02424 - Assignment 2

Spring 22, 02424
Assignment 2
April 1, 2022

Nicolaj Hans Nielsen, s184335
Anton Ruby Larsen, s174356

Danmarks
Tekniske Universitet

Contents

A | An Ozone Model

A.1 Data Exploration

In this exercise, will work with the dataset described in table A.1. The data describes the ozone concentration measured in parts per million. To predict the ozone concentration, we are given 8 explanatory variables.

Variable	Domain	Description
Ozone	Continuous	Ozone concentration [ppm]
Temp	Continuous	Temperature [F°]
InvHt	Continuous	Inversion base height [feet]
Pres	Continuous	Daggett pressure gradient [mm Hg]
Vis	Continuous	Visibility [miles]
Hgt	Continuous	Vandenburg 500 millibar height [m]
Hum	Continuous	Humidity, percent
InvTmp	Continuous	Inversion base temperature [F°]
Wind	Continuous	Wind speed[mph]

Table A.1 – Description of data is found here

In table A.2, we see the first 20 data samples out of 330 samples in total. All variables except InvTemp are integers and all except Pres are positive. We also notice that it seems as multiple of the variables only attain values in a limit number of levels. In table A.3, we show the total number of unique levels for each variable. We see that many of the variables have few levels but especially Vis and Wind have very few levels. That said, we will not convert any of the variables to factors because they have a natural mutual rank directly proportional to their numeric value.

	Ozone	Temp	InvHt	Pres	Vis	Hgt	Hum	InvTmp	Wind
1	3	40	2693	-25	250	5710	28	47.66	4
2	5	45	590	-24	100	5700	37	55.04	3
3	5	54	1450	25	60	5760	51	57.02	3
4	6	35	1568	15	60	5720	69	53.78	4
5	4	45	2631	-33	100	5790	19	54.14	6
6	4	55	554	-28	250	5790	25	64.76	3
7	6	41	2083	23	120	5700	73	52.52	3
8	7	44	2654	-2	120	5700	59	48.38	3
9	4	54	5000	-19	120	5770	27	48.56	8
10	6	51	111	9	150	5720	44	63.14	3
11	5	51	492	-44	40	5760	33	64.58	6
12	4	54	5000	-44	200	5780	19	56.30	6
13	4	58	1249	-53	250	5830	19	75.74	3
14	7	61	5000	-67	200	5870	19	65.48	2
15	5	64	5000	-40	200	5840	19	63.32	5
16	9	67	639	1	150	5780	59	66.02	4
17	4	52	393	-68	10	5680	73	69.80	5
18	3	54	5000	-66	140	5720	19	54.68	4
19	4	54	5000	-58	250	5760	19	51.98	3
20	4	58	5000	-26	200	5730	26	51.98	4

Table A.2 – 20 samples of the given dataset.

Ozone	Temp	InvHt	Pres	Vis	Hgt	Hum	InvTmp	Wind
35	63	196	128	24	53	65	193	11

Table A.3 – Unique levels in each variable.

A.1.1 Covariance

We will now investigate the mutual correlation of the variables. In figure A.1, the correlations are plotted with a empirical kernel density and a general trend estimated with a spline. We see that ozone is very positively correlated with temperature and Hgt. In the other end of the spectrum, Ozone and InvHt are quite negatively correlated. Very noticeable is the ozone correlation with wind that is essentially 0. Hence we will probably see it get dropped unless it has some higher order effect on ozone.

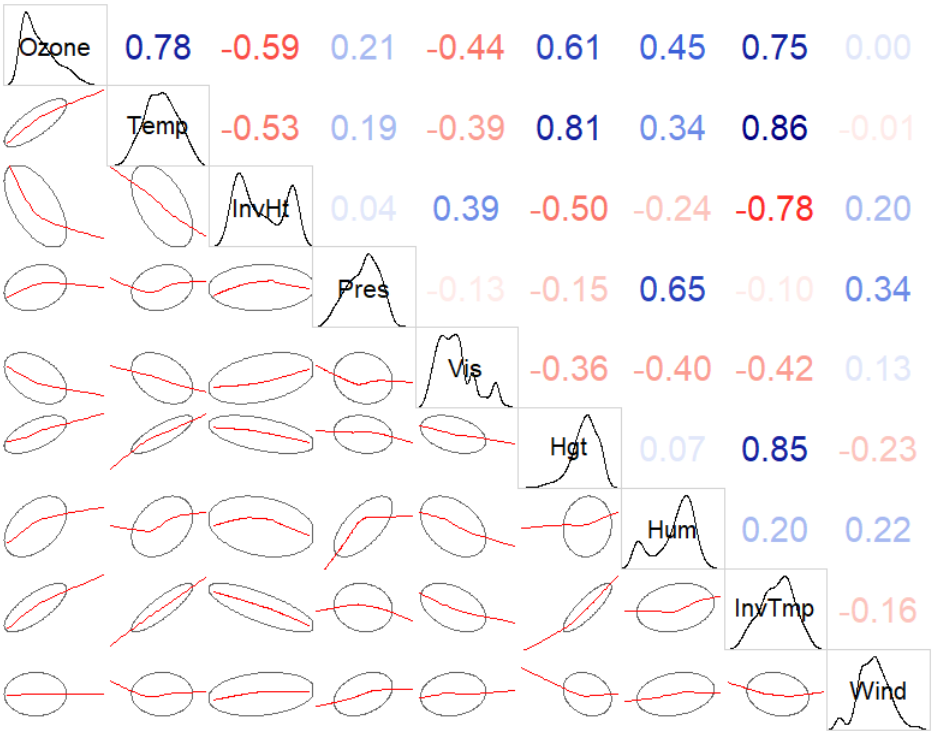

Figure A.1 – Mutual correlations in the provided dataset

A.2 Classic GLM

We will first fit a classic GLM, i.e., we assume normality, homoscedasticity and linearity as we did throughout all of assignment 1. We will only specify the model structure of the classic GLM in R notation here and refer to assignment 1 if one wants a more exhaustive mathematical notation. In a first attempt to formulate a sufficient model we will include up to all second order interactions as stated in eq. A.2.1.

$$\text{Ozone} \sim (\text{Temp} + \text{InvHt} + \text{Pres} + \text{Vis} + \text{Hgt} + \text{Hum} + \text{InvTemp} + \text{Wind})^2 \qquad \text{(A.2.1)}$$

As in assignment 1, we will perform a residual analysis to check if the model assumptions are met. We first consult the QQ-plot in figure A.2. We do not see anything alarming and hence proceed with residuals plotted against fitted and explanatory values presented in figure A.3. Here not all things are good. If we consult section 3.10 in [1], the fitted values, Pres, Temp, InvTemp and Hgt plotted against the residuals show a clear need of a transformation. Hence we consult page 88 in [1] where they suggest a log transformation to mitigate multiplicative effects. We will hence transform the response variable, Ozone.

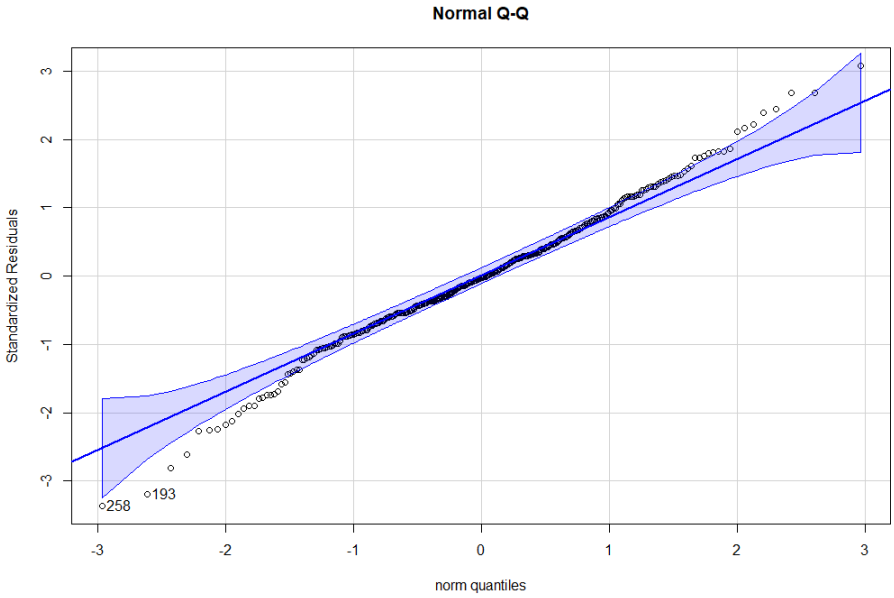

Figure A.2 – QQ plot for the model given in A.2.1

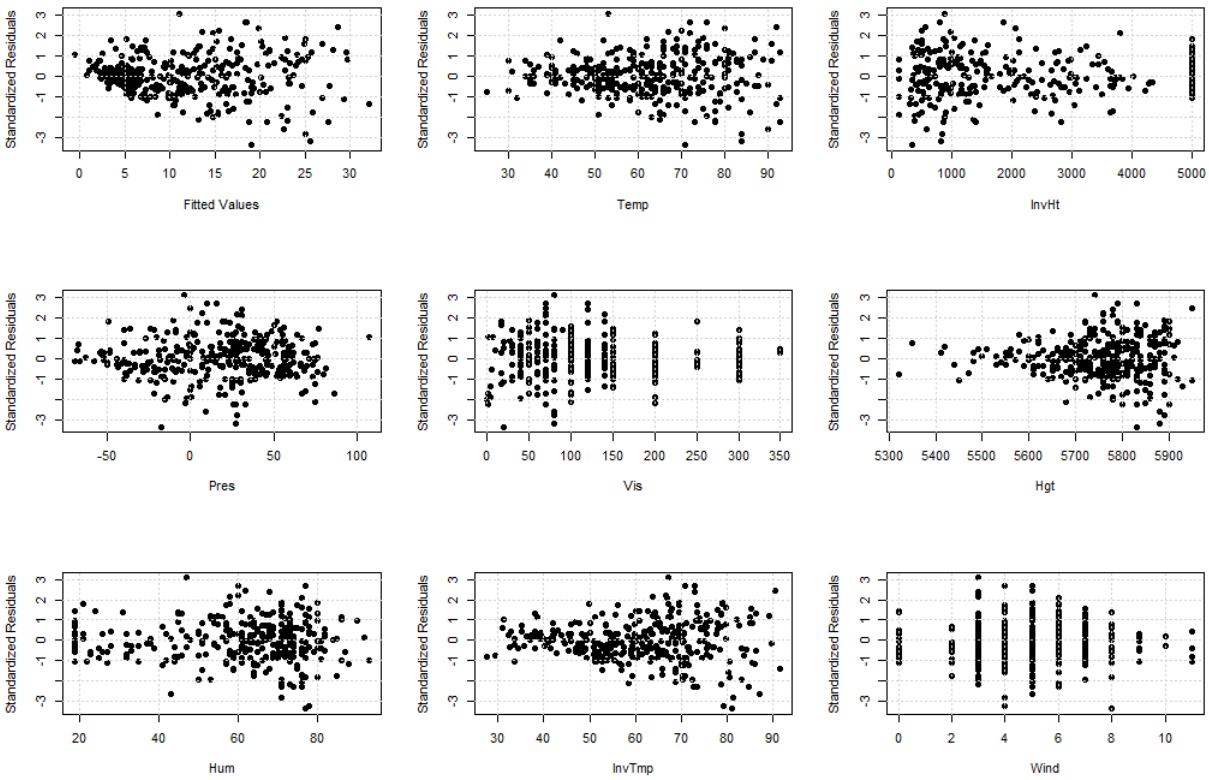

Figure A.3 – Residuals vs fitted and explanatory values for the model given in A.2.1

A.2.1 Transformed Response Variable

We fit a new model with a log transformation of the response variable.

$$\log(\text{Ozone}) \sim (\text{Temp} + \text{InvHt} + \text{Pres} + \text{Vis} + \text{Hgt} + \text{Hum} + \text{InvTemp} + \text{Wind})^2 \quad \text{(A.2.2)}$$

We again consult the QQ-plot and the residuals plotted against fitted and explanatory values. These are given in figure A.4 and A.5. We see that the QQ-plot is even better than for the untransformed data and in figure A.5 the cone shapes has disappeared. The only systematic behaviour we see in figure A.5 are straight lines which originates from the natural levels in the response variable. Hence we conclude our model meets sufficiency.

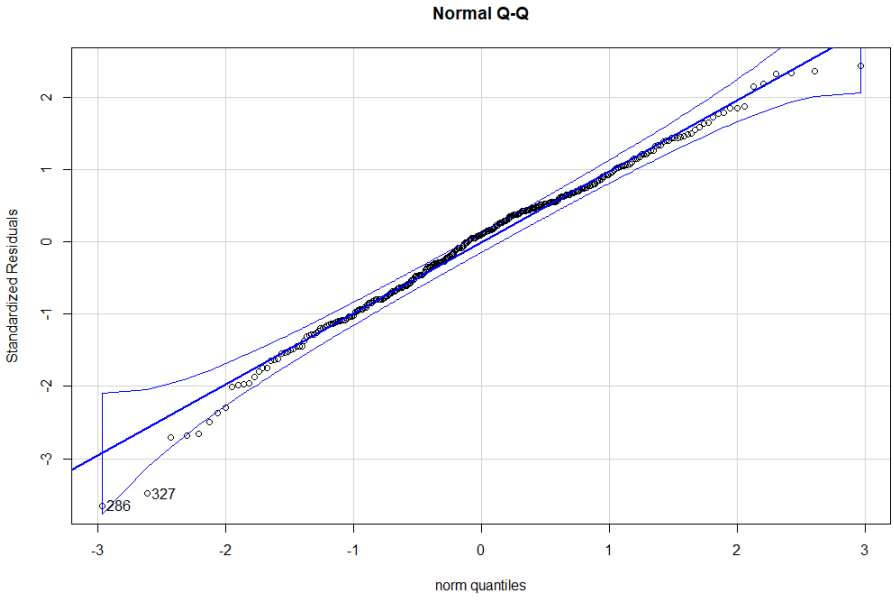

Figure A.4 – QQ plot for the model given in A.2.2

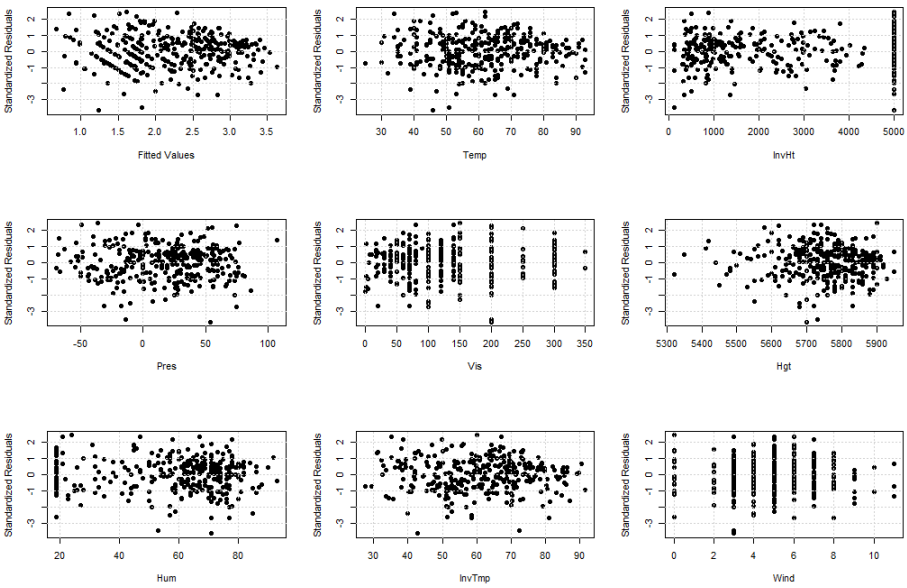

Figure A.5 – Residuals vs fitted and explanatory values for the model given in A.2.2

Backward selection

With the assumed sufficient model, we can now reduce the model until only necessary terms are left as we did in assignment 1. In this assignment, we will also use type II selection due to no natural chain of hypothesis. We will not display the interactions we remove here but refer to appendix D. Here we will just display the final model in equation A.2.3.

$$\begin{aligned}\log(\text{Ozone}) \sim \text{Temp} + \text{InvHt} + \text{Pres} + \text{Hgt} + \text{Hum} + \text{InvTmp} + \text{Pres:Hgt} \\ + \text{Pres:Hum} + \text{Pres:InvTmp} + \text{Hgt:Hum}\end{aligned} \tag{A.2.3}$$

Our final parameters with confidence interval in the response or identity domain are given in table A.4 and in figure A.6 we have chosen the three most significant parameters and the most significant second order interaction. To be able to plot then we lock all other variables at their mean and then only vary the parameter we are plotting.

	2.5 %	$\hat{\beta}$	97.5 %
(Intercept)	-6.3462960	6.7366553	19.1004854
Temp	0.0034518	0.0113302	0.0194410
InvHt	-0.0001163	-0.0000729	-0.0000296
Pres	0.1759629	0.3265077	0.4812166
Hgt	-0.0031714	-0.0009984	0.0012949
Hum	-0.4104574	-0.2371232	-0.0558516
InvTmp	-0.0048184	0.0063427	0.0173962
Pres:Hgt	-0.0000883	-0.0000599	-0.0000323
Pres:Hum	-0.0002419	-0.0001496	-0.0000618
Pres:InvTmp	0.0002586	0.0004710	0.0006887
Hgt:Hum	0.0000114	0.0000425	0.0000723

Table A.4 – Parameters for A.2.3 in the response or identity domain.

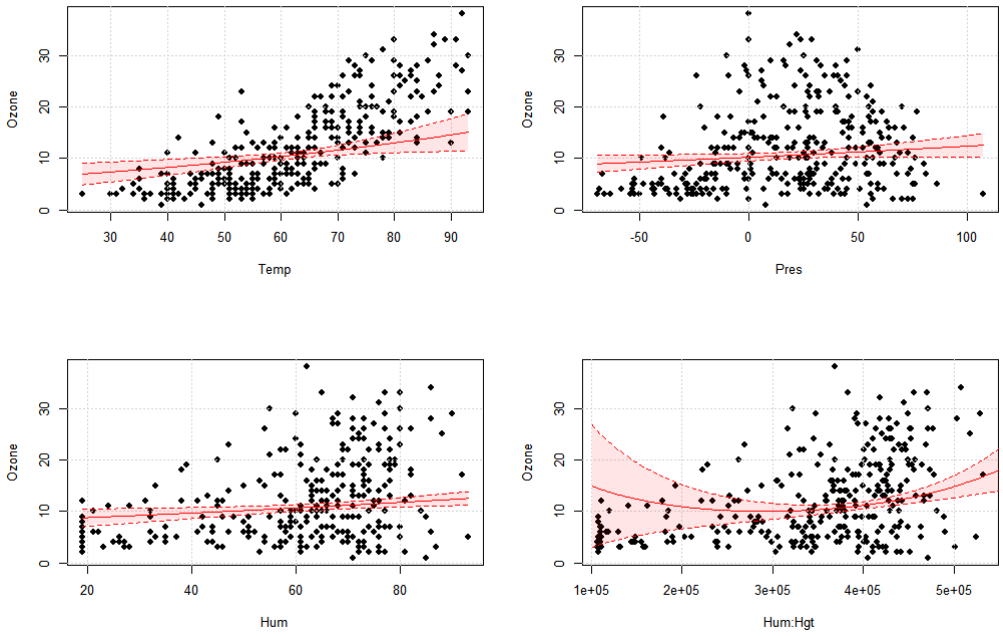

Figure A.6 – Some of the behaviour of our final model for the Classic GLM with a log transformation of the response variable.

A.3 General Linear Models

We will now expand our modeling framework and introduce generalized linear models. We refer to [1] and [2] for a thorough introduction to the framework of generalized linear models. In short, the idea is to extend the class of distribution considered, and formulate a linear model of a transformed version of the mean values.
In the following, we will list the specification needed to formulate a generalized linear model and relate them to the problem at hand.

Distributional assumption To specify a distributional assumption, we consider the type of data at hand. In our case, we have to consider ozone. The intrinsic data type would be continuous but it seems that the ozone measurements have been rounded to nearest integer which motivates a potential discrete treatment. This gives rise to the following distributional assumption.

Continues and positive Gamma and Inverse Gaussian

Discrete and positive Poisson and Negative Binomial

Link function The link function describes the relationship between the linear predictor η_i and the mean value parameter, [1], def 4.9:

$$\eta_i = g(\mu_i) \tag{A.3.1}$$

For each distributional assumption, there will be multiple link functions. A way to check if the link function is appropriate is to examine the working response[1] plotted against the linear predictor, see section 8.7.3 in [2]. The link function is appropriate if the plot with η_i against the working response is approximately linear.

Linear predictor Here we have to analyse the construction of the linear predictor under a chosen design.

$$\eta = \mathbf{X}\beta \tag{A.3.2}$$

The evaluation of a specific linear predictor is conducted by evaluating diagnostic residual plots as described in assignment 1. Worth noticing, is that now we should consider new scaled residuals which we will introduce below.

Precision Some models in the exponential family has a variance which is not related to the mean value. One example is the Gaussian distribution for which we estimate the variance, σ^2, separately from the mean value. The precision is just the inverse of the variance and when choosing a distribution we must be aware of how and if we need to estimate a precision separately from the mean. In table 4.2 in [1] one can see some precision parameters for different distributions.

A.3.1 Residuals

As with the general linear models, the residuals are the backbone of all the diagnostics which enables us to identify discrepancies between model assumptions and data, i.e. identify miss-specified link functions, poor linear predictors, overall goodness-of-fit, outliers and so on.
Before with the general linear model, we could use the reponse residuals r_R directly. Let y_i be the i'th observed response and $\hat{\mu}_i$ be the mean value parameter of a specified model, then the response residuals are defined as:

$$r_i^R = y_i - \hat{\mu}_i \tag{A.3.3}$$

In general linear models, the entire model diagnostic is built under the assumption of normality of these residuals. For a generalized linear model, the variance will in general depend on the mean which makes the response residuals inappropriate [1]. A way to handle the non-constant variance is by using the person residuals, def 4.16 in [1].

[1] The working response is defined as $z = r_W + \eta$ where η is the linear predictor and r_W is the working residual.

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/\omega_i}} \tag{A.3.4}$$

where $V(\cdot)$ is the variance function under the distribution assumption. Here ω_i is the i'th weights on the residual if it has been introduced. The Pearson residuals have the neat property that they are asymptotically normal under the true model if the Central Limit Theorem holds, section 7.5, [2]. The same property holds for the deviance residuals if the saddle point approximation holds, section 7.5, [2]. They are defined in definition 4.15 in [1] and given as,

$$r_i^P = \text{sign}(y_i - \hat{\mu}_i)\sqrt{\omega_i d(y_i, \mu_i)} \tag{A.3.5}$$

where sign$(\cdot)$ is the sign function and $d(y_i, \mu_i)$ is the unit deviance, def. 4.4 [1]. The deviance residuals have the additional property that they represent the difference between the log likelihood of the saturated and fitted model.
If r_P and d_D are asymptotically normal, it enables us to use all the diagnostics from the general linear model and χ^2_{n-p} GoF tests. However, if the response variable is far from normal and especially if it is distributed on a limited set of values, then r_P and r_D would be far from normal [2] [3]. This disintegrates the standard diagnostic apparatus as they rely on the assumption of normality. Therefore, we introduce the quantile residuals p. 302, [2].

$$r_i^Q = \Phi^{-1}\{F(y_i; \hat{\mu})\} \tag{A.3.6}$$

where F is the cumulative distribution function, CDF, of the distribution assumption with mean value parameter μ and Φ^{-1} is the CDF of a standard normal distribution. Essentially, we invert the fitted distribution function for each y_i and then find the corresponding quantile of the normal distribution. It has been shown that r_Q is normal under a correctly specified model [4].
The last residual to be introduced is the working residual that we use to identify if the link function is appropriate. This is defined as

$$r_i^W = (y_i - \hat{\mu}_i)\frac{\partial \hat{\eta}_i}{\partial \hat{\mu}_i} \tag{A.3.7}$$

where η_i is the linear predictor.

A.3.2 Negative Binomial

We have now defined our diagnostic framework and can now fit our first model. As discussed our response variable can both be interpreted as a count variable and a continuous variable. We will start by fitting the Negative Binomial distribution which assumes count data. As an initial model we assume a log link with the design given in A.3.8.

$$\texttt{Ozone} \sim (\texttt{Temp} + \texttt{InvHt} + \texttt{Pres} + \texttt{Vis} + \texttt{Hgt} + \texttt{Hum} + \texttt{InvTemp} + \texttt{Wind})^2 \tag{A.3.8}$$

We start by checking the sufficiency of our distribution, remark 4.22, [1]. In table A.5 both p values for the Pearsons and Deviance residuals are stated. We see that both p values are above 0.05 and hence there is no evidence contradicting the model assumption of a Negative Binomial distribution.

	GoF Statistic	Df	p Value
Deviance	310.13	293	0.24
Pearson	306.71	293	0.28

Table A.5

We continue with testing our link function by plotting the working response against the linear predictor. We see from figure A.7 that the general trend is linear and hence nor our link function is violating model assumptions.

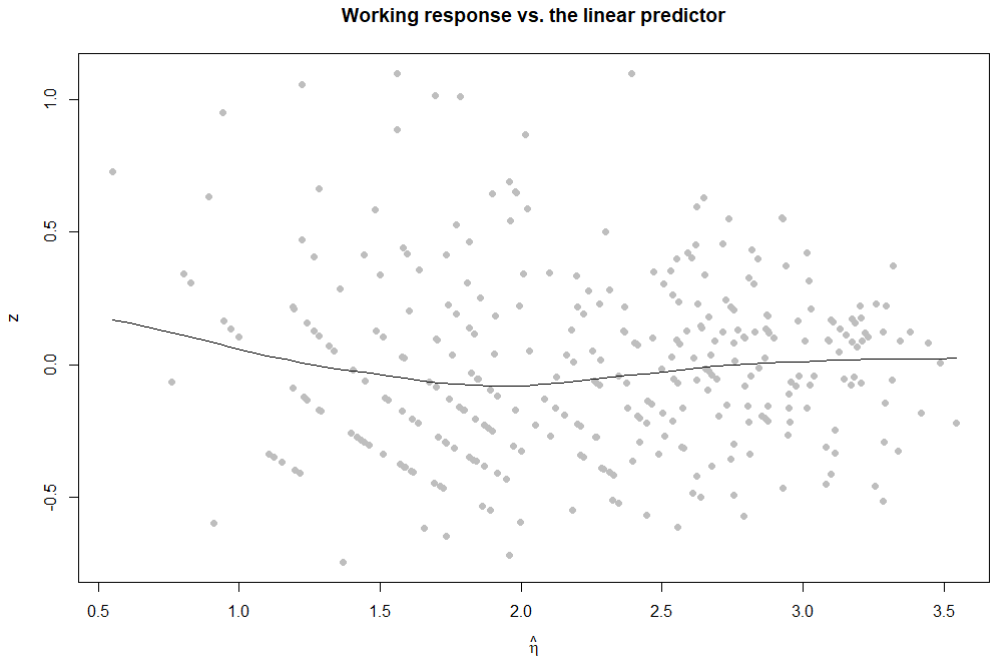

Figure A.7 – Working response plotted against the linear predictor for the Negative Binomial distribution with log link.

To test the sufficiency of our model design, i.e. the linear predictor, we will consult the QQ plot and the quantile residuals plot with the mean value parameter and explanatory variables. In figure A.9 and A.8, we see that again no problems and our linear predictor is not violating the model assumptions.

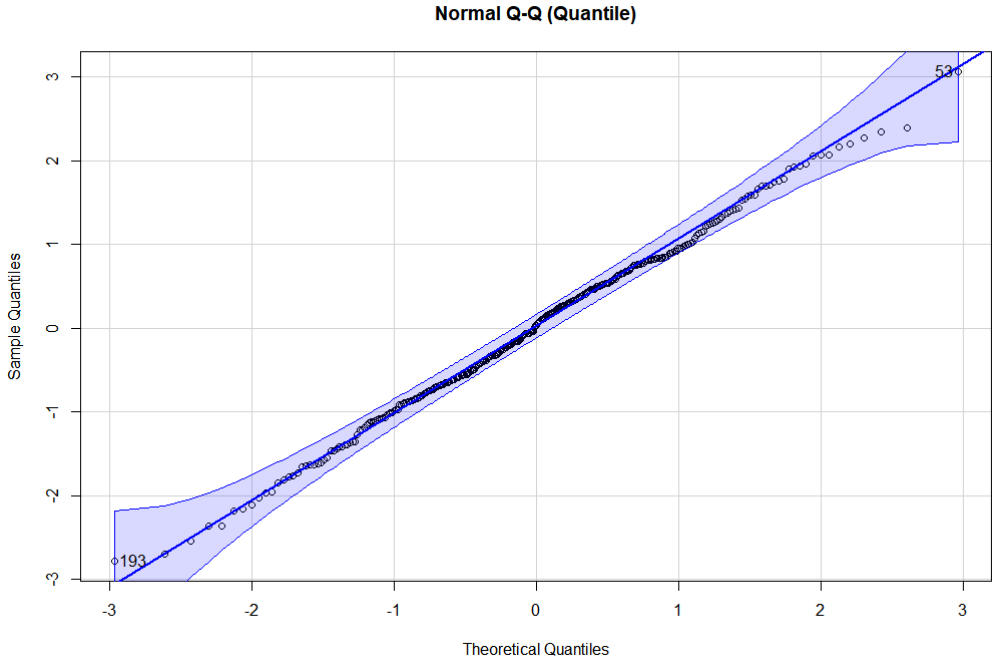

Figure A.8 – QQ plot for the quantile residuals of the Negative Binomial distribution with log link given in A.3.8

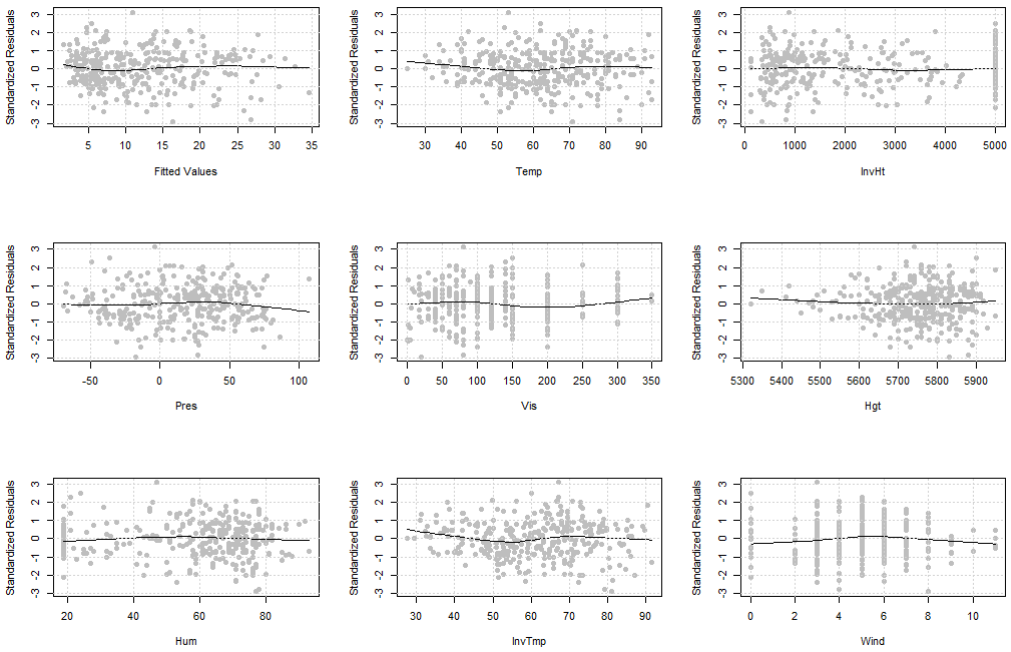

Figure A.9 – Quantile residuals plotted against mean value parameter and explanatory variables for the Negative Binomial distribution with log link.

We have now tested all model assumptions given in the start of section A.3 and remark 4.11 in [1], and hence we can proceed with a type II backward selection. We will not show all the term reductions here but refer to table D.2 in appendix D. We will only state the

final model here in equation A.3.9.

$$\begin{aligned}\texttt{Ozone} \sim \texttt{Temp} + \texttt{InvHt} + \texttt{Pres} + \texttt{Hgt} + \texttt{Hum} + \texttt{InvTmp} + \texttt{Vis} + \texttt{Temp:Pres} \\ + \texttt{Temp:Hgt} + \texttt{Pres:Hgt} + \texttt{Pres:Hum} + \texttt{Pres:InvTmp} + \texttt{Hgt:Hum}\end{aligned} \tag{A.3.9}$$

Our final parameters with confidence interval are given in table A.6 and in figure A.14 we have chosen the three most significant parameters and the most significant second order interaction. To be able to plot then we have locked all other variables at their mean and then only varied the parameter we are plotting.

	2.5 %	$\hat{\beta}$	97.5 %
(Intercept)	-0.4391871	15.5337789	31.6778260
Temp	0.1022613	0.2493992	0.3940310
InvHt	-0.0000642	0.0000525	0.0001685
Pres	0.2185415	0.3746808	0.5278469
Hgt	-0.0060783	-0.0030314	-0.0000088
Hum	-0.8901121	-0.6007359	-0.3098642
InvTmp	0.0163345	0.0478249	0.0791131
Vis	-0.0002840	0.0022648	0.0048176
Temp:Hgt	-0.0000655	-0.0000405	-0.0000150
InvHt:Hum	-0.0000041	-0.0000022	-0.0000002
Pres:Hgt	-0.0000962	-0.0000679	-0.0000390
Pres:Hum	-0.0002626	-0.0001902	-0.0001168
Pres:InvTmp	0.0002864	0.0004885	0.0006847
Hgt:Hum	0.0000570	0.0001123	0.0001673
Hum:InvTmp	-0.0010901	-0.0005858	-0.0000784
InvTmp:Vis	-0.0000928	-0.0000486	-0.0000042

Table A.6 – Coefficients for Negative Binomial

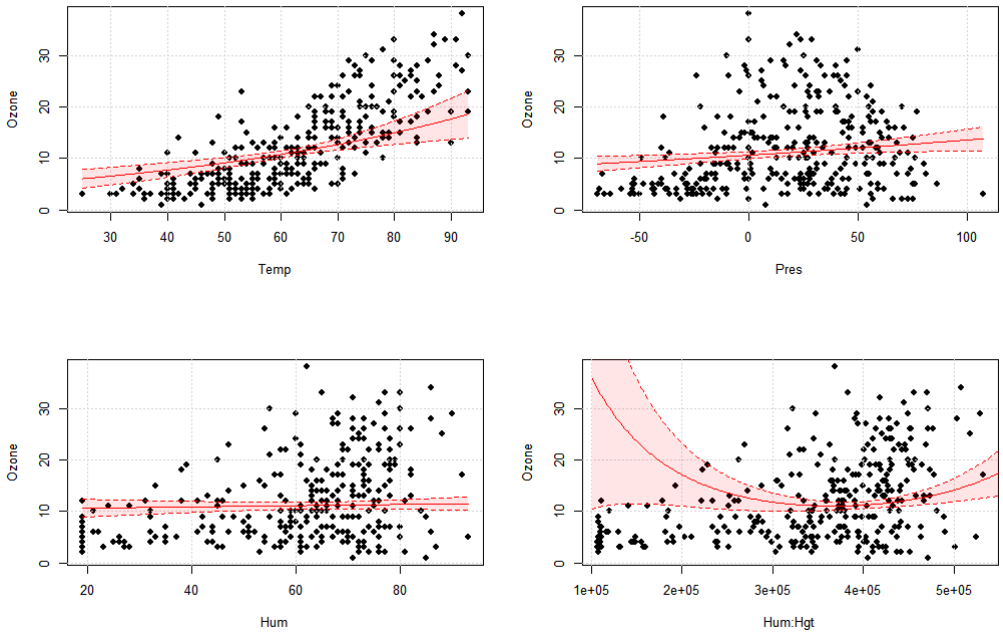

Figure A.10 – Some of the behaviour of our final model for the Negative Binomial distribution with log link.

A.3.3 Poisson

We continue with another distribution which assumes count data, namely the Poisson distribution. As an initial model we assume a log link with the design given in A.3.10.

$$\texttt{Ozone} \sim (\texttt{Temp} + \texttt{InvHt} + \texttt{Pres} + \texttt{Vis} + \texttt{Hgt} + \texttt{Hum} + \texttt{InvTemp} + \texttt{Wind})^2 \tag{A.3.10}$$

As for the Negative Binomial we will start by testing our distributional assumption. In table A.7 we see that the Poisson distribution fails both on the Deviance and Pearson residuals. One problem could be that the data is over or under dispersed and the variance of the Poisson is locked to be the same as the variance. Hence we will try to fit a Quasipoisson which allows for a separately fitted variance.

	GoF Statistic	Df	p Value
Deviance	342.91	293	0.02
Pearson	339.63	293	0.03

Table A.7 – Test for model sufficiency of the Poisson distribution.

In table A.8 we see that when we allow the variance to be fitted separately there is no problem. Hence we will proceed with the Quasipoisson but not change the log link assumption and the design given in A.3.10.

	GoF Statistic	Df	p Value
Deviance	342.91	293	0.44
Pearson	339.63	293	0.49

Table A.8 – Test for model sufficiency of the Quasipoisson distribution.

We continue with testing our link function by plotting the working response against the linear predictor. We see from figure A.11 that the general trend is linear and hence nor our link function is violating model assumptions.

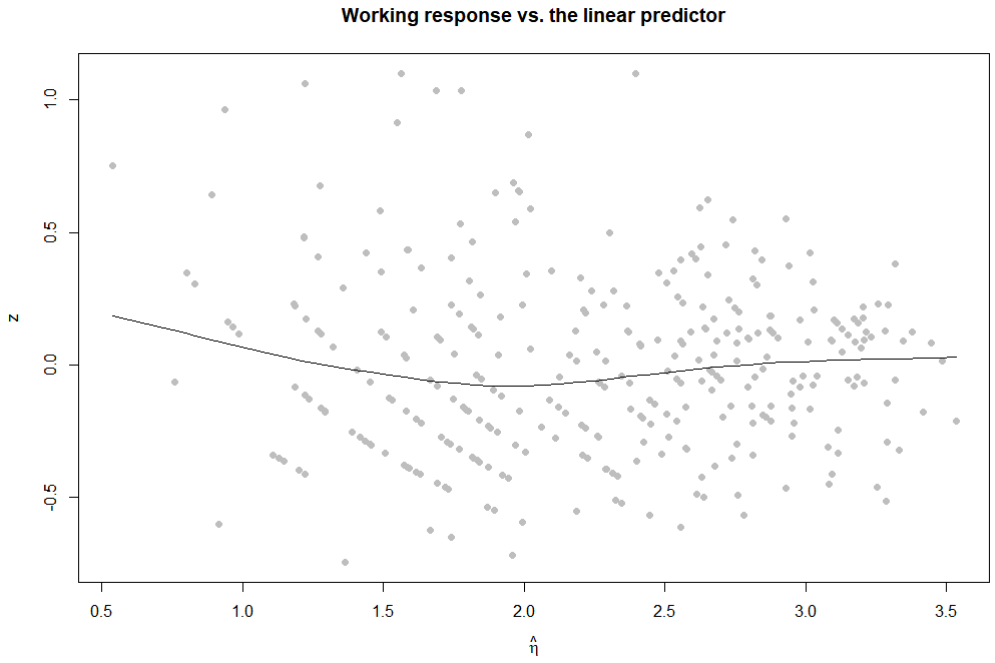

Figure A.11 – Working response plotted against the linear predictor for the Quasipoisson distribution with log link.

To test the sufficiency of our model design, i.e. the linear predictor, we will consult the QQ plot and the quantile residuals plotted against mean value parameter and explanatory variables. We see in figure **??** that again we see no problems and nor our linear predictor is violating model assumptions

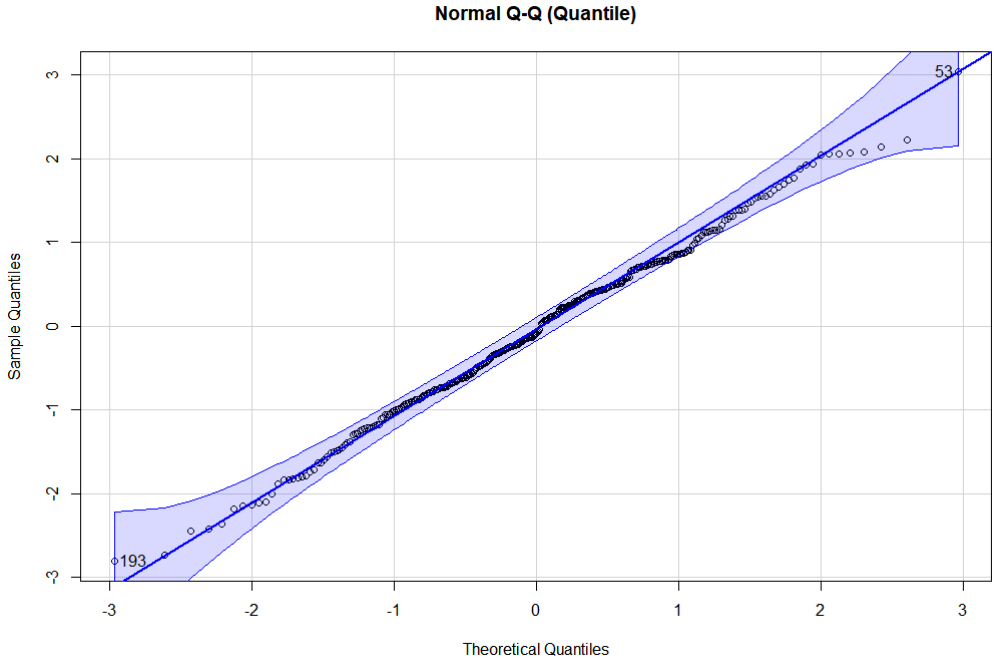

Figure A.12 – QQ plot for the quantile residuals of the Quasipoisson distribution with log link given in A.3.10

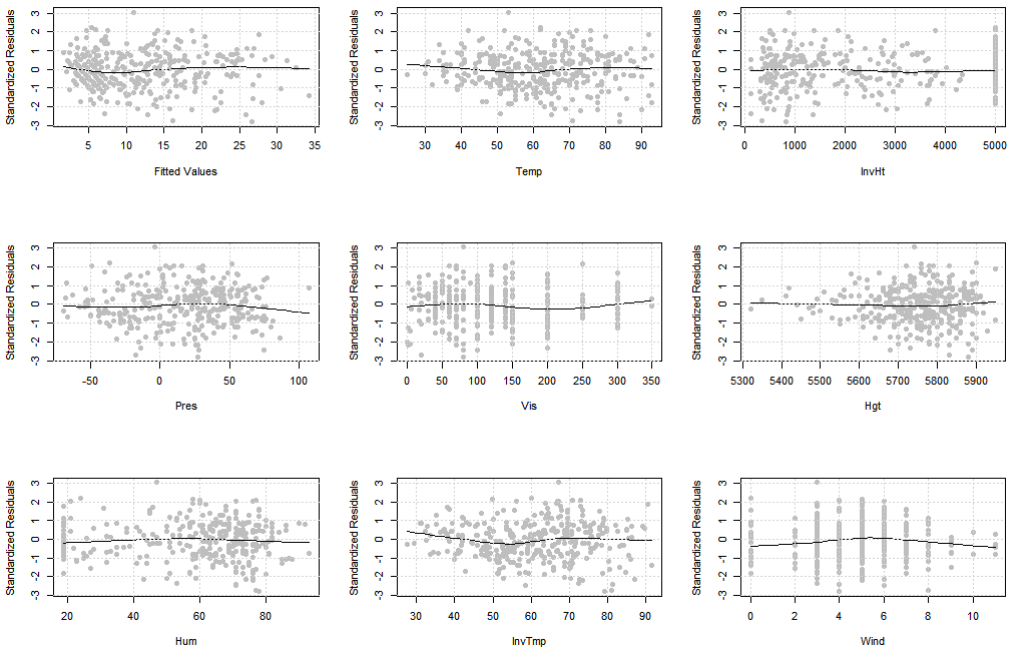

Figure A.13 – Quantile residuals plotted against mean value parameter and explanatory variables for the Quasipoisson distribution with log link.

We have now tested all model assumptions given in the start of section A.3 and remark 4.11 in [1], and hence we can proceed with a type II backward selection. We will not show all the term reductions here but refer to table D.3 in appendix D. We will only state the

final model here in equation A.3.11.

$$\begin{aligned}\texttt{Ozone} \sim \texttt{Temp} + \texttt{InvHt} + \texttt{Pres} + \texttt{Hgt} + \texttt{Hum} + \texttt{InvTmp} + \texttt{Vis} + \texttt{Temp:Pres} \\ + \texttt{Temp:Hgt} + \texttt{Pres:Hgt} + \texttt{Pres:Hum} + \texttt{Pres:InvTmp} + \texttt{Hgt:Hum}\end{aligned} \tag{A.3.11}$$

Our final parameters with confidence interval are given in table A.9 and in figure **??** we have chosen the three most significant parameters and the most significant second order interaction. To be able to plot then we have locked all other variables at their mean and then only varied the parameter we are plotting.

	2.5 %	$\hat{\beta}$	97.5 %
(Intercept)	-6.6289112	4.8337195	16.0246983
Temp	0.1176700	0.2614326	0.4093776
InvHt	-0.0001109	-0.0000690	-0.0000269
Pres	0.0988300	0.2500776	0.4007949
Hgt	-0.0026792	-0.0007153	0.0012940
Hum	-0.5541465	-0.3833646	-0.2134806
InvTmp	-0.0054388	0.0059121	0.0172191
Vis	-0.0012589	-0.0006647	-0.0000773
Temp:Pres	-0.0005050	-0.0003035	-0.0001035
Temp:Hgt	-0.0000670	-0.0000417	-0.0000170
Pres:Hgt	-0.0000726	-0.0000449	-0.0000171
Pres:Hum	-0.0002195	-0.0001324	-0.0000461
Pres:InvTmp	0.0004021	0.0006031	0.0008077
Hgt:Hum	0.0000379	0.0000673	0.0000968

Table A.9 – Coefficients for Quasipoisson

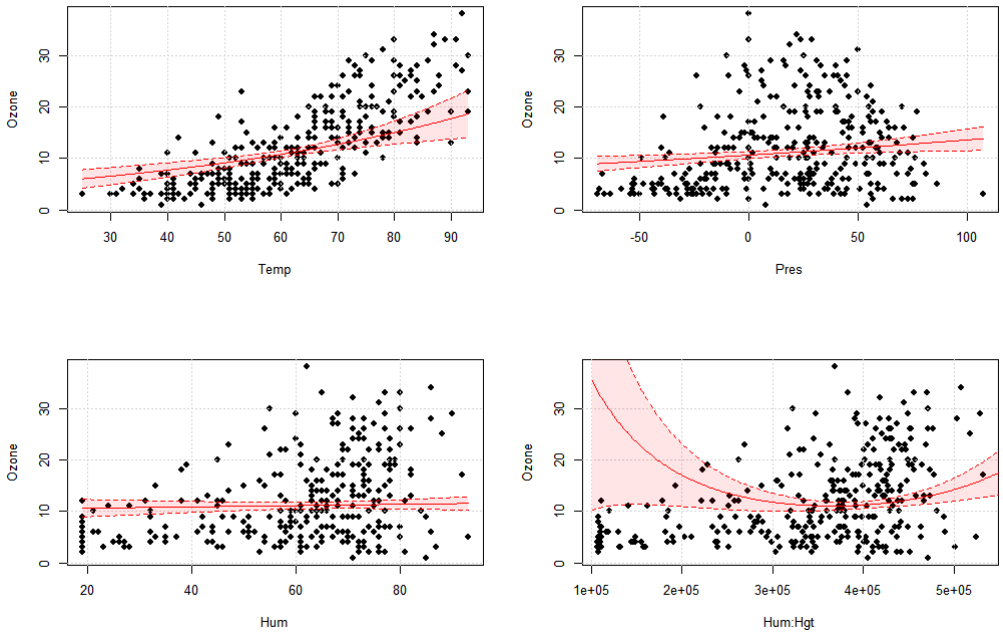

Figure A.14 – Some of the behaviour of our final model for the Quasipoisson distribution with log link.

A.3.4 Gamma

We now switch type of distribution and assumes a Gamma distribution which assumes the data is positive and continuous. As an initial model we assume a log link with the design given in A.3.12.

$$\texttt{Ozone} \sim (\texttt{Temp} + \texttt{InvHt} + \texttt{Pres} + \texttt{Vis} + \texttt{Hgt} + \texttt{Hum} + \texttt{InvTemp} + \texttt{Wind})^2 \qquad \text{(A.3.12)}$$

As for the two other distributions we will start by testing our distributional assumption. In table **??** we see that both p values are above 0.05 and hence there is no evidence contradicting the model assumption of a Gamma distribution.

	GoF Statistic	Df	p Value
Deviance	37.66	293	0.25
Pearson	35.68	293	0.49

Table A.10

We continue with testing our link function by plotting the working response against the linear predictor. We see from figure A.15 that the general trend is linear and hence nor our link function is violating model assumptions.

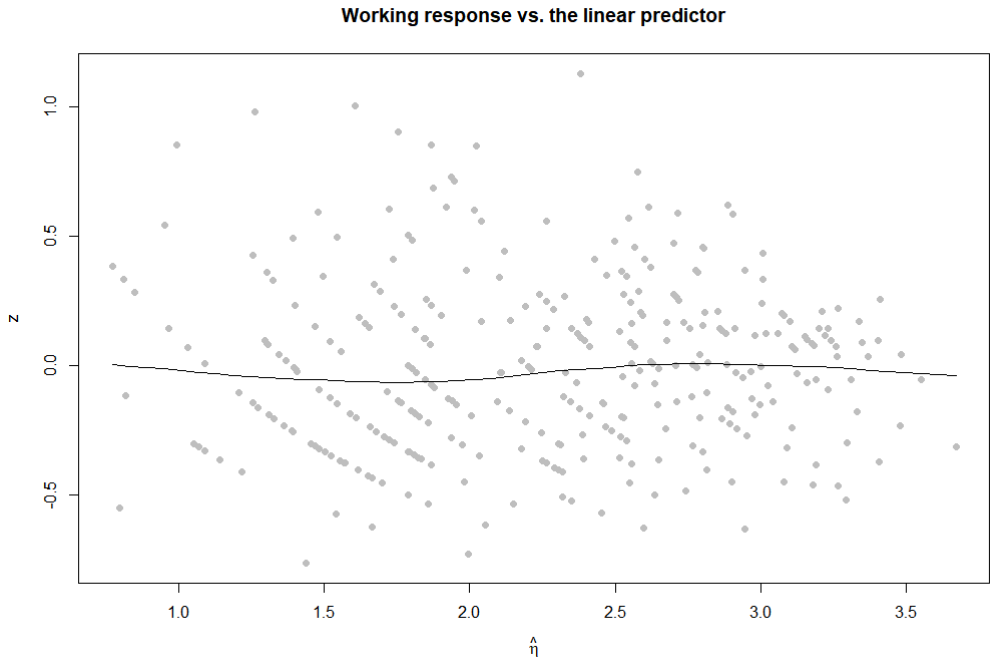

Figure A.15 – Working response plotted against the linear predictor for the Gamma distribution with log link.

To test the sufficiency of our model design, i.e. the linear predictor, we will consult the QQ plot and the quantile residuals plotted against mean value parameter and explanatory variables. We see in figure A.16 and A.17 that again we see no problems and nor our linear predictor is violating model assumptions

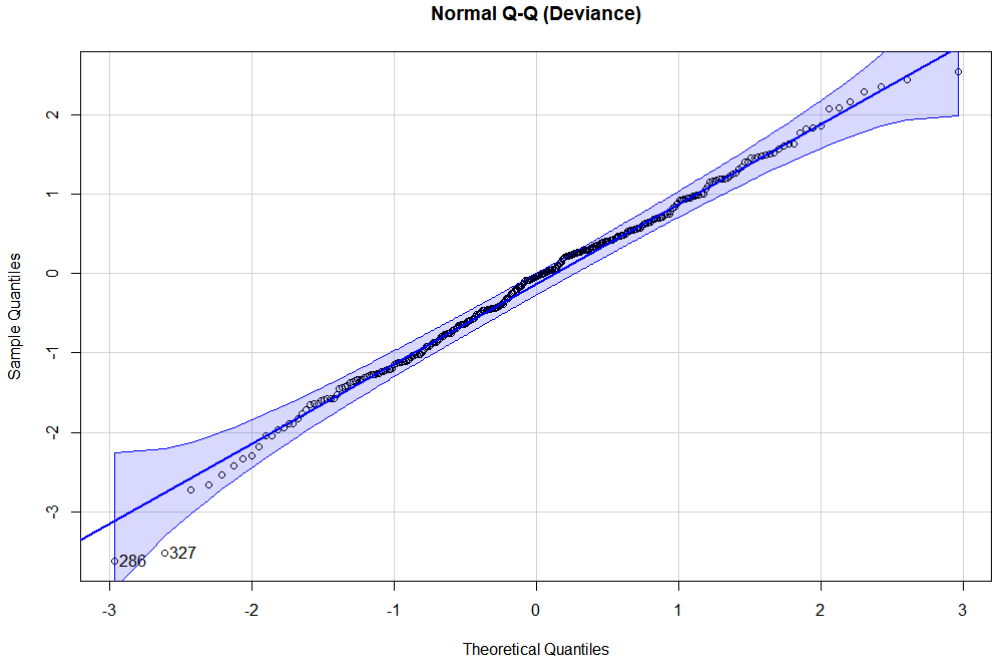

Figure A.16 – QQ plot for the quantile residuals of the Gamma distribution with log link given in A.3.12

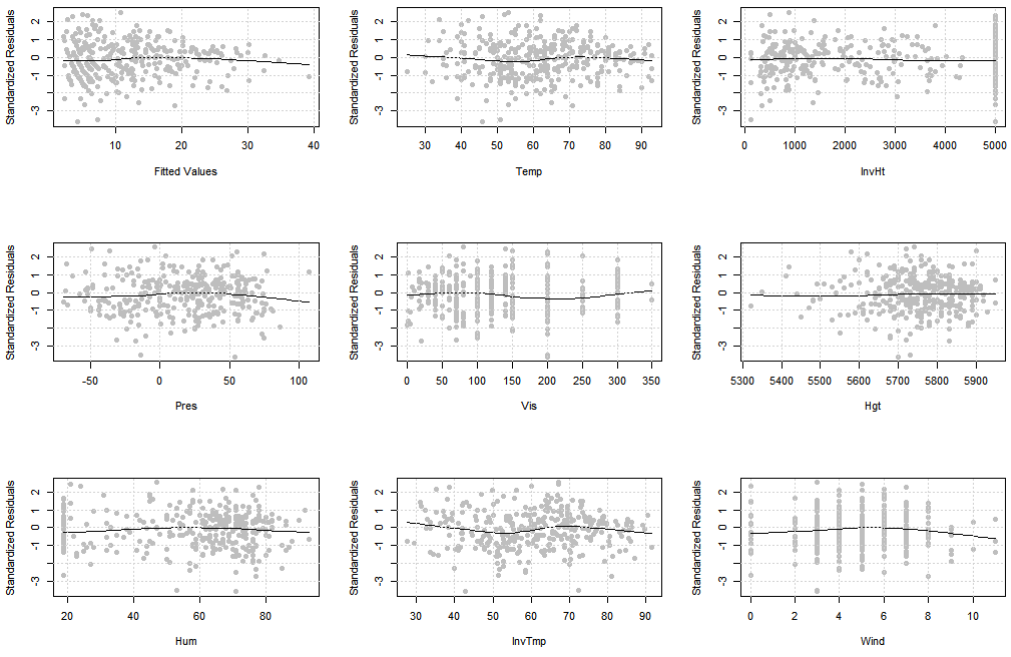

Figure A.17 – Quantile residuals plotted against mean value parameter and explanatory variables for the Quasipoisson distribution with log link.

We have now tested all model assumptions given in the start of section A.3 and remark 4.11 in [1], and hence we can proceed with a type II backward selection. We will not show all the term reductions here but refer to table D.4 in appendix D. We will only state the

final model here in equation A.3.13.

$$\begin{aligned}\texttt{Ozone} \sim\ & \texttt{Temp} + \texttt{InvHt} + \texttt{Pres} + \texttt{Hgt} + \texttt{Hum} + \texttt{InvTmp} + \texttt{Vis} + \texttt{Temp:Pres} \\ & + \texttt{Temp:Hgt} + \texttt{InvHt:Hum} + \texttt{Pres:Hgt} + \texttt{Pres:Hum} \\ & + \texttt{Pres:InvTmp} + \texttt{Hgt:Hum} + \texttt{Hum:InvTmp} + \texttt{InvTmp:Vis}\end{aligned} \tag{A.3.13}$$

Our final parameters with confidence interval are given in table A.11 and in figure A.18 we have chosen the three most significant parameters and the most significant second order interaction. To be able to plot then we have locked all other variables at their mean and then only varied the parameter we are plotting.

	2.5 %	$\hat{\beta}$	97.5 %
(Intercept)	-0.4391871	15.5337789	31.6778260
Temp	0.1022613	0.2493992	0.3940310
InvHt	-0.0000642	0.0000525	0.0001685
Pres	0.2185415	0.3746808	0.5278469
Hgt	-0.0060783	-0.0030314	-0.0000088
Hum	-0.8901121	-0.6007359	-0.3098642
InvTmp	0.0163345	0.0478249	0.0791131
Vis	-0.0002840	0.0022648	0.0048176
Temp:Hgt	-0.0000655	-0.0000405	-0.0000150
InvHt:Hum	-0.0000041	-0.0000022	-0.0000002
Pres:Hgt	-0.0000962	-0.0000679	-0.0000390
Pres:Hum	-0.0002626	-0.0001902	-0.0001168
Pres:InvTmp	0.0002864	0.0004885	0.0006847
Hgt:Hum	0.0000570	0.0001123	0.0001673
Hum:InvTmp	-0.0010901	-0.0005858	-0.0000784
InvTmp:Vis	-0.0000928	-0.0000486	-0.0000042

Table A.11 – Coefficients for the the Gamma distribution with log link

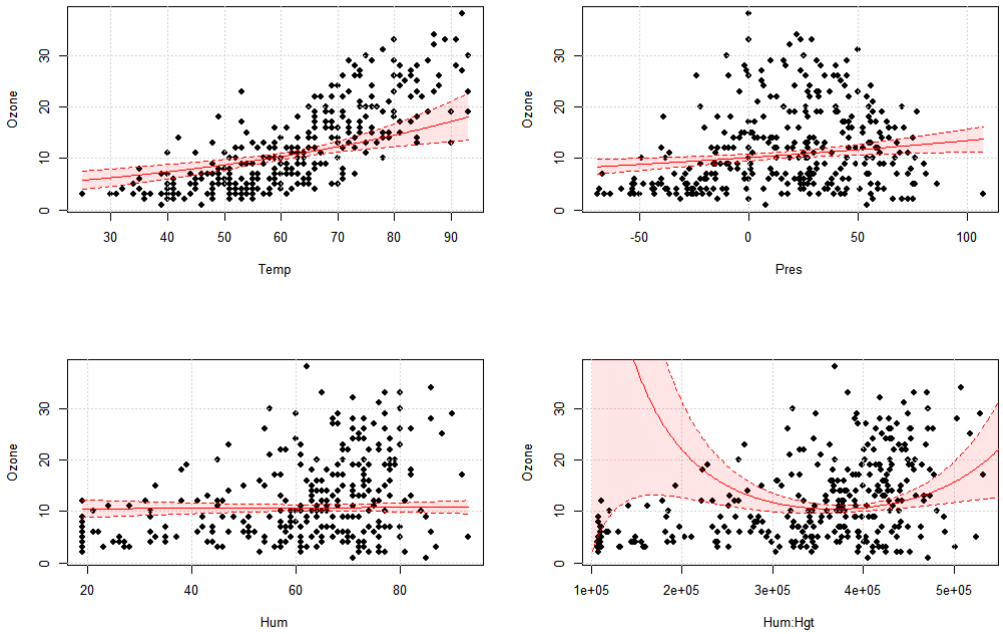

Figure A.18 – Some of the behaviour of our final model for the Gamma distribution with log link.

A.4 Comparison

To decide which GLM to use, we will first inspect some quantitative measures. In table A.12 all the fitted GLMs are stated with number of parameters, AIC and BIC. We see that the Quasipoisson is best on BIC and the Negative Binomial is the best on AIC. Hence we just choose the Quasipoisson.

Distribution	Link	Number of parameters	AIC	BIC
Gaussian	Log	11	1852.98	1898.56
Negative Binomial	Log	14	1730.55	1787.53
Quasipoisson	Log	14	1732.40	1785.59
Gamma	Log	16	1737.98	1802.57

Table A.12 – Quantitative measures for all the fitted GLMs.

One problem with only using quantitative to decide upon which model to select is that it only considers in-sample knowledge. If one wants to extrapolate out from the fitted region a model which violates the underlying nature of the data can produce some really wierd results.
Our response variable is a concentration which is only given in whole numbers. Hence we have also fitted count models but there is nothing wrong with a concentration which is not a integer. Therefore it is probably a poor choice of model to select a model only defined in the integers. We will hence go with the Gamma GLM even though it quantitatively is a worse model than both the Quasipoisson and the Negative Binomial.

A.5 Final model

We have hence decided to use the Gamma distribution for which parameters, model design and some plots of parameters are shown in section A.3.4

B | Clothing insulation level

In this exercise, we are to use the dataset from assignment 1 again but this time `clo` contains the number of times that each subject changes clothing insulation level during a day. To predict `clo` we are provided with the duration of time in which observation were measured (`time`), the number of observations during the day (`nobs`), the sex of the subject (`sex`), and average outdoor and indoor operating temperature (`tOut`, and `tInOp`), during a day.

The response is discrete hence the distributions like the Binomial, Poisson and negative Binomial should be considered. We will first investigate the Binomial GLM.

B.1 Binomial

To model the given data with the Binomial distribution, we need to consider `clo` as a proportion of the total number of observations, `nobs`, i.e. $Y = \frac{Z}{n} = \frac{\texttt{clo}}{\texttt{nobs}}$. To model this we consider the following model which is also given in slide 12, lecture 6.

$$f_Y(n, \mu) = \binom{n}{yn} \mu^{yn} (1-\mu)^{n-yn} \tag{B.1.1}$$

To use B.1.1, we need to decide upon a link function to map between the mean value parameter, μ, and the linear predictor, η. To do this, we will consider the 4 different link functions given in table B.1.

Link	$\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
logit	$\log(\mu/(1-\mu))$	$\exp(\eta)/[1+\exp(\eta)]$
cauchit	$\tan(\pi(\mu-0.5))$	$\frac{\pi - 2\arctan(\eta)}{2\pi}$
probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
cloglog	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\eta))$

Table B.1 – The 4 different link functions considered for the Binomial distribution.

We now want to develop a generalized linear model using the Binomial distribution, a link functions and a linear predictor of the available inputs `sex`, `tOut`, and `tInOp`. Hence

$$g(\mu) = \eta = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma) \tag{B.1.2}$$

As a first sufficient model we will use the same design matrix $\mathbf{X}$ as in section A3 in assignment 1. We will only give the R notation for the interactions here.

$$\frac{\texttt{clo}}{\texttt{nobs}} \sim \texttt{tOut} * \texttt{tInOp} * \texttt{factor(sex)} \tag{B.1.3}$$

Before moving on to investigate the Binomial distribution with the different link functions we will just make a short note on the precision parameter λ. In remark 4.11 in [1] it is stated that one need to specify a distribution, a link function, a linear predictor and a precision parameter. From table 4.2 in [1], we know that the precision parameter for the Binomial distribution is given as $\lambda = n$. Hence it will just equal the variable `nobs`.

B.1.1 Diagnostics

All the following diagnostic tools are from chapter 8 in [2]. Red and blue will respectively indicate female and male in all the following plots.

We will first fit four models, each with one of the link functions given in table B.1. In table B.2 we see p values for two tests of model sufficiency as explained in remark 4.22 in [1]. One is based on the pearson residuals and one is based on the deviance residuals. We must though remember that deviance and pearson residuals are not well suited for discrete distributions as mentioned in section A.3. From section 7.5 in [2] we see that for the Binomial distribution, if $\min(ny) < 3$, we cannot trust the deviance residuals and if $\min(ny) < 5$, we cannot trust the pearson residuals. This is the case for our data and hence we cannot really trust these p values. We will therefore not claim that the Binomial distribution is contradicting the model assumptions based on these p values.

The table B.2, we see that the cauchit link function has the best performence in BIC and AIC.

Link	p value(Deviance)	p value(Pearson)	AIC	BIC
Logit	0.0126	0.0036	279.9572	303.2584
Probit	0.0122	0.0040	280.2212	303.5224
cloglog	0.0127	0.0035	279.9055	303.2067
Cauchit	0.0169	0.0029	277.7715	301.0727

Table B.2 – Sufficiency test for the Binomial distribution, AIC and BIC under different link functions.

Before deciding on one of the link functions we will investigate the working response plotted against the linear predictor plot as described in section A.3. In figure B.1 we see that none of the link functions are performing well but the cauchit link is the one which is closest to being linear. Therefore, we will continue the diagnostics only with the cauchit link function.

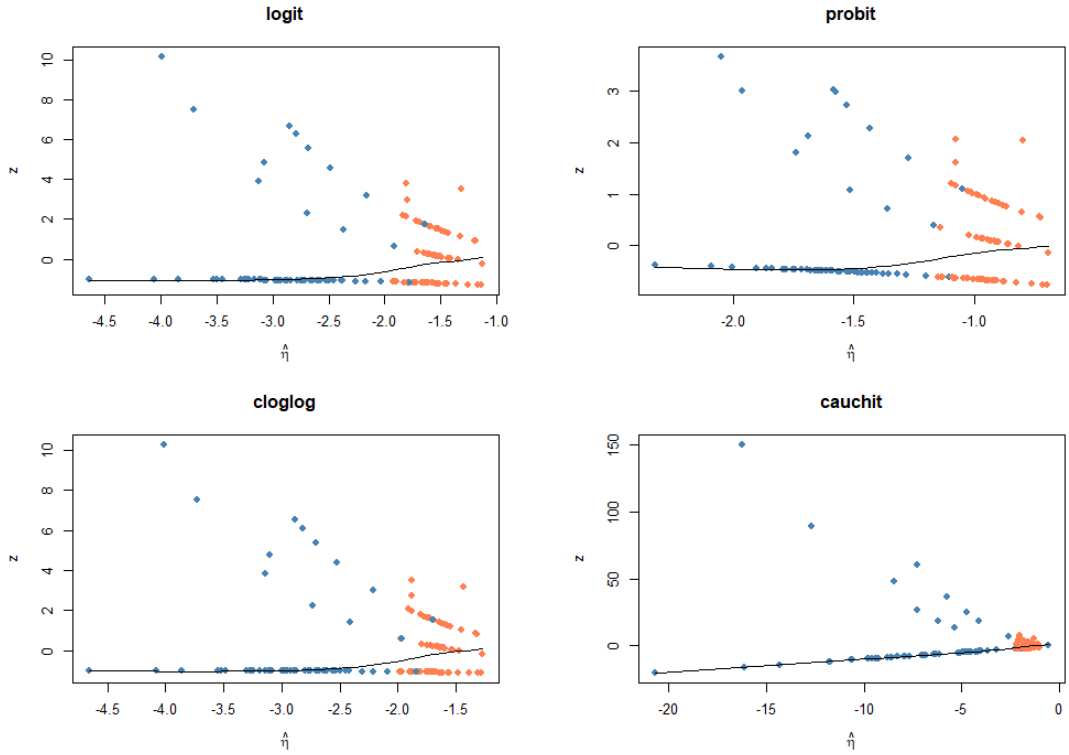

Figure B.1 – Working response vs linear predictor

We will now consider the QQ plot, based on the quantile residuals, to inspect the linear predictor. To show how the deviance and pearson residuals fail with the given model we will also plot a QQ plot based on the deviance residuals. In figure B.2, we see that the QQ plot for the deviance residuals has discrete levels due to the discrete nature of the response variable. This behavior is also explained in [3] where they claim that even for large amount of data, GLMs describing discrete data due not have approximate normal distributed Pearson or deviance residuals. Hence we will only consider the QQ plot for the quantile residuals which looks.

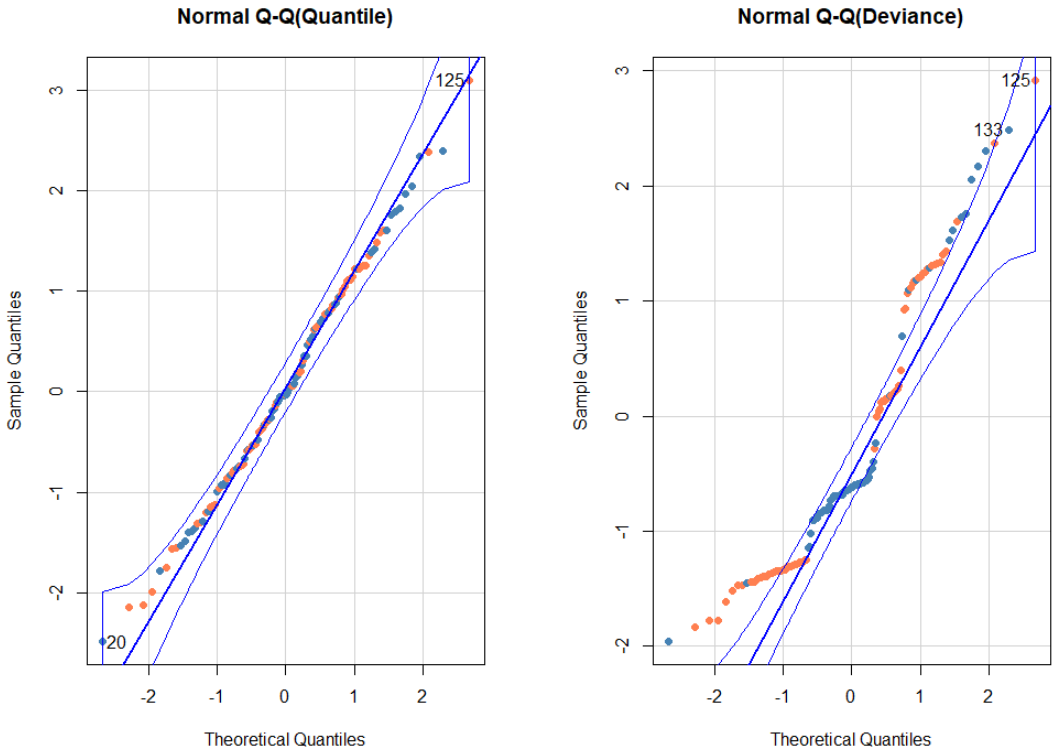

Figure B.2 – qq plot considering quantile and deviance residuals.

We now consider the quantile residuals plotted against the mean value parameter and the explanatory variables to see if any structural problems are present in the model. This analysis is similar to the analysis described in section 3.10 in [1] for classic GLMs. In figure B.3, we see that no significant patterns are visible for the plots with the explanatory variables but in the plot with the mean value parameter we see that females and males are split into two almost pure groups.

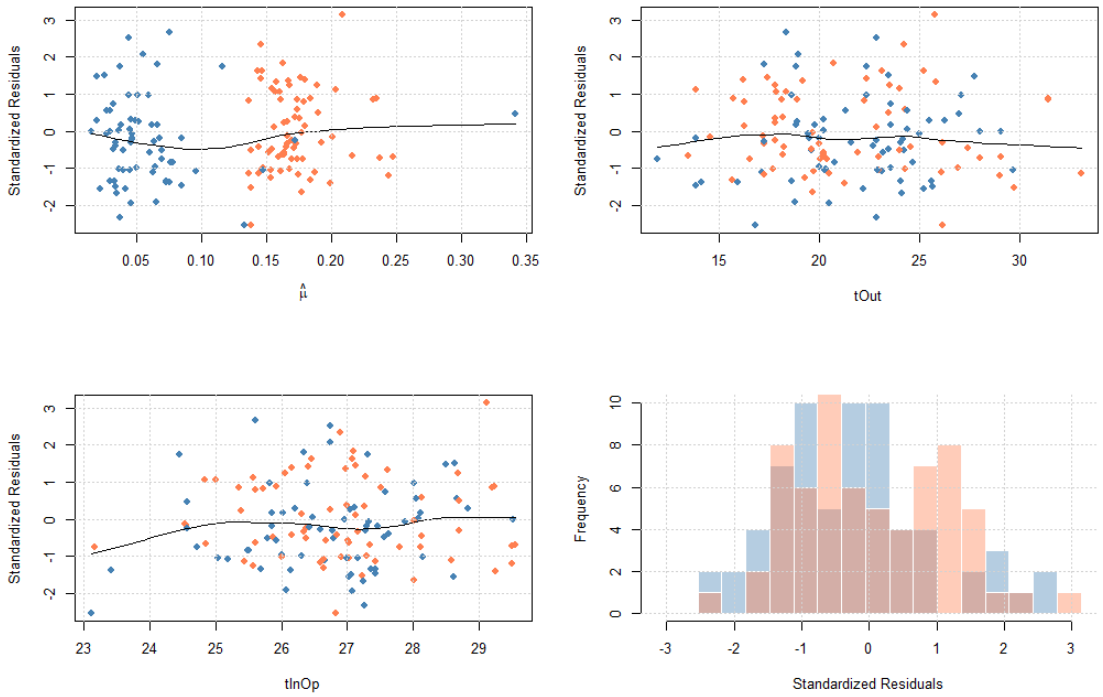

Figure B.3 – Quantile residuals plotted against mean value parameters, $\hat{\mu}$, and explanatory variables.

We have now checked the distribution, the link function and the linear predictor. We did not find any very alarming problems and no points indicated that we should perform an outlier analysis. We will hence now proceed with a type II backward selection. We could not find any theory on goodness of fit based on quantile residuals but in [3] they though state that deviance residuals are a bit better than pearson residuals for discrete distribution and hence we have used these. The goodness of fit test we have used is based on theorem 4.3 in [1] and the result is shown in table B.3.

Iteration	Interaction	Deviance	p value
1	factor(sex):tOut:tInOp	167.965	0.054
2	tOut:tInOp	168.497	0.466
3	factor(sex):tOut	169.400	0.342
4	tOut	169.449	0.824
5	factor(sex):tInOp	172.506	0.080
6	tInOp	172.572	0.798

Table B.3 – Type II backward selection of the Binomial GLM given in B.1.3

We end up with a linear predictor, η, where only the sex matters. Hence the full design in

B.1.2 simplifies to

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}. \tag{B.1.4}$$

This means we will have one mean value parameter for each `sex` and `nobs`. The found parameters are shown in table B.4.

	2.5 %	$\hat{\beta}$	97.5 %
`factor(sex)female`	-2.22	-1.67	-1.26
`factor(sex)male`	-8.52	-5.26	-3.47

Table B.4

Because we only have two intercepts in our model we can actually calculate the CDF of the process corresponding to the found parameters values for males and females. It should though be noted that we need to fix `nobs` to a specific value. In our dataset 94% of the observations have `nobs` = 5 and hence we will only consider these. In figure B.4 we have plotted the obtained processes and their confidence interval with the empirical CDF obtained from data. We see that the rate for which males change cloth during the day is heavily underestimated and oppositely the rate for which females change cloth during the day is heavily overestimated. This could indicate that the data is zero-inflated but we will get back to this latter.

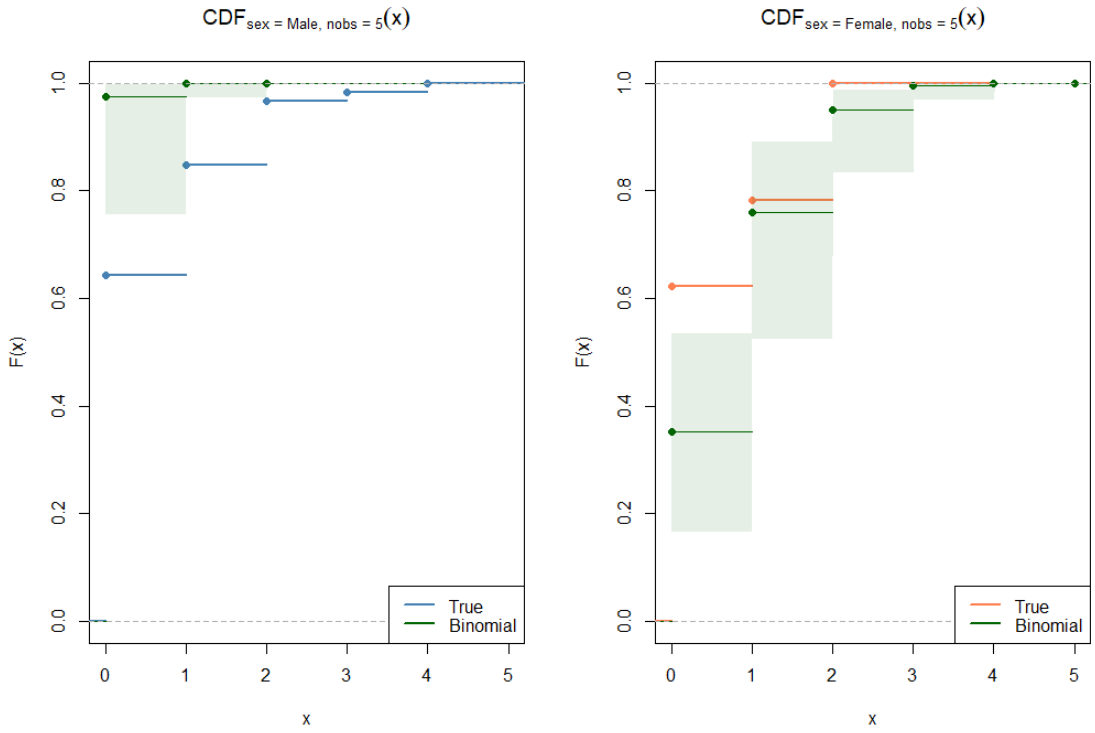

Figure B.4 – The CDFs and confidence interval for males and females plotted with the empirical CDF obtained from data.

B.2 Poisson

We now will consider the Poisson distribution given in B.2.1.

$$f_Y(y;\mu) = \frac{\mu^y e^{-\mu}}{y!} \tag{B.2.1}$$

Again, we want to predict the number of times during a day a subject changes clothing level but for different subject we have different number of observations and different duration of time in which we collect the data. Hence we can define a model which assumes that the frequency which one changes cloth with is proportional to the number of times a subject is observed. Hence we can model our mean value parameter, μ, as

$$\mu = \gamma \cdot \texttt{nobs}, \tag{B.2.2}$$

An alternative formulation would be to assume that the frequency which one changes cloth the duration of time a subject is observed

$$\mu = \gamma \cdot \texttt{time}. \tag{B.2.3}$$

These formulation gives raise to two different offset value as described in example 4.7 in [1]. We see this if we choose the log link for the Poisson.

$$\eta = \mathbf{X}\beta + \log(\texttt{nobs}) \tag{B.2.4}$$
$$\eta = \mathbf{X}\beta + \log(\texttt{time}) \tag{B.2.5}$$

If one chose to use `nobs` as offset, one implicitly assumes that the number of times a person changes cloth is at max the number of times we observe the person. This essentially means that we have proportion data and a better distribution choice would then be the Binomial. One the other hand, if we choose `time` as an offset we assume the underlying process is a counting process with no limit and hence the Poisson is a correct choice of distribution [2]. We will hence choose `time` as offset.

This gives the following linear predictors written in R notation.

$$\texttt{clo} \sim \texttt{tOut} * \texttt{tInOp} * \texttt{factor(sex)} + \texttt{offset}(\log(\texttt{time})) \tag{B.2.6}$$

Lastly before moving into diagnostics we observe in table 4.2 in [1] that for the Poisson distribution we cannot distinguish the precision parameter, λ, from the mean value parameter. Hence we do not need to estimate it.

B.2.1 Diagnostics

As in section B.1 all the following diagnostic tools are from chapter 8 in [2] and red and blue in all the following plots will respectively indicate female and male.

As for the Binomial, [2] also treat the Poisson distribution in section 7.5 and state that if $y < 3$ then the deviance residuals do not make sense and if $y < 5$ then Pearson's residuals do not make sense. Hence we must take the p values in table B.5 with a grain of salt. We though see that both the deviance and Pearson's goodness of fit approves the Poisson distribution.

	GoF Statistic	Df	p Value
Deviance	143.626	128	0.163
Pearson	153.605	128	0.061

Table B.5 – Sufficiency tests for the Poisson distribution.

To test the sufficiency of the link function we will plot the working response against linear predictor as explained in section A.3. We see from figure B.5 that the log link is not really a great link function. That said we could not really find a greater link and we must also take into account that also the working response for discrete distribution is bad.

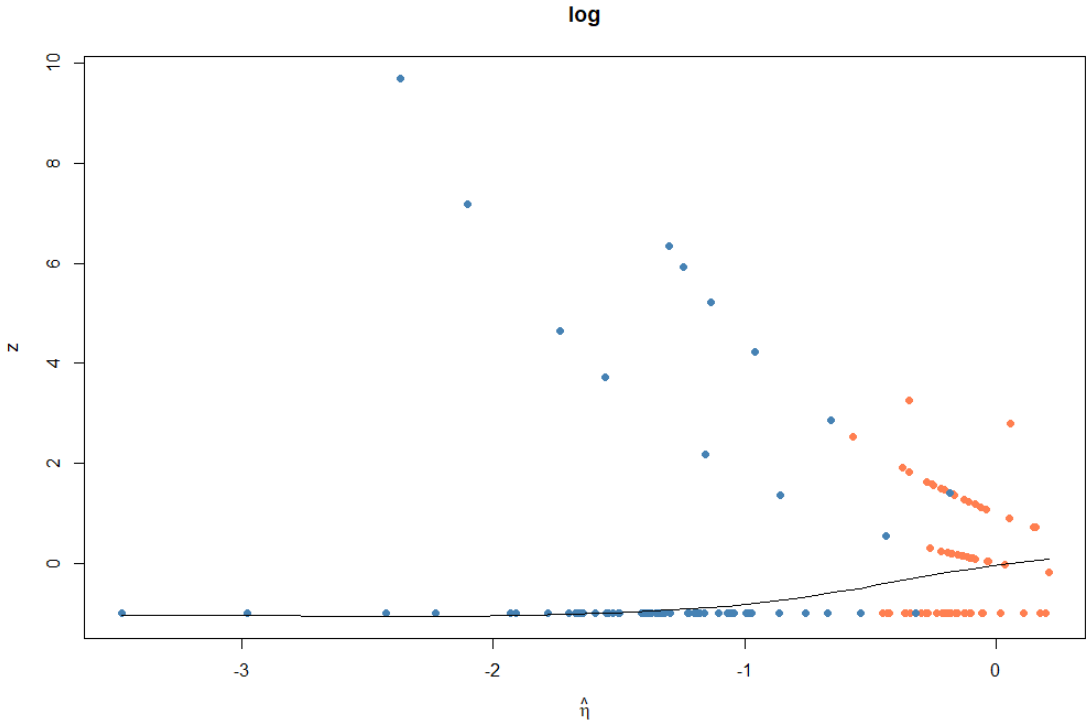

Figure B.5 – Working response vs linear predictor

We now move onto testing the linear predictor by consulting the QQ plot and the quantile residuals vs the mean value parameter and the explanatory variables. In figure B.6, we see the QQ plot for equation B.2.6 based on the quantile residuals. It looks fine, so we proceed with the quantile residuals vs the mean value parameter and the explanatory variables.

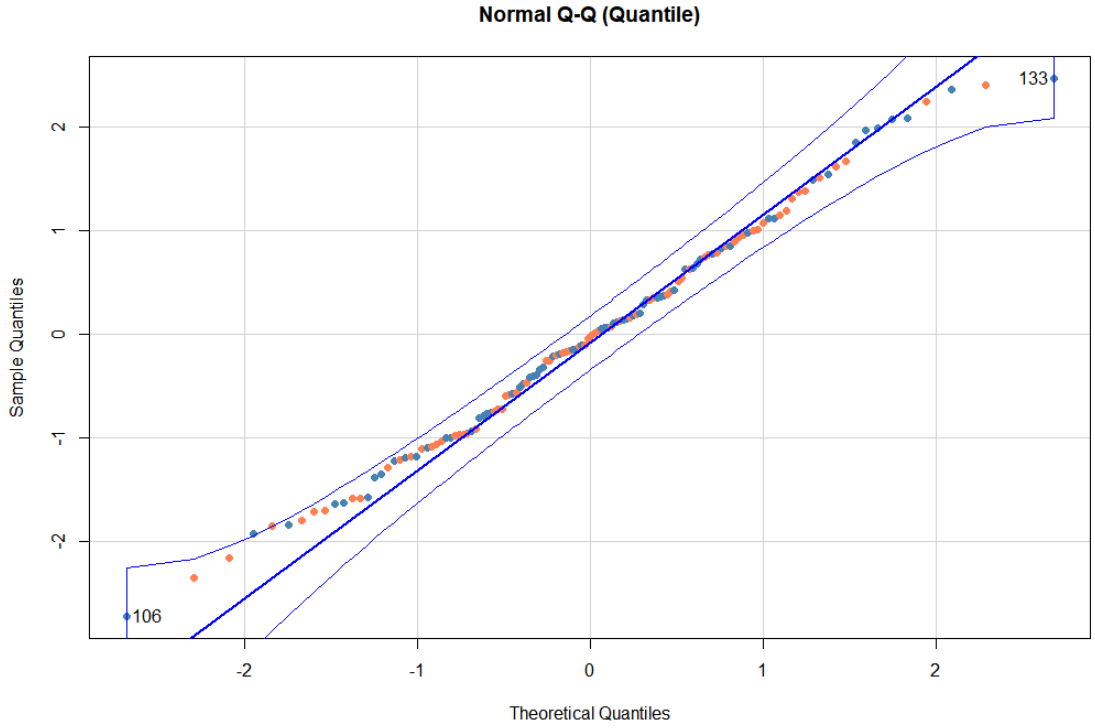

Figure B.6 – qq plot for B.2.6 considering quantile residuals.

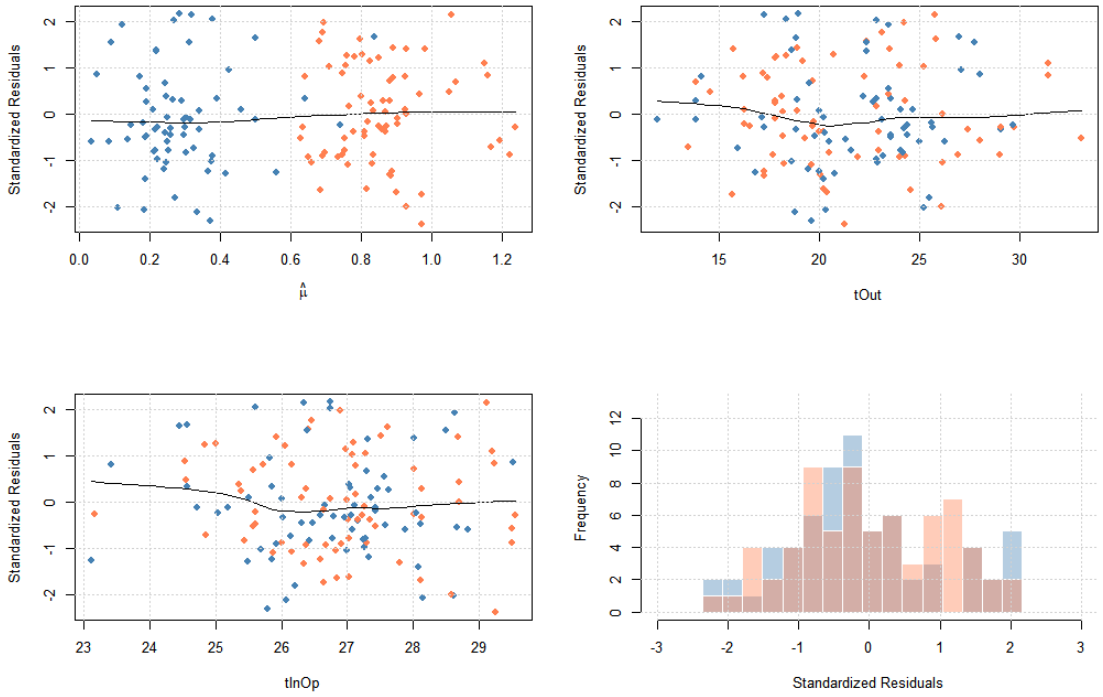

Figure B.7 – Quantile residuals plotted against mean value parameters, $\hat{\mu}$, and explanatory variables.

We see do not see any alarming problems from figure B.7 and hence we can continue with a type II backward selection.

Iteration	Interaction	Deviance	p value
1	factor(sex):tOut:tInOp	145.093	0.226
2	tOut:tInOp	145.330	0.626
3	factor(sex):tOut	145.680	0.554
4	tOut	145.759	0.779
5	factor(sex):tInOp	149.012	0.071
6	tInOp	149.145	0.715

Table B.6 – Type II backward selection of the Poisson GLM given in B.2.6

We see from table B.6 that the only necessary term is sex. We hence again conclude at the design given in equation B.1.4. The found parameters are presented in table B.7.

	2.5 %	$\hat{\beta}$	97.5 %
factor(sex)female	-2.32	-2.06	-1.81
factor(sex)male	-3.61	-3.12	-2.70

Table B.7

Because our offset variable has a lot of different offset values it is not as straight forward to visualize the underlying process as for the Binomial GLM. We can though take an average over all times in the dataset and use this as a common offset. If we do this we obtain processes shown in figure B.8. It is maybe a bit better than the Binomial model but it still has a tendency to underestimated how much males change cloth during a day and oppositely overestimate how much females change cloth during a day.

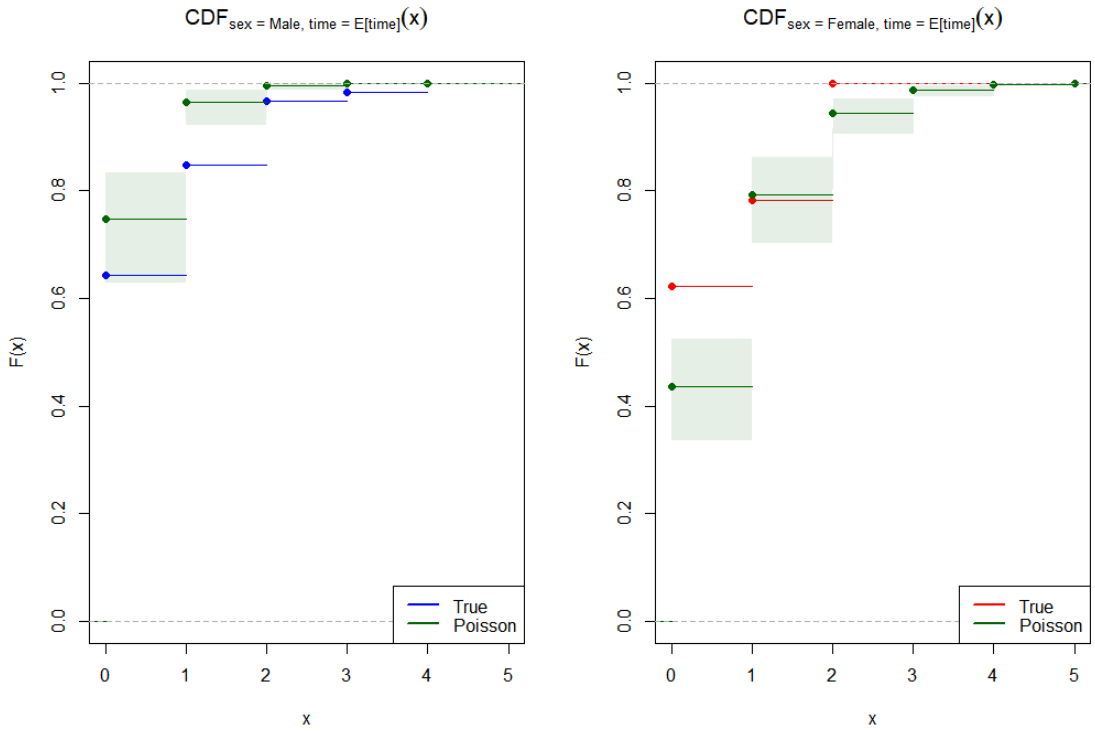

Figure B.8 – The CDFs and confidence interval for males and females plotted with the empirical CDF obtained from data.

B.3 Interpretation of the two models

In this section, we will dive into how the Poisson and Binomial distributions suits the structure of the data. Though before going into qualitative differences between the two distributions, we will asses how quantitative measures rates the two models.

Distribution	Link	Number of parameters	AIC	BIC
Poisson	Log	2.00	270.13	275.96
Binomial	Cauchit	2.00	274.10	279.93

Table B.8 – Some numbers on the final models.

In table B.8 we see that the Poisson distribution is better on both AIC and BIC. One should though still take into account that these numbers are only on in-sample data. If the distribution violates the nature of the data then predictions on out-of-sample data can show strange behaviour.

We start by considering how the Binomial distribution behaves. It assigns probability mass to all integers in the interval $[0, n]$ where n is the size of the considered interval, here `nobs`. Hence if the nature of the data is proportions, then this distribution is a very good distribution because the underlying process would never be able to produce an outcome larger than the size of the interval.

The Poisson distribution is defined on all positive integers, i.e. $\mathrm{N} \cup \{0\}$. Hence this distribution is suited for an underlying count process without an upper limit. It can be used for proportions if the upper limit is very large and very improbable, which is also the reason why the Binomial distribution can be used as a approximation for the Poisson in such cases.

Considering this we should be able to figure out which one of them fit the nature of the data better. We have a process which describes the number of times a person changes clothing level during a day. This is not a bounded process and hence one should think that the Poisson is a better choice. We could though also interpret the formulation of the exercise such that we aren't actually counting the number of times a person changes cloth but how many times the person changes cloth out the number of times we observe the person. This interpretation would be better suited with the Binomial distribution. Hence it is very important to be aware of what you want to describe before choosing a distribution.

We choose to understand the process as a counting process without a limit and hence the Poisson is the better fitting distribution of the two, but now new considerations arise. We saw in both figure B.8 and B.4 that the underlying data was zero-inflated and in such case one should consider paring the Poisson distribution with a Bernoulli such that the Bernoulli can predict if we have a zero or not and then the Poisson can predict in the positive integers. Another problem one can encounter using the Poisson is over- or underinflation. Because the Poisson is restricted to having the same variance and mean it will not be able to fit over- or underinflation. In such situation one can use a Quasipoisson or a Negative Binomial which is not restricted to having the same mean and variance.

C | Fan Speed

In this section, we will work with a dataset which containts information on fan speeds. In the table C.1 below, we will describe the available variables:

Variable	Domain	Description
TSV	Count Data	Themal Sensation Vote on an ordered 3-level scale
fanSpeed	Count Data	Fan Speed on a 3-level scale
fanType	Count Data	Type of fan with categories *downstream* and *upstream*
Subject Id	Factor	Identifier for the subject

Table C.1 – Description of Fan Speed data.

We will not work with Subject Id in our modelling. The information in the remaining variables is graphically displayed in figure C.1. For an overall understanding of the response variable, consider the proportion of observations that fall in each TSV level:

	0	1	2
PMF	0.578	0.312	0.110
CDF	0.578	0.890	1.000

Table C.2 – The probability mass function, PMF, and the cumulative distribution function, CDF.

In table C.2, we see that most often the thermal sensational vote is 0, a great deal answers 1 and only a few thermal sensational votes on 2.

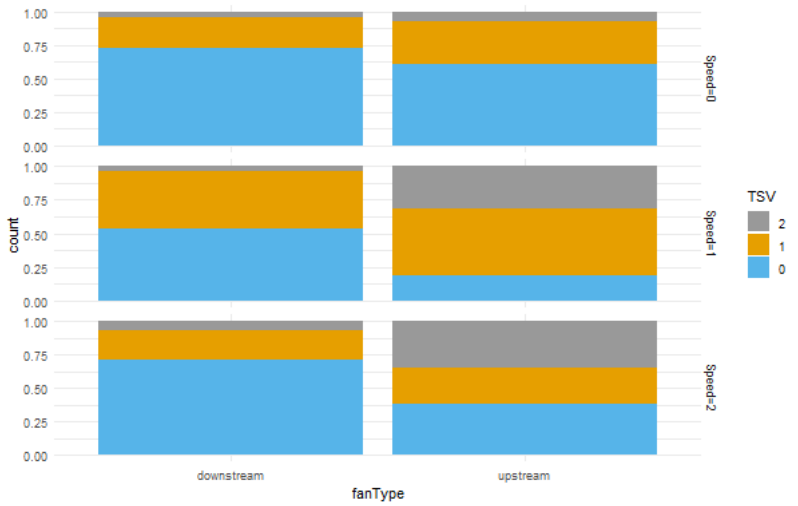

Figure C.1 – The number of counts for each level in TSV grouped by fan speed and fan type.

In figure C.1, we note that TSV seems to be lower, when the fan type is downstream. We see this when we compare the two fan-type-columns where the downstream column is more blue than the upstream. When we consider the rows, we see that the test subjects tend to give a higher TSV when we increase the fan speed. Especially when the speed is 0, we see larger potions of subjects rate the TSV as 0.
A different graphical representation can be seen in figure below

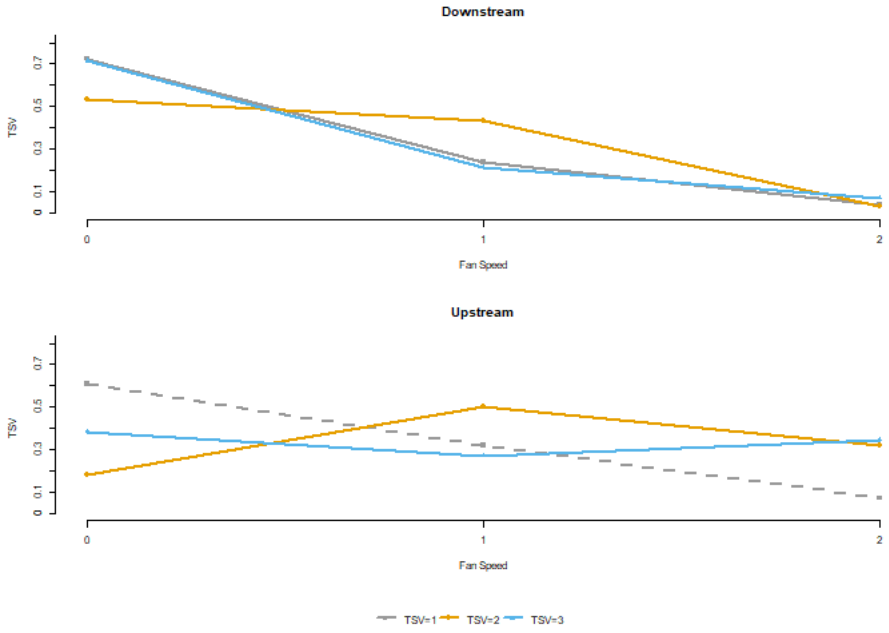

Figure C.2 – A different representation to asses the data.

In figure C.3, it seems that when the fanType is downstream, the TSV is decreasing in some non-linear way. We will comment this below.

C.1 Contingency Table and Test of Invariance

In the following, we will focus on TSV and fan speed. Below, we show the contingency table for the TSV and fanSpeed:

		Fan Speed			
		0	1	2	Row Total
TSV	0	97	20	20	137
	1	40	24	10	74
	2	8	8	10	26
Column Total		145	52	40	237

Let F^O denote the observed frequencies in the table above s.t. $F^O_{(0,0)} = 97$. Introduce F^E as the expected frequencies for each cell. Under the hypothesis that the TSV and fanSpeed are independent, we find $F^E_{i,j} = \frac{\text{rowsum}_i \cdot \text{colsum}_j}{Table\ total}$ e.g. $F^E_{(0,0)} = \frac{137 \cdot 145}{237} = 83.82$.

Under assumption of independence the test statistic $\chi^2_{obs} = \sum_i \sum_j = \frac{\left(F^O_{(i,j)} - F^E_{(i,j)}\right)^2}{F^E_{(i,j)}}$ follows a χ^2-distribution with $(r-1)(c-1)$ degrees of freedom. In our case $r = c = 2$ as r is the number of levels in TSV and c is the number of levels in fanSpeed. In our case:

Df	χ^2_{obs}	p-value
4	22.715	0.00014

Table C.3 – Peasons χ^2 test for test of independece between TSV and fan Speed

With a p-value of 0.00014 we strongly reject the hypothesis that TSV and fan Speed are independent.

C.2 Test for Independence with a Different Approach

The TSV consists of ordinal-scale observations with level in the range from 0 to 2. This type of ordinal scale is often denoted as a Likert-scale and we will model this data type using cumulative link models, clm [5].
Let Y_i be the observed TSV which can fall in the 3 levels. With the data available, we assume Y_i follows a multinomial distribution with parameters $\pi_{i,j}$ where i is the i'th observation and j is the TSV level. We now introduce the cumulative probability for Y_i as:

$$\gamma_{i,j} = P(Y_i \leq j) = \pi_{i,0} + \cdots + \pi_{i,j}, \qquad \sum_{j=0}^{2} \pi_{i,j} = 1 \tag{C.2.1}$$

i.e. $\gamma_{i,j}$ is the probability that Y_i falls into TSV level j or lower. With a link function g, we can introduce the linear predictor $\eta_{i,j}$ such that

$$g(\gamma_{i,j}) = \eta_{i,j}, \qquad \text{for } i = 1, \ldots, N \text{ and } j = 0, \ldots, 1 \tag{C.2.2}$$

In our case, with a multinomial distribution, g is the logit link function such that

$$\log \frac{\gamma_{i,j}}{1-\gamma_{i,j}} = \eta_{i,j} \qquad \text{for } i = 1, \ldots, N \text{ and } j = 0, \ldots, 1 \tag{C.2.3}$$

We can now move to the next step and consider different ways to construct the linear predictor η. It is very important that we first split the linear predictor into thresholds and regression variables [5]

$$\eta_{i,j} = \theta_j + \mathbf{x}_i^\intercal \beta \tag{C.2.4}$$

Of special interest are the θ_js. They are called thresholds[1] and are strictly ordered $-\infty < \theta_0 < \theta_1 < \infty$. They are quite intuitive. Consider therefore a model with TSV then only:

$$g(\gamma_{i,j}) = \theta_j \qquad \text{for } j = 1, 2 \tag{C.2.5}$$

Here we obtain the thresholds $\theta = [0.31, 2.09]^\intercal$. If we convert them back to the response domain, we get $\gamma = [0.58, 0.89]$ which corresponds directly to the CDF of the TSV values presented in table C.2 hence we have indeed created the desired model.
We can now introduce fanSpeed to the model and make an anova test for independence:

model name	symbolic	**R**-syntaks
M_{threds}	$g(\gamma_{i,j}) = \theta_j$	TSV ~ 1
M_{fanS}	$g(\gamma_{i,j}) = \theta_j - \beta_1$ (fanSpeed)	TSV $\sim$ fanSpeed

Table C.4 – Nested models to test independence

We construct the models using the package *Ordinal*. With the fitted models, we can test for independence between TSV and fanSpeed with an ANOVA test:

Df	χ^2_{obs}	p-value
2	15.86	0.00036

Table C.5 – Anova test between M_{fanS} and the nested model M_{threds}.

The ANOVA table in C.5 has a p-value of 0.00687 hence we cannot reject the effect of M_{fanS}, i.e. they are not independent.

C.3 Fit and develop a model for TSV

In section C.2, we introduced a framework to model the data with a multinomial distribution. We now consider the most complex model, we can built which is:

$$\text{TSV} \sim \text{fanSpeed*fanType} \tag{C.3.1}$$

now we perform a type II selection and find that we should remove the interaction term shown in table C.6.

Iteration	Interaction	LRT	p value
1	factor(sex):tOut:tInOp	4.3466	0.1138

Table C.6 – Type II backward selection of the model C.3.1

[1] See section 2 in [5]

This results in the model given in equation C.3.2

$$\text{TSV} \sim \text{fanSpeed} + \text{fanType} \tag{C.3.2}$$

Explicitly, let g be the logit function, and let $Y_i \in \{0, 1, 2\}$ be the levels for the TSV, then we have created the model:

$$\begin{aligned} g\left(P(Y_i \leq j)\right) = \theta_j - \beta_1 \mathbb{1}_{\{\text{fanSpeed}_i=1\}} - \beta_2 + \mathbb{1}_{\{\text{fanSpeed}_i=2\}} \\ - \beta_3 \mathbb{1}_{\{\text{fanType}_i=\text{upsteam}_i\}} \quad \text{for all } i = 1, \ldots, N \quad j = 0, 1, 2 \end{aligned} \tag{C.3.3}$$

where the parameters and provided confidence intervals are given in table C.7

	2.5 %	β	97.5 %
β_1	0.6075105	1.2294689	1.8592320
β_2	0.1392765	0.8619515	1.5789356
β_3	0.4651458	0.9943946	1.5366490

Table C.7 – Parameters estimates and 95& confidence intervals for the final model

and the thresholds are

	θ
θ_1	1.2294689
θ_2	0.8619515

Table C.8 – Thresholds for the final model

In this modelling approach we have taken two modelling choices:

1. It seems that the TSV is decreasing with the fanSpeed level. In figure C.3, it seems to do that in a non-linear fashion. This motivates a potential numeric treatment of the variable. We cannot be sure that this is the case and it doesn't seem useful to be able to extrapolat the votes to other fanSpeeds aas we don't know if higher fanSpeeds even exits.

2. We experimented with other link functions and saw slightly reduced AIC for Cauchit:

Link	AIC
Logit	417.425
Cauchit	417.013

Table C.9 – Some numbers on the final models.

As mention in [5] logit is the most common to use, and it allows for a better interpretation of the parameters. Therefore, we trade a bit of model performance for model interpretability and stick with the logit function.

C.4 Present the fitted model

First we give an interpretation of the parameters. In section C.2, we already related the thresholds θ_i to the CDF of the TSV for the simple model. With the introduced

β parameters, we have been a bit more careful. Consider the model in equation C.3.3 and consider first the case where all indicator parameters are 0 i.e. an observation where fanSpeed is 3 and fanType is upsteam. When we transform them back to the response domain, we obtain $\gamma = [0.726, 0.952]$. Compare this to the upper left proportion of TSV in figure C.1, to see that this indeed corresponds to the observed proportions.
Consider now that we introduce upstream and the value of β_3. It is shown in [5] that we can interpret the $\exp(\beta_3) = 2.70$ as an odds ratio of being rated level $j \in \{0, 1\}$ or above. We can again consult figure C.1 and C.3 to see that indeed it seems that when we are upstream, there seems to be a larger proportion of votes in higher levels of TSV. An alternative interpretation is that when the fanType is upstream, then we shift the threashold a constant amount of 0.99 [5]. The change in the distribution of proportions when the fanType is upstream can be seen in the figure below:

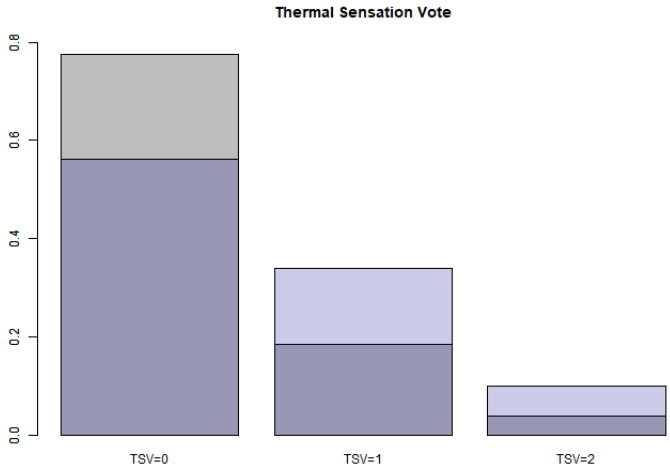

Figure C.3 – The distribution of proportions. In grey the distribution with fanType downstream and fanSpeed equal to 0. In blue the fanType is upstream. We see a gradual shift.

The parameters β_1 and β_2 also has this interpretation as odds ratios. Specifically the odds ratios are 3.42 and 2.37 for votes to be in level j or higher. We can consult C.1 to see that indeed there seems to be votes towards the larger levels when the speed is 1 and 2. As the odds ratio suggest, especially when the speed is 1.

D | Appendix

Appendix A

Iteration	Interaction	F value	p value
1	InvHt:Pres	0.018	0.892
2	InvTmp:Wind	0.071	0.790
3	Hgt:InvTmp	0.084	0.773
4	Temp:Wind	0.254	0.614
5	Hgt:Wind	0.056	0.813
6	Hgt:Vis	0.205	0.651
7	InvHt:InvTmp	0.321	0.572
8	Pres:Wind	0.433	0.511
9	Temp:InvTmp	0.806	0.370
10	InvHt:Hgt	0.953	0.330
11	Temp:InvHt	0.785	0.376
12	Vis:Wind	0.944	0.332
13	InvHt:Wind	0.777	0.379
14	InvHt:Vis	0.952	0.330
15	Hum:Wind	3.100	0.079
16	Wind	0.009	0.924
17	Hum:Vis	2.220	0.137
18	Pres:Vis	1.590	0.208
19	Temp:Vis	1.969	0.162
20	Temp:Hum	2.169	0.142
21	Temp:Pres	1.637	0.202
22	InvTmp:Vis	3.484	0.063
23	Vis	2.366	0.125
24	Hum:InvTmp	3.169	0.076
25	InvHt:Hum	1.342	0.248
26	Temp:Hgt	3.792	0.052

Table D.1 – Type II backward selection for the classic GLM given in equation A.2.2

Appendix for B

Iteration	Interaction	Deviance	p value
1	Hgt:InvTmp	310.129	0.996
2	Temp:InvHt	310.149	0.886
3	Hgt:Vis	310.338	0.854
4	Hgt:Wind	310.311	0.763
5	Temp:InvTmp	310.165	0.763
6	InvHt:Pres	310.148	0.656
7	Vis:Wind	310.327	0.651
8	InvHt:Vis	310.422	0.588
9	Temp:Wind	310.987	0.514
10	InvTmp:Wind	310.332	0.848
11	InvHt:Wind	310.828	0.459
12	Pres:Wind	311.802	0.289
13	Hum:Wind	310.880	0.277
14	Wind	310.552	0.610
15	Pres:Vis	311.630	0.241
16	Hum:Vis	312.064	0.408
17	InvHt:Hgt	313.250	0.192
18	InvHt:InvTmp	311.577	0.330
19	Temp:Vis	311.924	0.193
20	InvTmp:Vis	311.990	0.245
21	Temp:Hum	313.736	0.272
22	InvHt:Hum	313.629	0.460
23	Hum:InvTmp	314.569	0.358

Table D.2 – Type II backward selection for the Negative Binomial GLM given in equation A.3.8

Iteration	Interaction	Deviance	p value
1	Hgt:InvTmp	342.908	0.987
2	Temp:InvHt	342.913	0.942
3	Hgt:Vis	342.965	0.820
4	Hgt:Wind	343.088	0.725
5	Temp:InvTmp	343.211	0.726
6	InvHt:Pres	343.411	0.655
7	Vis:Wind	343.645	0.628
8	InvHt:Vis	343.916	0.603
9	Temp:Wind	344.439	0.470
10	InvTmp:Wind	344.483	0.833
11	InvHt:Wind	345.056	0.449
12	Pres:Vis	346.440	0.239
13	Temp:Vis	347.468	0.310
14	Temp:Hum	348.710	0.265
15	InvHt:Hum	349.169	0.498
16	Hum:InvTmp	349.441	0.602
17	Pres:Wind	350.872	0.232
18	InvTmp:Vis	352.431	0.212
19	Hum:Vis	354.179	0.186
20	Hum:Wind	355.458	0.258
21	Wind	355.641	0.669
22	InvHt:Hgt	358.052	0.120
23	InvHt:InvTmp	359.617	0.211

Table D.3 – Type II backward selection for the Poisson GLM given in equation A.3.10

Iteration	Interaction	Deviance	p value
1	Hgt:Wind	37.659	0.997
2	Hgt:Vis	37.659	0.981
3	InvHt:InvTmp	37.659	0.977
4	InvTmp:Wind	37.661	0.894
5	Temp:Wind	37.663	0.911
6	Hgt:InvTmp	37.667	0.856
7	InvHt:Pres	37.678	0.760
8	Pres:Wind	37.699	0.675
9	Temp:InvTmp	37.760	0.473
10	Vis:Wind	37.831	0.436
11	InvHt:Hgt	37.908	0.421
12	InvHt:Vis	37.990	0.402
13	Temp:InvHt	38.106	0.321
14	InvHt:Wind	38.255	0.260
15	Temp:Vis	38.494	0.155
16	Pres:Vis	38.700	0.186
17	Hum:Vis	38.905	0.187
18	Hum:Wind	39.109	0.188
19	Wind	39.122	0.742
20	Temp:Hum	39.330	0.183
21	Temp:Pres	39.602	0.128

Table D.4 – Type II backward selection for the Gamma GLM given in equation A.3.12

Bibliography

[1] H. Madsen and P. Thyregod, *Introduction to general and generalized linear models*. CRC Press, 2010.

[2] P. K. Dunn and G. K. Smyth, *Generalized Linear Models With Examples in R*. Springer science & business media, 2018.

[3] L. L. Cindy Feng and A. Sadeghpour, "A comparison of residual diagnosis tools for diagnosing regression models for count data," *BMC Medical Research Methodology*.

[4] P. K. Dunn and G. K. Smyth, "Randomized quantile residuals," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 236–244, 1996.

[5] R. H. B. Christensen, "Cumulative link models for ordinal regression with the r package ordinal," *Submitted in J. Stat. Software*, vol. 35, 2018.