Linear and Generalized Linear Models

Ligaya Breemer & Iris Hoekstra

30/12/2020

Abstract

This report contains our case study about the ShipAccidents dataset. We performed four analyses to try and explain the number of damage incidents from ship type, construction period, and operation period, using the aggregate number of service months as an offset. We reduced the number of factor levels of the construction variable and removed some influential points from our dataset in order to find the best model. We found that a quasi-Poisson model including type, construction period, and their interaction as predictors, and using service months as an offset, was the best.

All members of the case study group contributed equally.

Introduction

This report presents an analysis of a dataset of ship accidents. The dataset contains information about the number of damage incidents and aggregate number of service months for different kinds of ships. The ships are categorized according to three factors: the type of ship with 5 levels; the period in which the ship was constructed, with 4 levels; and the period during which the kind of ship operated, with 2 levels. The number of damage incidents will be described in terms of type, construction period, and operation period, and months of service was used as an offset variable for the number of incidents.

Exploring the data

After the initial preparation of the dataset, it appeared that some ships were constructed *after* their operation period. This meant that those ships were not in service and thus did not have any incidents. It is unclear what this information meant in a practical sense, so these cases were left out of the analysis. To describe the data, some descriptives and frequencies were evaluated. When looking at the summary of our data we see that both of the numeric variables, incidents and service, have a noticeable difference between their mean and median, indicating that they are not symmetrically distributed. This was of no surprise, since both variables represent counts, which are generally Poisson distributed.

type	construction	operation	service	incidents
A:7	1960-64: 8	1960-74:14	Min. : 45	Min. : 0.00
B:7	1965-69:10	1975-79:20	1st Qu.: 371	1st Qu.: 1.00
C:7	1970-74:10		Median : 1095	Median : 4.00
D:7	1975-79: 6		Mean : 4811	Mean :10.47
E:6			3rd Qu.: 2223	3rd Qu.:11.75
			Max. :44882	Max. :58.00

The frequency table (Table 1) show that the design is unbalanced and quite a couple of cells are empty (necessarily so because some ships were built after 1974 and could not operate before then). This means that Type II or Type III tests needed to be used for our model rather than Type I tests, in this case Likelihood Ratio Tests. We will elaborate on this later.

Table 1: Frequencies of construction period and ship type by operation period

	Operation 1960-1974				Operation 1975-1979			
	1960-64	1965-69	1970-74	1975-79	1960-64	1965-69	1970-74	1975-79
A	1	1	1	0	1	1	1	1
B	1	1	1	0	1	1	1	1
C	1	1	1	0	1	1	1	1
D	1	1	1	0	1	1	1	1
E	0	1	1	0	0	1	1	2

To further investigate the distributions of the variables, some graphical displays were made. While taking a look at these figures, it is important to keep in mind that the total sample size and the group sizes were low. Because of this, individual cases may have a great influence on the distribution of the data and many points may be considered outliers by the measures used to create these plots. We started by looking at the histogram (Figure 1), which presents the number of damage incidents and suggests that the incidents are indeed Poisson distributed.

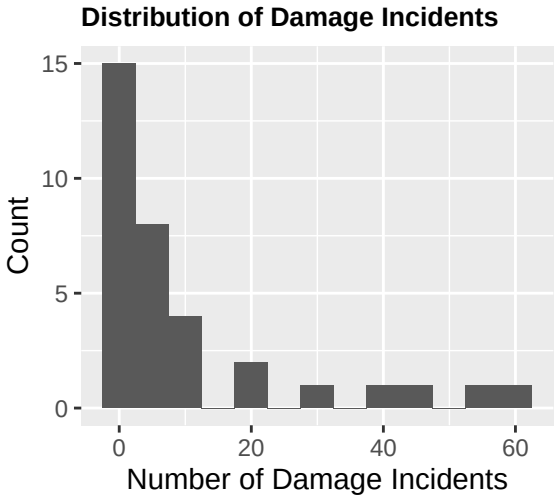


Figure 1: Histogram presenting the distribution of damage incidents

The next two figures show that ships of Type B generally have more months of service and more damage incidents. The differences between the other types are much smaller. The boxplot (Figure 2) indicates there might be a slight difference both in medians and variance between the groups. We found it interesting to look at these differences in the absence of Type B, since the difference in scale between Type B and the other types made it harder to spot differences on a smaller scale.

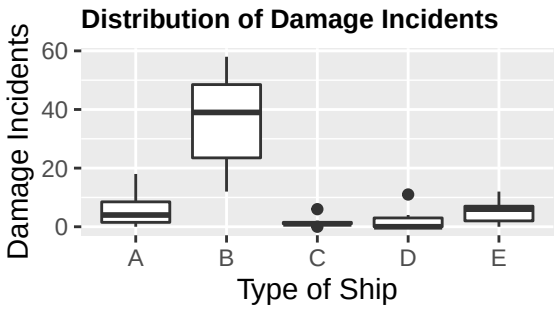


Figure 2: Distribution of damage incidents for each type of ship

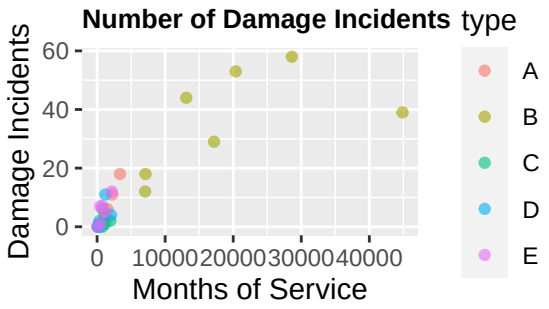


Figure 3: Number of damage incidents by months of service

Since type B has such large values, we removed type B to get a closer look at the other ships, see Figure 4. We observed that ships of type A and E have a higher median and more variation than ships of type C or D. Furthermore, the updated scatterplot (Figure 5) gives us a clearer image of the degree of variation between the smaller values.

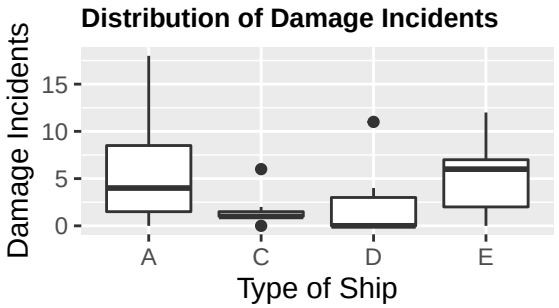


Figure 4: Distribution of damage incidents for each type of ship, excluding type B.

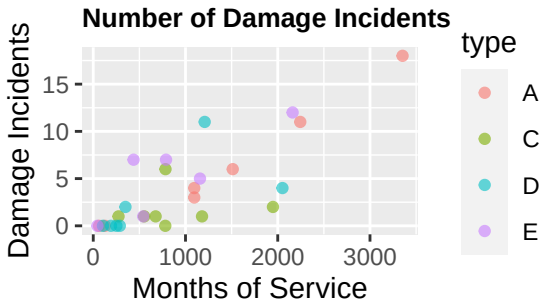


Figure 5: Number of damage incidents by months of service and ship type, excluding type B.

Next, the relationship between construction period and incidents was graphically investigated. Figure 6 shows some indication of differences in distribution of damage incidents for the different construction periods. Particularly, there seem to be some differences in both medians (skewness) and dispersion.

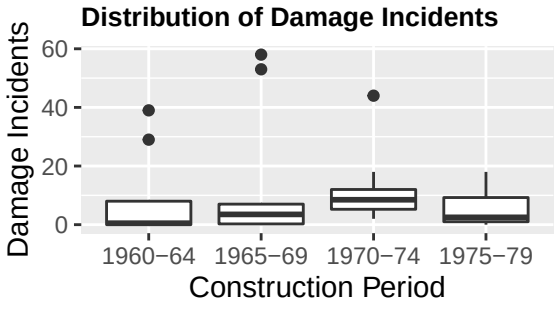


Figure 6: Distribution of damage incidents for each construction period.

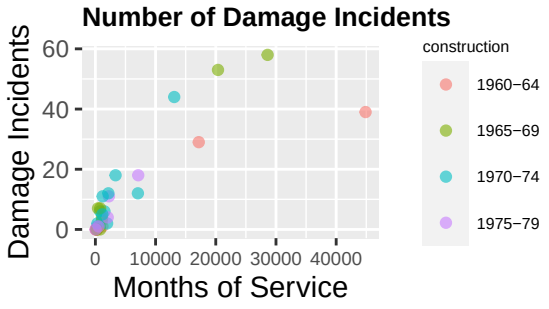


Figure 7: Number of damage incidents by months of service for each construction period.

Finally, we looked at some plots displaying the relationship between the number of damage incidents and the operation period. Here, Figure 8 shows that there is no indication of a median difference between the two operation periods, however there is a clear difference in dispersion, which is also visible in Figure 9. The distribution of damage incidents for ships operating between 1960 and 1974 is more dense around the median than the distribution for ships operating between 1975 and 1979. The sample sizes are rather small, though, (14 and 20, respectively) and their ranges are almost identical.

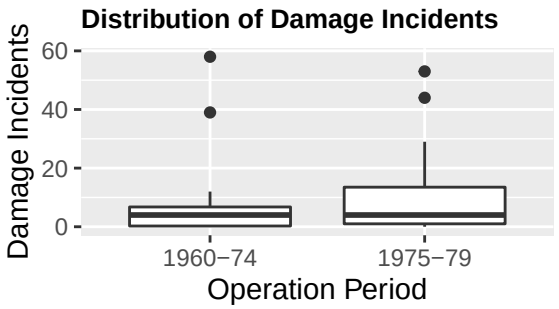


Figure 8: distribution of damage incidents for each operation period.

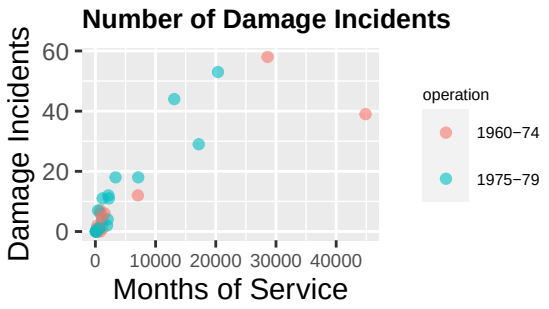


Figure 9: Number of damage incidents by months of service for each operation period.

Generalized linear models

By looking at the descriptives of our data, we have seen that our response variable is a count following the Poisson distribution. This meant that we could not use a "normal" linear model, but we had to use a generalized linear model from the Poisson family, otherwise called a count regression. A generalized linear model is used to examine different types of data, for example the count data which we see in our dataset. A generalized linear model (GLM) consist of three main components (Fox, 2016):

1. Random component: This specifies the conditional distribution of explanatory variables in the model. Which is on our dataset, the Poisson distribution.
2. Linear predictor: A function of regressors: $\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}$. The regressors X_{ij} are prespecified functions of the explanatory variable and the structure of the linear predictor is the familiar structure of the linear model, which makes it easier to work with.
3. Link function: This link function g(.)transforms the expectation of the response variable to the linear predictor. $g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}$. For our Poisson distributed dataset, we use the log link function.

Considering these components resulted in the following (effects) model, including all main effects and first-order interactions:

$\log \mu_{ijkl} = \eta_i = \log(service_i) + \mu + \alpha_j + \beta_k + \gamma_l + \delta_{jk} + \zeta_{jl} + \theta_{kl}$
With μ the overall mean (or intercept);
α_j the effect for ship type j;
β_k the effect for construction period k;
γ_l the effect for operation period l;
and δ_{jk}, ζ_{jl}, and θ_{kl} the interaction effects.

Deviance

The residual deviance of a Poisson GLM is the difference between the fit of the saturated model (with all data points as separated explanatory variables), and the current model. Deviance can be used as a measurement of the goodness of fit, and can be used to compare nested models.

(Over)dispersion

If Y is Poisson distributed with mean $\mu > 0$, then: $P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}$, with $y = 0, 1, 2, ...$ Since Y is a count variable, it is not possible to get values smaller than zero. According to the Poisson distribution, the $E(Y) = var(Y) = \mu$, meaning that the expectation and variance of a Poisson are both equal to the mean. So, we do not have a parameter that fits the variability, thus consequently we expect that the variability increases as the mean increases. If the observed variance is higher than the mean, this indicates that the data is overdispersed. If the observed variance is lower, our data is underdispersed. The dispersion parameter

(ϕ) can be estimated using:
$\hat{\phi} = \frac{X^2}{n-p} = \frac{\sum_i (y_i - \hat{\mu_i})^2 / \hat{\mu_i}}{n-p}$

In the regular Poisson case we expect $\phi = 1$ which means the $\mu = var(Y)$. Overdispersion means that $\phi > 1$ and underdispersion means $\phi < 1$.

Analysis 1

In our dataset, the number of observed incidents is related to the size of the service period. For example, one ship category was only used for 63 months and another ship category for 44882 months, which highly influenced the number of incidents that could have happened. In order to correct for this in our model, we have performed a log transformation on the service variable and used this as our offset. To check whether the service parameter could indeed be used as an offset, we checked whether the estimated coefficient of service was 1.

We saw that the estimated coefficient for months of service was almost 1 (0.906), which suggested that we could use service as a offset and that we were dealing with a rate model. Next up, we fitted a new model with service as an offset and checked whether there was a significant difference between the models. A χ^2-test comparing the old model (service as variable) the new model (service as offset) showed us that there was no significant difference ($p = 0.3621$), further confirming that we could use service as offset variable.

We were interested in finding out whether there are any first-order interactions between our predictors. This would be quite reasonable, since improvements in quality between construction periods might be different for the different types of ships (maybe they fixed many construction issues with ships of type A but not of type B), and 'newer' ships might be more popular than 'older' ships, which could explain an interaction between construction period and operation period. We fitted a model with all interactions and perform type II likelihood ratio tests (LRTs). We removed the non-significant terms one by one based on p-values and deviances. Only one of the interaction terms reduced the deviance significantly based on the type II LRTs (corrected for the other interactions), which was the interaction between type and construction.

```
## glm(formula = incidents ~ type + construction + operation + type:construction,
##     family = poisson, data = ShipAccidents, offset = log(service))
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: incidents
##                   LR Chisq Df Pr(>Chisq)
## type                23.573  4  9.725e-05 ***
## construction        31.401  3  6.998e-07 ***
## operation           10.621  1   0.001118 **
## type:construction   24.216 11   0.011852 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To confirm and interpret our findings we made interaction plots, shown in Figures 10 through 12. In Figure 10, you can see an interaction effect between type and construction, which confirms what we found in the output. More specifically, we found that the difference between type B and the other ship types is mediated by construction period. For the newest ships, the differences in damage incidents between sip types are relatively small, whereas for ships built between 1965 and 1969 the differences are much more extreme. In Figure 11, you see no interaction between type and operation. In Figure 12 however, it seems that there is some interaction happening, but according to our model this was not significant ($p = 0.435103$).

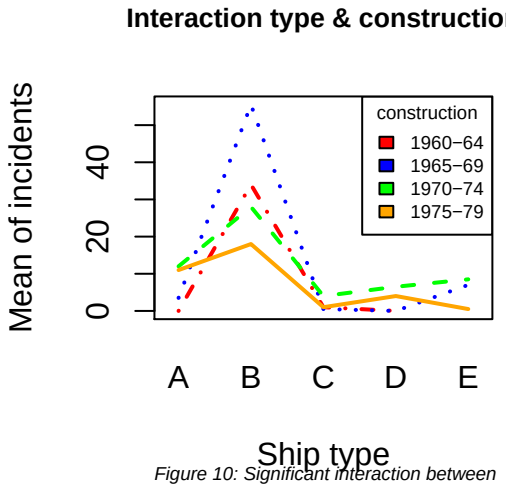


Figure 10: Significant interaction between type & construction

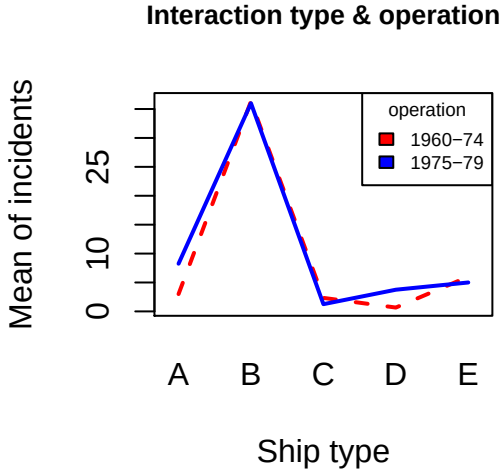


Figure 11: No interaction between type & operation

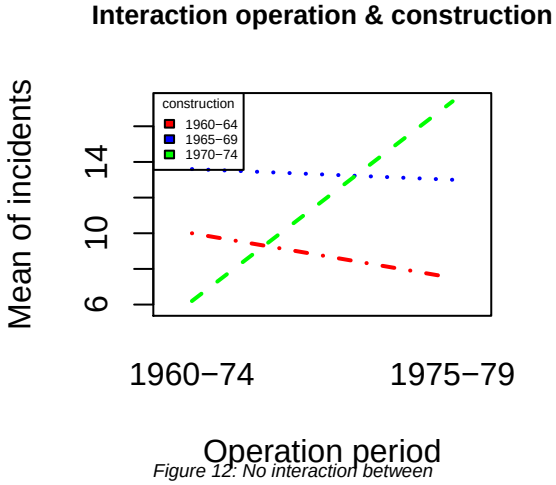


Figure 12: No interaction between operation & construction

Next up, we looked at the goodness-of-fit. The goodness-of-fit can be checked using Pearson's χ^2, which is defined as the following:
$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$ with y_i the observed counts, $\hat{\mu}_i$ the expected counts and $\chi^2 \sim \chi^2_{n-p}$ under the null hypothesis that the model fits well. The null-hypothesis was not rejected, with $p = 0.2329$.
We examined the deviance as well, and it looked good (deviance = 14.746 on 14 degrees of freedom), as did the deviances of the individual terms. The deviance was very close to the residual degrees of freedom, meaning there was no indication of overdispersion. The Wald-tests are almost exclusively insignificant, which might be due to our sparse data, which could have resulted in small z-values. We have also examined the deviance-R^2 of our model, which can be calculated by:
$R^2 = 1 - \frac{deviance_{residual}}{deviance_{null}}$.

Our $R^2_{Dev} = 0.899$, which is quite high, meaning that this model fits well. We selected this model and continued with the diagnostics.

Diagnostics

Our diagnostics indicate how well our model fits the data and if there are any outliers, unusual observations, or if our data violates our assumptions. We first looked at the Pearson residuals, which are comparable to standardized residuals used for linear models. In Figure 13 you see one observation which might be an outlier, as it does not fall within two standard deviations from the mean. Other than that, the Pearson residuals looked like they were spread out relatively evenly over the y-axis, which is good. Furthermore, Figure 14 shows that the predicted values were close to the observed values.

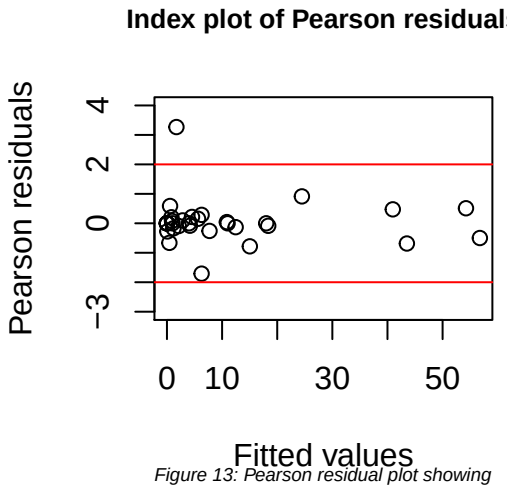


Figure 13: Pearson residual plot showing one unusual observation

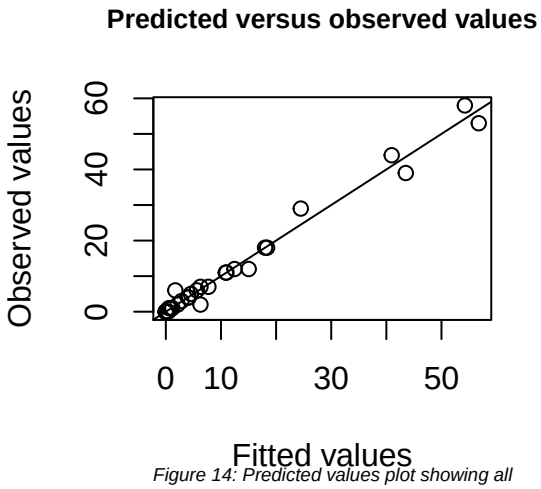


Figure 14: Predicted values plot showing all predicted values close to the line

We continued the diagnostics by looking at the half-norm and log-log plot. In the half-norm residual QQ-plot in Figure 15 we see that observations 19 and 20 might have been outliers. These points actually correspond to cases 21 and 22, because cases 7 and 15 have been removed when preparing the data. According to the log-log plot, the variance was smaller than the mean. There also are two clear clusters of points that are very different from the other points. One of the clusters has fitted values close to zero (since their logs are close to -20) with even smaller variance (logs of -40), and the other cluster has a much smaller variance, almost zero (since their logs are close to -40), with predicted values larger than one it seems (positive log values). We were not interested in the first cluster, since the fitted values and errors are probably zero when rounded. The second cluster was more interesting to us. It could represent a well-defined cluster with characteristics that the model could accurately distinguish from other cases.

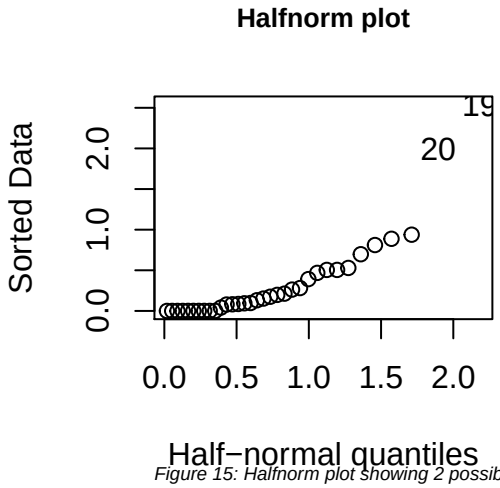


Figure 15: Halfnorm plot showing 2 possible outliers

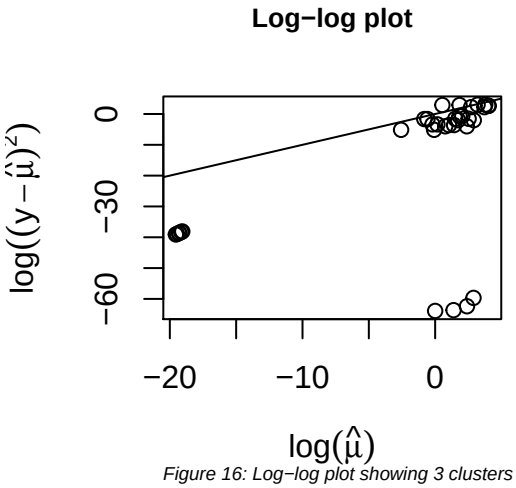


Figure 16: Log−log plot showing 3 clusters of observations

Lastly, we looked at the Cook's distances and checked for multicollinearity. We found one clear influential point, case 22, which had a Cook's distance value of 2.60. This was an order of magnitude larger than the second largest distance. In this output we can see the 10 largest Cook's Distance points. We calculated the Cook's D cut-off point, which was:
$D_c > \frac{4}{n-k-1} = \frac{4}{\text{residual df}} = \frac{4}{14} = 0.28571.$

We found only one case which Cook's distance exceeded this point, which was case 22.

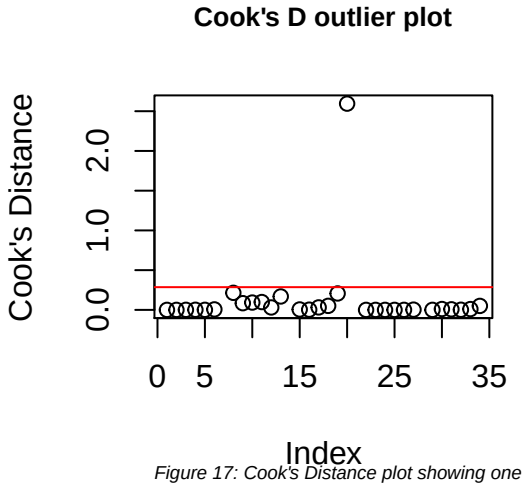


Figure 17: Cook's Distance plot showing one clear influential point

We checked for multicollinearity by calculating the generalized variance inflation factors. All of them were close to 1, which indicated no multicollinearity.

```
##                   GVIF Df GVIF^(1/(2*Df))
## construction 1.524160  3        1.072766
## operation    1.184960  1        1.088559
## type         1.341127  4        1.037370
```

So conclude, even though we saw some potential outliers on the halfnorm plot and the residuals, we do not see them as influential point by looking at the Cook's Distance. The only observation that stands out is case 22.

Interpretation

Since we used dummy coding, our model would look rather large when written down in effect model notation or mean model notation, one might just as well look at the `R` output provided earlier. To aid interpretation, we created a table (Table 2) with coefficients for each individual ship category, i.e. each combination of type, construction period, and operation period, and including the intercept:

Table 2: Estimated sum of coefficients

	Operation 1960-1974				Operation 1975-1979			
	1960-64	1965-69	1970-74	1975-79	1960-64	1965-69	1970-74	1975-79
A	-23.99	-5.96	-5.59		-23.60	-5.57	-5.21	-5.32
B	-6.94	-6.27	-6.15		-6.55	-5.88	-5.77	-5.98
C	-6.90	-7.48	-6.12		-6.52	-7.10	-5.74	-5.61
D	-24.59	-24.96	-5.10		-24.20	-24.57	-4.71	-6.24
E	-25.05	-4.63	-5.54		-24.66	-4.24	-5.16	-6.38

The table should be read as follows: for a ship of type A, constructed between 1960 and 1964, and which operated between 1960 and 1974, the expected value of damage incidents was: $e^{\log service - \beta_{111}} = e^{\log service - 23.99} = \frac{service}{e^{23.99}}$. Since the maximum observed value of aggregate service months was 44882, which is much smaller than $e^{23.99} \approx 2.62 * 10^{10}$, the expected value was approximately zero, which matches the observed value. For a ship of type A, constructed between 1965 and 1969, which operated between 1960 and 1974, the expected number of incidents was $e^{\log service - \beta_{121}} = e^{\log service - 5.96} = \frac{service}{e^{5.96}}$. We have observed 1095 service months, so the predicted value was $\frac{1095}{e^{5.96}} \approx 3$ (rounded to whole numbers since incidents is discrete), which also matches our data.
You can immediately see that some categories got an expected value of zero regardless of the number of service months, since their coefficients were relatively extreme (in the sense that they result in extreme denominators in the formula shown above). Another observation is that a column is missing in the first table, necessarily so, because ships built between 1975 and 1979 could not have operated between 1960 and 1974. Table 3 present the predicted values for the number of incidents for each category of ship. In this table you can see that we predicted a ship from type A, which was built between 1960-1974 and operated between 1960-1964 to have no incidents. We predicted that a ship of type B, which was built and operated during the same period as the previous example, has 44 incidents.

Table 3: Predicted number of incidents

	Operation 1960-1974				Operation 1975-1979			
	1960-64	1965-69	1970-74	1975-79	1960-64	1965-69	1970-74	1975-79
A	0.00	2.83	5.63		0.00	4.17	18.37	11.0
B	43.52	54.24	15.03		24.48	56.76	40.97	18.0
C	1.18	0.44	1.72		0.82	0.56	6.28	1.0
D	0.00	0.00	2.14		0.00	0.00	10.86	4.0
E		7.72	4.54			6.28	12.46	0.5

We compared our predicted values to the observed values and the differences can be found in the tables below (Table 4). Approximately two thirds of the predictions were correct (after rounding the residuals to whole numbers). We can also see that the residuals were higher for ships of type B. As we have seen in Figure 1, these ships had more incidents than the other ships. For a Poisson distributed variable, the variation is proportional to the mean, such that groups of cases with more damage incidents are likely to have a larger error when we try to predict them.

Table 4: Difference between observed and predicted number of incidents

	Operation 1960-1974				Operation 1975-1979			
	1960-64	1965-69	1970-74	1975-79	1960-64	1965-69	1970-74	1975-79
A	0.0000	0.1656	0.3651		0.0000	-0.1656	-0.3651	0
B	-4.5218	3.7593	-3.0324		4.5218	-3.7593	3.0324	0
C	-0.1848	-0.4401	4.2819		0.1848	0.4401	-4.2819	0
D	0.0000	0.0000	-0.1357		0.0000	0.0000	0.1357	0
E		-0.7177	0.4606			0.7177	-0.4606	0

Lastly, we calculated the confidence intervals around our predicted values. These 95% confidence intervals can be found in Appendix A. The method we chose was to manually transform the values using the inverse link function. The advantage of this is that we get no confidence intervals containing negative values. A drawback, however, is that the predictions with value zero have no upper bound, due to inflated standard errors.

Analysis 2: without outliers

After having looked at the results and diagnostics we decided to investigate the effect of leaving out the large influential observation we found. We decided to exclude the case (22) from out dataset, which had a Cook's Distance much larger than the cut off point ($D_{22} = 2.56$), and also much larger than the Cook's Distances of the other cases. These were the observed values for case 22:

	type	construction	operation	service	incidents
22	C	1970-74	1975-79	1948	2

Next, we repeated the entire process again. We initially found that the same model may be used as before. However, the residual deviance was much lower than the degrees of freedom, so we tested for underdispersion. The dispersion parameter was taken to be $\phi = 0.184$. We rejected the null hypothesis ($p < 0.001$), which meant we were dealing with underdispersion. To fix this issue we fitted a quasi-Poisson model with all main effects and interaction terms. Non-significant terms were removed one by one based on their p-values and in accordance with the Principle of Marginality. We ended up with our final model including all main effects: type, construction, and operation (all with $p < 0.001$) and we included the interaction between type and construction ($p < 0.001$) and between construction and operation ($p = 0.0418$). So, only the interaction between type and operation period was removed ($p = 0.69$). A Pearson's χ^2-test was performed on our final model and the null hypothesis that the model fits well was accepted ($p = 0.998$). We also found that $R^2_{Dev} = 0.983$, which meant our model fitted very well.

The interaction plots look similar to those we have seen before for our previous analysis, we therefore included them in an appendix, see Appendix B. Figure 18 shows the interaction between construction and operation period, which was significant in this analysis but not in our first one. It makes sense that our models are similar, since we only removed one observation. According to our model, the strongest interaction was between the different ship types and construction periods.

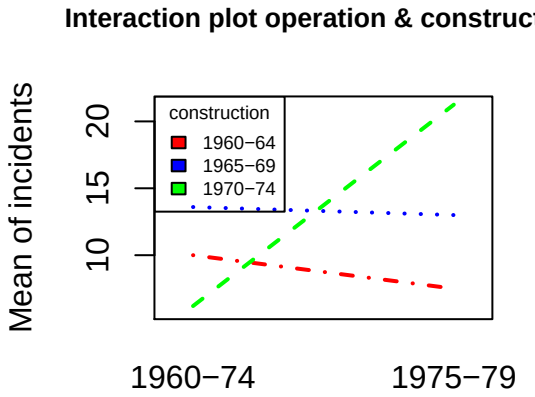


Figure 18: Interaction plot showing significant interaction

Diagnostics

We started by looking at our Pearson residuals (Figure 19). Interestingly, we now see that all residuals are within one standard deviation from the mean. Removing the outlier thus had a strong influence on how well our model fitted the data. In Figure 20 we see that all cases are on or close to the line, which suggests a good fit as well.

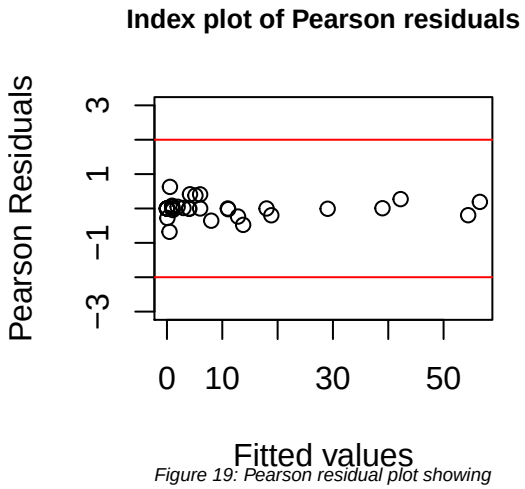


Figure 19: Pearson residual plot showing no indication of outliers

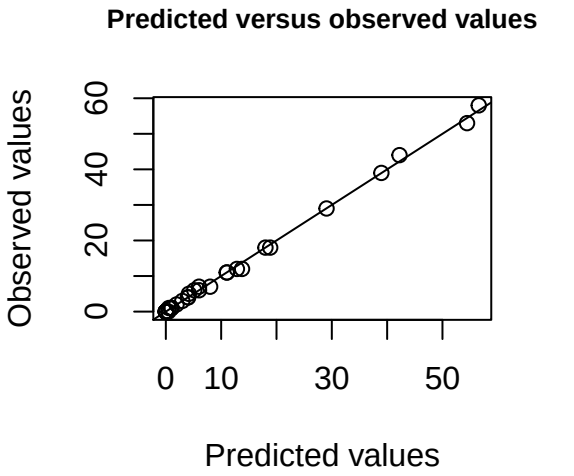


Figure 20: Predicted values plot showing good fit o

Furthermore, we looked at the half-norm plot (Figure 21) and saw that point 17, and possibly point 18, might have been outliers if only looked at the graph. But when we checked the y-axis we saw the highest value is around 0.9, which is still good, since we want our Pearson-residuals to be smaller than 2. In the log-log plot (Figure 22) we found a similar pattern as before, with two clusters quite far removed from the line and all points below the line.

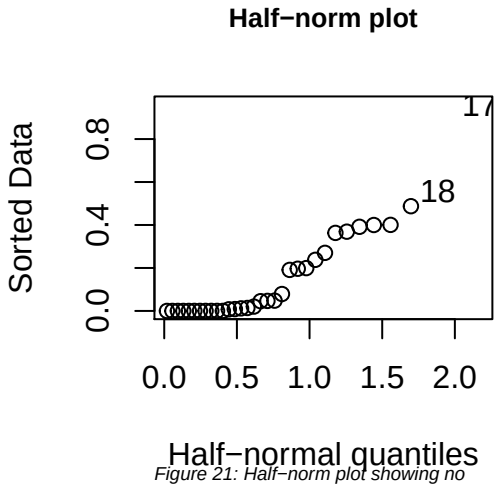


Figure 21: Half−norm plot showing no abnormalities

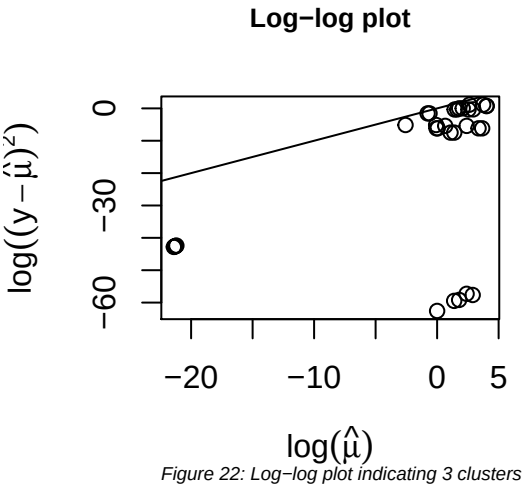


Figure 22: Log−log plot indicating 3 clusters of observations

Lastly, we looked at the Cook's Distance again to look for influential observations. The cut-off point for the Cook's distance was: $D_c > \frac{4}{n-k-1} = \frac{4}{11} = 0.364$. We found three points that exceeded this cut-off point. This indicated that those points were influential points in our data. The plot with predicted versus observed values showed that this most likely did not have a negative effect on the fit of our model, though, since all predicted values lay on or close to the line.

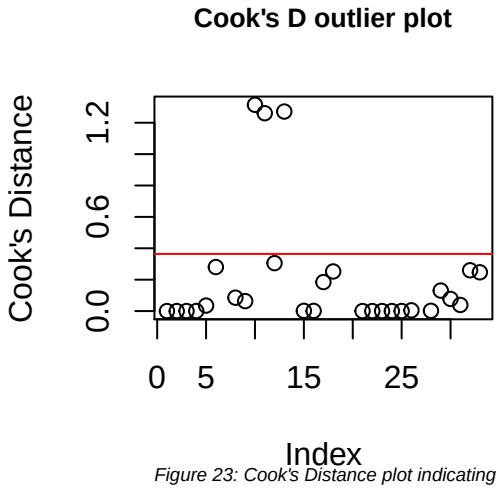


Figure 23: Cook's Distance plot indicating 3 influential points

Interpretation

Finally, we interpreted the results of our model. Table 5 shows the estimated coefficients, which can be interpreted the same way as the coefficients of the previous analysis. The coefficients were similar to those of the first analysis, tough the extreme coefficients were slightly more extreme for this analysis.

Table 5: Estimated sums of coefficient

	Operation 1960-1974				Operation 1975-1979			
	1960-64	1965-69	1970-74	1975-79	1960-64	1965-69	1970-74	1975-79
A	-26.13	-5.91	-5.68		-25.46	-5.61	-5.18	-5.32
B	-7.05	-6.23	-6.24		-6.38	-5.92	-5.74	-5.98
C	-7.03	-7.44	-4.87		-6.36	-7.13	-4.37	-5.61
D	-26.73	-26.91	-5.20		-26.06	-26.61	-4.69	-6.24
E	-27.19	-4.59	-5.63		-26.52	-4.29	-5.13	-6.38

In Table 6 we see the table of predicted values of the number of incidents, structured by ship type, construction period and operation period. We can see, for example, that we expected a ship of type A, which was built in 1960-1964 and operated between 1960-1974 to have no incidents. But for a ship of type B, which was built during the same time and operated in the same time, we expected to have 39 accidents.
The differences between the real values and the predicted values are presented in Table 7. We immediately noticed that the predicted values were more accurate, since the differences were smaller. This is what we expected as well, since the fit of our model improved.
Lastly, the confidence intervals for the predicted values may again be found in Appendix A.

Table 6: Predicted number of incidents

	Operation 1960-1974				Operation 1975-1979			
	1960-64	1965-69	1970-74	1975-79	1960-64	1965-69	1970-74	1975-79
A	0.00	2.98	5.14		0.00	4.02	18.86	11.0
B	38.95	56.56	13.77		29.05	54.44	42.23	18.0
C	1.05	0.46	6.00		0.95	0.54		1.0
D	0.00	0.00	1.93		0.00	0.00	11.07	4.0
E		8.01	4.16			5.99	12.84	0.5

Table 7: Difference between observed and predicted number of incidents

	Operation 1960-1974				Operation 1975-1979			
	1960-64	1965-69	1970-74	1975-79	1960-64	1965-69	1970-74	1975-79
A	0.0000	0.0237	0.8587		0.0000	-0.0237	-0.8587	0
B	0.0459	1.4425	-1.7686		-0.0459	-1.4425	1.7686	0
C	-0.0459	-0.4608	0.0000		0.0459	0.4608		0
D	0.0000	0.0000	0.0670		0.0000	0.0000	-0.0670	0
E		-1.0055	0.8429			1.0055	-0.8429	0

Analysis 3: Reducing construction

So far we have two models with many parameters but only a few observations. This is not ideal, since the goal of our case study is to explain the number of damage incidents in a model that is accurate, yet simple. Therefore, we decided to collapse the construction period variable into 2 levels instead of 4. We now have a group from 1960-1969 and a group from 1970-1979. Since we had found a significant interaction between type and construction period in our first two analyses, we expected that reducing the number of levels for

construction period would greatly increase our degrees of freedom. Furthermore, it made mores sense to reduce the number of levels for this factor than for the others. Operation period already only had two levels, and the grouping of years used to create the factor levels (for construction period) in the first place was rather arbitrary. Besides, reducing the number of levels for ship type would have inevitably resulted in unequal group sizes.

The first thing we noticed is that the estimated coefficient of service went down a bit to 0.83, but we found that our model did not differ significantly to the model where we included service as an offset ($p = 0.08$), so we used the offset.

This time we found significant effects of type ($p =< 0.001$), construction ($p = 0.0306$), operation ($p = 0.0046$) and the interaction between type and construction ($p = 0.0056$). We started by removing the interaction between construction period and operation period as it had the lowest deviance ($D = 0.238$). We found that the interaction between type and operation period was still not significant ($p = 0.13$), so we excluded this term as well. The deviance of our model was 47.13, which is much higher than the deviances of our previous analyses (14.746 and 2.473, respectively). We had an adequate $R^2_{Dev} = 0.68$, though. However, a Pearson's χ^2-test rejected the null hypothesis that the model fits well ($p = 0.001$). Despite the seemingly large difference between our deviance and the degrees of freedom, the dispersion test was not significant with p = 0.096. We then looked at the diagnostics to see whether there were any signs of outliers, influential points, or (multi)collinearity that could explain the lack in fit.

Interestingly, when looking at the interaction plot in Figure 24, we found an indication that there might be an interaction between construction period and operation period, even though this term had the lowest deviance in our full model. This difference might be caused by the fact that the interaction plots did not control for the other variables present in the model. The other interaction plots can again be found in Appendix B.

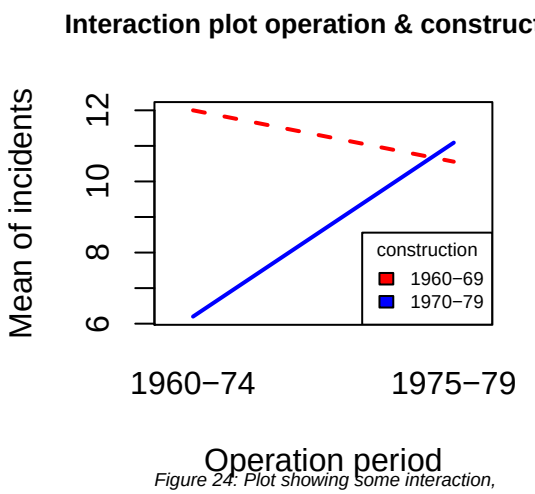


Figure 24: Plot showing some interaction, but not significant

Diagnostics

We looked at the diagnostics again. First of all, Figure 25 shows there were 4 Pearson residuals outside the two standard deviation range, which might suggest we have some outliers. When we looked at the predicted versus observed values plot in Figure 26, we found some observations quite far from the line as well. Especially the higher values seemed to fit less well.

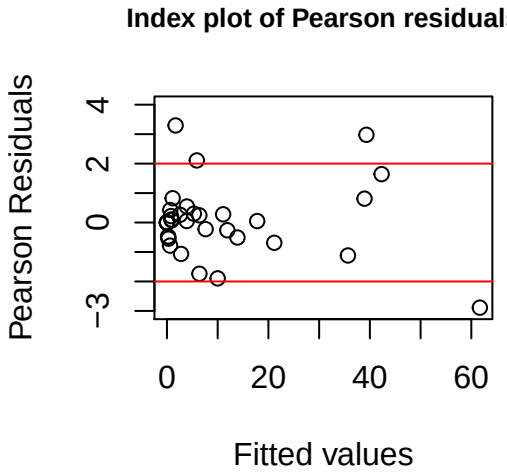


Figure 25: Pearson residual plot indicating some ou

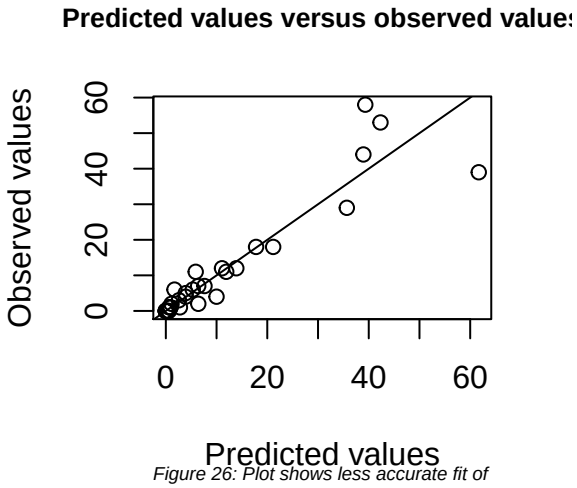


Figure 26: Plot shows less accurate fit of the model for larger values

In our half-norm plot (Figure 27) we saw 2 values which might be outliers, but if we carefully look at the scale of the y-axis it seems that there are more values exceeding the 2 standard deviations of the Pearson residuals. This is in accordance to our findings in Figure 25. When we looked at our log-log plot (Figure 28), we noticed only 2 clusters, where before we noticed 3 clusters. This seemed to be an improvement since more values were close to the line. We also saw that there were more values above the line than before, which indicated that the mean and variance were more similar.

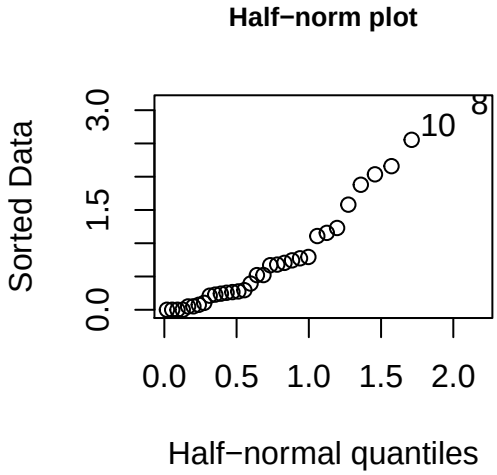


Figure 27: Plot indicates quite some outliers

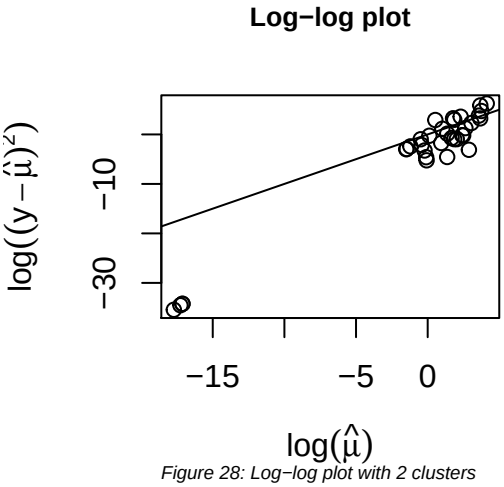


Figure 28: Log−log plot with 2 clusters of observations

Furthermore, we looked at the Cook's Distance for influential points. We calculated the cut-off point $D_c = 4/23 = 0.173913$. We found more influential points for this models, with 7 points exceeding our cut-off point. These points corresponded to the 7 points with the highest residuals as well.

```
##        22         9        32        11        30        21        12
## 2.4021104 1.6021466 1.1335219 0.5736297 0.3280351 0.3196459 0.3164911
```

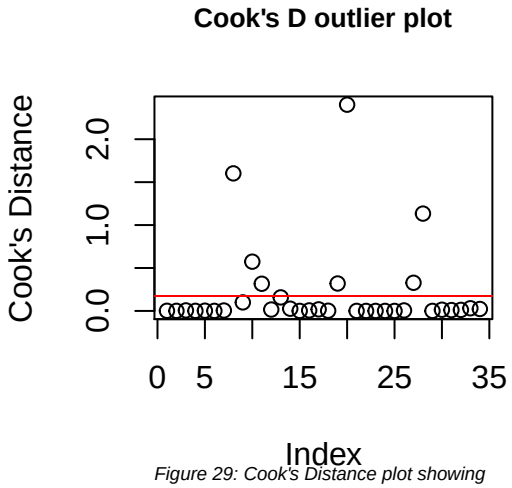


Figure 29: Cook's Distance plot showing many influential points

Finally, we checked for (multi)collinearity again, since we changed the factor levels of the construction variable. Again, we found no indication of (multi)collinearity, with the highest generalized $VIF = 1.48$.

	GVIF	Df	GVIF^(1/(2*Df))
construct2	1.476211	1	1.214994
operation	1.182987	1	1.087652
type	1.322630	4	1.035571

Interpretation

First of all, we examined the tables with the estimated coefficients (Table 8). These tables are interpreted the same way as the previous coefficient tables. One important difference to note is that because construction period now has two levels, we could not account for the ship categories that were constructed after the operation period. In the other coefficient tables we were able to remove those coefficients. Needless to say this is not possible anymore now that ships built between 1970 and 1979 have all been coerced to one level. One consequence of reducing the number of levels of the construction variable was that there were fewer extreme coefficients. In the tables we only saw two extreme values, as opposed to eight.

Table 8: Estimated sums of coefficients

	Operation 1960-1974		Operation 1975-1979	
	1960-69	1970-79	1960-69	1970-79
A	-6.05	-5.65	-5.64	-5.24
B	-6.59	-6.23	-6.18	-5.82
C	-7.15	-6.13	-6.74	-5.72
D	-22.78	-5.74	-22.37	-5.32
E	-4.64	-5.69	-4.23	-5.27

Table 9 presents the crosstabulations containing the predicted values. We noticed nothing unusual here. More interestingly, we can look at Table 10 which shows the differences between the predicted values and the observed values. The table reflects what we found earlier in Figures 25 and 26, namely that the residuals for this analysis were larger than for our first two analyses. We only got correct predictions for about half of the

categories. However, we also noticed that the errors or residuals were quite small compared to our very first analysis, as becomes evident when comparing Table 4 to Table 10.
The 95% confidence intervals for the predicted values can be found in Appendix A again.

Table 9: Predicted number of incidents

	Operation 1960-1974		Operation 1975-1979	
	1960-69	1970-79	1960-69	1970-79
A	1.44	5.31	2.06	14.85
B	50.50	13.90	39.00	30.05
C	0.77	1.70	0.73	3.65
D	0.00	1.12	0.00	7.94
E	7.62	3.92	6.38	4.69

Table 10: Difference between observed and predicted number of incidents

	Operation 1960-1974		Operation 1975-1979	
	1960-69	1970-79	1960-69	1970-79
A	0.06095	0.6926	-0.06100	-0.3463000
B	-2.00415	-1.8962	2.00415	0.9481000
C	-0.27045	4.2983	0.27045	-2.1491500
D	0.00000	0.8751	0.00000	-0.4376000
E	-0.62080	1.0781	0.62080	-0.3593667

Analysis 4: without outliers

We wanted to improve the fit slightly, however we did not want to remove too many points because we had a small dataset already. We decided to remove the 3 datapoints (10% of our dataset) with the highest Cook's D. These were the following observations:

	type	operation	service	incidents	construct2
9	B	1960-74	44882	39	1960-69
22	C	1975-79	1948	2	1970-79
32	D	1975-79	2051	4	1970-79

We fitted a new model with all the main effects to see whether we could use service as offset again. We found that the coefficient of log(service) was 1.179. A χ^2-test showed us that service months may be used as an offset, since the models did not differ significantly ($p = 0.1408$). After a process of backwards elimination, we found one significant interaction, which is between type and construction (p = < 0.001), and we found that the main effect of operation period was no longer significant. When we checked for overdispersion we found a small and significant dispersion parameter ($\phi = 0.724, p = 0.029$) which means we had underdispersion. We therefore fitted a quasi-Poisson model and repeated the process of backwards elimination. Again, we found the interaction between type and construction to be significant ($p < 0.001$), however we also had all significant main effects now. Our model had a deviance R^2 of 0.84 and a Pearson's χ^2 of 0.94, which indicated that our model fitted the data well.

We examined our interaction plots again and noticed that the interaction between type and construction, and type and operation, did not change much compared to Analysis 3. These figures can be found in Appendix

B. However, we did find a difference when plotting the operation and construction interaction (Figure 30). In Figure 24 we noticed that the lines were not parallel and even going in opposite directions, indicating a strong interaction. In this analysis, however, we observed (in Figure 30) that only the strength of the effect seemed to vary for different construction periods and there was no significant interaction at play. Apparently, we have removed cases which influenced this interaction.

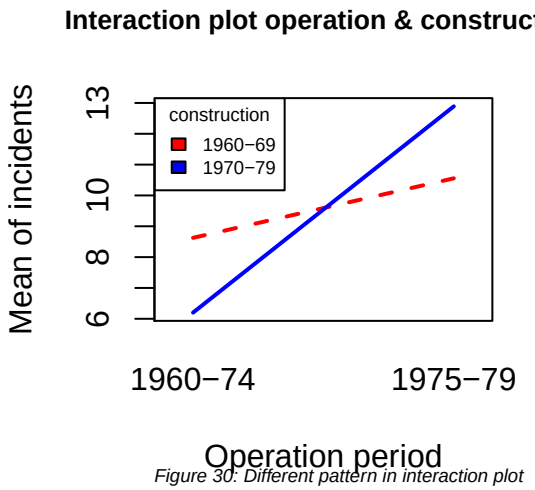


Figure 30: Different pattern in interaction plot compared to previous analyses

Diagnostics

We investigated the diagnostics of the fourth analysis using the same procedure as before. In Figures 31 and 32 you see that this model fitted much better than the previous one. All Pearson residuals fell within two standard deviations from zero and in general the predicted values were closer to the observed values. We still noticed that the residuals were higher for higher numbers of incidents, which was expected considering that we were working with a Poisson model.

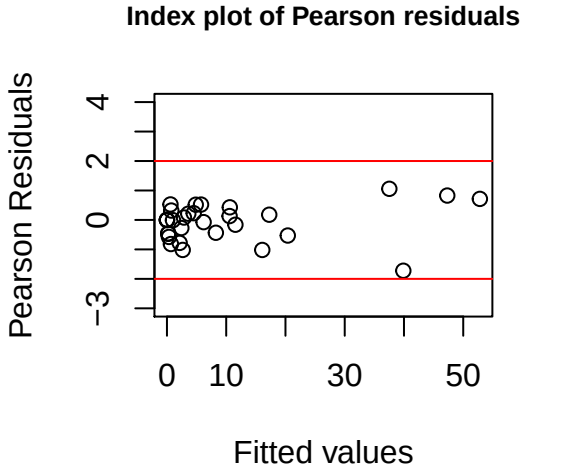


Figure 31: Pearson residual plot showing no outlier

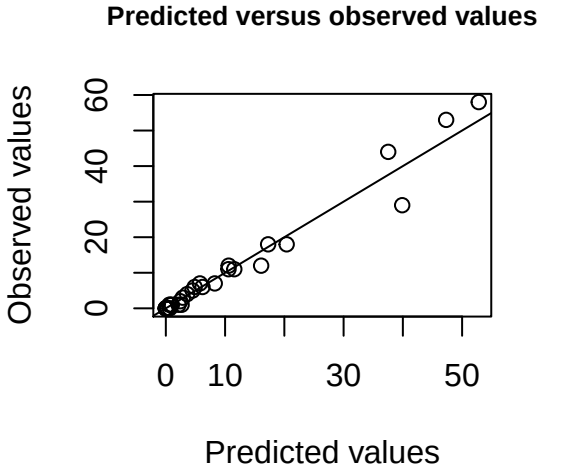


Figure 32: Predicted values plot indicating a good f

The half-norm plot (Figure 33) looked adequate, there seemed to be one potential outlier which is outside the line of the plot and might exceed the 2 standard deviations. Other than that the plot looked good with all observations in one line. The log-log plot in Figure 34 looks similar to Figure 28, showing 2 clusters of points.

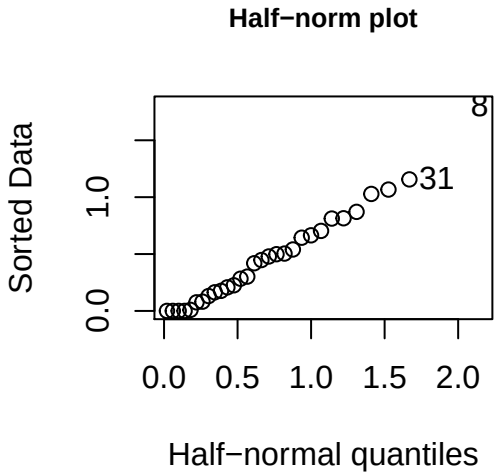


Figure 33: Half−norm plot not indicating outliers

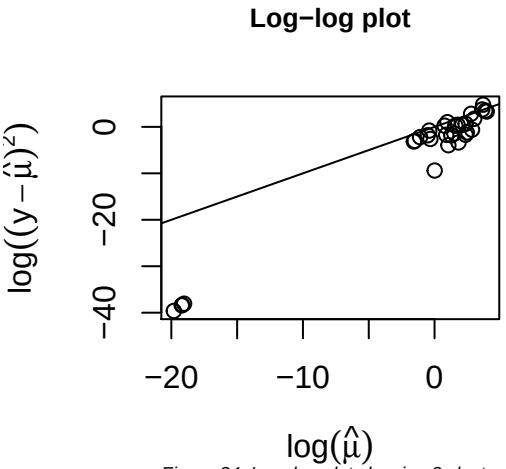


Figure 34: Log−lop plot showing 2 clusters of observations

The Cook's Distance cut-off point was $D_c = 4/20 = 0.2$. We found 4 observations that exceeded the cut-off point and could have been influential points. Interestingly, these points did not (all) match the four points that exceeded the cut-off point in Analysis 3, suggesting that our fourth model fits certain categories better or worse than our third model, rather than having a global increase in fit.

```
##        11        10        14        21
## 0.6482718 0.4610926 0.4479621 0.3408781
```

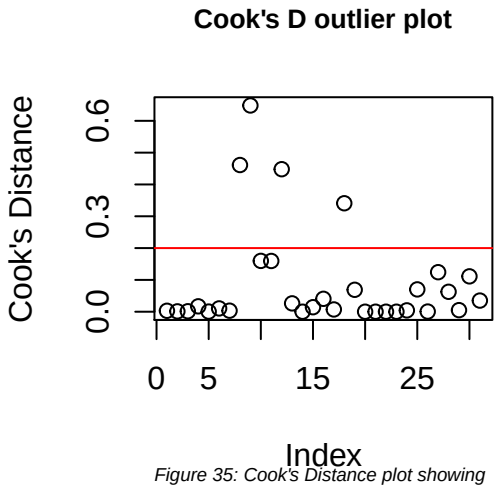


Figure 35: Cook's Distance plot showing some influential points

Interpretation

Compared to the coefficients of Analysis 3, this analysis resulted in similar coefficients with some minor differences (Table 11). The two rather extreme coefficients have gotten even more extreme, whereas the other coefficients have generally gotten lower.

Table 11: Estimated sums of coefficients

	Operation 1960-1974		Operation 1975-1979	
	1960-69	1970-79	1960-69	1970-79
A	-5.95	-5.50	-5.72	-5.27
B	-6.29	-6.08	-6.07	-5.86
C	-7.06	-5.08	-6.83	-4.85
D	-24.69	-4.97	-24.46	-4.74
E	-4.56	-5.55	-4.33	-5.32

When we looked at the predicted number of incidents (Table 12), we noticed nothing unusual and all the predicted values seemed be normal. But when we looked at the differences between the predicted and observed values (Table 13), we found something interesting. On the one hand, there were more "correct" predictions now (differences that round to zero) compared to the previous analysis. On the other hand, the predictions that were not correct had much higher residuals. For all three analyses performed before this one, the highest (absolute) residual could be rounded to a 5. Here we had multiple (absolute) residuals larger than that, with the highest value being 19. This suggested that our model did not fit very well after all.
Again, the 95% confidence intervals for the predicted values can be found in Appendix A.

Table 12: Predicted number of incidents

	Operation 1960-1974		Operation 1975-1979	
	1960-69	1970-79	1960-69	1970-79
A	1.60	6.19	1.90	14.41
B	67.82	16.08	43.60	28.96
C	0.84	4.86	0.66	8.68
D	0.00	2.43	0.00	14.26
E	8.25	4.51	5.75	4.50

Table 13: Difference between observed and predicted number of incidents

	Operation 1960-1974		Operation 1975-1979	
	1960-69	1970-79	1960-69	1970-79
A	-0.09650	-0.1866	0.09655	0.09330
B	-19.32475	-4.0847	-2.59680	2.04230
C	-0.33880	1.1398	0.33880	-7.17640
D	0.00000	-0.4277	0.00000	-6.76125
E	-1.25060	0.4868	1.25060	-0.16230

Conclusion & Discussion

In this report we examined the ShipAccidents dataset by using Poisson regression. We tried to explain the number of damage incidents in terms of ship type, the operation period, and the construction period of the ship. Moreover, we investigated the effects of including and excluding influential points, and we examined the trade-off between degrees of freedom and deviance that resulted from reducing the number of factor levels of one of our predictors (and in doing so, categories).

Generally, we found that ships of type B have more damage incidents than other ship types, though they also have more service hours. The variation in damage incidents for ships of type B can be explained by construction period. When we look at construction period as a factor with four levels, we see that for ships of type B, those ships constructed between 1965 and 1969 had the most incidents. For the last two analyses, in which construction period only had two levels, we saw that there were more incidents for ships of type B constructed between 1960 and 1969, which is in line with the findings of the first two analyses. We also found that ships constructed after 1969 generally had more incidents when they operated between 1975 and 1979 than between 1960 and 1974. The interactions between operation period and the other variables were not included in every model, though.

In all four analyses, the service variable was used as an offset. We also found one significant interaction that all models had in common, which was the interaction between ship type and construction period. In our first and third model, we found no under- or overdispersion, while in the second and fourth model we had to correct for underdispersion by using a quasi-Poisson model. We found good R^2_{dev} values in all our models, varying from 0.68 to 0.98. The Pearson's χ^2 was not significant for our first, second, and fourth model, suggesting these models were a good fit. We found a significant Pearson's χ^2 for our third model. This model also had the lowest R^2_{Dev}, although it was still moderately good. When looking at the diagnostics of our models, we found that the third model had the most outliers and influential observations, which reinforces the idea that the model does not fit well. The diagnostics of the other models looked good, with only a few potential outliers or influential points. After removing the influential points, our models (Analyses 2 and 4) showed good fitted values and low residuals and Cook's Distances. A drawback of our study is that we could not compare the analyses statistically, since they use different datasets (because we removed outliers) and/or different distributions (Poisson versus quasi-Poisson). Therefore, we can only compare the models based on our own judgment.

The second model seemed to fit best based on the residuals and R^2_{Dev}. We found a significant interaction between construction period and operation period in our second model, which we did not find in the other models. By including this interaction into our model, we ended up with few degrees of freedom and a very low residual deviance, which suggest that the model overfitted our data. Unfortunately, we did not have enough observations to split the data into a testing and training set so we could check this. Our small sample size was a serious shortcoming in this study. Besides making us unable to test if our models overfit the data, the sample size also left us with few degrees of freedom and low power.

Furthermore, even though we only used three predictors, we had many parameters due to the fact that our predictors are factorial and we included interactions. Even though we are now able to describe ship accidents in terms of ship type, construction period, and operation period, the models including interactions are so complex that their usefulness may be questioned. Additionally, since we found significant interactions and we have a very small dataset, we could have found significant results which were not actually meaningful. This is also called Freedman's Paradox (Freedman, 1983) and is a common occurrence when the number of variables is similar to the number of data points.

We tried to deal with this by reducing the number of levels of the construction variable. This resulted in more degrees of freedom and easier interpretation, but it had some drawbacks too. First of all, the fit was undeniably worse as the R^2_{Dev} decreased from 0.98 to 0.68 and the Pearson's χ^2 goodness-of-fit test informed us our new model did not fit adequately (Analysis 3). Additionally, there were many more indications of outliers or influential points compared to our first two analyses. After removing some of these points, the fit increased to an acceptable level, though ($R^2_{Dev} = 0.84$, Analysis 4). There still were many points that had a high Cook's Distance, indicating that they were influential points that could have strongly affected the fit of the model. However, the Pearson residuals looked fine. One possible explanation for why there were so many potential influential points is the small sample size and, as an extension of this issue, that most strata of our design only had a single observation.

In conclusion, which model is the best depends on how you balance fit and simplicity. The model form Analysis 2 had the highest fit (0.98), but also the most parameters, making interpretation difficult. The last model, from Analysis 4, had a moderately high fit (0.84), but nine parameters fewer. We think that the reduction in fit is worth the gain in degrees of freedom.

References

Fox, J. (2015). Applied Regression Analysis and Generalized Linear Models. Thousand Oaks, Canada: SAGE Publications. Freedman, David A. (1983). "A Note on Screening Regression Equations". The American Statistician. 37 (2): 152–155. doi:10.1080/00031305.1983.10482729.

Freedman, David A. (1983). "A Note on Screening Regression Equations". The American Statistician. 37 (2): 152–155. doi:10.1080/00031305.1983.10482729

Appendix A: Confidence Intervals of Predicted Values

Analysis 1:

	predicted	lower	upper
1	0.0000	0.0000	Inf
2	0.0000	0.0000	Inf
3	2.8344	1.3340	6.0221
4	4.1656	1.9741	8.7902
5	5.6349	3.6369	8.7307
6	18.3651	12.2639	27.5015
8	11.0000	6.0917	19.8630
9	43.5218	33.8273	55.9945
10	24.4782	18.4922	32.4018
11	54.2407	43.4950	67.6412
12	56.7593	45.6421	70.5844
13	15.0324	11.0001	20.5430
14	40.9676	31.2972	53.6259
16	18.0000	11.3407	28.5697
17	1.1848	0.2953	4.7527
18	0.8152	0.2025	3.2821
19	0.4401	0.0617	3.1381
20	0.5599	0.0787	3.9854
21	1.7181	0.8391	3.5177
22	6.2819	3.1359	12.5841
24	1.0000	0.1409	7.0993
25	0.0000	0.0000	Inf
26	0.0000	0.0000	Inf
27	0.0000	0.0000	Inf
28	0.0000	0.0000	Inf
29	2.1357	1.1990	3.8041
30	10.8643	6.2999	18.7357
32	4.0000	1.5012	10.6578
33	0.0767	0.0108	0.5442
35	7.7177	4.5240	13.1661
36	6.2823	3.6636	10.7727
37	4.5394	2.7395	7.5217
38	12.4606	7.7150	20.1254
40	0.9233	0.1301	6.5551

Analysis 2:

	predicted	lower	upper
1	0.0000	0.0000	Inf
2	0.0000	0.0000	Inf
3	2.9763	2.1424	4.1347
4	4.0237	2.9106	5.5627
5	5.1413	4.0912	6.4609
6	18.8587	15.8072	22.4993
8	11.0000	8.5364	14.1746
9	38.9541	34.0723	44.5353
10	29.0459	24.8820	33.9065
11	56.5575	50.8046	62.9617
12	54.4425	48.8130	60.7213
13	13.7686	11.4637	16.5371
14	42.2314	37.3856	47.7052
16	18.0000	14.7635	21.9460
17	1.0459	0.5726	1.9105
18	0.9541	0.5215	1.7456
19	0.4608	0.1980	1.0723
20	0.5392	0.2319	1.2536
21	6.0000	4.2565	8.4577
24	1.0000	0.4313	2.3186
25	0.0000	0.0000	Inf
26	0.0000	0.0000	Inf
27	0.0000	0.0000	Inf
28	0.0000	0.0000	Inf
29	1.9330	1.4540	2.5698
30	11.0670	8.7495	13.9984
32	4.0000	2.6269	6.0907
33	0.0767	0.0331	0.1777
35	8.0055	6.3397	10.1090
36	5.9945	4.7165	7.6188
37	4.1571	3.2368	5.3390
38	12.8429	10.4177	15.8328
40	0.9233	0.3982	2.1408

Analysis 3:

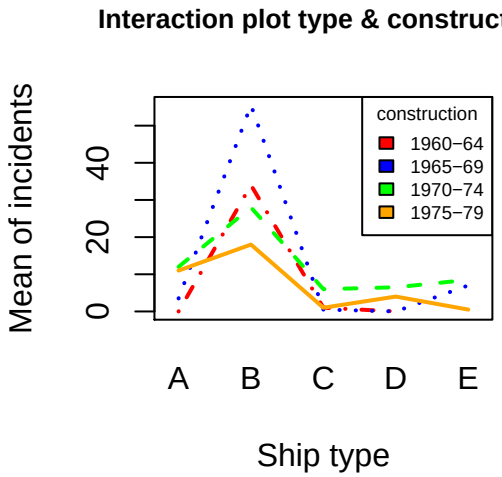
	predicted	lower	upper
1	0.2991	0.1408	0.6353
2	0.2243	0.1063	0.4733
3	2.5790	1.2142	5.4779
4	3.8977	1.8468	8.2262
5	5.3074	3.6097	7.8035
6	17.7880	12.7478	24.8210
8	11.9046	8.5315	16.6115
9	61.6872	51.6259	73.7092
10	35.6785	29.3116	43.4286
11	39.3211	32.9078	46.9842
12	42.3132	34.7623	51.5044
13	13.8962	10.3378	18.6793
14	38.9444	30.8814	49.1127
16	21.1594	16.7786	26.6841
17	0.9269	0.2973	2.8902
18	0.6559	0.2102	2.0464
19	0.6140	0.1969	1.9145
20	0.8032	0.2574	2.5061
21	1.7017	0.8622	3.3587
22	6.3983	3.3242	12.3154
24	0.9000	0.4676	1.7322
25	0.0000	0.0000	Inf
26	0.0000	0.0000	Inf
27	0.0000	0.0000	Inf
28	0.0000	0.0000	Inf
29	1.1249	0.6671	1.8968
30	5.8844	3.6571	9.4680
32	9.9908	6.2093	16.0752
33	0.2305	0.1448	0.3669
35	7.6208	4.4659	13.0045
36	6.3792	3.7217	10.9344
37	3.9219	2.3873	6.4430
38	11.0709	6.9558	17.6204
40	2.7767	1.7446	4.4194

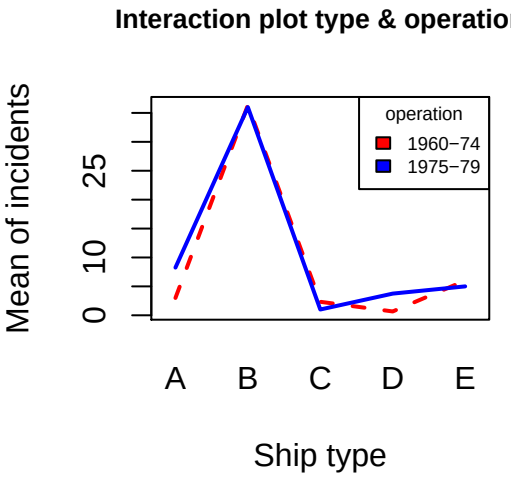
Analysis 4:

	predicted	lower	upper
1	0.3318	0.1888	0.5833
2	0.2071	0.1181	0.3631
3	2.8612	1.6278	5.0290
4	3.5998	2.0535	6.3105
5	6.1866	4.6245	8.2764
6	17.2613	13.4387	22.1711
8	11.5521	8.9939	14.8381
9	82.8431	69.9503	98.1122
10	39.8881	34.5976	45.9875
11	52.8064	44.5882	62.5393
12	47.3055	41.0313	54.5392
13	16.0847	12.8616	20.1154
14	37.5264	31.4910	44.7184
16	20.3890	17.1098	24.2966
17	1.0091	0.4306	2.3648
18	0.5944	0.2531	1.3963
19	0.6685	0.2852	1.5665
20	0.7280	0.3099	1.7100
21	4.8602	2.7822	8.4902
22	15.2130	8.6070	26.8894
24	2.1398	1.2106	3.7822
25	0.0000	0.0000	Inf
26	0.0000	0.0000	Inf
27	0.0000	0.0000	Inf
28	0.0000	0.0000	Inf
29	2.4277	1.5729	3.7470
30	10.5723	7.0255	15.9098
32	17.9502	11.9282	27.0124
33	0.2209	0.1558	0.3131
35	8.2506	5.5323	12.3045
36	5.7494	3.8262	8.6393
37	4.5132	3.1090	6.5516
38	10.6059	7.4800	15.0379
40	2.6601	1.8761	3.7717

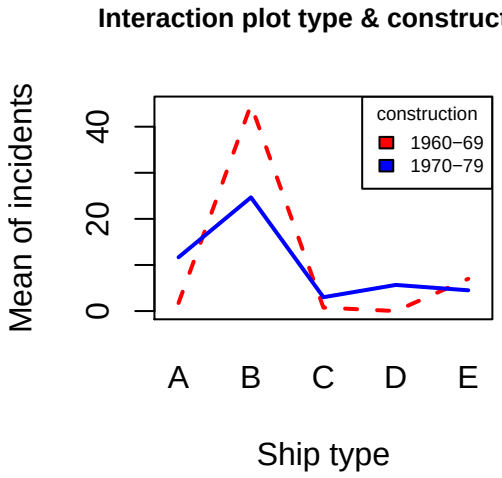
Appendix B: Additional Interaction Plots

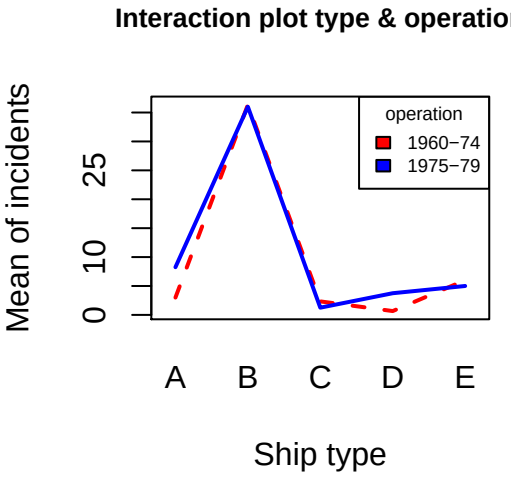
Analysis 2



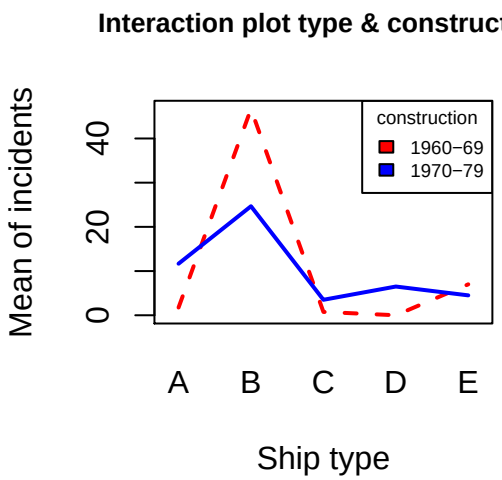


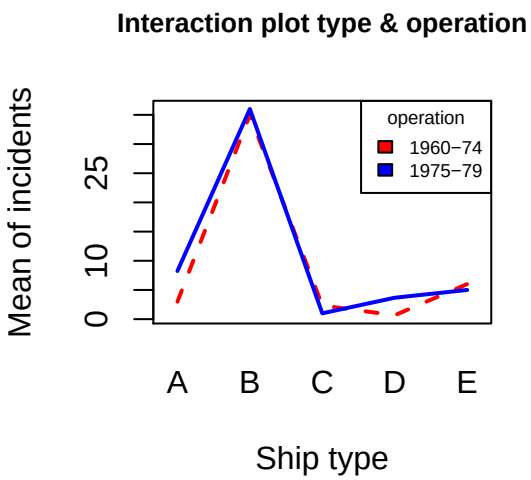
Analysis 3





Analysis 4


Interaction plot type & construction
Mean of incidents
0
20
40
construction
1960−69
1970−79
A
B
C
D
E
Ship type


Interaction plot type & operation
Mean of incidents
0
10
25
operation
1960−74
1975−79
A
B
C
D
E
Ship type