Part 1

Assessing the dataset

The dataset comprises of 9 NA values for *price*, which is the response variable, so those observations were removed. Independent variable *parker* has 18 NAs. Once those NAs are removed, it is inferred that *parker* is correlated (fig1) with many other independent variables such as *year*, *s.temp* and *h.temp*. Although *h.temp* also has some correlations but *parker* seems to be the most dominating. However, removing those 18 observations attenuated the dataset further; given the observed correlation of *parker* with other variables, removing it as an independent variable seemed necessary.

	year	price	h.rain	s.temp	w.rain	h.temp	parker
year	1	-0.21	-0.30	0.41	0.19	0.44	0.41
price	-0.21	1	0.08	0.36	0.03	0.33	0.67
h.rain	-0.30	0.08	1	0.08	-0.45	-0.45	-0.19
s.temp	0.41	0.36	0.08	1	-0.27	0.48	0.50
w.rain	0.19	0.03	-0.45	-0.27	1	0.49	0.28
h.temp	0.44	0.33	-0.45	0.48	0.49	1	0.58
parker	0.41	0.67	-0.19	0.50	0.28	0.58	1

fig 1: correlation matrix

After pre-processing the dataset, the univariate distribution of *price* and the bivariate distributions of *price* with respect to all the independent variables were assessed (Appendix1) and there seemed to be some linearity between price and most variables. The histogram (fig2) shows that the data points are clustered towards a specific range of prices with a few observations falling out of this range. The data appears to be right-skewed and approximates a Gamma distribution, moreover, prices are strictly positive $(0,\infty)$ and here they are continuous. Manning et al (2002), modelled costs most robustly by a Generalised Gamma Regression compared to OLS with log link and OLS with the identity link. Because *price* is reflected as a percentage of 1961 mean prices, it has been scaled. However, the actual 1961 mean price is unknown and if it is assumed to be $100, then the observed percentage of prices in the dataset can be assumed to be equal to the actual mean prices.

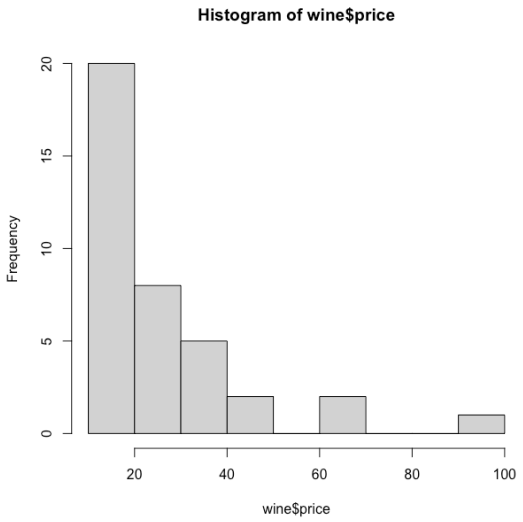


fig 2: univariate distribution of price

Other models such as Binomial and Poisson were considered but since our dataset is neither binary nor count, they were ruled

out. Hence, the initial model to be considered is a gamma GLM with 5 predictors: *year, h.rain, s.temp, w.rain,* and *h.temp.*

Choosing the models

Model1:

$$y_i = price \sim Gamma(\mu_i, v)$$
$$\log(\mu_i) = \beta_0 + \beta_1 year + \beta_2 h.rain + \beta_3 s.temp + \beta_4 w.rain + \beta_5 h.temp$$

In this model, all the coefficients except *h.temp* were statistically significant at the 5% level as *h.temp* had a p-value of 0.1335 and the least t-statistic of 1.540. With that in consideration, ANOVA on Model1 was run and addition of *h.temp* did not reduce the residual deviance by a large number as it went down from 3.7626 to 3.5104. Further, VIF value for *h.temp* was 1.658209, which could imply a certain degree of multicollinearity. However, the residual deviance cannot actually be used for goodness of fit here since the dispersion parameter (ϕ) is unknown.

The residual plots of this model (Appendix2) revealed that the deviances were not normally distributed as the points did not align well on the theoretical line. Further, the deviances were not uniformly scattered around 0, showing signs of heteroscedasticity. However, all points were well within the Cook's Distance so there were no alarming outliers. The AIC here is 265.41 but another model without *h.temp* was examined.

Model2:

$$y_i = price \sim Gamma(\mu_i, v)$$
$$\log(\mu_i) = \beta_0 + \beta_1 year + \beta_2 h.rain + \beta_3 s.temp + \beta_4 w.rain$$

After removing *h.temp*, the significance of all other independent variables was preserved and in fact all of their p-values had reduced further making *h.rain* to the 0.001 level as opposed to the 0.05 level in Model1. The AIC of this model increased but only very minutely to 266.09 and the residual deviance also increased minutely to 3.7626. The Number of Fisher Scoring iterations in this model and in Model1 were the same, 6, which means there were no convergence issues.

The residual analysis (Appendix3) is key to the comparison, and in this model, the points on the Q-Q plot align better on the theoretical line implying that the deviances are more towards a normal distribution than in Model1. Further, the deviances are more evenly distributed around 0 than those in Model1, and there are no points outside

Cook's Distance. Although a slight increase in the AIC, the residual plots showed some improvement and those weighted against the AIC, Model2 maintains more parsimony and a better fit. In order to ascertain Model2's fit, two other potential models were assessed before finalising the approach.

Model3:

$$y_i = price \sim Gaussian(\mu_i, \sigma_i^2)$$
$$\log(\mu_i) = \beta_0 + \beta_1 year + \beta_2 h.rain + \beta_3 s.temp + \beta_4 w.rain$$

This model assumes that *price* follows a Gaussian distribution and has a log link function to ensure values are strictly positive just as prices should be. This model, although with significant coefficients, has a much higher AIC of 302.81 with the residual deviance of 4686.9 on 33 degrees of freedom.

There is one point outside Cook's Distance here, which is the wine that the highest price of $100 - this variance in price has not been accommodated by this model, and since prices can vary differently, Gamma GLM did indeed fit better. The residuals are more heteroscedastic than the other models. However, the residuals gravitate towards a more normal distribution here (Appendix4)

Model4:

$$y_i = \log(price) \sim Gaussian(\mu_i, \sigma_i^2)$$
$$\mathrm{E}(\log(y_i)) = \beta_0 + \beta_1 year + \beta_2 h.rain + \beta_3 s.temp + \beta_4 w.rain$$

The model uses the log-transformed *price* variable in order to assess if the variance in *prices* can be improved. The response is assumed to be lognormal, and the identity link is used. However, the model fits the log(*price)* and not the actual *price* so the interpretation will not be the same. The coefficients are significant, and the AIC dropped down to 30.628, the best across all the other models.

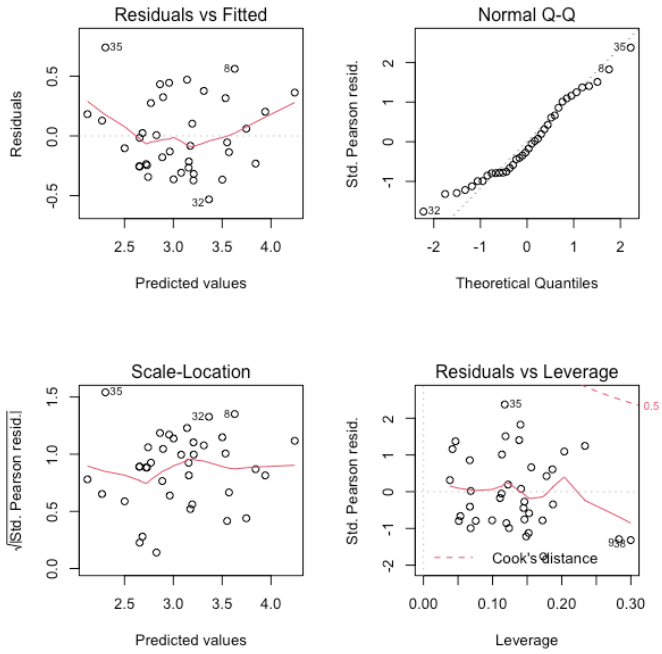


fig 3: Model 4

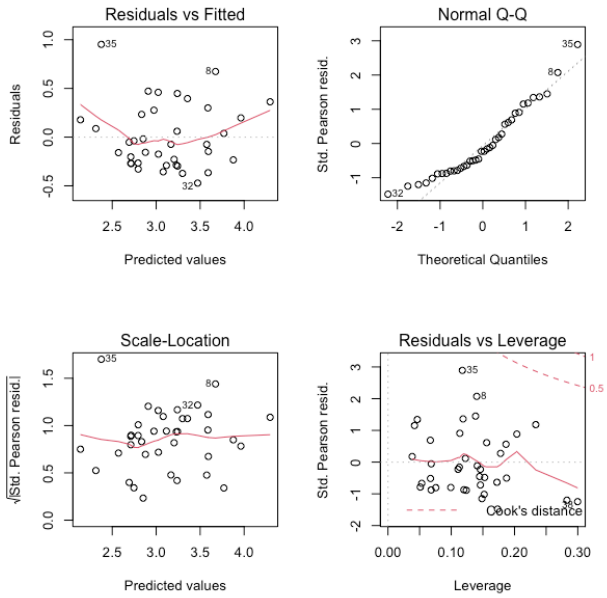


fig 4: Model 2

All the points are within the Cook's Distance and the residuals show an equally uniform distribution around 0 as Model2 (fig3 and fig4). The Q-Q plot suggests a slightly better approximation to normal distribution here though. Except for the AIC, all other comparison points differ closely. However, the biggest limitation is the interpretability of the coefficients as the fit as the mean of log(*price*) cannot be exponentiated to get the mean of *price*. Whereas, in the Gamma GLM with log link, the log of mean of *price* can be exponentiated to get the mean of *price*. Lastly, the *price* reflects an average value of wines in a given year, and log link renders the mean. Therefore, Model2 is the chosen regression model.

Interpretation of the parameters:

$$y_i = price \sim Gamma(\mu_i, \nu)$$
$$\log(\mu_i) = 53.6 - 0.03(year) - 0.0028(h.rain) + 0.5122(s.temp) + 0.001(w.rain)$$

	Estimate	Standard Error	t-value	Pr(> | t |)
Intercept	**53.6007984**	**10.6668902**	**5.025**	**1.71e-05**
year	**-0.0300501**	**0.0055522**	**-5.412**	**5.46e-06**
h.rain	**-0.0028365**	**0.0009288**	**-3.054**	**0.00444**
s.temp	**0.5122724**	**0.0834042**	**6.142**	**6.36e-07**
w.rain	**0.0010464**	**0.0004528**	**2.311**	**0.02722**

Intercept: The expected value of *price* when all other parameters are set to 0 is $53.6.

Coefficients: Due to the log link, the coefficients were exponentiated to calculate the percentage change in the expected value with respect to a unit change in the variable(s).

i. year: 1 year increase in production year will result in a reduction of 3% in the expected value of price.
ii. h.rain: 1 mm increase in the rain in harvest month will cause a 0.28% reduction in the expected value of price.
iii. s.temp: 1 degree increase in average summer temperature of the preceding year will increase

the expected value of price by 67%.

iv. w.rain: 1 mm increase in rain in the winter preceding harvest will increase the expected value of price by 0.105%.

The magnitude of affect some of these parameters have on price is quite small, especially *w.rain* and *h.rain.*

Standard Errors, t-values, and p-values:

The standard errors are calculated to assess how the estimates differ from the population values and the t-values measure how the variance compares. If the coefficient estimates do not affect the expected value of *price,* then t-value would be small in magnitude and the associated p-value would be large. The parameters of the model are all statically significant to be able to reject the null hypothesis that they do not affect the *price* at all.

Confidence Intervals (95%):

i. intercept: Between 32.90 and 74.37
ii. year: Between -4% and -1.2%
iii. h.rain: Between -0.46% and -0.09%
iv. s.temp: Between 42% and 96%
v. w.rain: Between 0.01% and 0.2%

Analysis Limitations

Firstly, the model comprises of a very few numbers of observations and to approximate a better fit, more observations are important. There were many NA values which resulted in the removal of *parker,* and that prevented from assessing the value that could have added. The analysis overlooks the dispersion parameter, which is unknown hence it is unable to comment on the residuals. There are no interaction terms included that may have helped understand how certain independent variables affect other independent variables' impact on the dependent variable. For example, how temperature over the preceding summer could have affected the rain in the winters, which could have caused changes in prices. Some other models such as Generalised Additive Models have not been assessed for their suitability to fit the data.

With regards to the model's performance, the residual plots can be improved. The Q-Q plot did not show a perfectly normal distribution, although it was better than some other models, it must further be improved to make deviances more normal and the model more accurate. There are some deviances with non-uniform and higher magnitudes of variances than others, but more observations can help ascertain the level of heteroscedasticity. Lastly, the

model showed a very high AIC value, as compared to Model2 so it is being penalised for the number of parameters and thus the quality of fit. All the coefficient estimates lie within the confidence intervals but *s.temp* has a broad interval and that might question the ability of the model.

Conclusion and Recommendations

The analysis aimed to study if and how certain features such as weather conditions and time affect wine prices; this can be extrapolated to study how climate change might impact wine quality and prices (Asimov, 2019). The summer temperatures preceding harvest have a large impact on wine prices, it aligns with the common belief that warmer regions indeed produce pricier wines if we consider Spain, Chile, or Argentina as popular wine producing countries. Interestingly, it was notice that rains in the winter preceding harvest have a significant but not a substantial impact on wine prices. A more aged wine did prove to demand a higher price as studied in the regression. Further, rains in the harvest month are not favourable.

Initially, *parker* ratings showed a high significance and it had shadowed other features and it was dropped because of many missing values. However, it showed that the ratings actually encapsulated many of the features. Perhaps the ratings were based on similar metrics as the other independent variables. Even harvest temperature presented some correlation.

An important advantage of the Gamma GLM model was that it captured information from observations that were regarded as high influence and high leverage outliers by the Gaussian GLM. Prices are strictly positive, and the nature of price distribution showed that many values were concentrated close to zero and very few spread towards the right (Siegel, 2016).

For more robust analysis, it is recommended to gather more observations and record them along more features. Alcohol content, pH levels, grapes as factor variables, and soil properties may render a deeper insight into factors that strongly affect wine prices (Puckett, 2020). Although the model used rain and temperatures of the harvest year, it may be of value to analyse how these may affect the grapes in the next year by making time series adjustments. Robert Parker ratings is one such metric, but other ratings may predict prices differently. Lastly, other models such as Generalised Additive Models or distribution models may be explored and compared to overcome the limitations presented by the Gamma GLM.

References

Asimov, E. (2019) *How Climate Change Impacts Wine (Published 2019)*, *Nytimes.com*. Available at: https://www.nytimes.com/interactive/2019/10/14/dining/drinks/climate-change-wine.html (Accessed: 14 December 2021).

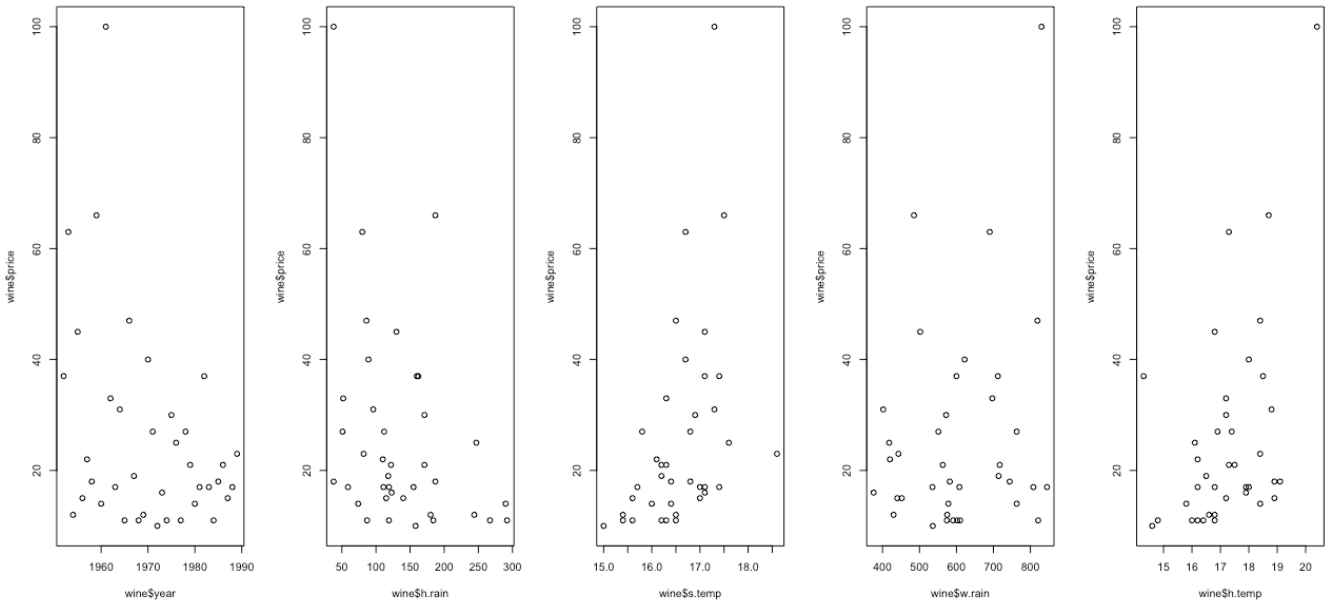
Manning, W., Basu, A. and Mullahy, J. (2002) *Modeling Costs with Generalized Gamma Regression*. Chicago: The University of Chicago. Available at: https://www.researchgate.net/publication/228930311_Modeling_Costs_with_Generalized_Gamma_Regression (Accessed: 9 December 2021).

Puckett, M. (2020) *Five Traits of the World's Most Expensive Wines | Wine Folly*, *Wine Folly*. Available at: https://winefolly.com/tips/5-traits-of-the-worlds-most-expensive-wines/ (Accessed: 10 December 2021).

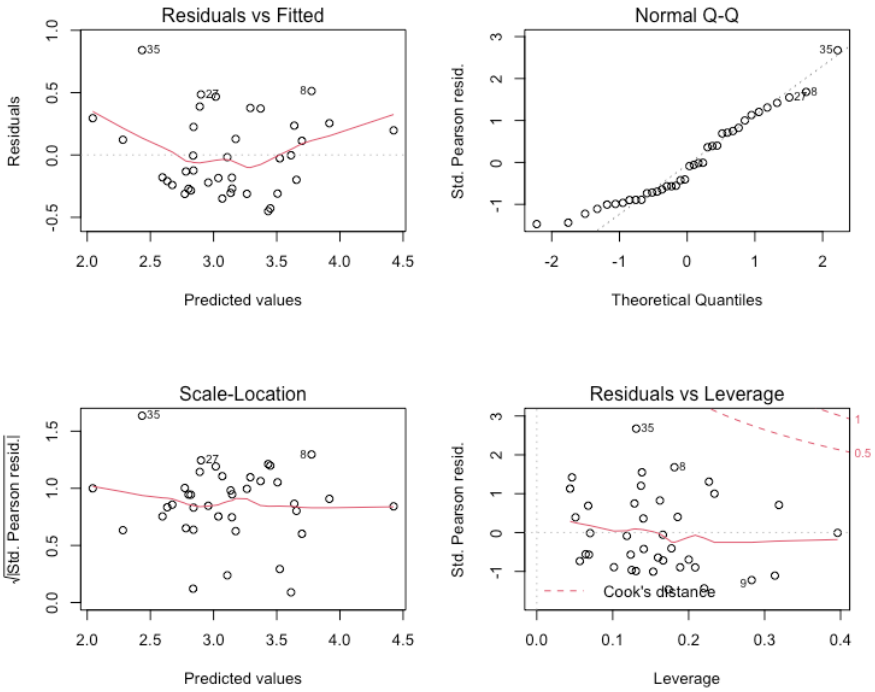
Siegel, A. (2016) "Histograms", *Practical Business Statistics*, pp. 41-70. doi: 10.1016/b978-0-12-804250-2.00003-1.

Appendix

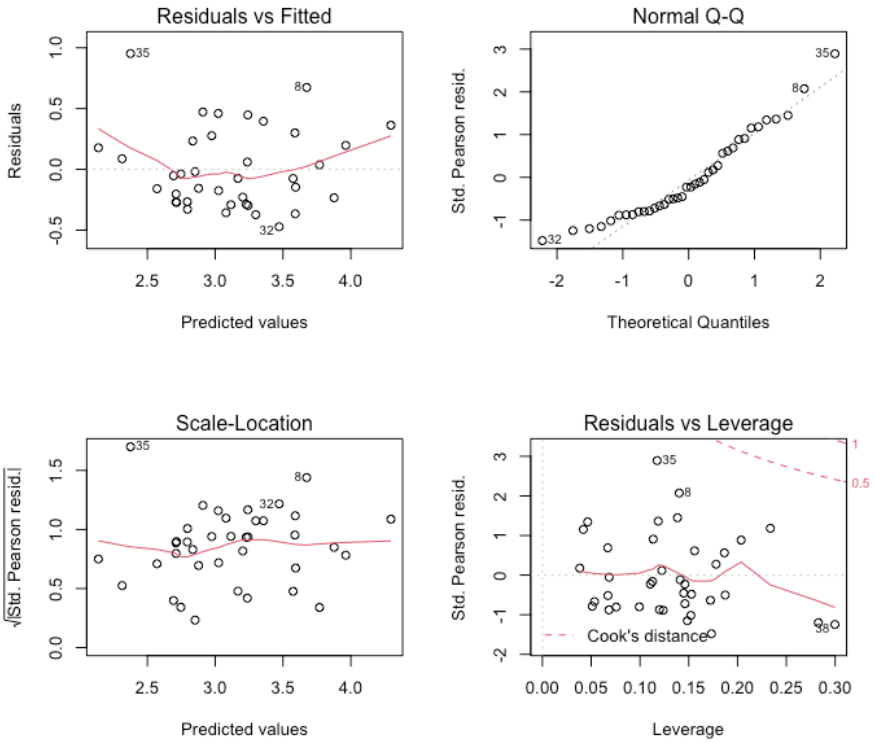
1. *price* plotted against independent variables.



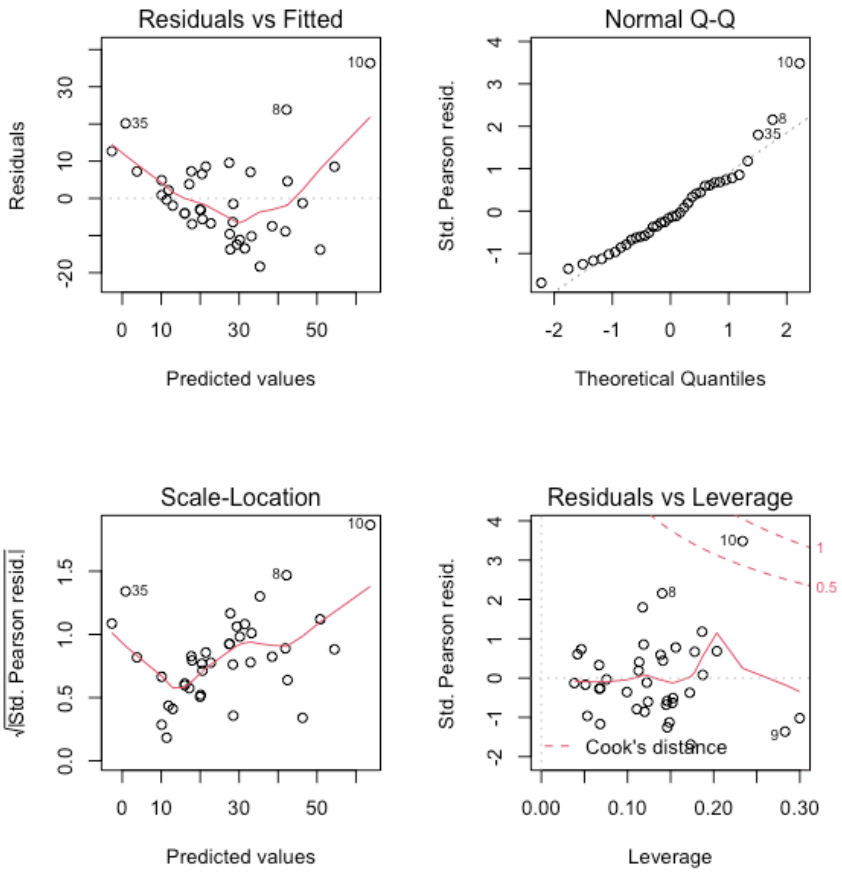
2. Model1 residual plots



3. Model2 residual plots



4. Model3 residual plots



Part 2:

(a)

The first plot shows that the marginal increment in the number of mistakes becomes greater by a unit increase in the speed. As can be seen at speed 2 the mistakes are 7 and at speed 3 the mistakes are more than 25. Similarly, the jump in mistakes from speed 4 (at 47) to speed 5 (121) is very high. The variance here is very unstable as the points follow a sharp curve shape. The right plot is of the same mistakes and speed, but the mistakes have been log transformed. The transformations have scaled down the mistakes to a different unit which approximate a more linear trend. The original data may not be normally distributed.

Hence, the data may be fitted using a Gaussian model with a log-transformed response variable. In this case, the log of mistakes will be used to improve the linearity (as seen in the figure). However, this model will be assuming that log(mistakes) is normally distributed and not the mistakes itself. Moreover, the speed will be a predictor for log(Mistakes) and not for the actual mistakes. This consideration must be kept in place whilst interpreting the coefficients.

Another possibility is to use the Poisson model because the mistakes are positive integers that can be categorised as a count variable. If the data is actually not normally distributed, and shows the characteristics of a Poisson distribution, then especially with the log-link as the canonical link, Poisson model may be suitable to fit the data. If the operatives work at a different rate, the offset argument of the model can accommodate this as well.

However, in order to robustly ascertain suitability of models, it is paramount to study previous modelling of similar data and to compare the residual plots, summary statistics, and goodness of fit parameters for all the potential models.

(b)

The model that has been fitted is a generalised linear model (GLM) with the assumption that Mistakes have a Poisson distribution. In order to have only positive estimates of the mean of mistakes, the canonical link is a log-link. This model is generally used to model a response with a positive count distribution.

$$y_i = Mistakes \sim Poisson(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_0 + \beta_p x_{ip}$$
$$\Rightarrow \log(\mu_i) = 0.27143 + 0.90598(Speed)$$

Under the model, the mean and variance of Mistakes is assumed to be equal. One unit change in the Speed, will increase the expected value of Mistakes by 147%; because of the log-link, this inference is derived by exponentiating the coefficient of speed.

(c)

The Fisher scoring iterations is an algorithmic approach to estimating a GLM. The iterative method aims to locate the model parameters that maximise the fit. The iterations continuously move towards the direction of the parameters where fit keeps improving. Finally, they converge to a solution when another iteration does not add any valuable improvement. The number of Fisher scoring iterations is essentially the number of iterations the algorithm took to deliver the parameters of a given model. A high number of iterations may imply some problems in the convergence process and in the parameters themselves. In this scenario, the algorithm converged to a solution after 3 attempts and another attempt would not have maximised the fit as well.

(d)

Firstly, the very small p-value of the coefficient of Speed is statistically significant in having an impact on the response. With respect to the goodness of fit, since the response assumes a Poisson distribution (phi is known), the residual deviance chi-squared test can be done to compare the maximised log likelihood of the current model with that of the saturated model. The null hypothesis here holds that the current model shows the same behaviour as the saturated model. The p-value of residual deviance of 2.0163 calculated at 3 degrees of freedom is 0.57. This is much greater than the 0.05 level. Therefore, the null hypothesis cannot be rejected, and it implies that the current model has a good fit.

The AIC is relatively low at 29.828 and the Number of Fisher Scoring Iterations are 3, implying that the algorithm easily converged.

The Deviance Explained = 0.9907985. This implies that the model is able to explain almost 99% of the deviance in the data. However, the model only has 5 observations, hence the high percentage. More observations can better help understand the model's fit. However, from the apparent statistics, the model seems to have a good fit, especially so, if deviance residuals can prove to be normally distributed and homoscedastic.

(e)

The Null deviance compares the likelihood of the null model with the likelihood of the saturated model. The null hypothesis is that the null model behaves in the same way as the saturated model. The null model is the model with just the intercept and the saturated model is the one with as many parameters as the number of observations i.e., the best fit model.

A low null deviance can imply that the intercept itself is able to explain the model and thus, using fewer predictors will be in best interest. Here, the null deviance of 219.1271 is high. The p-value from the chi-square test on 4 degrees of freedom is 0 which means the null hypothesis can be rejected and the null model behaves differently from the saturated model, therefore, adding more parameters than the intercept may help fit the data better.

Another way to test the same null hypothesis is the ANOVA, which compares nested models. A null model by definition is a subset of a saturated model since the saturated model encapsulates the intercept. As per ANOVA, if the p-value returned is smaller than 0.05, the null hypothesis is then rejected in favour of the saturated model implying that the difference between the two models is statistically significant.

(f)

The model that has been fitted is a linear regression model with the log-transformed response variable (Mistakes). The transformation is done to improve the linearity of the variance. However, the response variable has changed from Mistakes to log(Mistakes).

$$\log(y_i) \sim Gaussian(\mu_i, \sigma_i^2)$$

$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$E(\log(y_i) = \ \beta_0 + \beta_1 x_i$$

$$\Rightarrow E(log(Mistakes) \ = \ -0.13208 \ + \ 1.01095(Speed)$$

According to the model log(Mistakes) has been assumed to follow a normal distribution and a unit change in predictor (Speed) will increase the expected value of log(Mistakes) by 1.01095.

(g)

The first line of the code creates a scatter plot with Speed on the x-axis and logMistakes on the y-axis. Then, the second one draws a straight line on the same plot at the intercept -0.13208 and slope 1.01095 (model llm). The third line of code adds another straight dashed line which takes the intercept and slope from model speed.glm. Finally, the last line creates a legend box at coordinates (1,4.5) with the given descriptions and line types.

(h)

The figure is of the plot of observed change in logMistakes with respect to Speed. The two lines represent the two models fitted to the observed data. As can be seen, the Poisson estimates are quite accurate at some points, especially for Speed 5 and Speed 4. However, the normal linear model line is much closer to the observed points overall, especially at Speeds 1 and 2. As for Speed 3, the closeness to the point is similar for both the lines. Ultimately, simply observing the given plot, the normal linear model with logMistakes seems to fit the data better than the Poisson GLM with log link. Certainly, this cannot be ascertained using only 5 data points and by just the given visualisation.