Robust Statistics: Exam Assignment

Othman El Hammouchi

May 31, 2023

Question 1

The sensitivity curves corresponding to the M-estimators of location for Huber's ρ with $b = 1.5$ and Tukey's bisquare with $c = 4.68$ are shown (in red) in Figs. 1 and 2. Also plotted (in black) are their influence functions. For a general M-estimator, these are given by

$$\mathrm{IF}(x, T, F) := \frac{\psi(x)}{\int \psi'(y)\, dF(y)}\,, \tag{1}$$

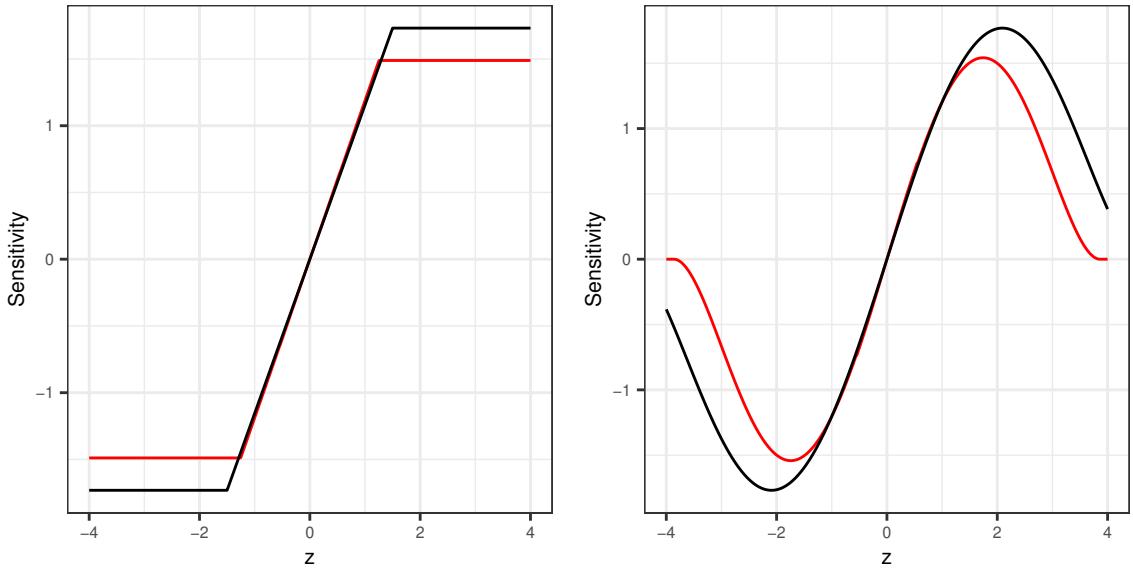
as explained on p. 12 of Part 2 of the course slides. In the case of Huber's ρ, the denominator is given by

$$\int \psi'(y)\, dF(y) = \int_{-b}^{b} dF(y) = F(b) - F(-b)\,, \tag{2}$$

which for $F = \Phi$ is ≈ 0.866. Similarly, we have

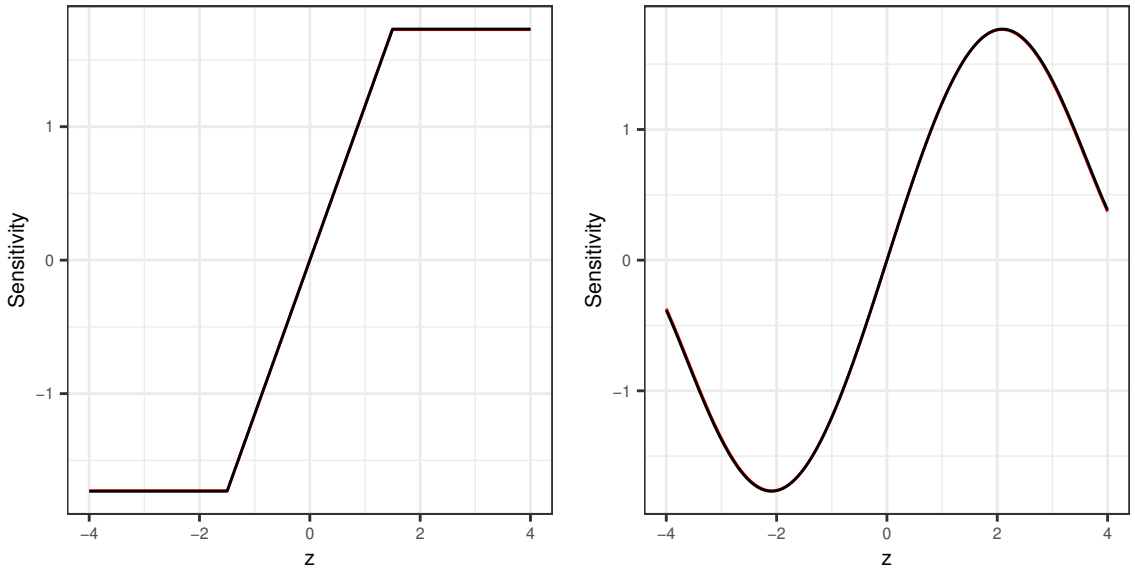
$$\int \psi'(y)\, dF(y) = \frac{1}{\sqrt{2\pi}} \int_{-c}^{c} \left(\left(1 - \frac{y^2}{c^2}\right)^2 - \frac{4y^2 \cdot \left(1 - \frac{y^2}{c^2}\right)}{c^2} \right) e^{-y^2/2}\, dy \approx 0.7573 \tag{3}$$

for the Tukey bisquare with $F = \Phi$. As we can see by comparing Fig. 1 with Fig. 2, $\mathrm{SC}(x, T_n, X_n) \to \mathrm{IF}(x, T, F)$ as the sample size goes to infinity, and the two are almost identical for $n = 1e6$..



(a) Huber **(b)** Tukey

Figure 1: Sensitivity curves (red) based on a symmetric standard normal sample with $n = 100$ together with the corresponding influence functions (black)



(a) Huber **(b)** Tukey

Figure 2: Sensitivity curves (red) based on a symmetric standard normal sample with $n = 1e6$ together with the corresponding influence functions (black)

Question 2

In the first part of this section, we will analyse the distribution of the `Copper_mcg` variable from the food dataset. A QQ-plot against the theoretical quantiles of the standard normal distribution indicates that it is strongly right-skewed, as is often the case when the data are constrained to be positive. Examination of the farthest points in the tail of the distribution reveals that these are all measurements for liver products, and a cursory glance at online resources on nutrition does indeed indicate that this food type is rich in copper (see [1]). Consequently, it seems plausible that the copper content in the catalogued food items simply follows a fat-tailed distribution.

Fitting the λ-parameter for the Box-Cox and Yeo-Johnson transforms using the robust methods implemented in the `transfo` function from the `cellWise` package [2], we obtain estimates of -0.1061973 and 0.3051917, respectively. The resulting QQ-plots are shown in Fig. 4. Both transforms produce a conspicuous distortion at the lower end of the distribution the right tail, ostensibly in their attempt to capture extreme values in the right tail. Investigation of the horizontal section reveals that it contains points for which the original observations have the same value of $1e-5\,\mu g$, which is presumably not an exact measurement but rather a cut-off value. If this is deemed to justify discarding or unifying these observations, then the Yeo-Johnson transform would yield a reasonable approximation to normality, while the Box-Cox transform preserves the right tail. Depending on the interpretation of the extreme points as either outliers or regular observations from a fat-tailed distribution, one result will be preferred over the other.

Next, we investigate the use of dimensionality reduction techniques for the description of the entire food dataset. Using the ROBPCA procedure implemented in the `PcaHubert` function of the

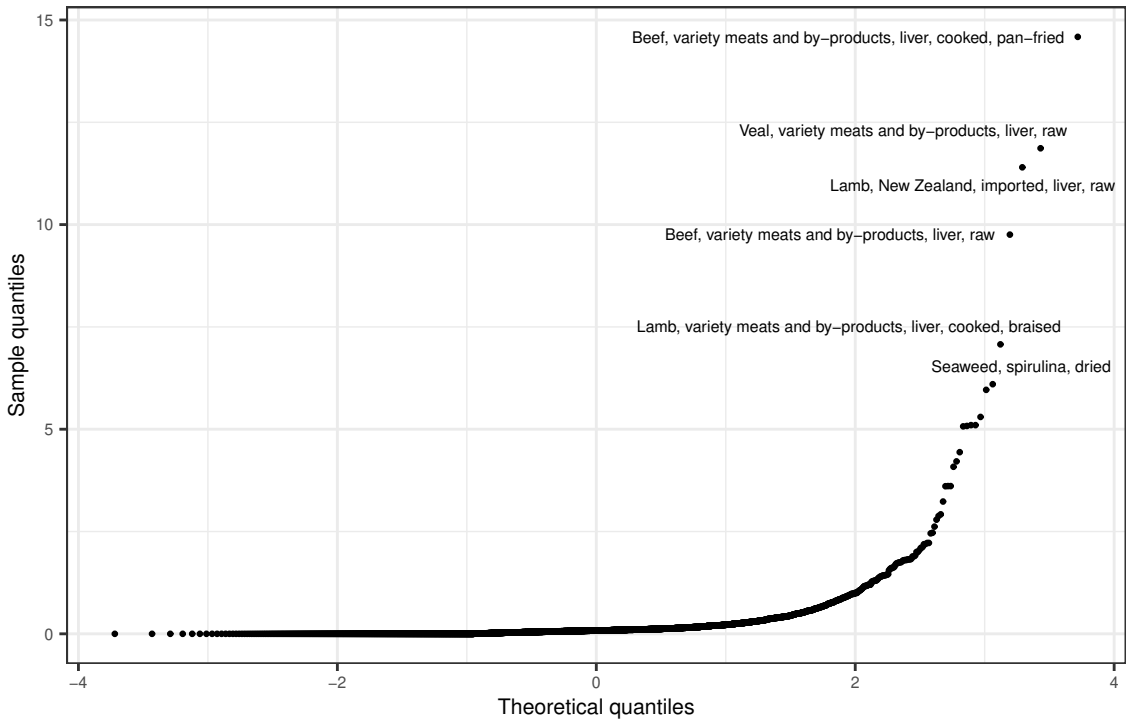
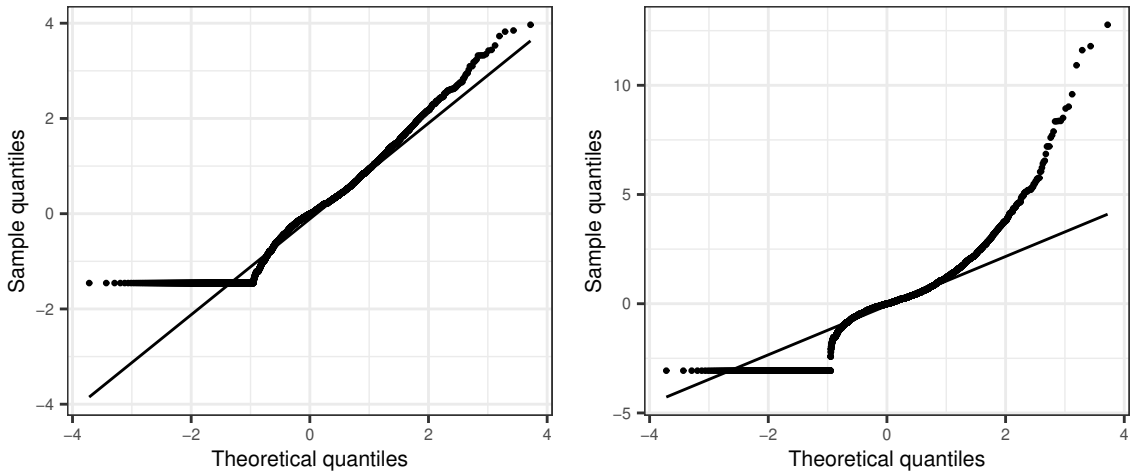


Figure 3: QQ-plot of `Copper_mcg` variable

`rrcov` package [3] with an initial number $k = 15$ of principal components gives us the screeplot in Fig. 5a, which seems to have an elbow at $k = 6$. Refitting the model with this number yields the loadings matrix shown in Fig. 5c. Given that the loading vectors constitute an orthonormal basis of the feature space $\mathbb{R}^p$, their Euclidean norm will always equal 1; hence, it follows that the average value of any loading is $1/\sqrt{p}$, so that ones which exceed this threshold indicate a salient correlation between the relevant component and the original variable. These are marked in bold in Fig. 5c.

Without expert knowledge from nutritional science, it seems difficult to derive a clear interpretation from these results. The first two principal components load onto a large number of variables, a number of which are common to both. Nevertheless, there do seem to be two blocks of variables for which there is less overlap, suggesting that these might be somehow related. The third and fourth component are clearly associated with a single one of the original variables (`VitA_mcg` and `Sugar_g`, respectively), although the latter also exhibits secondary correlations with the surrounding ones. Lastly, the penultimate component exhibits an opposite association of approximately equal strength with two variables, which strongly hints at the presence of an underlying common driver of both, and the final component correlates with many variables, but these notably cover the final block omitted by the PCs 1 and 2, which suggest that the three of them might represent global patterns in the data while the remainder account for more concentrated pockets of variation.

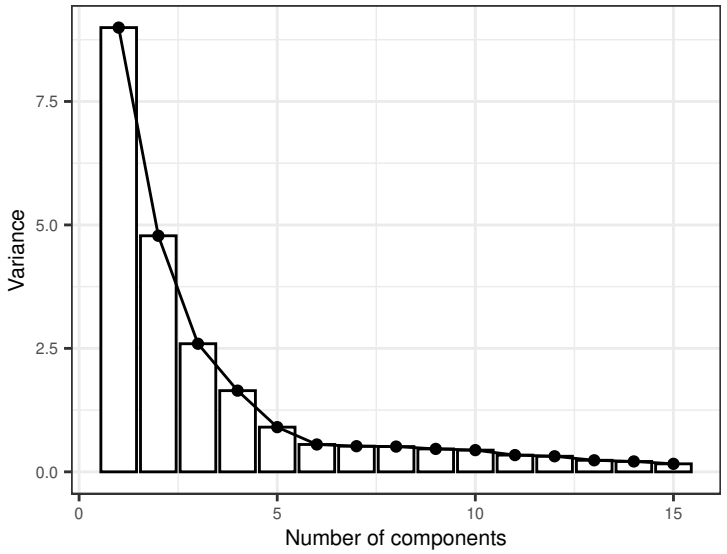
With regards to outliers, the plot in Fig. 5b indicates that all three types are present in this data. For increased interpretability, we augmented the original dataset to include information about the



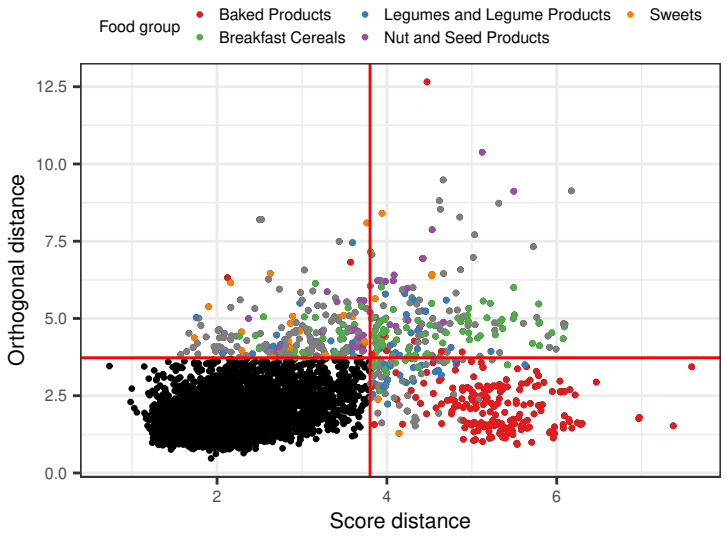
(a) Yeo-Johnson **(b)** Box-Cox

Figure 4: QQ-plots of the transformed variable

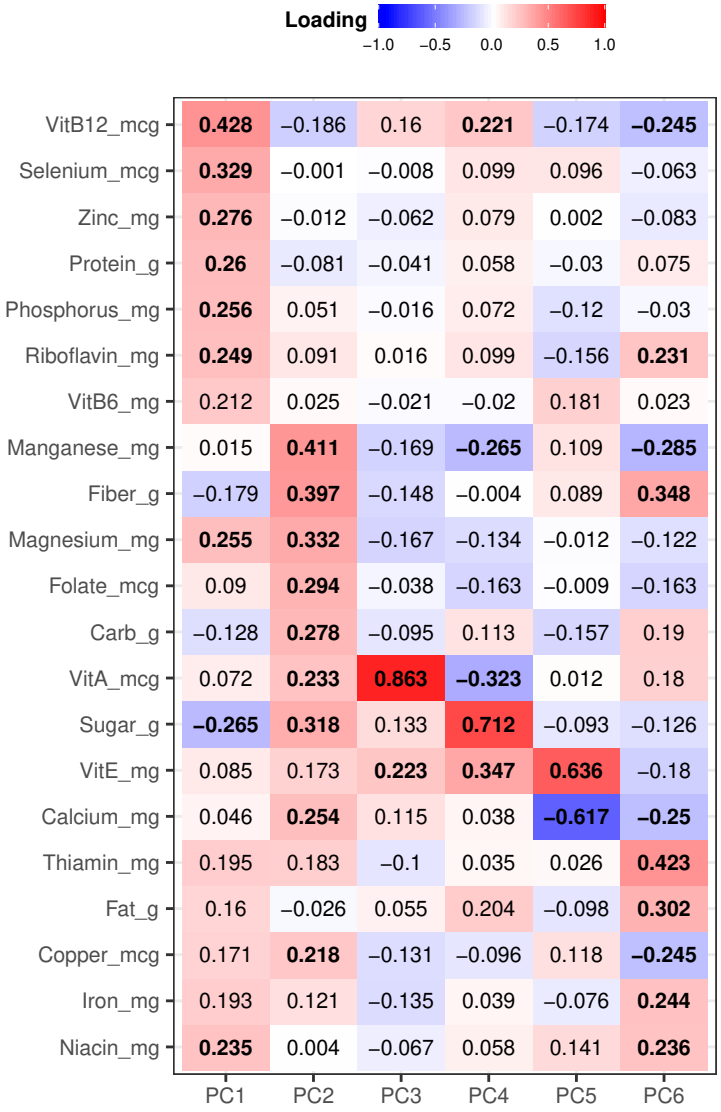
food group of the items using the `usdanutrients` package [4] and highlighted the largest groups for each outlier type in different colours. This reveals that the good leverage points are dominated by the category of Baked Products, with Legumes and Legume Products as a distant second. Breakfast Cereals constitutes the largest group within both orthogonal and bad leverage outliers. Notice that the largest groups of outliers consist of carbohydrate-rich foods, which might aid in interpreting the cause of their deviation.



(a) Screeplot



(b) Outlier plot



(c) Loadings

Figure 5: ROBPCA results

Question 3

Our aim in this section is to construct a normal linear model for predicting the price of a car based on the `Topgear` dataset. Table 1 shows summary statistics for this response variable. Notice in particular the striking discrepancy between the mean and median, which hints at the presence of outliers in the data.

Statistic	N	Mean	St. Dev.	Min	Median	Max
Price	242	51,170.54	106,879.40	7,995	26,050	1,139,985

Table 1: Summary statistics for the `Price` variable

The correlation structure of the continuous variables is visualised in Fig. 6. The top row indicates that cars with stronger engines (higher fuel consumption and power, lower acceleration time) generally tend to be more expensive, unsurprisingly. We also observe a high degree of interdependence between the variables of physical measurements (top left square outlined in black), which could lead to issues with multicollinearity.

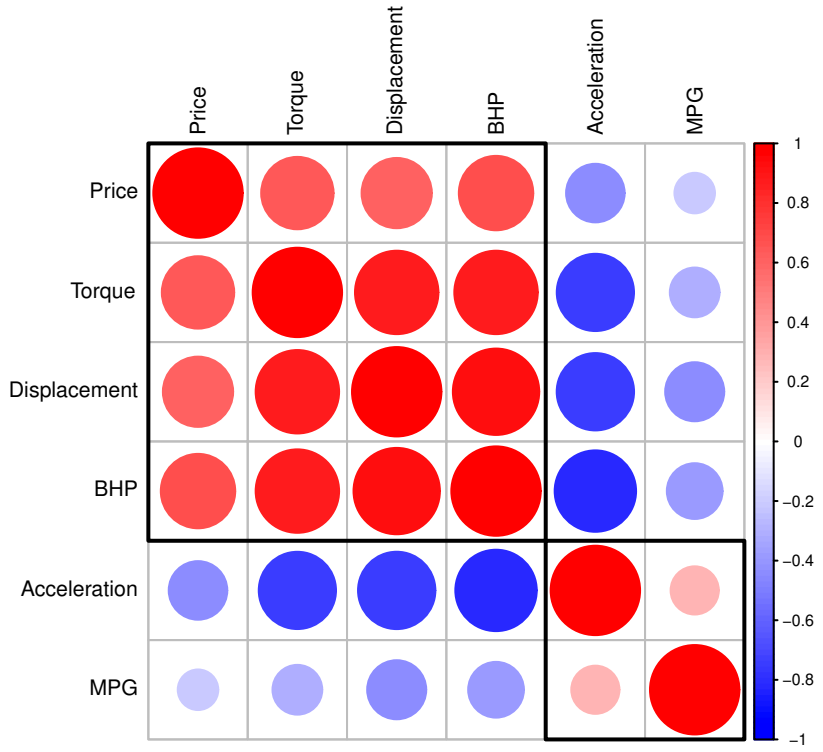
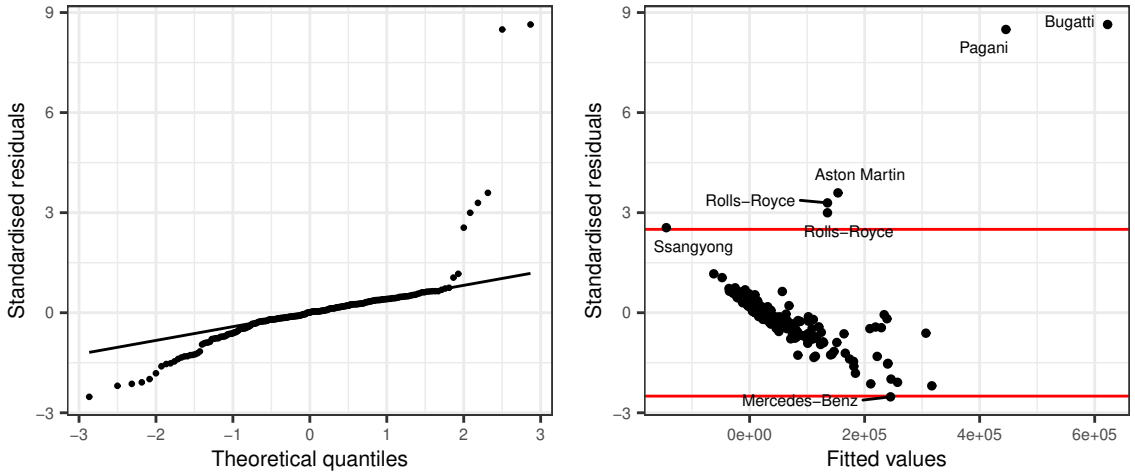


Figure 6: Visualisation of the correlation between the continuous variables in the `Topgear` dataset

To assess the appropriateness of the linear model for this dataset, we consider scatterplots of

the response against the other continuous variables. The model name being unique to each car, it plays the role of an identifier and is therefore not relevant in this context. The car maker, on the other hand, could very plausibly influence the price. As the number of brands in the dataset is rather high (48 on a total of 242 observations), using this variable as-is would not provide much insight, however. We therefore do a little feature engineering and cluster the makers on the mean price of their cars, dividing them into high, mid and low price categories, which can then be colour coded.

The resulting scatter plots are shown in Fig. 8a. The observations for the car makers Bugatti and Pagani stand out starkly in all of them, which makes sense, as these are hypercar manufacturers. The data clearly suffers from heteroscedasticity, with higher prices being more dispersed than lower ones. Moreover, while linearity could perhaps seem reasonable within each group separately, it is undeniable that the general slope of the observations changes as we move between groups, contrary to the parallel line pattern we expect from the linear model. Fitting the regression and inspecting the diagnostic plots confirms our suspicions: the graph of the standardised residuals against the fitted values shows a clear downward trend and the distribution of the residuals shows strong deviations from normality in the tails. These findings suggest that the model might benefit from transformation of the response.

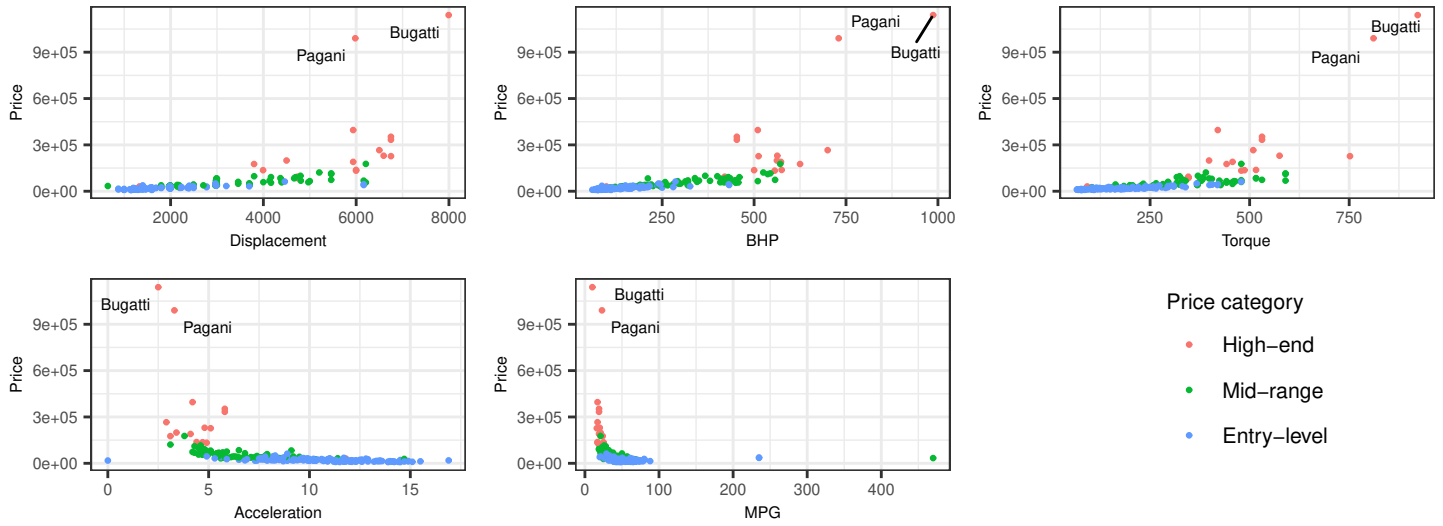


(a) Quantile-quantile plot of standardised residuals **(b)** Studentised residuals vs fitted values

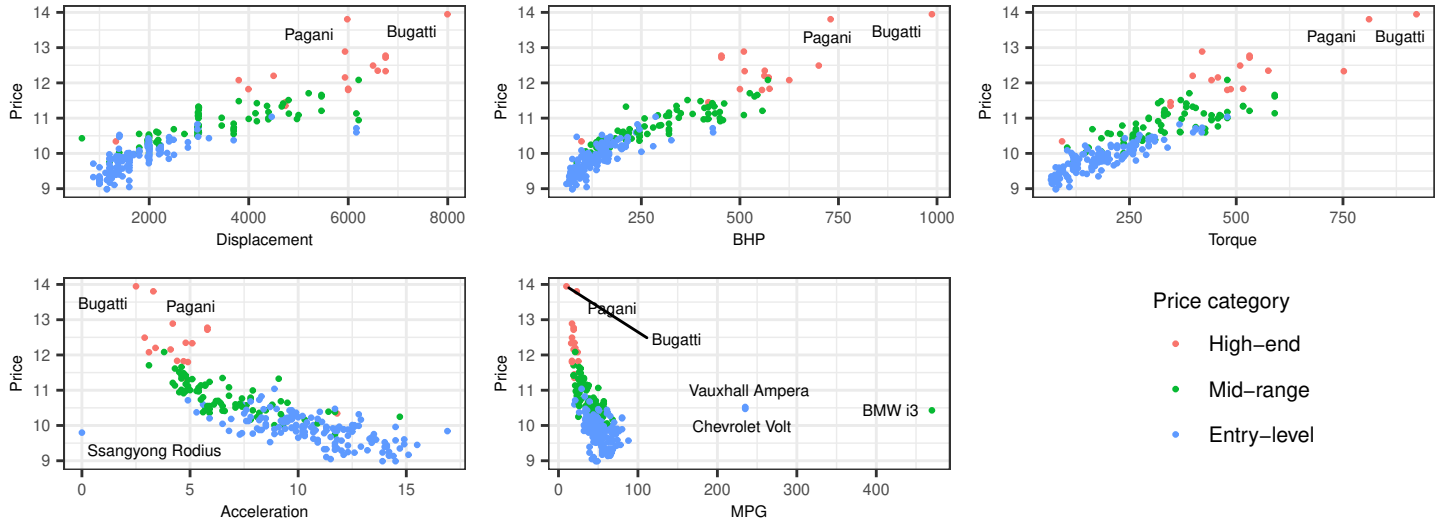
Figure 7: Diagnostic plots for the linear model

Applying a log-transform to the `Price` variable, we obtain the scatters plots shown in Fig. 8b. The linear model seems to fit the data much better now: the variance has been stabilised, and the previously observed pattern is gone from Fig. 9b. Moreover, the hypercar manufacturers which previously stood out like a sore thumb now appear to blend in with the general pattern. However, the number and regularity of outlying residuals in Figs. 9a and 9b can lead one to suspect the presence of fat tails in the price distribution, especially as this is often observed in economic data. Furthermore, we are confronted with some new extreme outliers in the scatterplots for the `Acceleration` and `MPG` variables.

The bad leverage of these observations has a severe impact on the linear fit; in particular, it



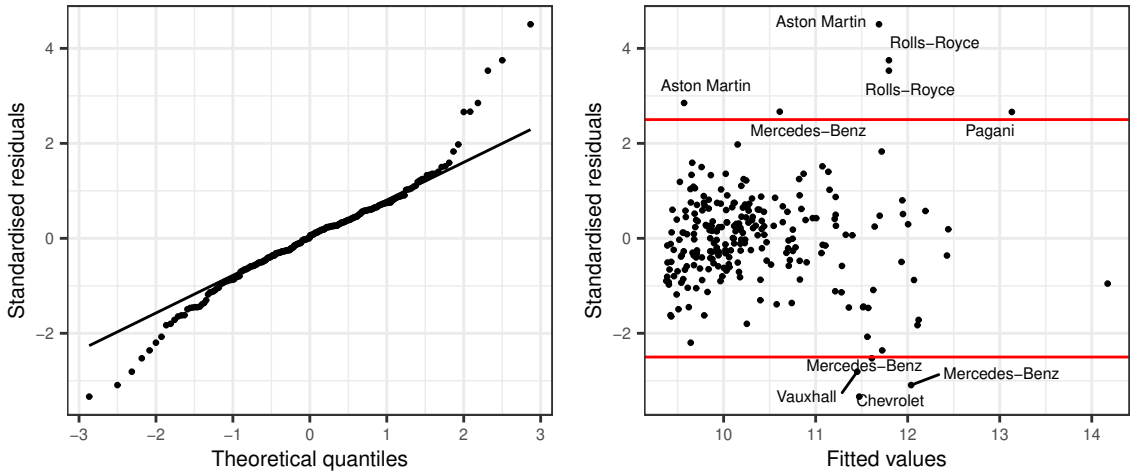
(a) Original



(b) Transformed

Figure 8: Scatter plots of the `Price` variable against the numerical features in the `Topgear` dataset. The colours indicate different price categories.

results in a positive coefficient estimate for the `MPG` variable, which is surprising in view of the expectations expressed earlier on the basis of Fig. 6. For the acceleration, closer inspection of the offending observation reveals that it corresponds to a car supposedly able to achieve a speed of 62 mph in 0 s, which is clearly an error. As for the `MPG` outliers, these correspond to the Chevrolet Volt and Vauxhall Ampera, which are plug-in hybrids, and the BMW i3, which is an electric car. After consulting a domain expert (automotive engineer), it appears that `MPG` measurements do not incorporate battery consumption for electric-powered vehicules, and are therefore not very meaningful. Based on these considerations, we have chosen to exclude these observations from the remainder of the analysis. Interestingly, neither the classical nor the robust fit were able to identify them as bad leverage points.



(a) Quantile-quantile plot of standardised residuals

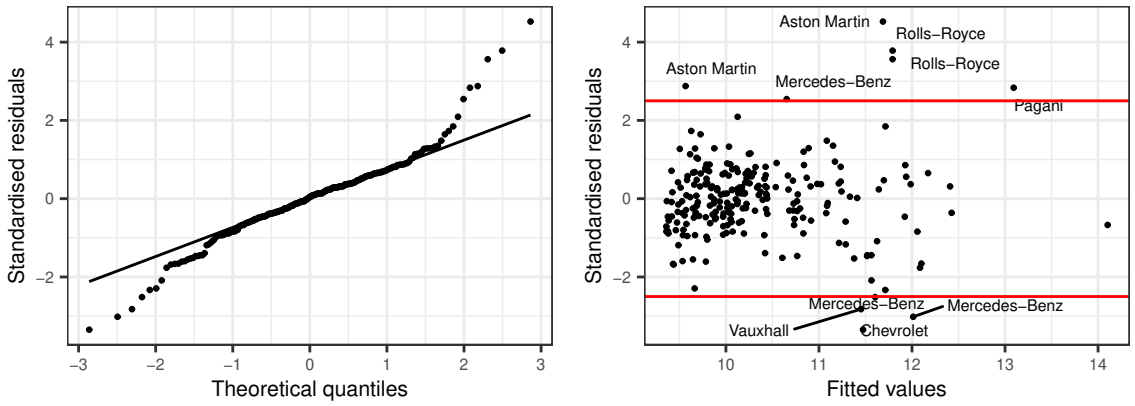
(b) Studentised residuals vs fitted values

Figure 9: Diagnostic plots for the normal linear model with log-transformed response.

Table 2 summarises the regression results for the transformed model under a classical least squares fit, excluding the aforementioned observations. As we can see, they are now consistent with Fig. 6, although the fat tails are still present in the residual diagnostics. This makes it difficult to trust parametric significance tests, and have therefore chosen not to include them here. The coefficients are dominated by the intercept term, which not surprising: regardless of its characteristics, a car cannot be sold profitably for few hundred pounds. Moreover, because the response has been log-transformed, these coefficients describe effects *on a multiplicative scale*, i.e. they express the *rate* at which the car price increases as a function of the predictors.

	(Intercept)	Displacement	BHP	Torque	Acceleration	MPG
Estimate	9.29	0.0000	0.003	0.002	-0.02	-0.001

Table 2: Regression results for the linear model with transformed response. The response has been log-transformed and the bad observations excluded.



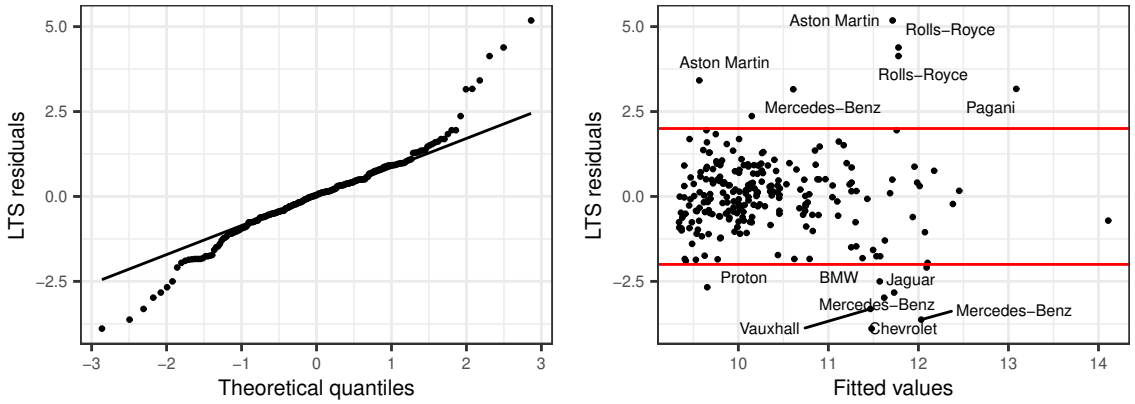
(a) Quantile-quantile plot of standardised residuals **(b)** Studentised residuals vs fitted values

Figure 10: Diagnostic plots for the linear model. The response has been log-transformed and the bad observations excluded.

Finally, we repeat the previous analysis using the robust LTS fit in place of the classical one. The model diagnostics are broadly the same, but the robust fit flags many more observations as good leverage outliers than the least squares one does. Moreover, the deviation in the tails is slightly more pronounced, as can be seen more clearly by comparing the outlier maps in Fig. 12. The persistence of these phenomena is worrying and leads us to suspect one of two things: either the Gaussian distribution must be replaced by a more appropriate one, or there is some pattern in the data which our model fails to capture. Looking back at the scatterplots in Fig. 8b, we conjecture that there might be groupwise effect between the different price categories which is unaccounted.

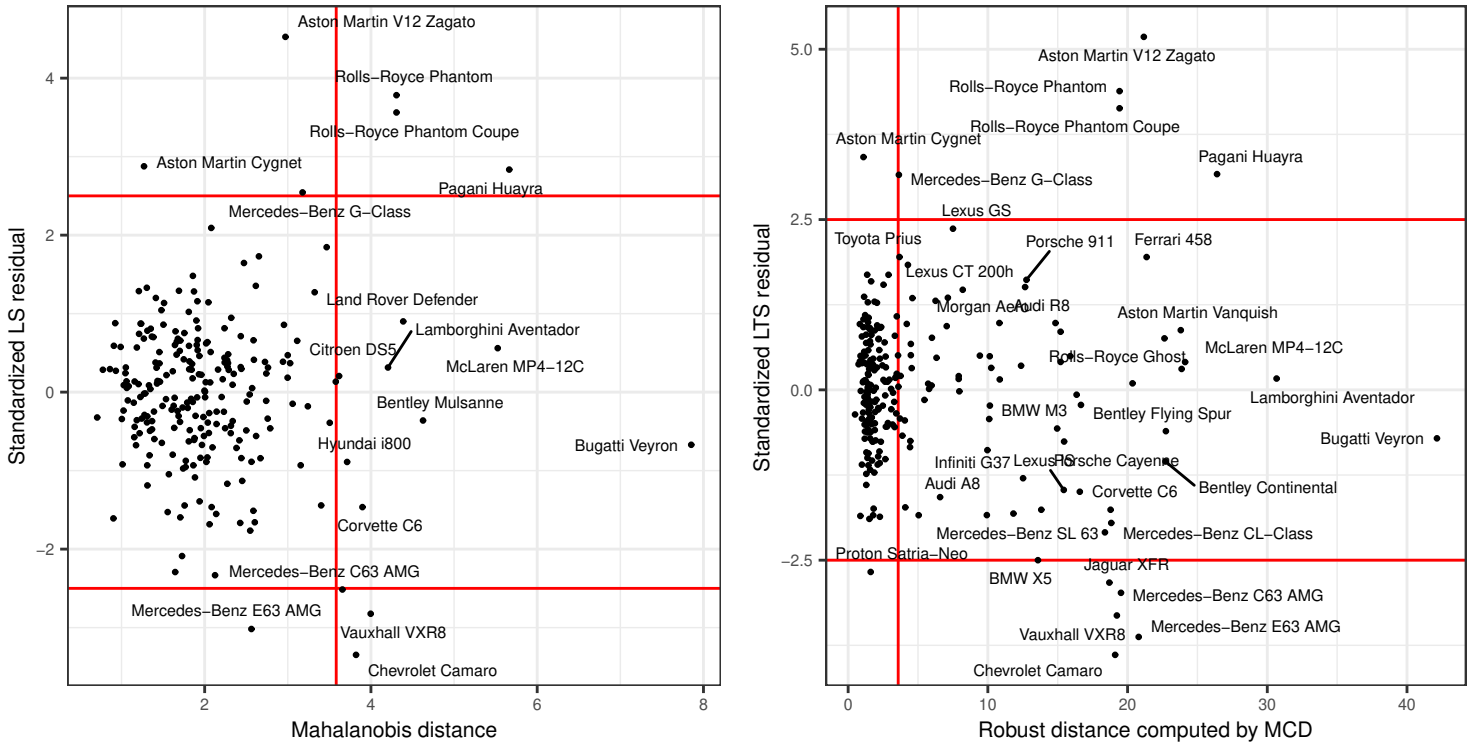
	Intercept	Displacement	BHP	Torque	Acceleration	MPG
Estimate	9.34	0.0000	0.003	0.002	-0.02	-0.0000

Table 3: Regression results for the linear model with robust fit. The response has been log-transformed and the bad observations excluded.



(a) Quantile-quantile plot of standardised residuals **(b)** Studentised residuals vs fitted values

Figure 11: Diagnostic plots for linear model with robust fit. The response has been log-transformed and the bad observations excluded.



(a) Classical LS fit

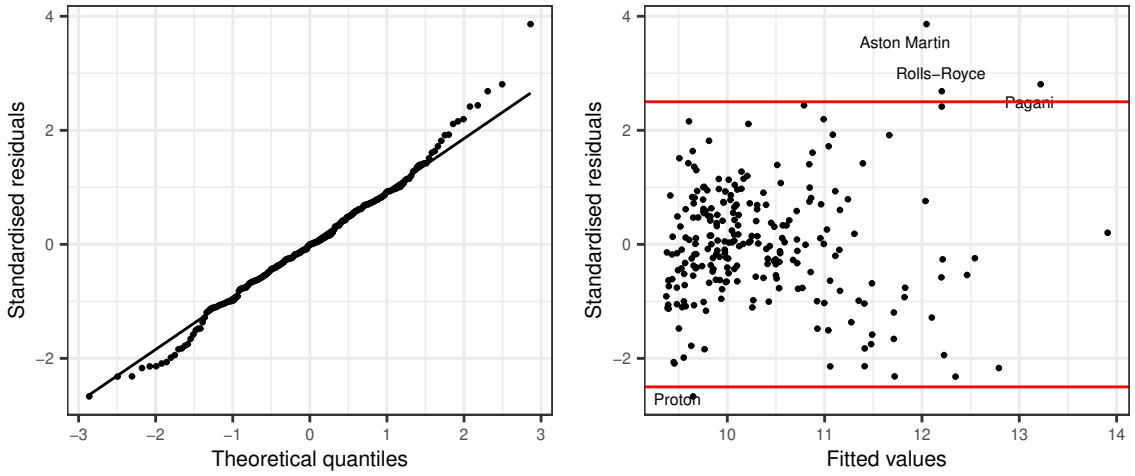
(b) Robust LTS fit

Figure 12: Outlier maps

Refitting the model a final time while incorporating the price category variable, we obtain the diagnostic plots displayed in Figs. 13a and 13b, which show notable improvement. While the residuals still exhibit heavier tails than would warranted under normality, the deviation does not appear to be so significant as to invalidate inferences made on this basis; in fact, both the Shapiro-Wilk and Anderson-Darling tests fail to reject normality of the residuals at a significance level of 0.05, though the former only narrowly.

	(Intercept)	Displacement	BHP	Torque	Acceleration	MPG	CatMid-range	CatEntry-level
Estimate	10	3e-05	0.0014	0.0022	-0.02	-0.0015	-0.61	-0.86
p-value	2.3e-132	0.35	0.00033	3.4e-18	0.039	0.34	5.4e-14	1.4e-19

Table 4: Regression results for the linear model with classical fit. The response has been log-transformed and the bad observations excluded.



(a) Quantile-quantile plot of standardised residuals

(b) Studentised residuals vs fitted values

Figure 13: Diagnostic plots for linear model with classical fit. The response has been log-transformed and the bad observations excluded.

Once again, the sign of the coefficients agrees with the correlations for the continuous variables, and makes logical sense for the newly introduced categorical one. Moreover, significant effects are only present for 2 out of 3 variables from the first feature group in Fig. 6 and 1 out of 2 from the second, which is consistent with the dependency structure observed there.

Question 4

The `warpbreaks` dataset contains observations of the number of breaks in yarn during weaving for different types of wool (A and B) and levels of tension (low, medium and high). In this section,

we will be interested in predicting the number of breaks as a function of the wool type and tension level. A boxplot indicates the existence of an inverse relation between the response and the latter of these variables.

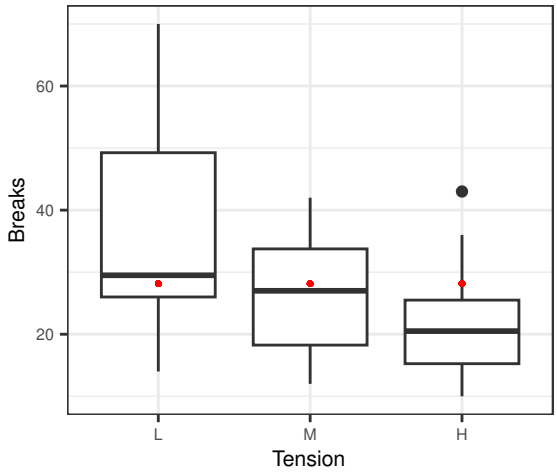


Figure 14: Boxplot of response at the different levels of tension. The mean is indicated in red.

As the response consists of counts, the data lends itself naturally to a Poisson GLM. The independent variables are categorical in nature and must therefore be converted using some kind of encoding scheme before the model can be fitted. The default one used in `R` is dummy encoding, where a binary vector is used to indicate which levels are present for a particular observation. Note that the indicator vector for a categorical variable with k levels will have only $k-1$ entries, otherwise the model would not be identifiable. The resulting design matrix will be an array with values in $\{0, 1\}$, of which the first few rows are shown in Table 5.

	(Intercept)	woolB	tensionM	tensionH
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0
6	1	0	0	0

Table 5: First few rows of the design matrix

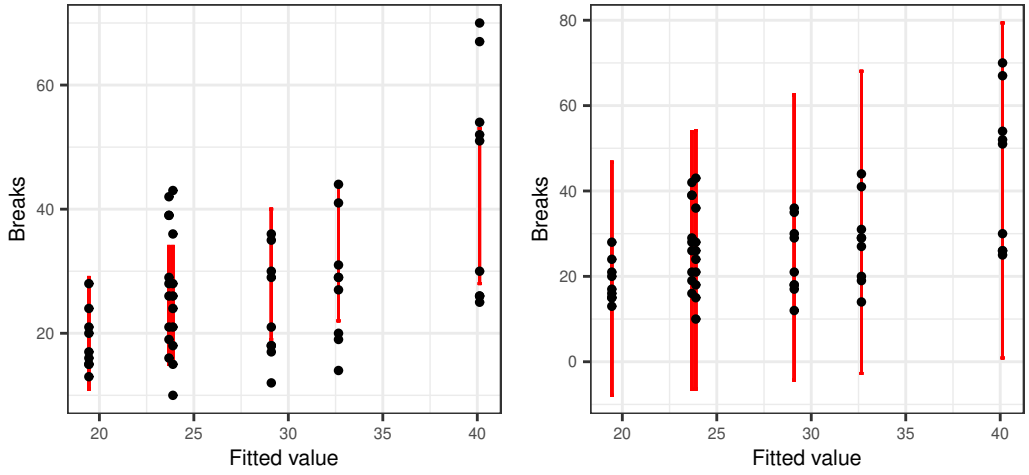
Table 6 shows the results of fitting the Poisson regression `breaks` $\sim$ `wool` + `tension` to the `warpbreaks` dataset. As expected, the coefficient estimate becomes more negative for higher levels of tension. It is important to bear in mind, however, that this model uses a log link, so that these numbers express effects on a multiplicative scale. All coefficient estimates are found to be significant, but the p-values can only be trusted if the model assumptions are fulfilled.

	Dependent variable:
	breaks
woolB	-0.206^{***} (0.052)
tensionM	-0.321^{***} (0.060)
tensionH	-0.518^{***} (0.064)
Constant	3.692^{***} (0.045)
Observations	54
Log Likelihood	-242.528
Akaike Inf. Crit.	493.056
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Results for the Poisson GLM

To check for the presence of overdispersion in the data, we consider the plot of response against the fitted value given in Fig. 15a. The bars show 95% confidence intervals under the Poisson GLM. A significant number of observations fall outside of this range, suggesting that the data exhibits more variability than is appropriate for the model. A formal test using the `dispersiontest` function from the `AER` package [5] (which is based on [6]) confirms our suspicion, rejecting the null hypothesis of unit dispersion with $p = 2.33\text{e} - 6$ The significance tests for the coefficients shown earlier are therefore unreliable.

To account for this, we adjust the GLM to use the quasi-Poisson family, which allows the dispersion parameter ϕ to be greater 1. The resulting diagnostic plot, shown in Fig. 15a, looks much better. Note that the bars represent a range of 3 estimated standard deviations around the fitted mean.



(a) Poisson **(b)** Quasi-Poisson

Figure 15: Response against fitted value

Finally, we repeat the same analysis using a robust estimator. As we were unable to find an implementation supporting the quasi-Poisson family, we will use the ordinary variant and limit ourselves to comparison of coefficients. As we can see in Table 7, these are very similar, increasing our confidence in the previous model.

	(Intercept)	woolB	tensionM	tensionH
Estimate	3.56	-0.15	-0.23	-0.45

Table 7: Results for the Poisson GLM with robust fit

References

[1] C. Mikstas. "Liver: Is it good for you?" (2023), [Online]. Available: https://www.webmd.com/diet/liver-good-for-you.

[2] J. Raymaekers and P. Rousseeuw, *Cellwise: Analyzing data with cellwise outliers*, R package version 2.5.0, 2022. [Online]. Available: https://CRAN.R-project.org/package=cellWise.

[3] V. Todorov, *Rrcov: Scalable robust estimators with high breakdown point*, R package version 1.7-2, 2022. [Online]. Available: https://CRAN.R-project.org/package=rrcov.

[4] H. Wickham. "Usda nutrients." (2014), [Online]. Available: https://github.com/hadley/usdanutrients.

[5] C. Kleiber and A. Zeileis, *Applied Econometrics with R*. New York: Springer-Verlag, 2008, ISBN 978-0-387-77316-2. [Online]. Available: https://CRAN.R-project.org/package=AER.

[6] A. Cameron and P. K. Trivedi, "Regression-based tests for overdispersion in the poisson model," *Journal of Econometrics*, vol. 46, no. 3, pp. 347–364, 1990, ISSN: 0304-4076. DOI: https://doi.org/10.1016/0304-4076(90)90014-K.