <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: 566.606  on 32  degrees of freedom
## Residual deviance:  69.462  on 26  degrees of freedom
## AIC: 175.82
##
## Number of Fisher Scoring iterations: 46
```

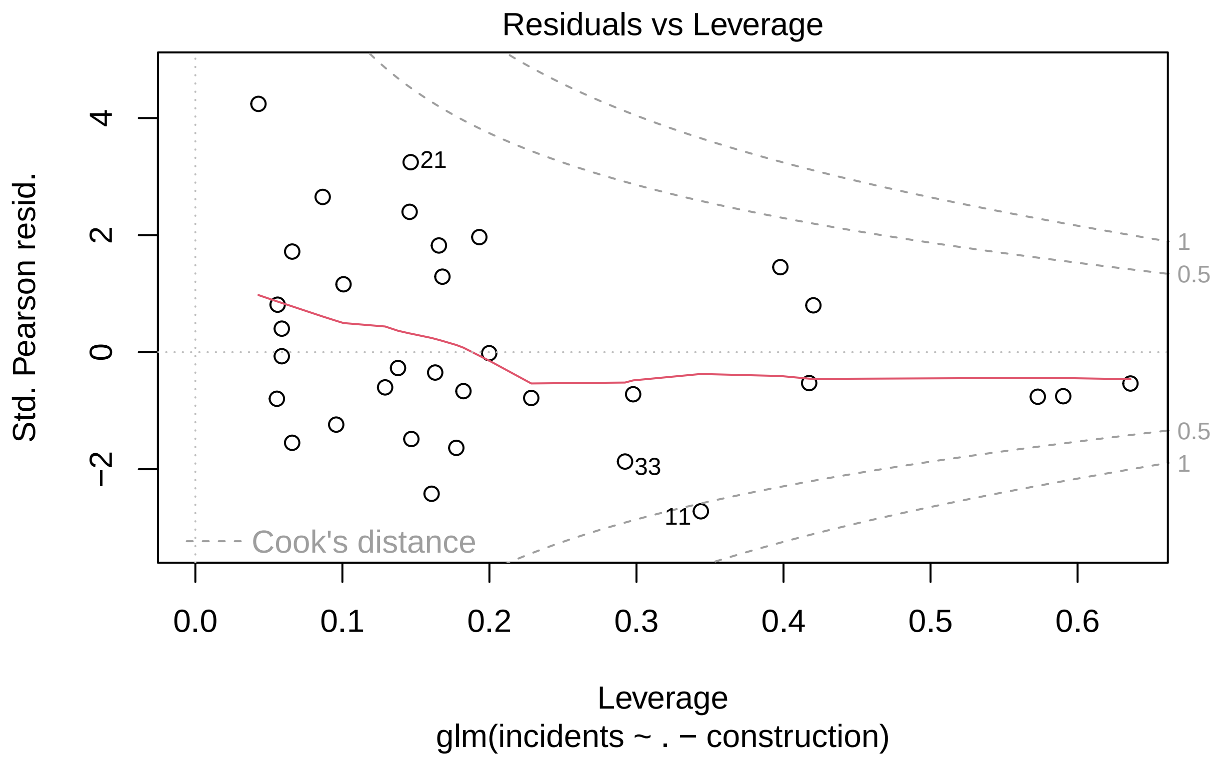
```
plot(poisson_ident_no_constr_no9)
```




Normal Q−Q
Std. Pearson resid.
4
2
0
−2
30
21
11
−2
−1
0
1
2
Theoretical Quantiles
glm(incidents ~ . − construction)



Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
30
21
11
0
20
40
60
80
Predicted values
glm(incidents ~ . − construction)



Predictions

Predicted number of incidents for every ship in the dataset.

```
(log_link_predictions = predict(poisson_no_constr_no9, type='response'))
```

```
##         1         2         3         4         5         6         8        10
##  3.953863  6.473586  4.253634  6.998127  4.389677  8.298822  7.632291 41.154606
##        11        12        13        14        16        17        18        19
## 59.300514 52.377210 11.658629 30.251227 19.257814  1.280121  2.008700  1.242229
##        20        21        22        24        25        26        27        28
##  2.027592  1.242416  2.231961  1.966980  1.712248  2.786131  1.717038  2.804491
##        29        30        32        33        35        36        37        38
##  1.724964  3.028073  3.227055  5.770457  3.710073  5.943783  3.814594  6.770005
##        40
##  5.991088
```

```
(ident_link_predictions = predict(poisson_ident_no_constr_no9, type='response'))
```

```
##          1          2          3          4          5          6          8
##  0.2476199  0.2320973  3.1182195  3.2924886  4.3548311  9.9885773  6.6998429
##         10         11         12         13         14         16         17
## 43.6415310 77.3717713 53.1133234 13.4801720 31.5512061 13.8116124  3.0581870
##         18         19         20         21         22         24         25
```

```
## 1.3730905 1.8779199 1.7408120 1.8838508 5.5129222 0.5486828 0.3630645
##        26        27        28        29        30        32        33
## 0.1043713 0.4727878 0.3623694 0.6536831 3.3753128 5.8752254 2.4712821
##        35        36        37        38        40
## 4.5033416 3.6337563 5.5946439 8.7462705 3.9451333
```

Cross validation and model comparison

Below: Poisson regression with the log link produced lower root mean squared error (RMSE) than the model with the identity link. With 10-fold cross validation, the identity link produces lower mean squared error. The residual diagnostic plots for the log link model look better than those for the identity link, but the Q-Q plot looks better for the model that uses the identity link function. Overall, since the log link model has more significant predictors than the identity link model, the log link seems preferable for these data.

```r
set.seed(1510)
### LOOCV

# R program to implement
# Leave one out cross validation

# defining training control
# as Leave One Out Cross Validation
train_control_loocv <- trainControl(method = "LOOCV")
```

```r
# training the model by assigning sales column
# as target variable and rest other column
# as independent variable
log_loocv <- train(incidents ~ .-construction,
               data=ShipAccidents[rownames(ShipAccidents)!=9,],
               method='glm',
               family=poisson(),
               trControl = train_control_loocv)

# printing model performance metrics
# along with other details
log_loocv
```

```
## Generalized Linear Model
##
## 33 samples
##  4 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 32, 32, 32, 32, 32, 32, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   6.275326  0.8387806  4.35623
```

```r
## THE GLM ALGORITHM ENCOUNTERS ISSUES SOMETIMES
ident_loocv <- train(incidents ~ .-construction,
```

```r
                 data=ShipAccidents[rownames(ShipAccidents)!=9,],
                 method='glm',
                 family=poisson('identity'),
                 start = coef(poisson_no_constr),
                 control = glm.control(epsilon = 1e-4, maxit=100),
                 trControl = train_control_loocv)
```

```
## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: glm.fit: algorithm stopped at boundary value

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds
```

```
## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: glm.fit: algorithm stopped at boundary value
```

```
ident_loocv
```

```
## Generalized Linear Model
##
## 33 samples
##  4 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 32, 32, 32, 32, 32, 32, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7.377884  0.8508288  4.278508
```

10-fold cross validation

```
train_control_10fold = trainControl(method='cv', number=10)
```

```
log_10fold <- train(incidents ~ .-construction,
               data=ShipAccidents[rownames(ShipAccidents)!=9,],
               method='glm',
               family=poisson(),
               trControl = train_control_10fold)
log_10fold
```

```
## Generalized Linear Model
##
## 33 samples
##  4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 30, 30, 29, 30, 30, 30, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   6.035802  0.7150489  4.720531
```

```
ident_loocv <- train(incidents ~ .-construction,
               data=ShipAccidents[rownames(ShipAccidents)!=9,],
               method='glm',
               family=poisson('identity'),
               start = coef(poisson_no_constr),
               control = glm.control(epsilon = 1e-4, maxit=100),
               trControl = train_control_10fold)
```

```
## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: glm.fit: algorithm stopped at boundary value
```

```
## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: step size truncated: out of bounds

## Warning: glm.fit: algorithm stopped at boundary value
```

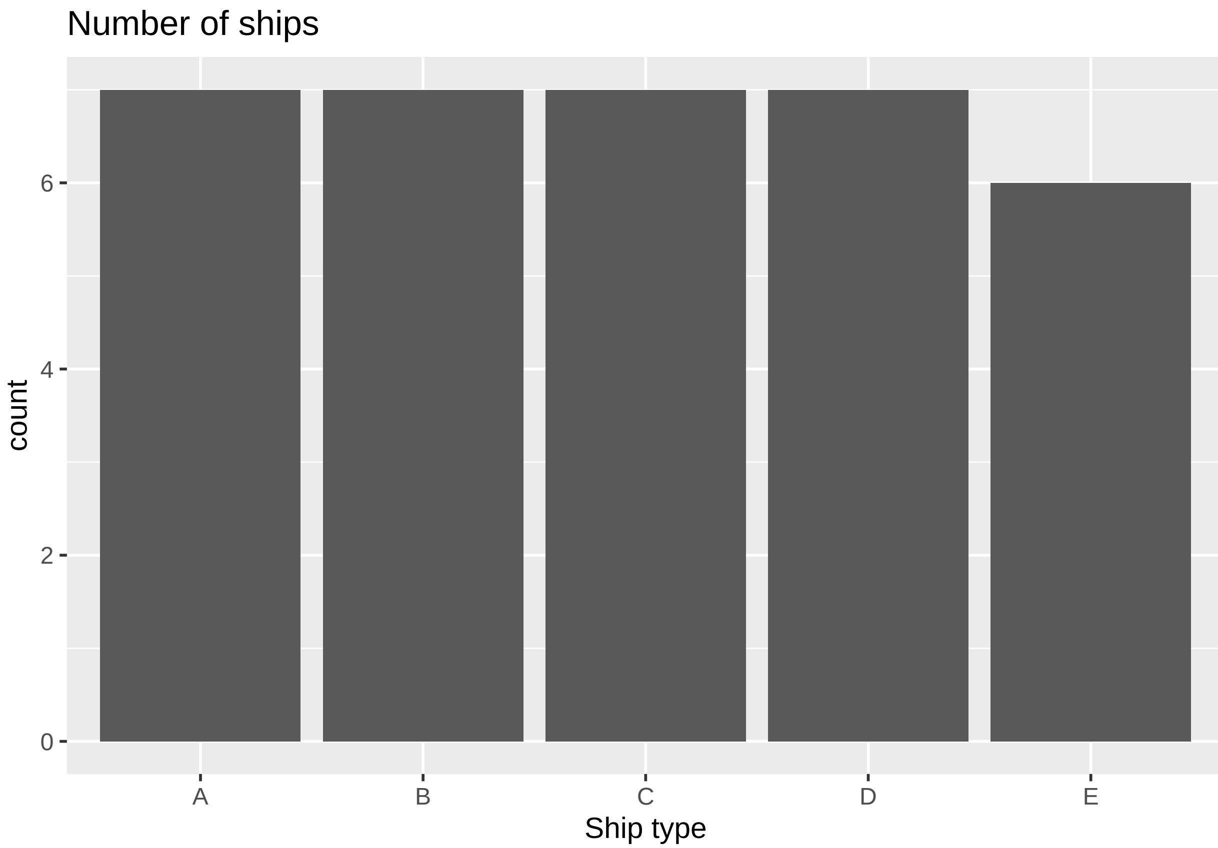
```
ident_loocv
```

```
## Generalized Linear Model
##
## 33 samples
##  4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 29, 30, 29, 30, 30, 31, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   5.350599  0.8739577  4.174536
```

Visualizations for data

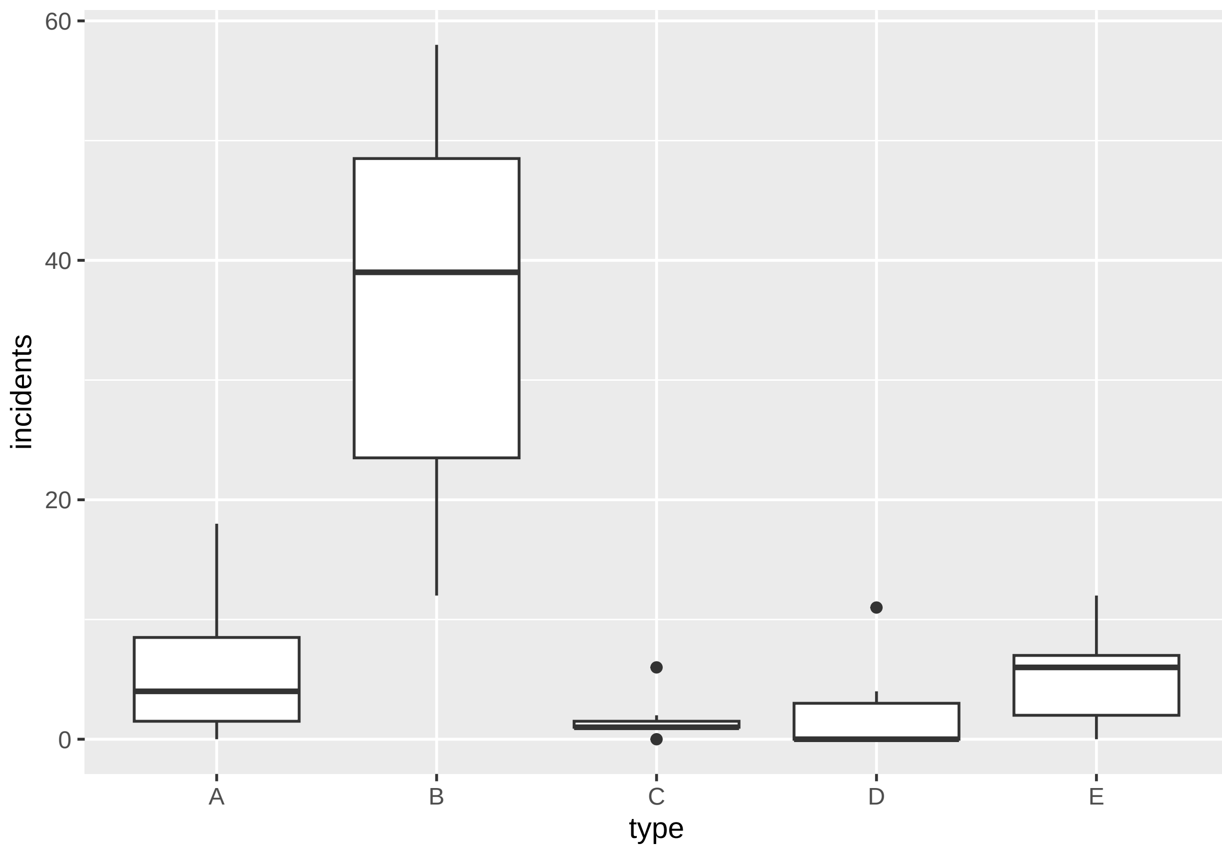
Plot below shows the number of ships of each type in the dataset. There are seven ships of each type, except for Type E, which has 6 ships.

```
ggplot(ShipAccidents, aes(x= type)) +
  geom_bar() +
  labs(title = 'Number of ships',
       x = 'Ship type')
```



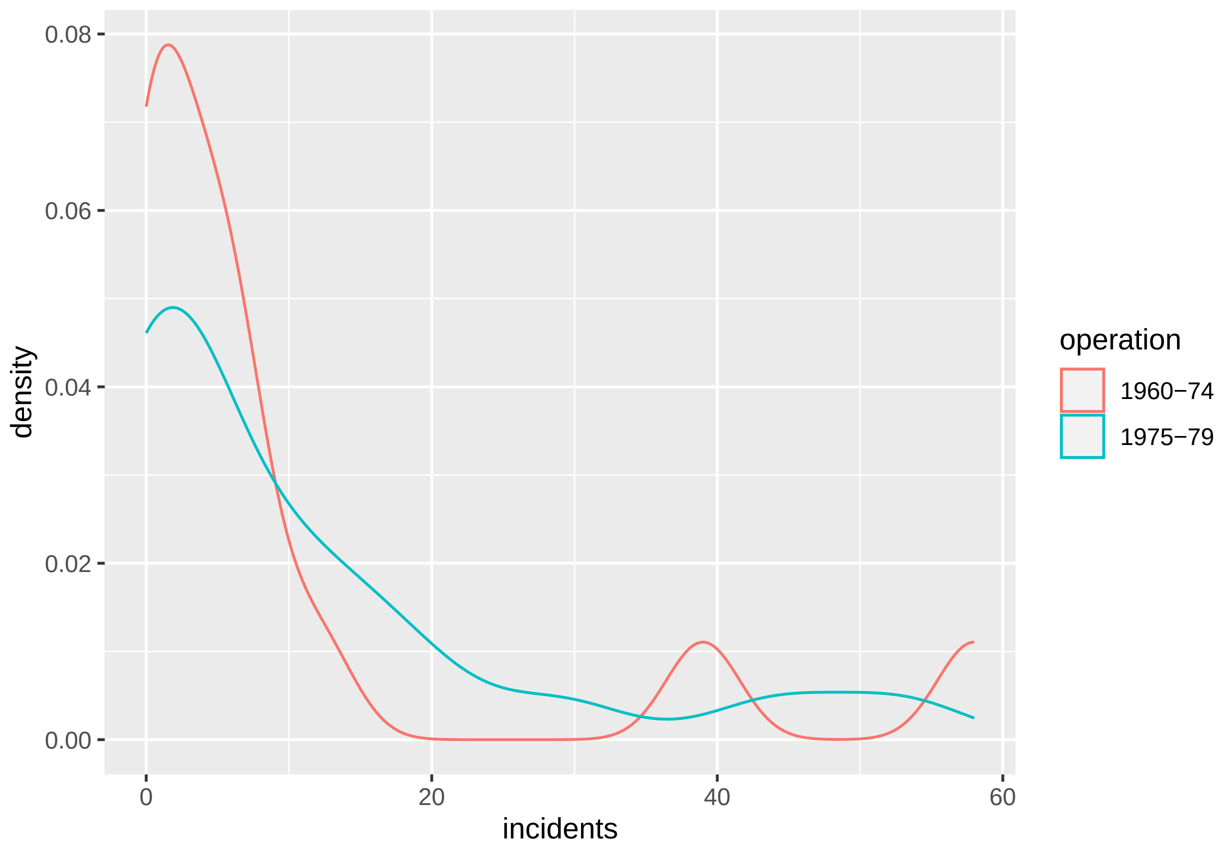
This boxplot shows the number of incidents by ship type. Type B had the greatest number of incidents by a wide margin, with most ships experiencing around 23-48 incidents. and Type C had the least. Many ships of Type D appear to have no incidents. Ships of Types A and E experienced moderate number of incidents, with a median around 3-6.

```
ggplot(ShipAccidents, aes(x= type, y=incidents)) +
  geom_boxplot()
```



Ships in operation between 1975 and 1979 tended to experience more incidents than those in operation between 1960 and 1974, although there were still a few ships in the earlier group experiencing a large number of incidents.

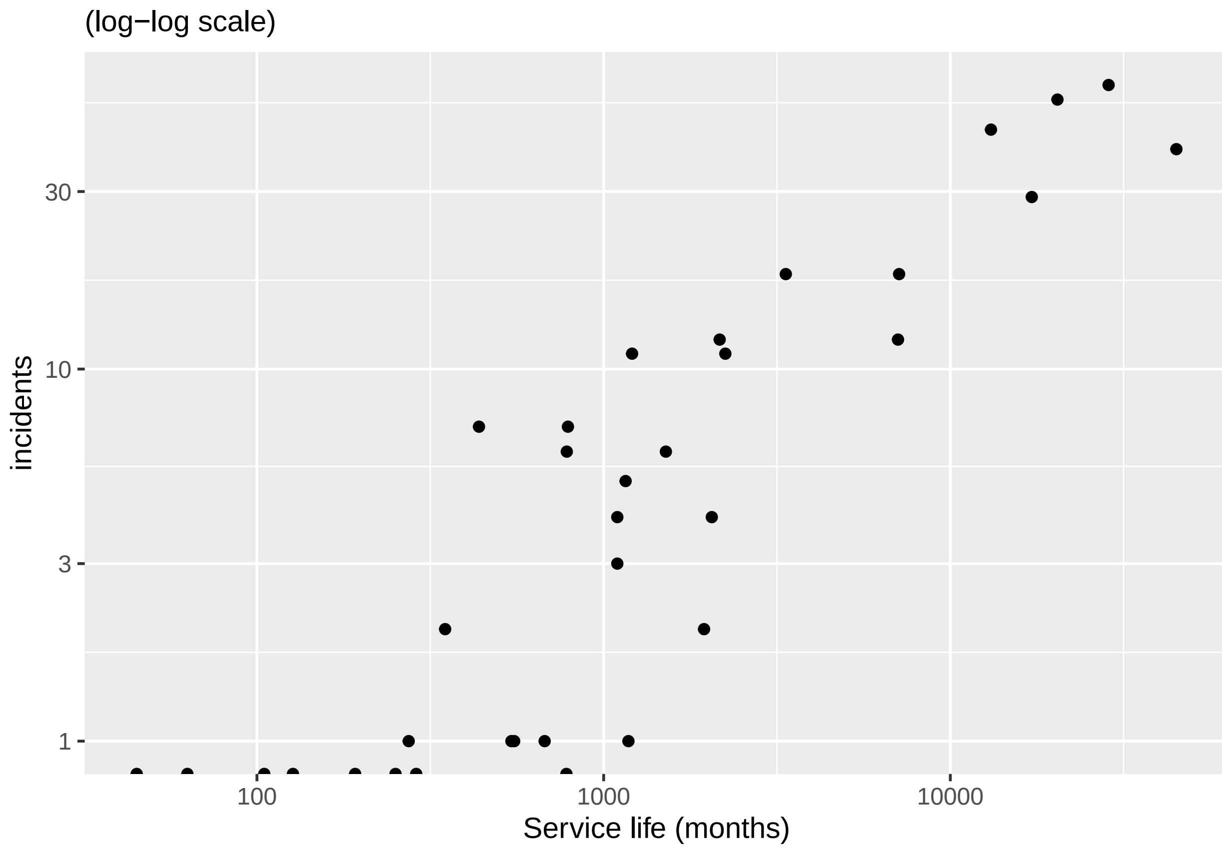
```
ggplot(ShipAccidents, aes(color= operation, x=incidents)) +
  geom_density()
```



The number of months a ship was in service has approximately a linear relationship to the number of incidents the ship was involved in, although there were several ships involved in no incidents at all.

```
ggplot(ShipAccidents, aes(x= service, y=incidents)) +
  geom_point() +
  scale_y_continuous(trans='log10') +
  scale_x_continuous(trans='log10') +
  labs(subtitle = '(log-log scale)',
       x = 'Service life (months)')
```

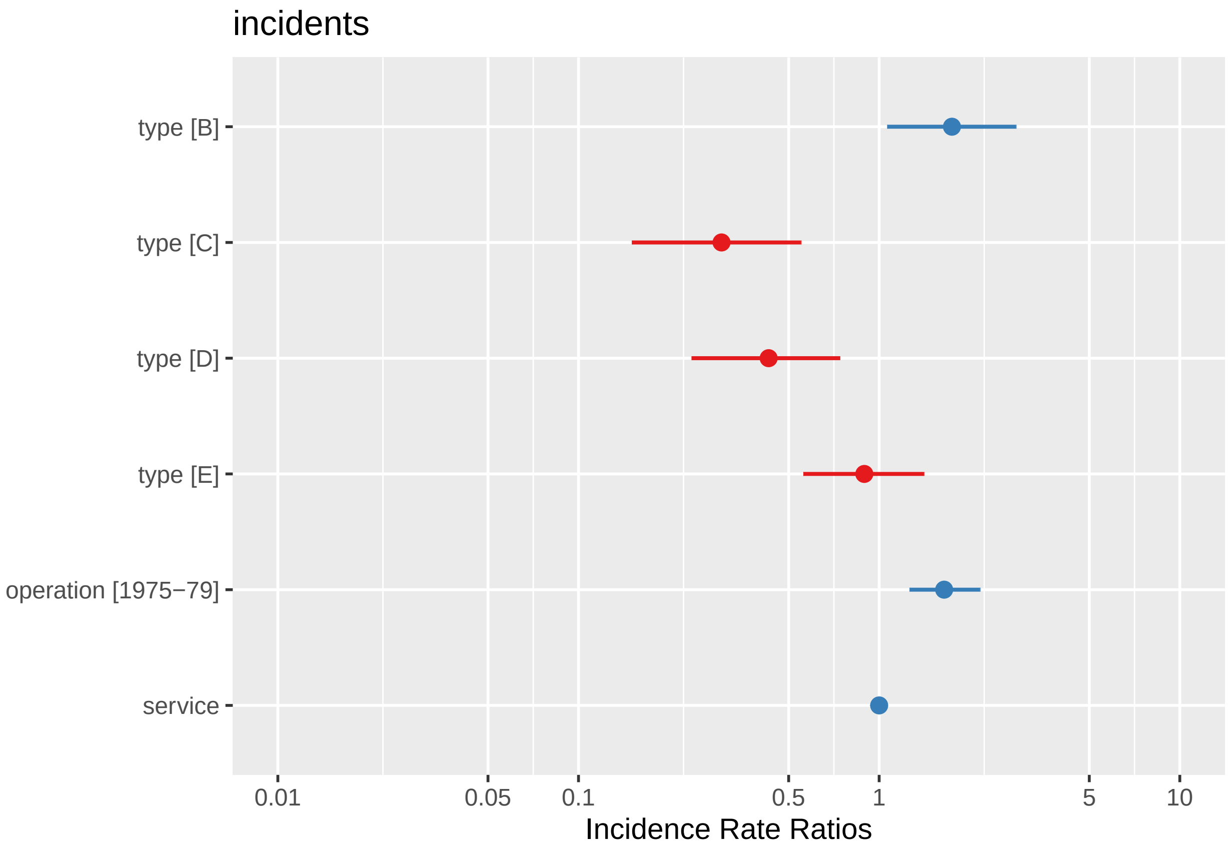
```
## Warning: Transformation introduced infinite values in continuous y-axis
```



This displays the estimated exponentiated coefficients of each of the predictor variables, along with their confidence intervals. As discussed before coefficients are statistically significant, except for the coefficient for Ship Type E. Type B experienced more incidents than Type A, and the rest of the ship types experienced fewer incidents. Ships in operation during the later period experienced more incidents than ships in operation during the earlier period.

In summary, ship type, operation year, and service duration all have a meaningful relationship to the number of incidents a ship was involved in.

```
plot_model(poisson_no_constr_no9)
```

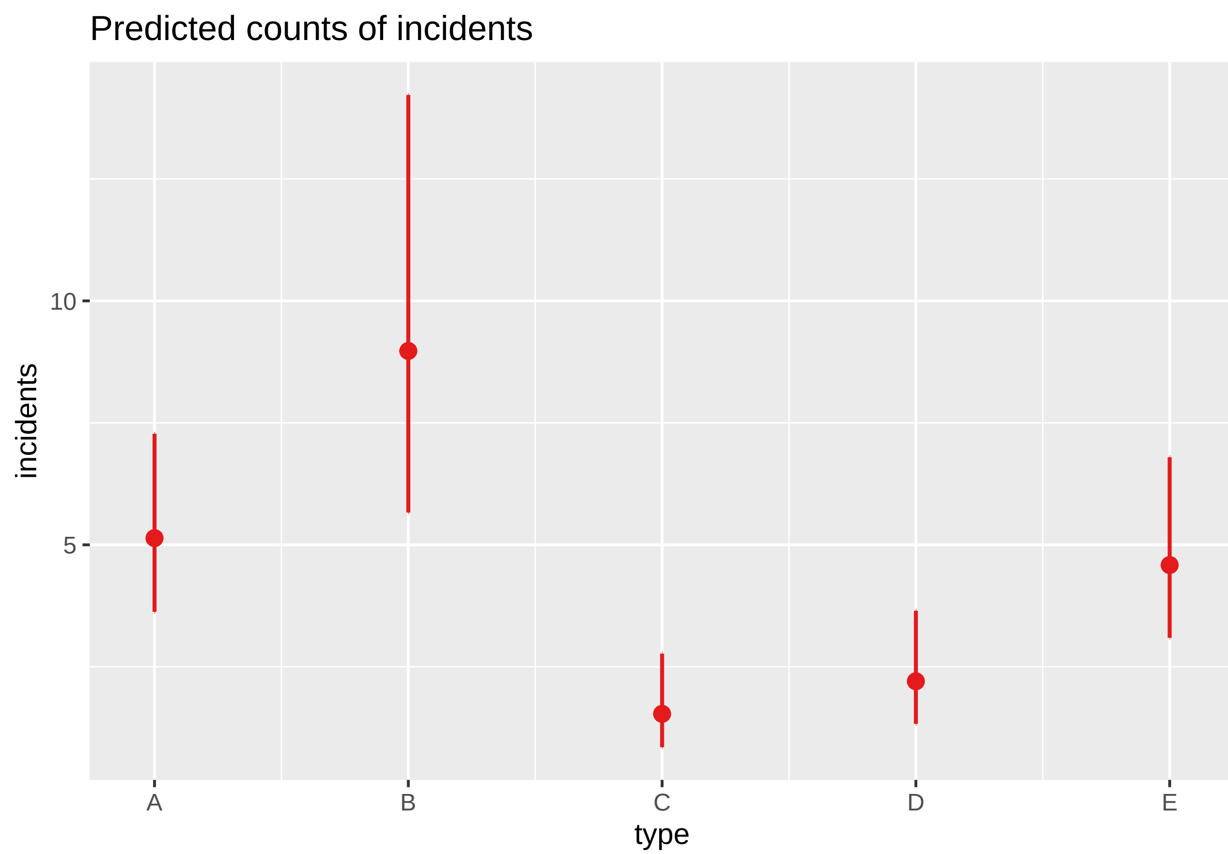


These plots show predicted number of incidents, based on the final log-link model, relative to predictor variables. (The plot of incidents relative to construction is meaningless.)

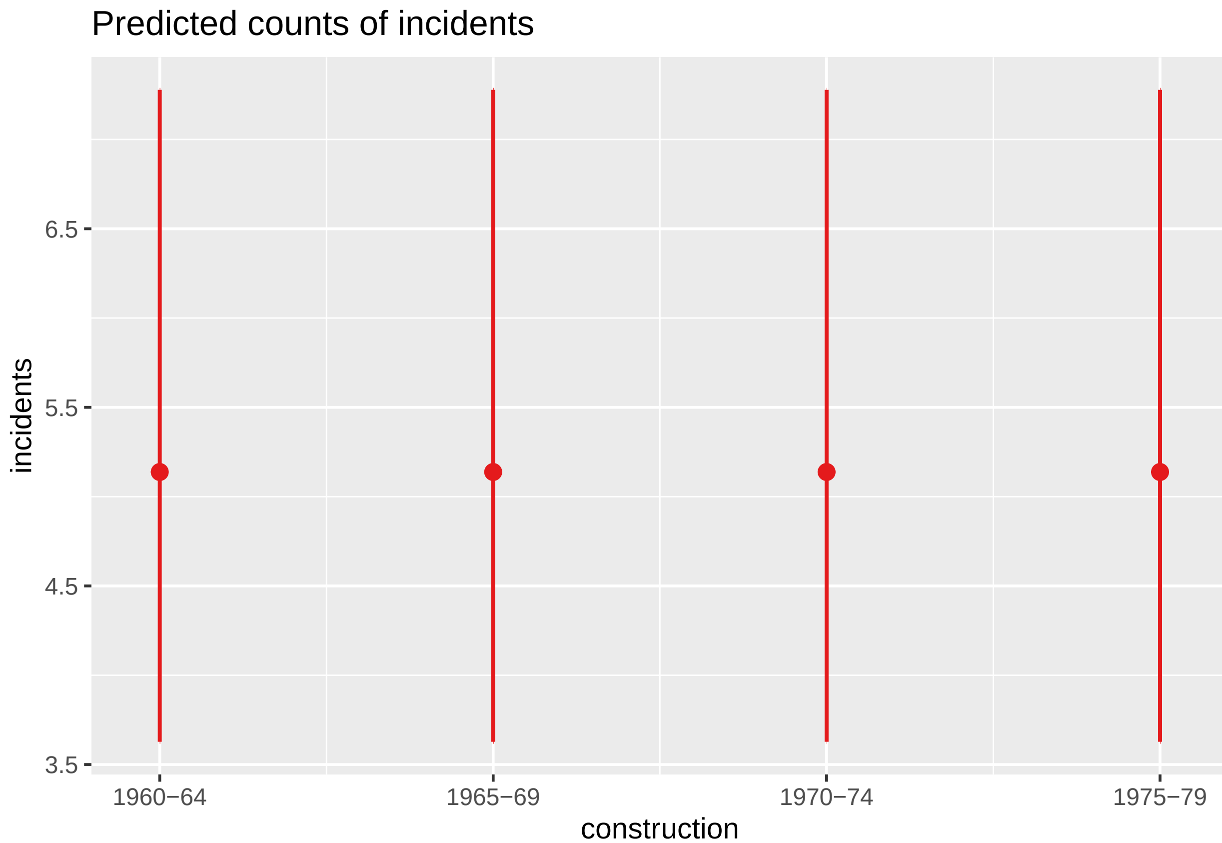
```
plot_model(poisson_no_constr_no9, type='eff')
```

```
## Package ‘effects‘ is not available, but needed for ‘ggeffect()‘. Either
##   install package ‘effects‘, or use ‘ggpredict()‘. Calling ‘ggpredict()‘
##   now.
```

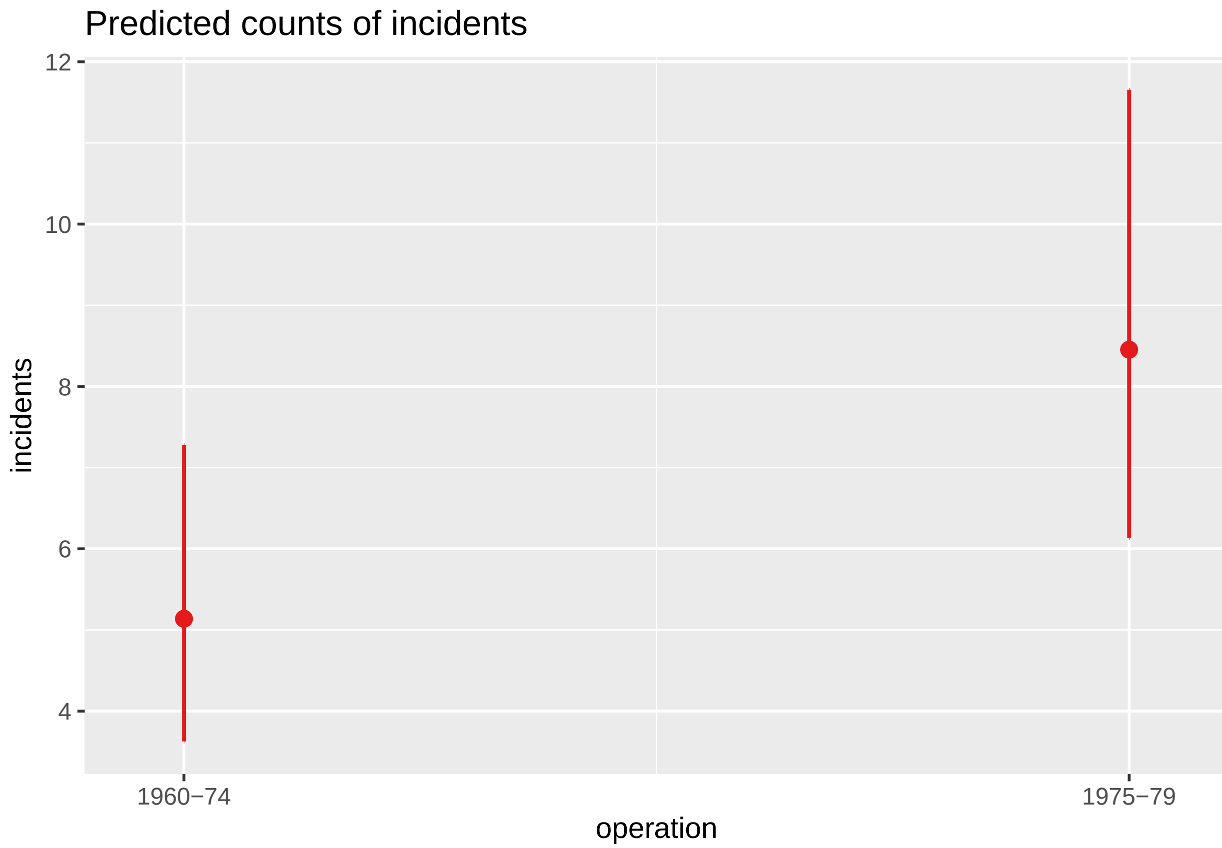
```
## $type
```



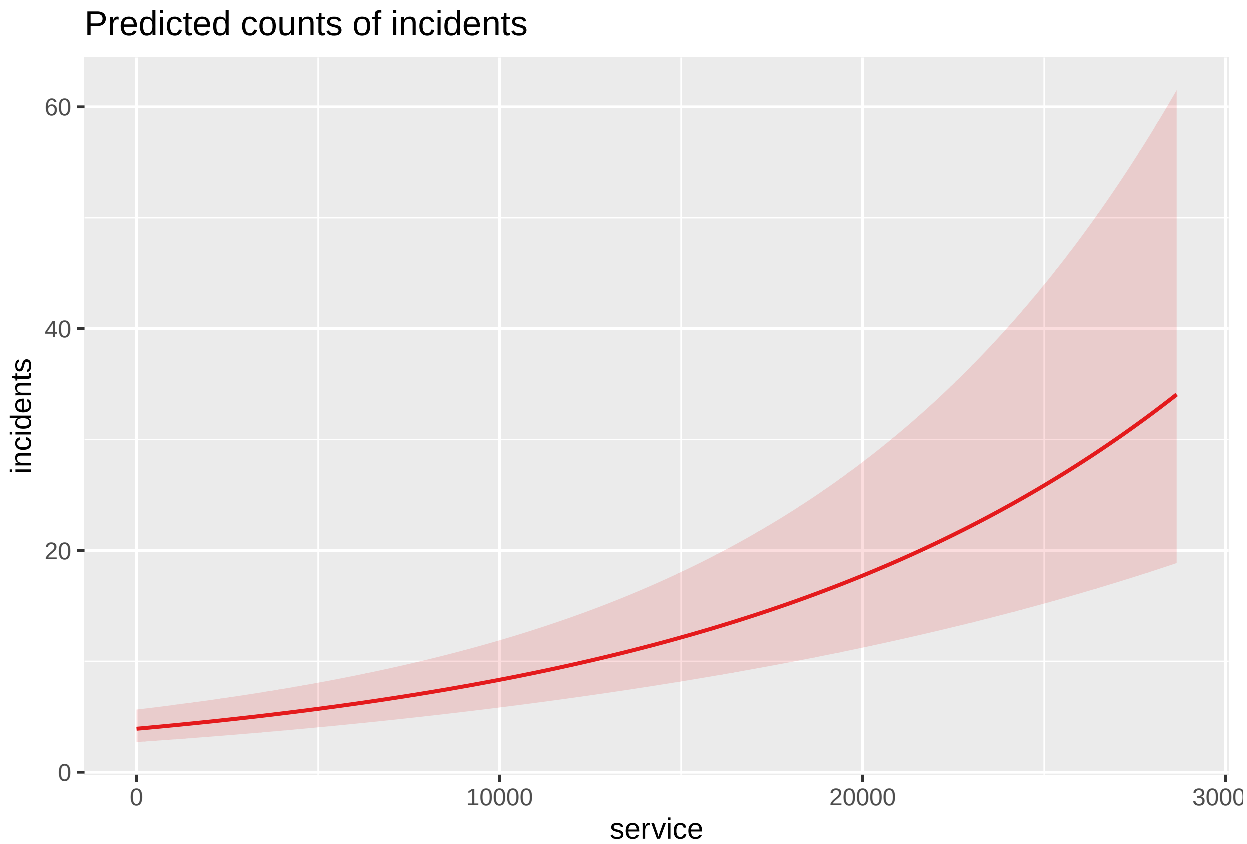

```
##
## $construction
```


Predicted counts of incidents
6.5
5.5
4.5
3.5
incidents
1960−64
1965−69
1970−74
1975−79
construction

```
##
## $operation
```


Predicted counts of incidents
12
10
8
6
4
incidents
1960−74
1975−79
operation

```
##
## $service
```


Predicted counts of incidents
incidents
60
40
20
0
0
10000
20000
3000
service