Stanwell-Fletcher Lake: Analysis of the sediment composition of samples taken from the lake basin.

160175125

Contents

Sediment	Percentages			Depth	Sediment	Percentages			Depth
no.	Sand	Silt	Clay	(m)	no.	Sand	Silt	Clay	(m)
1	77.5	19.5	3.0	10.4	21	9.5	53.5	37.0	47.1
2	71.7	24.9	3.2	11.7	22	17.1	48.0	34.9	48.4
3	50.7	36.1	13.2	12.8	23	10.5	55.4	34.1	49.4
4	52.2	40.9	6.6	13.0	24	4.8	54.7	41.0	49.5
5	70.0	26.5	3.5	15.7	25	2.6	45.2	52.2	59.2
6	66.5	32.2	1.3	16.3	26	11.4	52.7	35.9	60.1
7	43.1	55.3	1.6	18.0	27	6.7	46.9	46.4	61.7
8	53.4	36.8	9.8	18.7	28	6.9	49.7	43.4	62.4
9	15.5	54.4	30.1	20.7	29	4.0	44.9	51.1	69.3
10	31.7	41.5	26.8	22.1	30	7.4	51.6	40.9	73.6
11	65.7	27.8	6.5	22.4	31	4.8	49.5	45.7	74.4
12	70.4	29.0	0.6	24.4	32	4.5	48.5	47.0	78.5
13	17.4	53.6	29.0	25.8	33	6.6	52.1	41.3	82.9
14	10.6	69.8	19.6	32.5	34	6.7	47.3	45.9	87.7
15	38.2	43.1	18.7	33.6	35	7.4	45.6	46.9	88.1
16	10.6	52.7	36.5	36.8	36	6.0	48.9	45.1	90.4
17	18.4	50.7	30.9	37.8	37	6.3	53.8	39.9	90.6
18	4.6	47.4	48.0	36.9	38	2.5	48.0	49.5	97.7
19	15.6	50.4	34.0	42.2	39	2.0	47.8	50.2	103.7
20	31.9	45.1	23.0	47.0					

Figure 1: Shows the data for the proportions of sand, silt, and clay for 39 samples of sediment taken from the basin of the Stanwell-Fletcher Lake. Data provided by Coakley, J.P. & Rust, B.R. (1968) and adapted by Eleanor Stillman and Miguel Juarez of The University of Sheffield.

1 Introduction

Stanwell-Fletcher Lake is a 131 square mile glacial lake located within the Arctic Circle on Somerset Island in Northern-most Canada. It was formed during the Pleistocene when the lake basin was carved out by a glacier moving through a graben. The excavation then filled with sea water as the glacier receeded and over time the salt water was flushed out by freshwater flowing into the lake from the nearby mountains (Coakley, J.P. & Rust, B.R. (1970)).

This report is a statistical analysis of the sediment composition of the lake bed, and the question it tries to answer is; is the sediment composition dependent on the depth of the water in the lake? and if so, why? The report then attempts to present a possible explanation for the results of the analysis. The data used for the analysis is taken from Coakley, J.P. & Rust, B.R. (1968) and adapt by Eleanor Stillman & Miguel Juarez from The University of Sheffield. The data consists of 39 samples of sediment taken from the surface of the lake bed at different depths.

The data recorded of the sediment composition is broken down into 3 fundamental constituents, sand, silt, and clay. These constituents are primarily characterised by their particle sizes; sand has the largest particle size with a diameter ranging from 2mm down to 0.05mm, followed by silt which has a diameter ranging from 0.05mm down to 0.002mm, and then by clay which has a diameter less than 0.002mm (Stephen. (2015)).

Sand is created by erosion or weathering of rocks which break down into smaller pieces over time. This process is a consequence of many types of natural phenomena such as; waves, which forcefully impacts rocks and weakens them over time; frost, which causes rocks to break apart when the water freezes and expands inside cracks; and even glacial movement, where the underbelly of the glacier will grind against the rock underneath through a process called scouring. Silt on the other hand is mainly only created by processes that exploit weaknesses in the molecular lattices of sand sized particles, such as frost or chemical weathering (Wikipedia (2019)). Lastly, clay is formed by chemical weathering only, usually from particles breaking down in acidic solutions (Erik Painter (2015)).

There are many ways in which sediment can be transported across the environment, either being transported across the environment by fast winds or flowing water, and in general, the further the sediment has been transported, the smaller the grain size will be. Lakes are interesting places on land in regards to sediment transport as lakes will capture any sediment being transported across the environment which will accumulate on the lake bed (Abbas, A. / Gajula, N. (2016)).

Table 1: Shows various numerical summaries of the data for each constituent.

sand	silt	clay
Min. :0.0200	Min. :0.1950	Min. :0.0060
1st Qu.:0.0645	1st Qu.:0.4230	1st Qu.:0.1595
Median :0.1060	Median :0.4800	Median :0.3490
Mean :0.2419	Mean :0.4569	Mean :0.3011
3rd Qu.:0.4065	3rd Qu.:0.5240	3rd Qu.:0.4540
Max. :0.7750	Max. :0.6980	Max. :0.5220

The particle size of sand means that sand is generally transported by rolling along the surface of the environment. However if the air or water velocity is high enough, then the sand particles will be suspended within the fluid due to the turbulence created by the fast flowing fluid, the sand can then be transported rapidly across long distances. Silt behaves much like the sand in that if the velocity of the fluid is slow enough, the turbulence will not be sufficient to suspend the silt within the fluid and the silt will settle, but silt can be suspended in fluid moving much slower than the sand can, meaning silt will travel faster over distances than sand given the same condtitions. Clay behaves differently in that the smaller clay particles can be suspended within water for long periods of time as the particles are small enough to bounce off water molecules and remain suspended. Clay therefore has the possibility to travel most rapidly out of the three constituents over distances (Fondriest Environmental, Inc. (2014)).

2 Exploratory Analysis

2.2 Numerical Summaries

The numerical summaries for each of the constituents are displayed in **Table 1**. The means of the three constituents show that the average proportion of sand is lower than the proportion of clay, and much lower than the proportion of silt. We might expect then that sand is either proportionally less abundant than silt and clay, or high proportions of sand are concentrated at just a few depths and low in proportion everywhere else. The high mean for the silt proportion implies that the silt must be found in relatively high proportions

at most depths.

The difference between the mean and median proportions (calculated as: mean - median) for sand, silt, and clay are 0.136, -0.023, and -0.048 respectively. The distribution of the silt and clay proportions are slighty negatively skewed, but the distribution of the sand proportion has a large positive skew. This large skew suggests that the hypothesis that the sand proportion is high at a few depths, but low at the remaining depths is true. The correlation coefficients for sand, silt, and clay are -0.768, 0.444, and 0.837 respectively. The correlations tell us that the silt and clay proportion increase with depth, whereas the sand proportion decreases with depth. These numerical summaries suggest that the silt and the clay has a greater ability to migrate into the centre of the lake whereas the sand collects and settles close to the shore. This is most likely due to the tidal dynamics close to the shore where the water currents are the most powerful and the turbulence is high. As a consequence, the silt and clay are carried away from the shore while the sand is left behind.

The interquartile ranges for sand, silt, and clay are 0.342, 0.101, and 0.294. The interquartile range for sand is 238.6% larger than the interquartile range for silt. Likewise the interquartile range for clay is 191.6% larger than the range for silt, and the interquartile range for sand is 16.1% larger than the interquartile range for clay. So we expect that the spread in the sand proportion values is the highest of the 3 constituents, and the spread of the silt proportion values is the lowest.

The distance between the median, and the 1st & 3rd quantiles for sand are 0.041, and 0.3 respectively. The latter is 624.1% larger than the former showing the sand proportion is extremely asymmetrical in its distribution. Likewise for silt, the distances are 0.057 and 0.044. In this case the latter is 22.8% smaller than the former, showing the distribution of silt is largely symmetrical. Lastly, the distances for clay are 0.189, and 0.105. In this case the latter is 44.6% smaller than the former, so clay is more asymmetrically distributed than silt, but still much less asymmetrically distributed than sand.

2.3 Visual Summaries

The next task is to visually explore the relationship between depth and the constituents. **Figure 2** shows the proportion of sand measured in each of the 39 samples taken from the Standwell-Fletcher Lake basin. The structure of the data exhibits an exponential

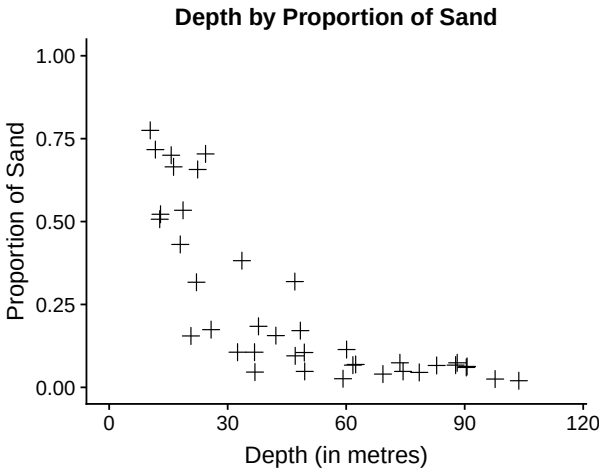

Figure 2: A plot showing the proportion of sand measured in 39 samples of sediment taken at varying depths under the Stanwell-Fletcher Lake. The proportion of sand looks to exponentially decay as the depth increases, as does the variation in the data.

decay, so it is likely that the sand is most abundant near the lake shore and becomes increasingly rare towards the centre of the lake where it is deeper. This matches what we might intuitively expect to happen, given that silt and clay can be suspended and transported in water flowing at much lower velocities than sand can, and the strong flows are likely to be nearer to the shore due to the tidal dynamics, then the currents carry away more of the silt and clay than the sand. Some sand may eventually make its way to the centre of the lake by rolling along the lake bed due to the underwater currents, but is found in relatively much smaller quantities than the silt and clay due to its inability to remain suspended in the water and transported efficiently.

The variation in the data is much higher for shallower depths which may imply a large difference in the strength of the currents where the samples were taken. Where the proportion of sand is relatively high, we may expect strong currents which take away the silt and clay, and where the proportion of sand is relatively low, we may expect weak currents in that area which result in deposition of the silt and clay there.

The measurements of the proportion of clay in each sample are plotted in **Figure 3**. The structure of the data is a logarithmic relationship this time where the proportion of clay increases as the depth increases. In the shallower water, the clay proportion is very low which suggests the hypothesis that the clay is

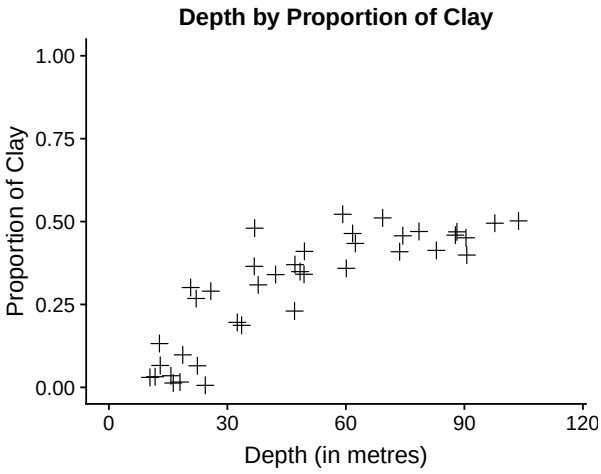

Figure 3: A plot showing the proportion of clay measured in 39 samples of sediment taken at varying depths under the Stanwell-Fletcher Lake. The data appears to follow a logarithmic relationship with a upper bound of the mean at around 0.5. The variation appears larger for the clay between depths of 20 and 50 metres.

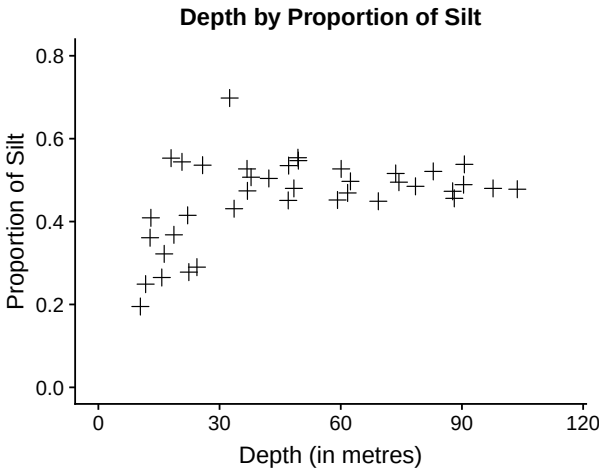

Figure 4: A plot showing the proportion of silt measured in 39 samples of sediment taken at varying depths under the Stanwell-Fletcher Lake. The proportion of silt looks to be independent of depth below 35 metres; the variance also becomes smaller below 35 metres. The relation looks to be logarithmic.

washed into the centre of the lake by tidal dynamics is likely to be true. Observing that for the deeper parts of the lake, greater than 60 metres, the sand and clay proportion do not appear to change much, which implies the silt proportion also does not change much either. We deduce then that the dynamics of the underwater currents are much weaker in the deeper parts of the lake. There are three ways then in which the silt and clay can be transported into the centre of the lake. The first is that the currents, while being comparitively weak compared with the currents found in the shallower depths, are still strong enough to keep the silt and clay suspended in the water which causes them to be distributed relatively evenly. The second way is that the silt and clay is deposited on the underwater banks of the lake bed, and periodically builds up to the point where the sediment avalanches down the slope, some of which is kicked up into the water and is transported. The final way is the currents close to the surface of the lake disperse the silt and clay above the deeper regions, where the silt and clay eventually sink down to the bottom.

The data for the silt proportion has a similar structure to the clay proportion data. **Figure 4** shows the measurements of the silt proportion in each sample. Like the clay proportion, the silt proportion also appears to be logarithmic. The main difference between the silt and the clay proportion is that the silt is found in higher proportions than the clay at shallow depths. This is likely due to the higher turbidity of the water needed to transport the silt relative to the clay. Note also the ouyling observation in the silt data. The proportion data corresponding to that observation does sum to 1, so the outlier is unlikely to be an error in the data input. This means that at the location the sample was taken from, there must be a buildup of silt, and it is likely not to be representative of the expected composition of the sediment at that specific depth, so there is an argument for discounting this particular observation.

3 Model Building

3.1 The Approach

The exploratory analysis highlighted that the composition of the sediment is indeed dependent on depth, so therefore there is an opportunity to model the proportions of the constituents by knowing only the depth. Since the data type is proportional data which bound to the standard unit interval [0,1], and since the variation of the proportions is not constant in every case, a natural choice for modelling this type of data is with beta regression, where the response is assumed to come from a beta distribution. (Cribari-

Neto, F. & Zeileis, A. (Date unspecified)). Due to the dependency of the constituents, where the sum of the proportions is always 1, the decided approach to the model building is to model each constituent independently where the response depends only on the depth of the water, and then assess which two of the three models (which will be referred to as the η models) infers the relationship of the third constituent most accurately (this inferred model will be referred to as the ζ model). A nice consequence of this approach is that the model fits can be assessed visually rather than having to rely only on statistical checks. One unfortunate reason for this approach is that it is not possible to construct a model that depends on one of the other constituents and produces a smooth line of best fit through the data due to the particular dependency of the constituents on each other, in fact trying this approach with a beta regression model results in extreme overfitting.

3.2 Model selection

Regarding the approach, the main difficulty with the model selection is that the models we want to compare were created on different subsets of the data so they cannot be compared directly with statistical tests. It is possible however to compare the fits indirectly through the assessments of the ζ and η model which fit the same data. The η models are the two models which are constructed with conventional model building methods, and are dependent on the depth. The ζ model is the set of fitted values obtained by subtracting the fitted values of the η models away from 1.

The linear predictor for the ζ model must then be of the form

$$\zeta_a = 1 - \eta_b - \eta_c,$$

where ζ_a is the linear predictor of the inferrred model, η_b, η_c are the linear predictors of the fitted models, and $a, b, c \in \{\text{sand}, \text{silt}, \text{clay}\}$ with $a \neq b \neq c$, i.e a model for each constituent must appear in the linear predictor equation. There exists 1 ζ model and 1 η model for each constituent.

The method of assessment for the ζ model is to calculate a quantity which we will name R (defined in §3.2.1). The value for R is defined such that the ζ model with the smallest R value is the model that is most alike the η model for that data. The R values, in conjunction with the diagnostics for the ζ models will then be used to choose the most suitable model, which subsequently implies the selection of the two η models we should choose.

3.2.1 Formula for the calculation of R

Let $\hat{y}_i$ be the fitted value of observation i for the η model, q_i, p_i be the values of the 0.975 and 0.025 quantiles of observation i respectively for the η model, $h_i = q_i - p_i$ be the size of the interval, and let $\hat{z}_i$ be the fitted value of observation i for the ζ model. Then define

$$S_1 = \frac{(\hat{y}_i - \hat{z}_i)^2}{(\hat{y}_i - p_i)^2}, \qquad S_2 = \frac{(\hat{y}_i - \hat{z}_i)^2}{(q_i - \hat{y}_i)^2},$$

If $\hat{y}_i > \hat{z}_i$, set $S_2 = 0$. If $\hat{y}_i \leq \hat{z}_i$, set $S_1 = 0$. Then

$$R = \sum_i^{39} (S_1 + S_2).$$

The ζ model that is closest to the fitted model will then have the lowest R score. Note that it is not possible to simply take the difference of the two models because $\hat{y}_i - \hat{z}_i = \eta_a - (1 - \eta_b - \eta_c)$ which is symmetric for sand, silt, and clay so produces the same number for each model. Hence why the difference must be normalised by the estimated spread of the data.

3.3 Model Construction

See the attached appendix for the model building.

3.4 The Formal Model

The formal model is the combination of the 2 chosen η models and the ζ model which is inferred by the chosen η models. The linear predictors and dispersion predictors for the two chosen η models from §3.3 are given by

$$\begin{aligned} \eta_{i,\text{silt}} &= -2.5615 - 0.013x_i + -0.013\log(x_i), \\ \phi_{i,\text{silt}} &= 2.1273 + 0.0461x_i, \end{aligned}$$

and

$$\begin{aligned} \eta_{i,\text{clay}} &= -7.4247 - 0.0172x_i + 1.9668\log(x_i), \\ \phi_{i,\text{clay}} &= 6.3735 + 0.0883x_i - 1.9361\log(x_i), \end{aligned}$$

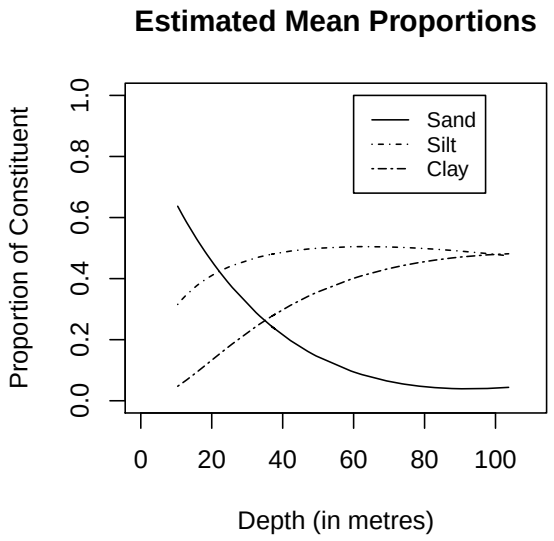

Figure 5: shows the estimated mean proportion for each constituent for various depths in the Stanwell-Fletcher Lake by the formal model. For the extremeties of the plot, only 2 of the 3 constituents are expected to be found. We would not expect to find clay near to the shore, and we would not expect to find sand in the deepest parts of the lake.

where both models use the `logit` μ link and `log` ϕ link.

The linear predictor for the inferred ζ model is then given by

$$\zeta_{i,\text{sand}} = 1 - \log\left(\frac{\eta_{i,\text{clay}}}{1-\eta_{i,\text{clay}}}\right) - \log\left(\frac{\eta_{i,\text{silt}}}{1-\eta_{i,\text{silt}}}\right).$$

The fit of the models on the data can be seen in **Figure 11**, §3 of the Appendix.

The mean proportions for the constituents estimated by the formal model are shown in **Figure 5**. The plot tells us that we expect to find only two of the three constituents at the extremes of the depth. We should not expect to find clay at the shoreline, and likewise we should not expect to find sand in the deepest parts of the lake. As hypothesised in the EDA we would expect to see silt in relatively high proportions in all parts of the lake.

4 Conclusions

The analysis has improved our understanding of how the sediment composition changes with water depth. We can now say that the composition does indeed depend on the water depth of the lake. We saw that the proportion of sand is inversely proportional to the increase in the silt and clay proportion combined.

The highest proportions of sand are found near to the shore of the lake, and the proportion quickly decays as we move towards the centre of the lake where the water is much deeper. This is likely due to the lack of turbidity needed to keep the sand suspended in the water and transported away from the shore efficiently. The clay is found in very low proportions near the shore of the lake and as we move towards the deeper areas, the clay proportion rises. This is likely because of the wave action near the shore providing sufficient turbidity and strong currents which carry the clay away from the shore. The currents then lose their strength in the deeper parts of the lake where the clay is deposited. The silt behaves more like the clay than the sand, meaning that the dynamics of the lake currents and turbulence of the water is powerful enough to also suspend the silt and transport it across the lake. We see however that the silt is also able to be deposited at all depths of the lake despite being found in smaller quantities near the shore. This is likely due to how the silt is defined. The silt is defined to have grain sizes spanning the gap between the grain size of sand and clay, so for silt with a grain size close to the grain size of the sand, we expect the silt to exhibit propoerties closer to that of sand in regards to transportation. Likewise we expect silt to behave more like clay when the grain size is closer to that of clay. Therefore the large silt particles will be deposited closer to the shore like the sand, while the smaller silt particles will be carried to the central part of the lake like the clay. This is why we find the silt in all parts of the lake.

5 Study Limitations

Overall the study achieves its main goal, which was to assess whether the sediment composition of the basin of the Stanwell-Fletcher Lake was dependent of the water depth of the lake. The study concluded that indeed this was the case. Also, with the information from the gathered sources, the study was able to determine possible explanations for why the sediment composition is dependent of the depth of the water

and provide a great deal of insight into the possible dynamics of the water within the lake.

The main limitation of the study is the formal models' inability to describe the variation of the sand proportion directly. However this can be indirectly understood from the modelled variation in the silt and the clay proportions.

6 Acknowledgements

A acknowledgement goes to Francisco Cribari-Neto and Achim Zeileis, whose comprehensive guide on performing beta regression was a great help in the making of this study.

7 References

Coakley, J.P. & Rust, B.R. (1970). Canadian Journal of Earth Sciences, 1970, 7(3): 900-911, https://doi.org/10.1139/e70-085. Accessed online March 2019.

Coakley, J.P. & Rust, B.R. (1968). Sedimentation in an Arctic lake J. Sed. Petrology, 38, 1290-1300. Accessed March 2019.

Stephen. (2015). How to Create Amazing Garden Soil from Clay, Silt or Sand. http://www.livingoffgridguide.com/gardening/how-to-create-amazing-garden-soil/. Accessed online March 2019.

Wikipedia (2019). Silt. https://en.wikipedia.org/wiki/Silt. Accessed online March 2019.

Erik Painter (2015). What is clay made out of?. https://www.quora.com/What-is-clay-made-out-of. Accessed online March 2019.

Fondriest Environmental, Inc. (2014). "Sediment Transport and Deposition." Fundamentals of Environmental Measurements. 5 Dec. 2014. Web. < https://www.fondriest.com/environmental-measurements/parameters/hydrology/sediment-transport-deposition/ >. Accessed online March 2019.

Cribari-Neto, F. & Zeileis, A. (Date unspecified). Beta Regression in R. https://cran.r-project.org/web/packages/betareg/vignettes/betareg.pdf. Accessed online March 2019.

Abbas, A. / Gajula, N. (2016). What can a sedimentary rock tell you about the environment in which it was formed?. https://www.quora.com/What-can-a-sedimentary-rock-tell-you-about-the-environment-in-which-it-was-formed (2016). Accessed online March 2019.

Appendix

1 η Model Analysis

1.1 Model for the sand proportion

The initial model chosen for the sand proportion, based on the findings of the EDA, an appropriate choice for the initial model takes the following form;

$$g_1(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 \log(x_i),$$
$$g_2(\phi_i) = \gamma_1 + \gamma_2 x_i,$$
$$i = 1, ..., 39, \qquad y_i \sim Be(\mu_i, \phi_i),$$

where x_i is the depth of observation i, ϕ_i is the dispersion parameter, g_1 is the `logit` μ link function, g_2 is the `log` ϕ link function, and β_j and γ_k are the regressor coefficients; $y_i = g_1^{-1}(\mu_i)$ is assumed to be beta distributed with probability density function

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi - 1}(1-y)^{(1-\mu)\phi - 1}.$$

The model for the mean will be referred to as the μ model, and the model for the dispersion will be referred to as the ϕ model.

```
M1.sand <- betareg(sand ~  log(depth) | depth)
summary(M1.sand)
```

```
##
## Call:
## betareg(formula = sand ~ log(depth) | depth)
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -2.0461 -0.7155 -0.0560  0.7272  2.1074
##
## Coefficients (mean model with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.1692     0.5822   8.879   <2e-16 ***
## log(depth)   -1.8254     0.1459 -12.513   <2e-16 ***
##
## Phi coefficients (precision model with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.121906   0.423392   2.650  0.00805 **
## depth       0.041157   0.007894   5.214 1.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 50.95 on 4 Df
## Pseudo R-squared: 0.7827
## Number of iterations: 18 (BFGS) + 1 (Fisher scoring)
```

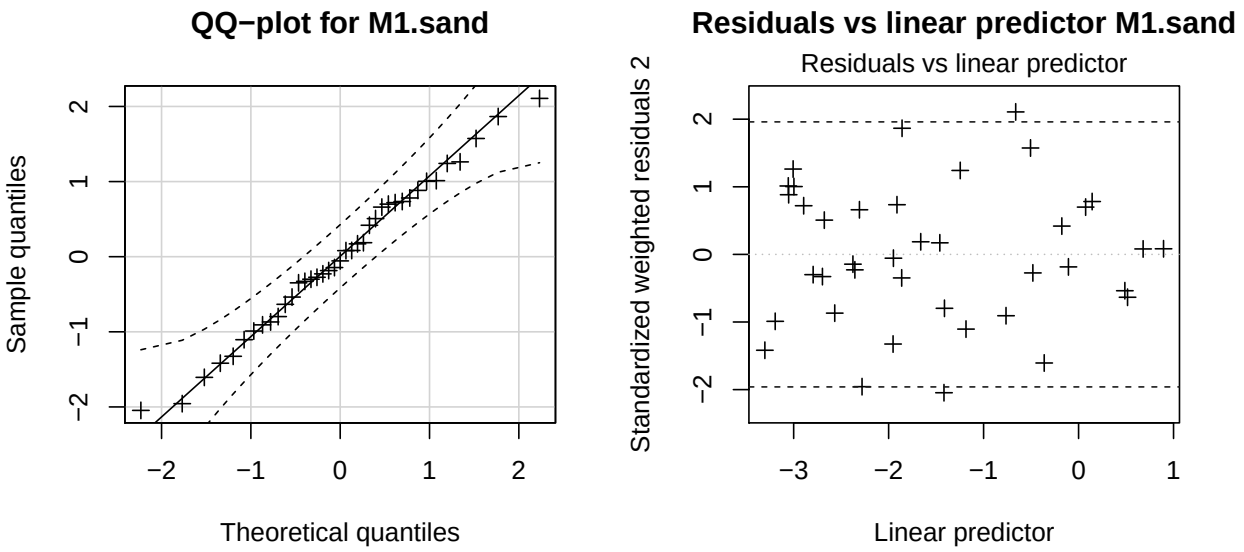

Figure 6: Shows relevant diagnostics for the M1.sand model. The QQ-plot shows the theoretical quantiles vs the sample quantiles for the standardised residuals. The residuals appear to be well approximated by a standard normal distribution. The residuals vs linear predictor plot shows the distribution of the standardised residuals. The residuals are well scattered about the horizontal axis, however there is evidence that the variance is not constant and is largest in the central region of the plot.

The linear term was removed from the `M1.sand` model as it is not significant (summary not shown). The summary for the updated `M1.sand` model now shows all the coefficients are significant and the pseudo R-squared is high. The residuals also look symmetrical and there are no extreme residual values so the model seems to be a good fit overall.

Figure 6 shows the relevant diagnostics for the `M1.sand` model. The Q-Q plot suggests that assumption that the standardised residuals are normally distributed is a good assumption. The residuals vs linear predictor plot shows evidence that the model could be improved to better capture the variation in the data as the variance looks non-constant, where it appears that the variance is larger in the central region of the plot.

To try and improve the ability of the model to model the variation in the data, various additions to the model were tested using different sets of link functions, checking the diagnostics every time a valid model was found, similar to the assessment of `M1.sand` (The processs is omitted from the appendix as it is largely uninteresting. The best model found at the end of the exploration is given by the following `M2.sand` model.

```
M2.sand <- betareg(sand ~ log(depth)| depth+log(depth), link="logit", link.phi="sqrt")
summary(M2.sand)
```

```
##
## Call:
## betareg(formula = sand ~ log(depth) | depth + log(depth), link = "logit",
##     link.phi = "sqrt")
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -2.1140 -0.6657 -0.0406  0.8537  1.6809
##
```

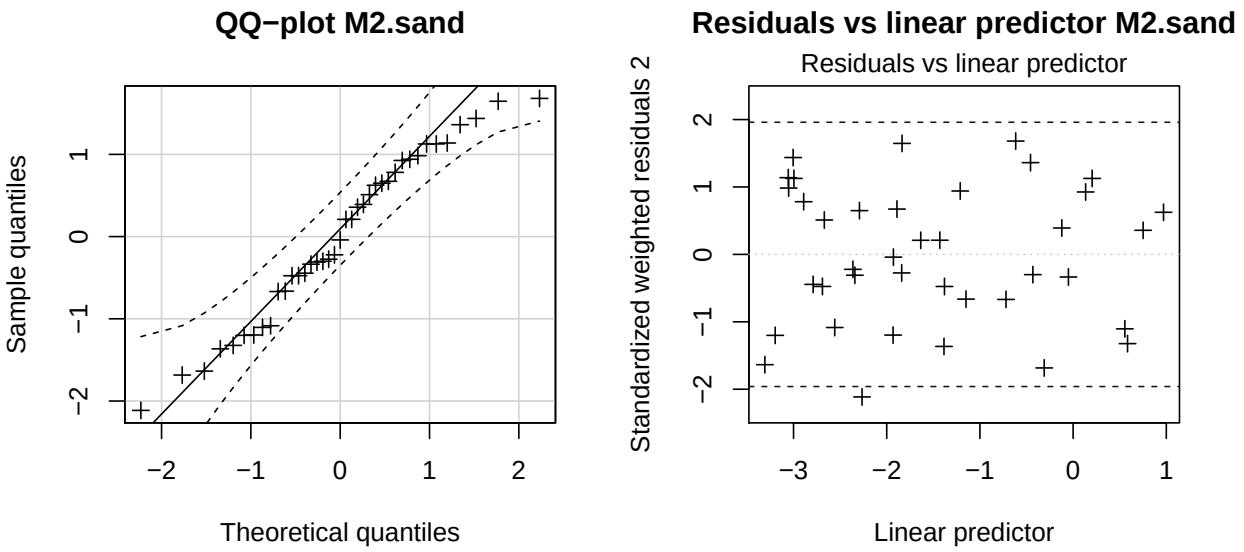

Figure 7: Shows the relevant diagnostics for the M2.sand model. The Q-Q plot indicates that a normal approximation is reasonable for the residuals. The residuals vs linear predictor plot has improved substantially and shows no concerning issues.

```
## Coefficients (mean model with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.3261     0.4100   12.99   <2e-16 ***
## log(depth)   -1.8601     0.1046  -17.78   <2e-16 ***
##
## Phi coefficients (precision model with sqrt link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 25.23388    7.99786   3.155 0.001605 **
## depth        0.36168    0.09733   3.716 0.000202 ***
## log(depth)  -9.91765    3.24288  -3.058 0.002226 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 54.46 on 5 Df
## Pseudo R-squared: 0.7827
## Number of iterations: 56 (BFGS) + 5 (Fisher scoring)
```

The summary shows that the log term is significant with the choice of the `sqrt` link function for the dispersion model.

Figure 7 shows the relevant diagnostics for the `M2.sand` model. The plots shows that the distribution of the residuals has improved significantly over the `M1.sand` model, and now look almost perfectly normally distributed.

```
AICc(M1.sand,M2.sand)
```

```
##         df      AICc
## M1.sand  4 -92.72933
```

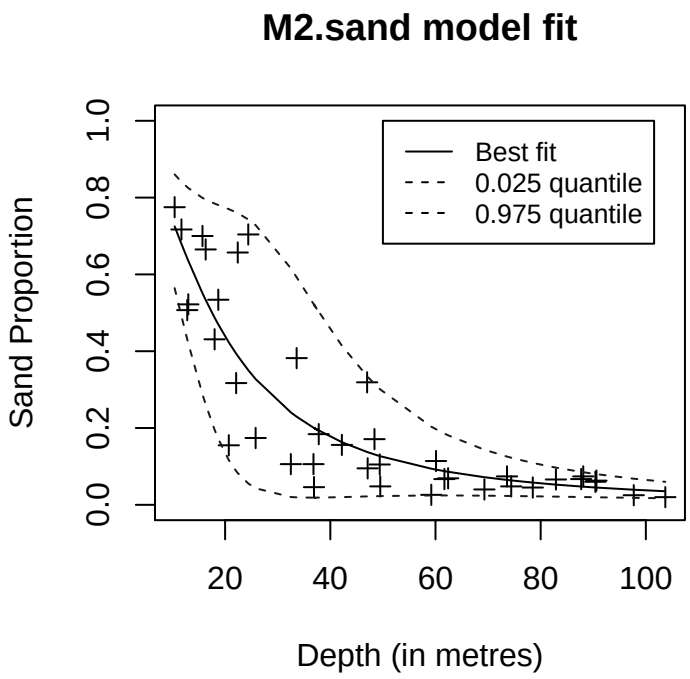

Figure 8: shows the best fit line and the 0.975 and 0.025 quantile boundaries for the M2.sand model. The model looks like it fits the data well very well.

```
## M2.sand  5 -97.10854
```

The AICc comparison shows that the `M2.sand` model outperforms the `M1.sand` model by a good margin. Note that the AICc will always be used since $39/K < 40$ where 39 is the number of observations and K is the number of model parameters where $K \geq 1$ in all cases (Cribari-Neto, F. & Zeileis, A. (Date unspecified)).

Figure 8 shows the best fit line and the 0.975 & 0.025 quantile boundaries for the `M2.sand` model. The model fits the data very well. Since the diagnostics show no problems with the fit, we take it as the best model to model the sand proportion.

```
sand.eta <- M2.sand
```

1.2 Model for the silt proportion

The EDA shows the silt proportion data has a logarithmic structure, therefore the initial model for the silt proportion takes the form identical to the form in §1.1.

```
M1.silt <- betareg(silt ~ depth+log(depth) | depth)
summary(M1.silt)
```

```
##
## Call:
## betareg(formula = silt ~ depth + log(depth) | depth)
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -1.6680 -0.7323 -0.0760  0.6201  3.0071
```

```
##
## Coefficients (mean model with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.561507   0.699546  -3.662 0.000251 ***
## depth       -0.013007   0.004213  -3.088 0.002018 **
## log(depth)   0.820669   0.235303   3.488 0.000487 ***
##
## Phi coefficients (precision model with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.127329   0.444221   4.789 1.68e-06 ***
## depth       0.046090   0.008076   5.707 1.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 57.53 on 5 Df
## Pseudo R-squared: 0.4793
## Number of iterations: 21 (BFGS) + 1 (Fisher scoring)
```

The summary shows that all the coefficients are significant and the pseudo R-squared is not high but it is ok. The residuals look quite symmetrical but the maximum positive residual is very large and needs to be investigated.

Figure 9 shows various diagnostics for the `M1.silt` model. The diagnostics for the model fit looks ok but there is room for improving the model to account for the variation in the data better. The largest outlying residual belonging to observation 12 was discussed in the EDA as a possible unusual location within the lake where the silt had built up. The observation is not thought to be representative of what we would expect to measure as the silt proportion at that particular depth, and rather the observation is thought of as a freak observation. Therefore the decision was taken to replace observation 12 with a value that better represents that data, with the intention that the variation is more accurately modelled by this change. The replacement value then is chosen such that the value of the proportion for observation 12 lies within, but close to the 0.975 quantile boundary in order to minimise the effect of changing the datapoint on the model fit. The replacement value is then decided to be 0.55. Observation 37 on the other hand is more representative of the data and so this observation was not removed; if it were to be removed it would likely cause the model to underestimate the variance of the observation even more than it is perhaps already doing so.

```
arctic.data$silt[14] <- 0.55
attach(arctic.data)
```

```
## The following objects are masked from arctic.data (pos = 3):
##
##     clay, depth, number, sand, silt, x1, x2, x3

## The following objects are masked from arctic.data (pos = 11):
##
##     clay, depth, number, sand, silt, x1, x2, x3
```

```
M1.silt <- betareg(silt ~ depth+log(depth) | depth)
```

Even with the adjustment of the the residual, the model should still be improvement to tackle the apparent heteroskedasticity within the residuals. As before with the sand models, a search was performed trying different combinations of links and additional complexity until a good improvement on the `M1.silt` model was found. Unfortunately however no such improvement could be found, therefore `M1.silt` was taken as the best model.

```
silt.eta <- M1.silt
```

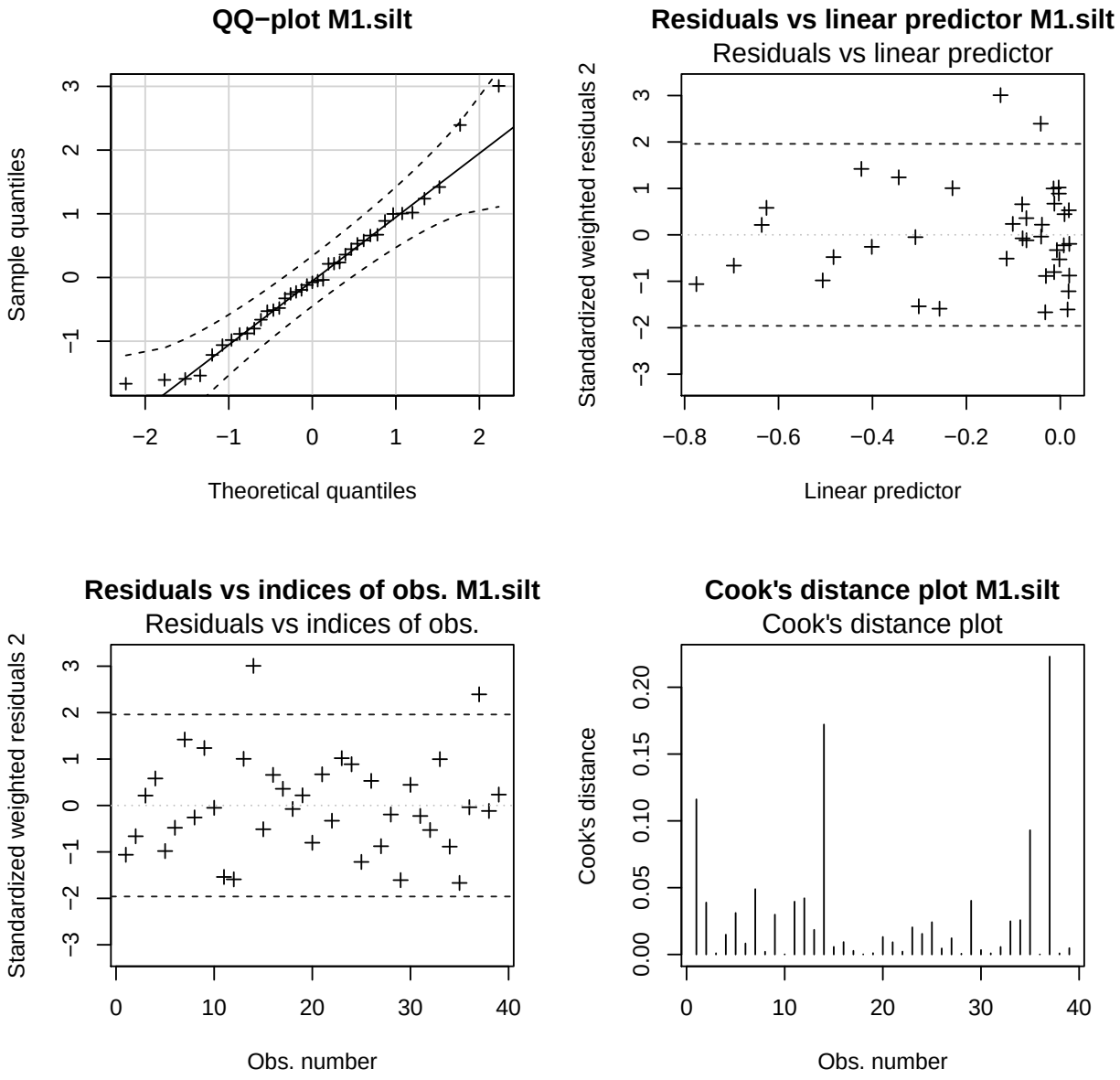

Figure 9: Shows various diagnostics for the M1.silt model. The QQ-plot shows the residuals can be sensibly assumed to come from a standard normal distribution, not counting the outliers. The residuals vs linear predictor plot shows a fairly evenly random scattering of the residuals implying that the fitted values are good, but there is some evidence of heteroskedasticity, and there is also the two large positive outliers which need to be addressed. The residuals vs observed values plot shows that the outlying residuals belong to observations 12 and 37. The Cook's distance plot shows observations 12 and 37 are highly influential on the model fit so can not be easily omitted.

1.3 Model for the clay proportion

The EDA showed that the clay proportion data was similar in structure to the silt proportion data, so the form of the intitial model is again identical to the form in §1.1.

```
M1.clay <- betareg(clay ~ depth+log(depth) | depth)
summary(M1.clay)
```

```
##
## Call:
## betareg(formula = clay ~ depth + log(depth) | depth)
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -3.9761 -0.6212  0.2504  0.7972  1.7799
##
## Coefficients (mean model with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.067537   1.390922  -5.081 3.75e-07 ***
## depth       -0.017008   0.007471  -2.276   0.0228 *
## log(depth)   1.883228   0.451870   4.168 3.08e-05 ***
##
## Phi coefficients (precision model with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.204389   0.468276   2.572   0.0101 *
## depth       0.047142   0.008325   5.663 1.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 52.73 on 5 Df
## Pseudo R-squared: 0.663
## Number of iterations: 19 (BFGS) + 1 (Fisher scoring)
```

The summary shows all the coefficients are significant and the pseudo R-squared is reasonably high. Note the extremely large residual.

Figure 10 shows the diagnostics for the `M1.clay` model. The residuals are clearly heteroskedastic, so extra complexity should be added to the ϕ model to attempt to aleviate this issue. The extreme residual corresponding observation 12 does not appear as influential in the Cook's distance plot. The reason is because that extreme residual corresponds to a proportion which is close to zero in value. Visually speaking the datapoint has no space to move to change the model fit (seen in **Figure 11**). However because the value for the proportion of this observation is close to zero, the residual distance is amplified by the way the standardised residuals are weighted, which produces the extreme value. Therefore it is safe to ignore this outlier. Observation 37 is left alone for the same reasons discussed in §1.2, in that it would likely cause the model to underestimate the variation of the observation.

Again, attempts to improve the model were explored thoroughly to solve the heteroskedasticity issue.

```
M2.clay <- betareg(clay ~ depth+log(depth) | depth+log(depth), link="probit", link.phi="log")
summary(M2.clay)
```

```
##
## Call:
## betareg(formula = clay ~ depth + log(depth) | depth + log(depth),
##     link = "probit", link.phi = "log")
##
```

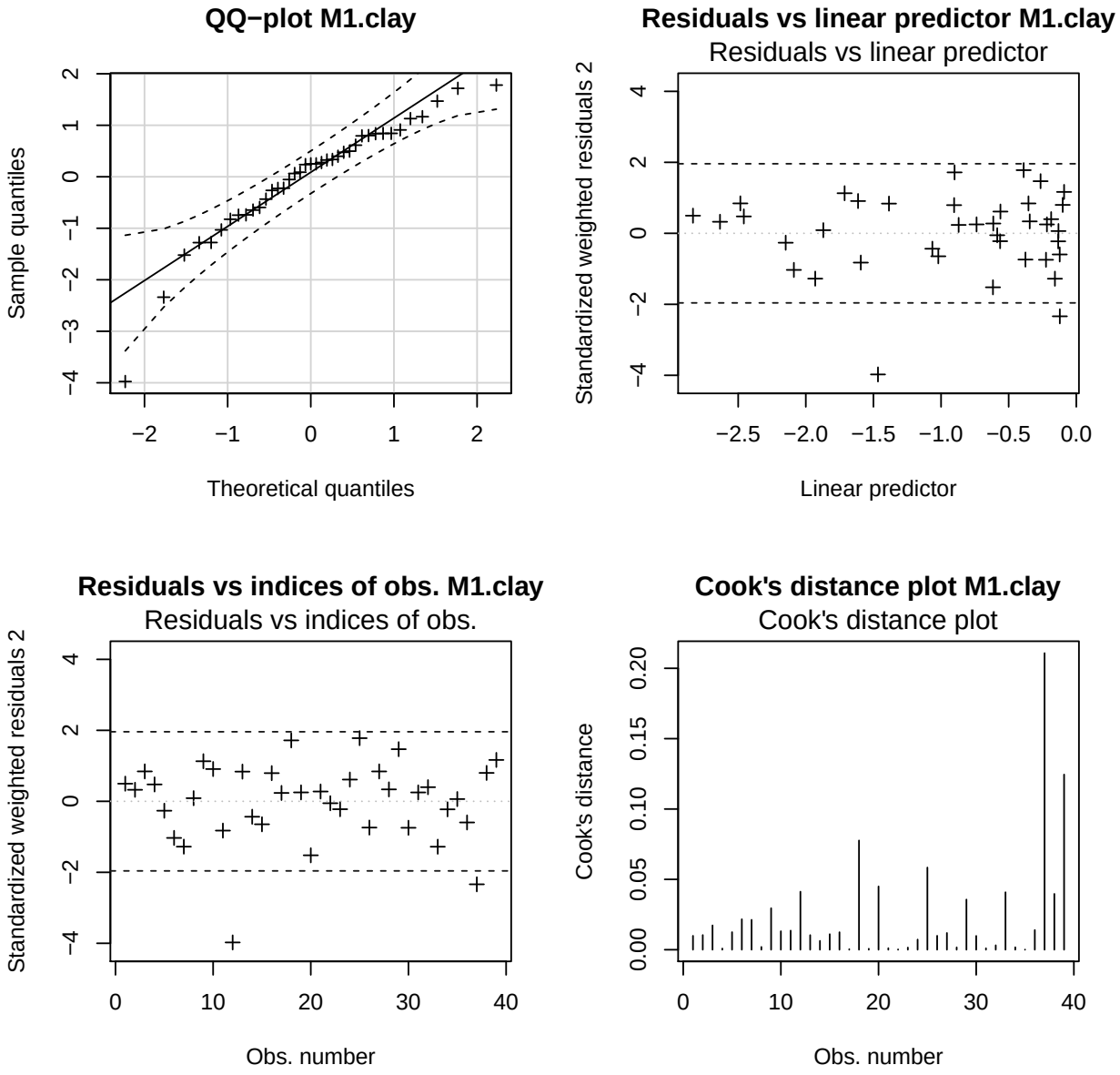

Figure 10: Shows various diagnostics for the M1.clay model. The QQ-plot shows evidence that the residuals may be reasonably approximated by a standard normal distribution, barring the extreme outlier. The residuals vs linear predictor plot shows clear heteroskedasticity which needs to be addressed. The only influential outlier is observation 37 which the Cook's distance plot shows is very influential on the model fit so cannot easily be omitted. The extreme outlier does not show up as influential so can be ignored without issue.

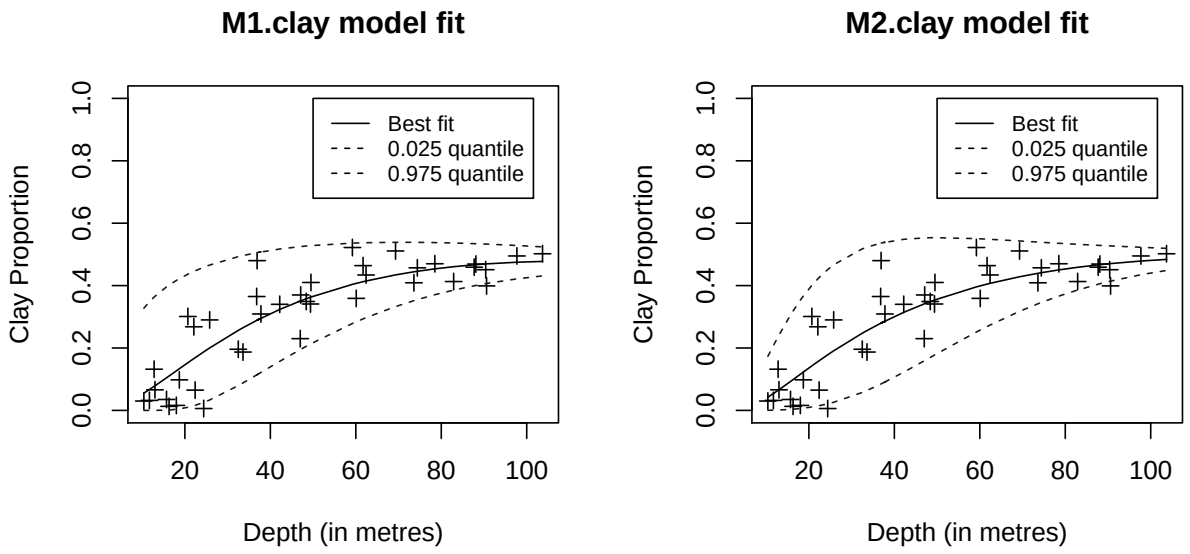

Figure 11: shows the best fit line and the 0.975 and 0.025 quantile boundaries for the M1.clay model and M2.clay model. The M2.model looks like it fits the variation in the data better than the M1 model.

```
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -3.6984 -0.4852  0.2954  0.7806  1.6224
##
## Coefficients (mean model with probit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.162026   0.680630  -6.115 9.66e-10 ***
## depth       -0.008574   0.003993  -2.147   0.0318 *
## log(depth)   1.079532   0.227246   4.751 2.03e-06 ***
##
## Phi coefficients (precision model with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.86577    3.14649   2.182  0.02911 *
## depth        0.09345    0.02896   3.227  0.00125 **
## log(depth)  -2.13714    1.20668  -1.771  0.07655 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 53.37 on 6 Df
## Pseudo R-squared: 0.7086
## Number of iterations: 27 (BFGS) + 14 (Fisher scoring)
```

The summary shows that the log coefficient in the dispersion model is borderline significant, but since no other improvement could be found, this model was compared with the `M1.clay` model.

Figure 11 shows the fits of the two models `M1.clay` and `M2.clay`. Visually speaking the `M2.clay` model looks to fit the variation in the data much better than the `M1.clay` model, especially for the most shallow depths.

```
AICc(M1.clay,M2.clay)
```

```
##          df      AICc
## M1.clay  5 -93.63481
## M2.clay  6 -92.11487
```

The AICc shows that the `M1.clay` is slightly better than the `M2.clay model`. However it is believed that this is not actually true, and that observation 37 is causing the issue since it is highly influential. The models were refit without this observation to test this hypothesis.

```
M1.clay <- update(M1.clay, subset=-37)
M2.clay <- update(M2.clay, subset=-37)
AICc(M1.clay, M2.clay)
```

```
##          df      AICc
## M1.clay  5 -95.49853
## M2.clay  6 -97.59503
```

We see that the `M2.clay` model is now fitting the data better. Given this the decision was made to accept the `M2.clay` model as the better model.

```
M2.clay <- betareg(clay ~ depth+log(depth) | depth+log(depth), link="logit", link.phi="log")
```

Figure 12 shows the diagnostics for the `M2.clay` model. The heteroskedasticity issue looks to have cleared up. Also note that the outlying residuals that are picked up by the standardised residuals are not apparent outliers when using the Pearson residuals which means the outliers are likely not indicating a problem with the model fit. We therefore choose `M2.clay` as the best model for the clay proportion.

```
clay.eta <- M2.clay
```

Figure 11 shows the model fit. The fit looks quite good overall and there are no significant issues which need to be addressed.

2 ζ Model Analysis

```
sand.zeta <- 1- fitted(silt.eta)-fitted(clay.eta)
silt.zeta <- 1- fitted(sand.eta)-fitted(clay.eta)
clay.zeta <- 1- fitted(silt.eta)-fitted(sand.eta)
```

Figure 13 shows the best fit lines of the best η model and the corresponding ζ model. By eye, the sand ζ model appears to be the closest to the η model. The clay model looks invalid for depths up to around 10 metres as the best fit line looks to imply negative values so we will eliminate this model as a potential candidate. The remaining two ζ models are compared numerically by calculating R to confirm which model is the best.

```
yhat <- fitted(silt.eta)
zhat <- silt.zeta
p <- predict(silt.eta, type = "quantile", at = c(0.025))
q <- predict(silt.eta, type = "quantile", at = c(0.975))
h <- q - p

s1 <- rep(0,39)
s2 <- rep(0,39)

for (i in 1:39){
```

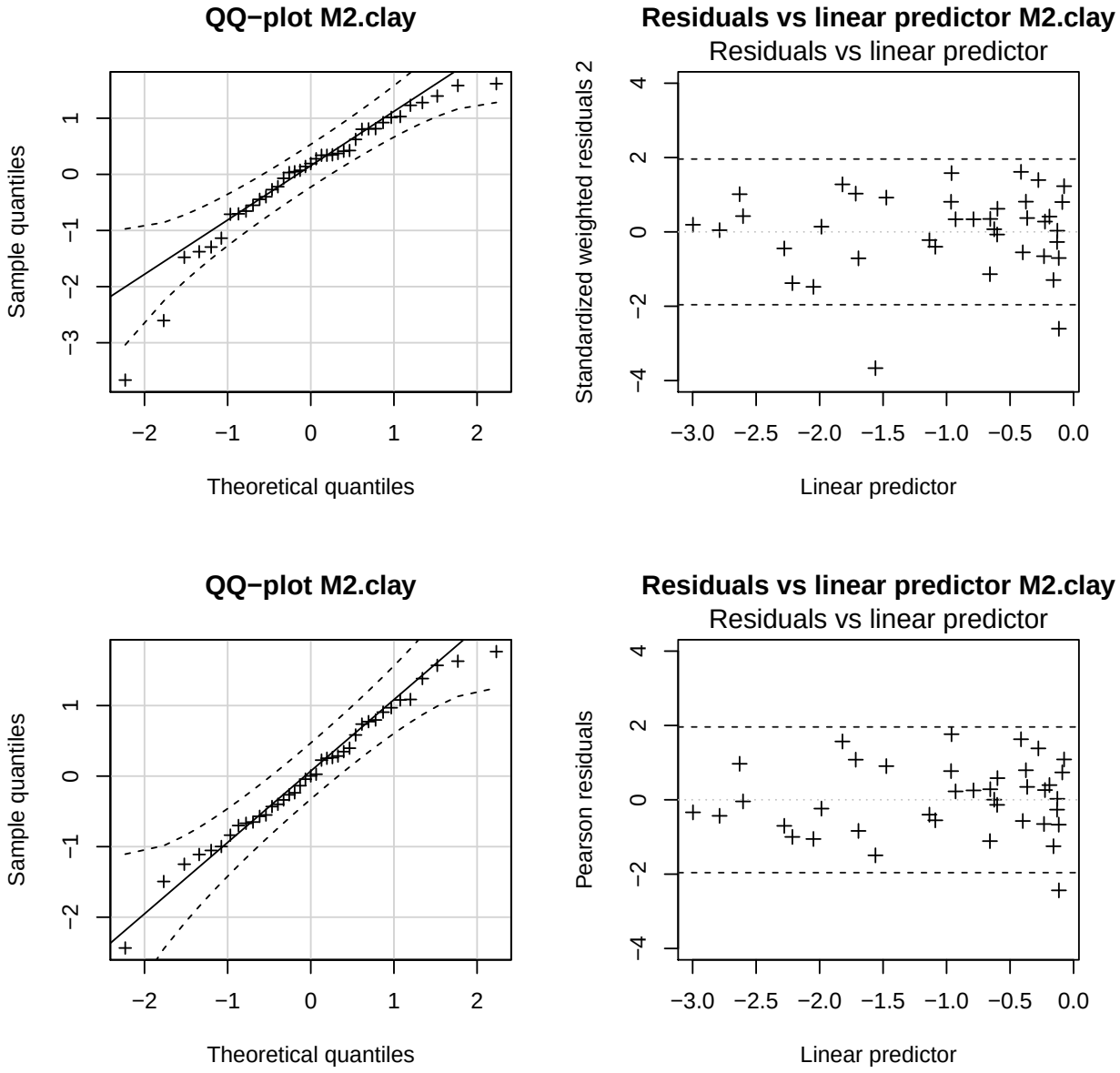

Figure 12: Shows various diagnostics for the M2.clay model. There are two QQ plots, the top left one is made using the standardised residuals, and the bottom left one is made using the Pearson residuals. We see that the outlying observations are likely problems caused by the type of residuals we are looks at rather than a problem with the data or the model fit itself. Likewise, the residuals vs linear predictor plots are made in the same fashion and the Pearson residual plot in the bottom right shows that the outliers are not a big problem, and the standardised residuals are over penalizing them. Also note that the heteroskedasticity situation has improved and is no longer a concern.

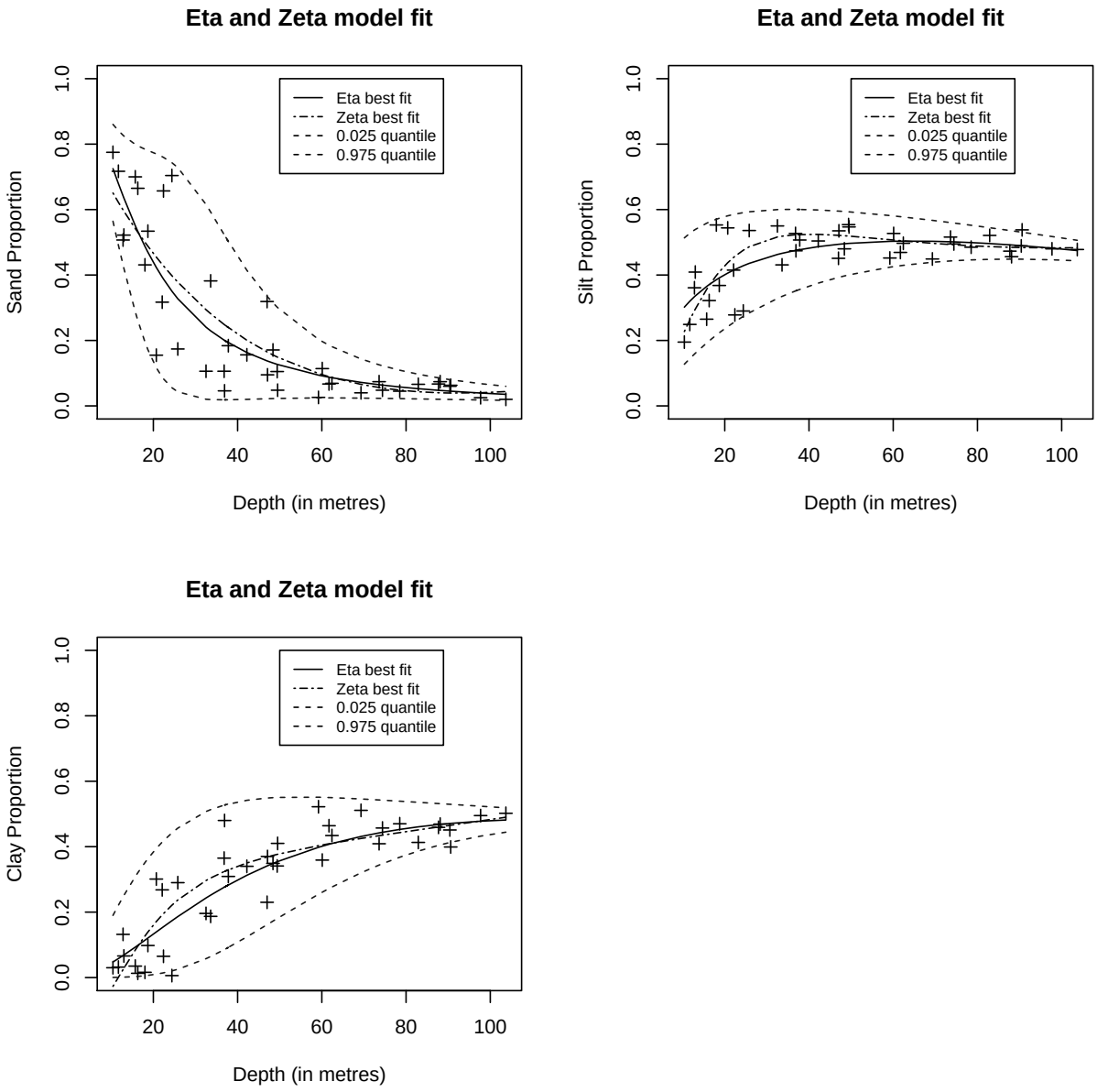

Figure 13: shows the best fit line for the eta model (solid black line) and the corresponding zeta model (dashed blue line), and also shows the fitted model 0.975 and 0.025 quantile boundaries . Visually speaking, the sand zeta model appears to be the closest to the fitted model. The clay zeta model looks to be invalid for depths above 10 metres which is a problem.

```
  if ((yhat[i]-zhat[i])>=0){
    s1[i] <- 1*(yhat[i] - zhat[i])^2/(yhat[i]-p[i])^2
  }

  if ((yhat[i]-zhat[i])<0){
    s2[i] <- 1*(yhat[i] - zhat[i])^2/(q[i]-yhat[i])^2
  }
}

R.silt <- sum(s1)+sum(2)


yhat <- fitted(sand.eta)
zhat <- sand.zeta
p <- predict(sand.eta, type = "quantile", at = c(0.025))
q <- predict(sand.eta, type = "quantile", at = c(0.975))
h <- q - p

s1 <- rep(0,39)
s2 <- rep(0,39)

for (i in 1:39){

  if ((yhat[i]-zhat[i])>=0){
    s1[i] <- 1*(yhat[i] - zhat[i])^2/(yhat[i]-p[i])^2
  }

  if ((yhat[i]-zhat[i])<0){
    s2[i] <- 1*(yhat[i] - zhat[i])^2/(q[i]-yhat[i])^2
  }
}

R.sand <- sum(s1)+sum(2)



sprintf("R for the sand model", R.sand)
```

```
## [1] "R for the sand model"
```

```
R.sand
```

```
## [1] 2.900893
```

```
sprintf("R for the silt model", R.silt)
```

```
## [1] "R for the silt model"
```

```
R.silt
```

```
## [1] 2.624659
```

The output shows surprisingly that the silt model is the closest match to its corresponding η model, although the models are close so further investigations are wise. Now to check the diagnostics to make sure there are no problems with either of the models. The Pearson residuals will be compared as to avoid over-penalized data points.

```
## [1] 35 39
```

```
## [1] 37 12
```

Figure 14 shows the diagnostics for the `sand.zeta` model and the `silt.zeta` model. The Q-Q plot for the sand model does not look fantastic, however the points do remain inside the boundaries so it could be assumed that the distribution of the residuals very roughly approximate a standard nromal distribution. If we were to assume this then the residuals vs linear predictor plot shows that the residuals are normally distributed around the horizontal axis indicating that the model is a good fit to the data. The Q-Q plot for the silt model shows the residuals can be sensibly assumed to come from a standard nromal distribution, however the residuals vs linear predictor plot shows that the residuals exhibit heteroskedasticity which is a problem.

```
par(mfrow=c(1,2))

p = predict(sand.eta,type="quantile",at=c(0.025))
q = predict(sand.eta,type="quantile",at=c(0.975))

plot(seq(1,39,1),rep(0,39) , pch=".", ylim=c(-0.1,1.1), xlab="Observation Number",
     ylab="Interval Height", main="Scaled Sand.zeta model fit")
abline(h=0)
lines(seq(1,39,1),(q-p)/(q-p))
lines(seq(1,39,1),(fitted(sand.eta)-p)/(q-p))
lines(seq(1,39,1),(sand.zeta-p)/(q-p), col="grey10", lty=2)
points(seq(1,39,1),(sand-p)/(q-p), pch=3)

p = predict(silt.eta,type="quantile",at=c(0.025))
q = predict(silt.eta,type="quantile",at=c(0.975))

plot(seq(1,39,1),rep(0,39), pch="." , ylim=c(-0.1,1.1), xlab="Observation Number",
     ylab="Interval Height", main="Scaled Silt.zeta model fit")
abline(h=0)
lines(seq(1,39,1),(q-p)/(q-p))
lines(seq(1,39,1),(fitted(silt.eta)-p)/(q-p))
lines(seq(1,39,1),(silt.zeta-p)/(q-p), col="grey10", lty=2)
points(seq(1,39,1),(silt-p)/(q-p), pch=3)
```

Figure 15 shows the best fit lines for the model on the data, where the data and model fit has been stretched such that the gap between the quantile boundaries is constant. This better illustrates how the ζ models differ from the η models. Visually speaking, the silt ζ model looks to be more erratic on the plot that the sand ζ model. From the evidence gathered, overall the `sand.zeta` model appears to be the best model to choose. The only problem with the fit is the assumption that the Pearson residuals may not be right.

Figure 16 shows the best fit lines of the final models and the confidence intervals for the η models.

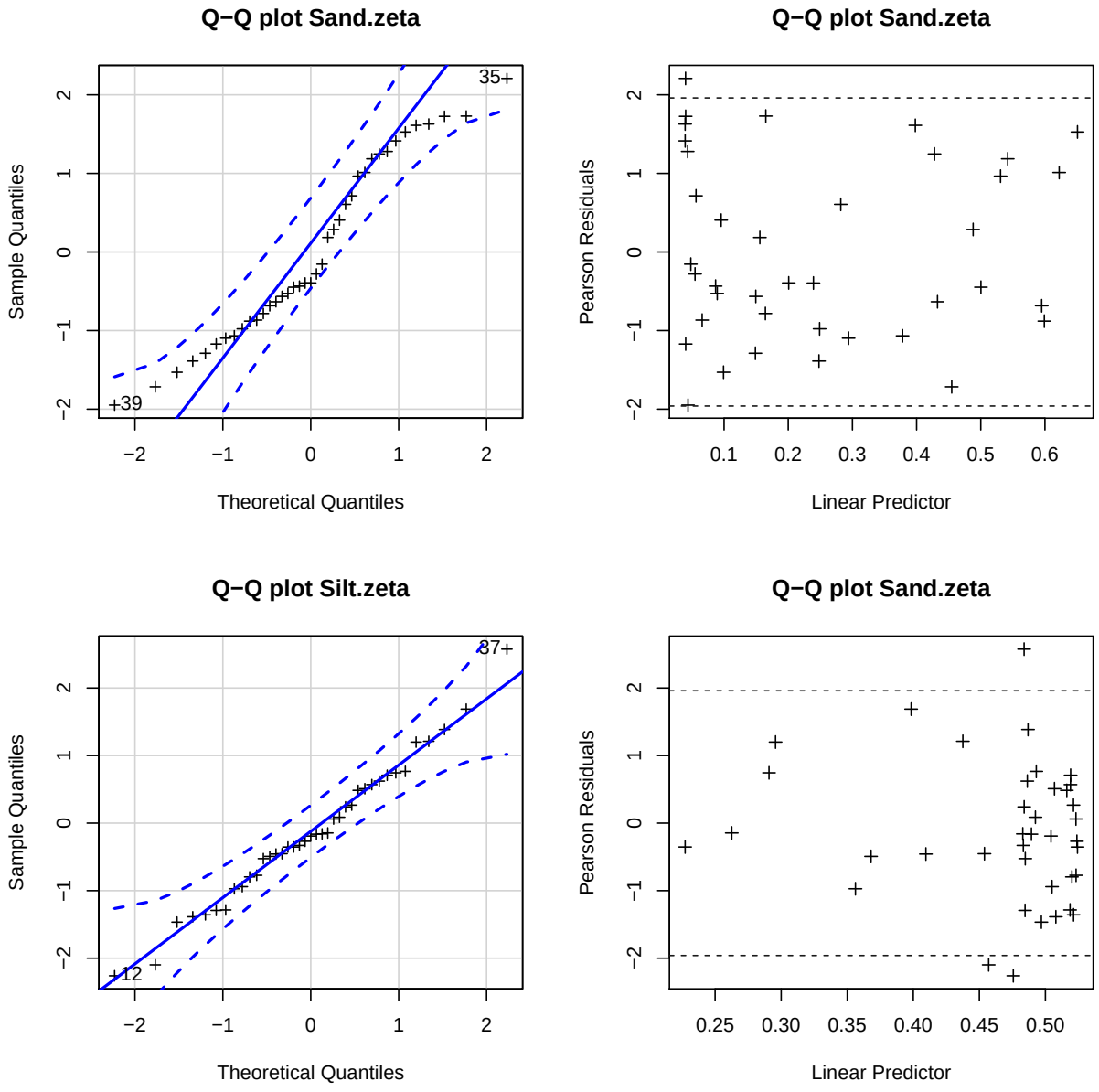

Figure 14: shows the diagnostics for the selected zeta models. The QQ-plot looks good and the residuals also looks good. Therefore we take sand.a as a valid model.

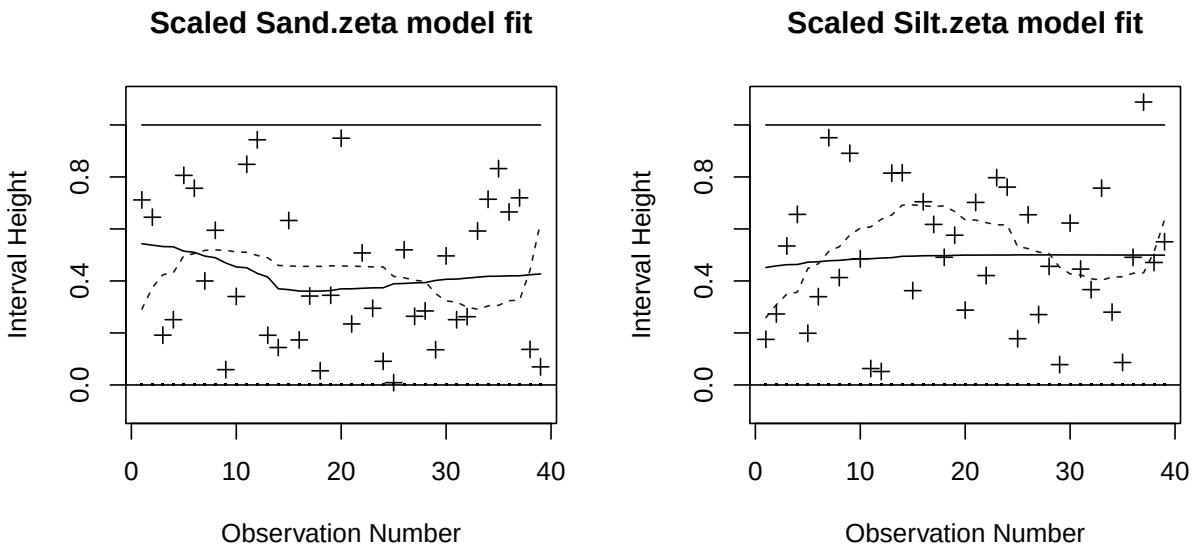

Figure 15: Shows the fits for the zeta models where the data and models have been scaled by the magnitude of the quantile interval such that the gap between the quantile boundaries is constant. The top and bottom horizontal black lines correspond to the 0.975 and 0.025 quantile boundaries respectively. The black curves correspond to the best fit lines of the eta models, and the dashed black lines correnspond to the best fit line of the zeta models. The crosses are the data. We see that out of the two models, the zeta model for sand looks slightly better than the model for slit.

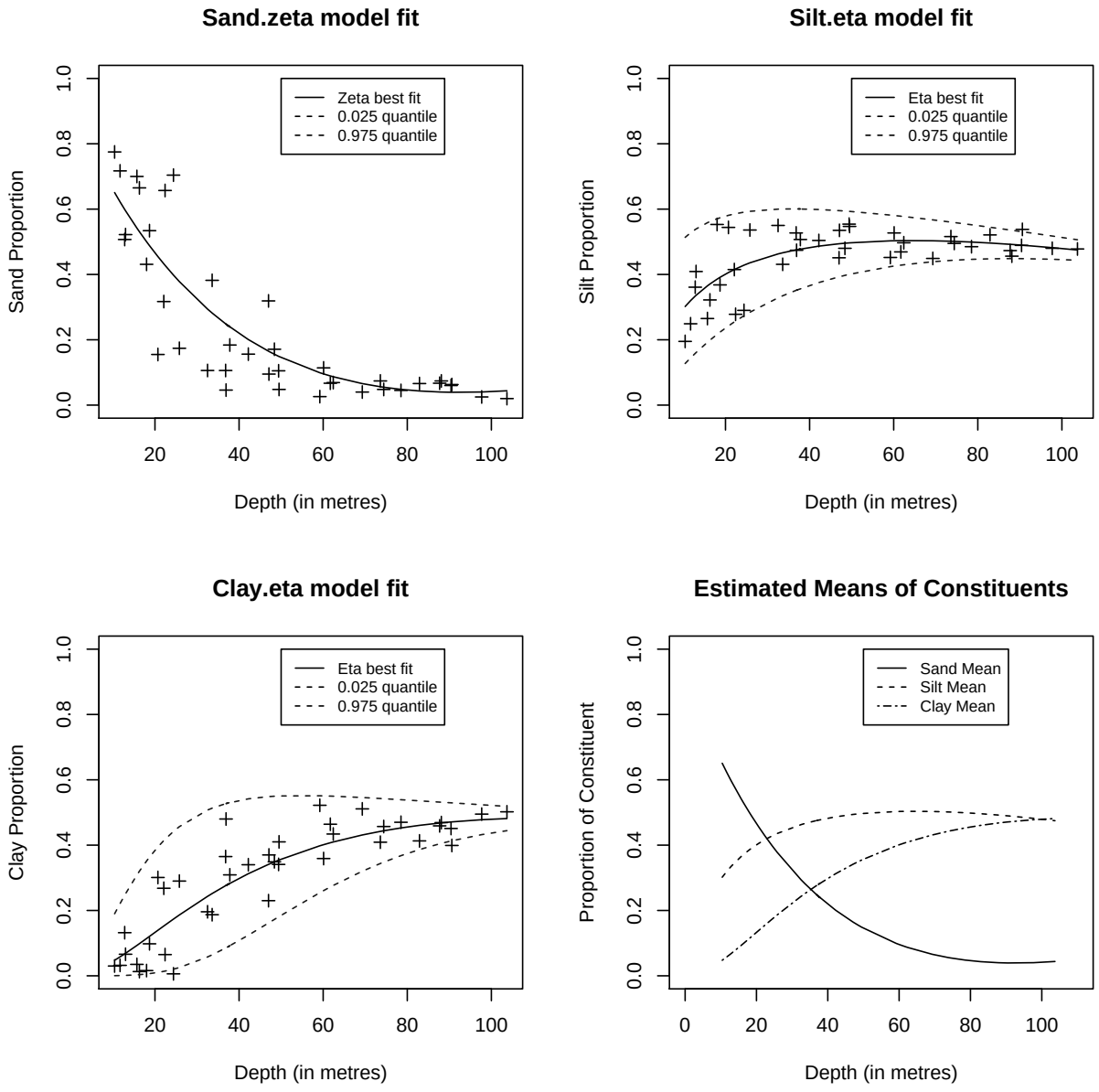

Figure 16: shows the fit of the chosen models. The bottom right plot shows the best fit lines only of the chosen models to see them more clearly.