MTH2006 Summative Coursework for Year 2, Term 2

Section 1

Exploring the Relationship between HRR_delta180 and MAS

There is a positive relationship between HRR_delta180 and MAS as show in $Figure\ 1$ below. The upward trend suggests that a greater difference between peak heart rate during exercise and heart rate measured 180 seconds after finishing (ie. larger HRR_delta180) correlates with a greater cardiorespiratory fitness (ie. higher MAS measured).

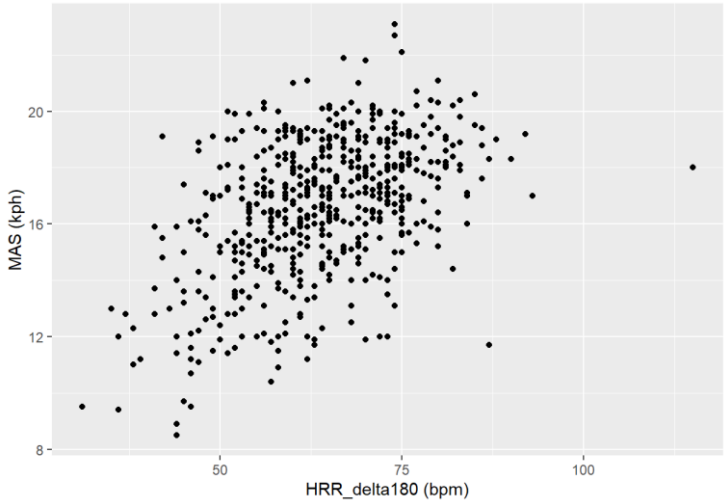

Figure 1: Scatter plot showing of MAS results against the calculated HRR_delta180

The data is concentrated within 50-80 bpm for HRR_delta180 and within 12-20 kph for MAS. One potential outlier would be the HRR_delta180 datapoint ≈ 115 bpm, while a linear model may work well, this particular outlier could skew the model and under-predict the slope. As shown in $Figure\ 2$, this outlier lies far from the otherwise relatively bell-shaped distribution of the HRR_delta180 data. In contrast, in $Figure\ 3$ the density plot for MAS shows overwhelming majority of the data lies below 20 kph and a negative skew (longer left-hand tail) indicating the data is spread more at lower values than higher values.

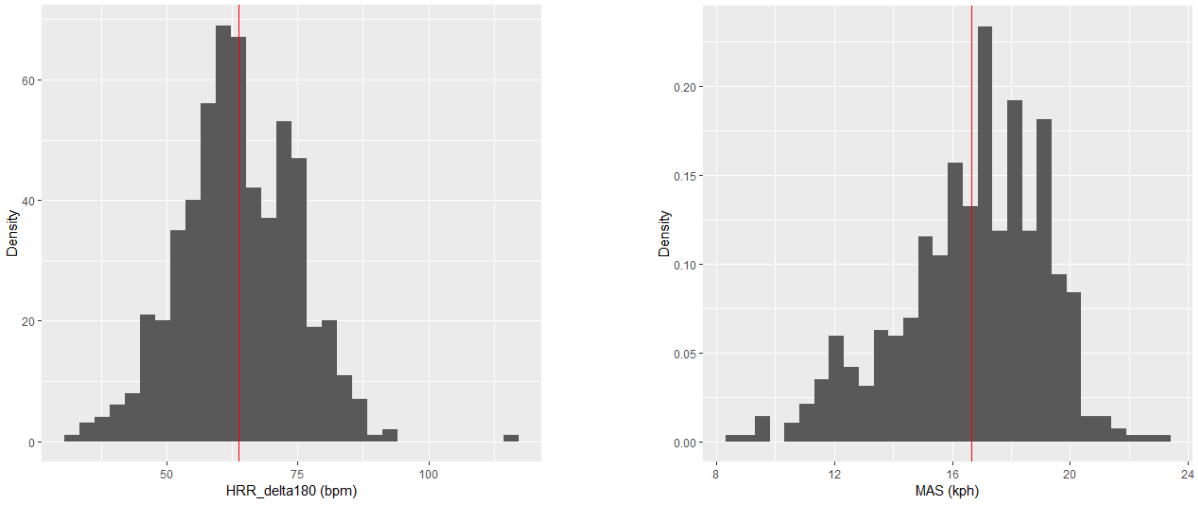

Figure 2 + 3: Histograms showing the distribution for HRR_delta180 and MAS data respectively

Applying Linear Regression to Examine this Relationship

Starting with a Simple Linear Model of the form:

$$MAS \sim \beta_0 + \beta_1(HRR\_delta180)$$

Coefficients of this Linear Model (to 4sf.):

	Estimate	*Std. Error*	*t-value*	*p-value*
Intercept	9.725	0.5394	18.03	< 2e-16
HRR_delta180	0.1085	0.008335	13.02	< 2e-16

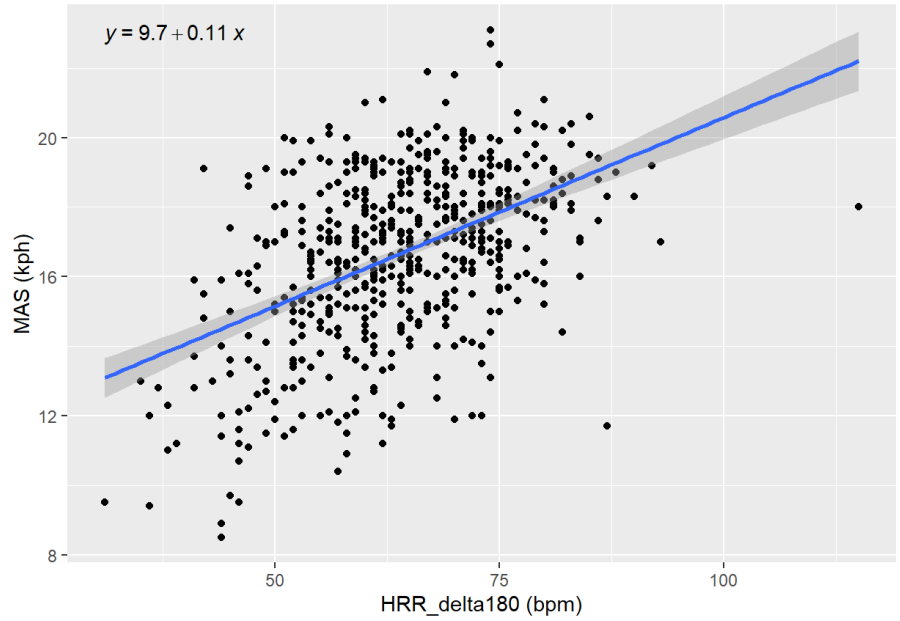

Figure 4: Simple Linear Regression Model with a 95% confidence band and eq. labelled

While the simple model does show a significant relationship between HRR_delta180 and MAS (p-value < 0.05) we can extend this model to account for potential confounding factors, namely: BMI, Age and Sex.

The following model can be constructed:

$$MAS \sim \beta_0 + \beta_1 BMI + \beta_2 Age + \beta_3 Sex + \beta_4(HRR\_delta180)$$

Coefficients of this Linear Model (to 4sf.):

	Estimate	*Std. Error*	*t-value*	*p-value*
Intercept	18.81	0.8639	21.77	< 2e-16
BMI	-0.2333	0.02672	-8.729	< 2e-16
Age	-0.02233	0.007984	-2.797	0.00533
Sex	-3.426	0.2071	-16.54	< 2e-16
HRR_delta180	0.07303	0.007071	10.33	< 2e-16

Conducting a Hypothesis Test at the 5% sig. level to test the significance of the relationship between MAS and HRR_delta180 under this model will have the following hypotheses:

$$H_0: \beta_4 = 0$$

$$H_1: \beta_4 \neq 0$$

Given that:

$$p - value < 2 * 10^{-16}$$

We have sufficient evidence to reject the null hypothesis as the p-value<0.05, therefore there is a meaningful relationship between HRR_delta180 and the MAS results, and HRR_delta180 does play a significant role in the linear model.

Similarly, the overall model can be considered a relatively good fit as the linear regression returns:

$$R^2 = 0.5025, \qquad {R^2}_a = 0.4989$$

supporting the positive correlation identified previously, improved from the results of the Simple Linear Model (without covariates):

$$R^2 = 0.2298, \qquad {R^2}_a = 0.2284$$

indicating a stronger correlation and a better fitting model when including BMI, Age and Sex as covariates.

Extending the Linear Model to include Interaction Effects

Accounting for Sex differences using interaction terms can be done with the following Linear Model form:

$$MAS \sim \beta_0 + \beta_1 BMI + \beta_2 Age + \beta_4 (HRR\_delta180) + \beta_3 Sex + \beta_5 (BMI)(Sex) + \beta_6 (Age)(Sex) + \beta_7 (HRR\_delta180)(Sex)$$

Coefficients of this Linear Model (to 4sf.):

	Estimate	*Std. Error*	*t-value*	*p-value*
Intercept	19.53	0.9322	20.95	< 2e-16
BMI	-0.2574	0.02935	-8.772	< 2e-16
Age	-0.01902	0.00866	-2.196	0.02851
HRR_delta180	0.06961	0.007863	8.852	< 2e-16
Sex	-7.798	2.401	-3.247	0.00123
BMI:Sex	0.1448	0.07475	1.937	0.05324
Age:Sex	-0.0113	0.02436	-0.464	0.6428
HRR_delta180:Sex	0.02253	0.01801	1.251	0.2116

Conducting a Hypothesis Test at the 5% sig. level to test whether there is evidence of interaction effects in the linear model between Sex and other variables will have the following hypotheses:

$$H_0: \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_1: at\ least\ one\ of\ \beta_5, \beta_6, \beta_7 \neq 0$$

Given that:

$$0.05324, 0.6428\ and\ 0.2116\ (p-values) > 0.05\ (sig.lvl)$$

There is not enough evidence to reject the null hypotheses as none of the p-values are critically significant, therefore the model does not give evidence for significant interactions between Sex and the other variables in present.

Assessing the Linear Model via Residuals

The assumptions about the true residuals of a linear model are as follows:

1) They are normally distributed
2) They have 0 expectation
3) They have constant variance
4) They are all independent from each other

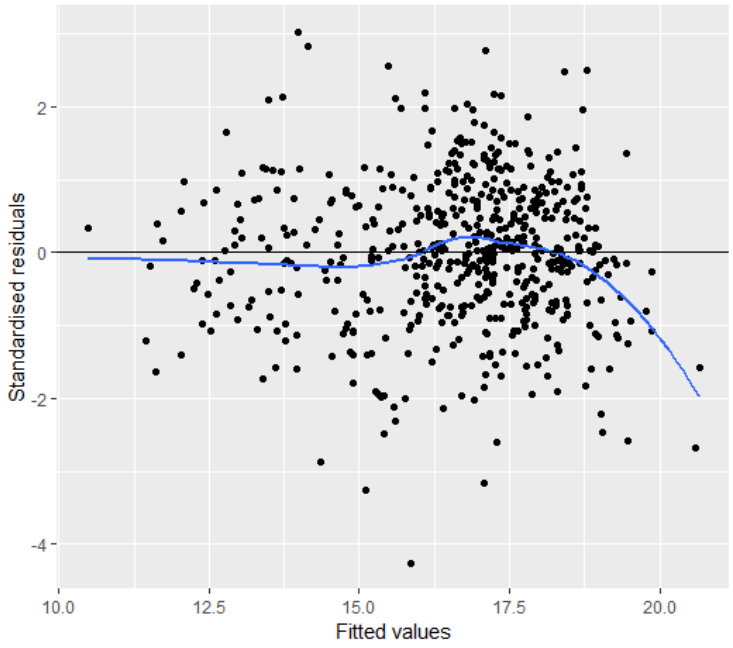

Figure 5: scatter plot demonstrating standardised residuals of the linear model vs. the fitted values

At the higher end, the residuals are lower than they should be as shown in $Figure\ 5$ – ie. the smoothing curve is far below the assumed mean of 0. The residuals tend to a mean of 0 up until MAS of ≈17.5 (kph) after which the mean deviates strongly indicating potential outlier(s) at the upper end of the HRR_delta180 data, in other words they have an unusually low MAS measurement (than expected by the linear model).

Overall, this shows the mean of the residuals differs from 0 depending on the fitted values and is not constant, potentially the assumption that mean ≈ 0 would be true if the possible outlier(s) are removed. There appears to be relatively constant variation in the residuals, but there is a slight jump from a spread of ≈ 4 std. (before ≈13 kph) to ≈7 std. (after ≈13 kph) which may be indicative of a non-constant variance in the residuals, another potentially poor assumption if shown to be significant.

On the other hand, the Q-Q plot in $Figure\ 6$ demonstrates the distribution of the standardised residuals of this linear model can accurately be assumed to follow a normal distribution as the data follows it very closely. Despite the slightly heavier left-hand tail, the normal distribution is still a good fit for overall, as shown in $Figure\ 7$. We can also evaluate overly influential observations which are negatively impacting the reliability of our model. This can be done by comparing each residual to its leverage (distance from the mean of the explanatory variable – indicating its capacity to affect the model) as shown in $Figure\ 8$ below.

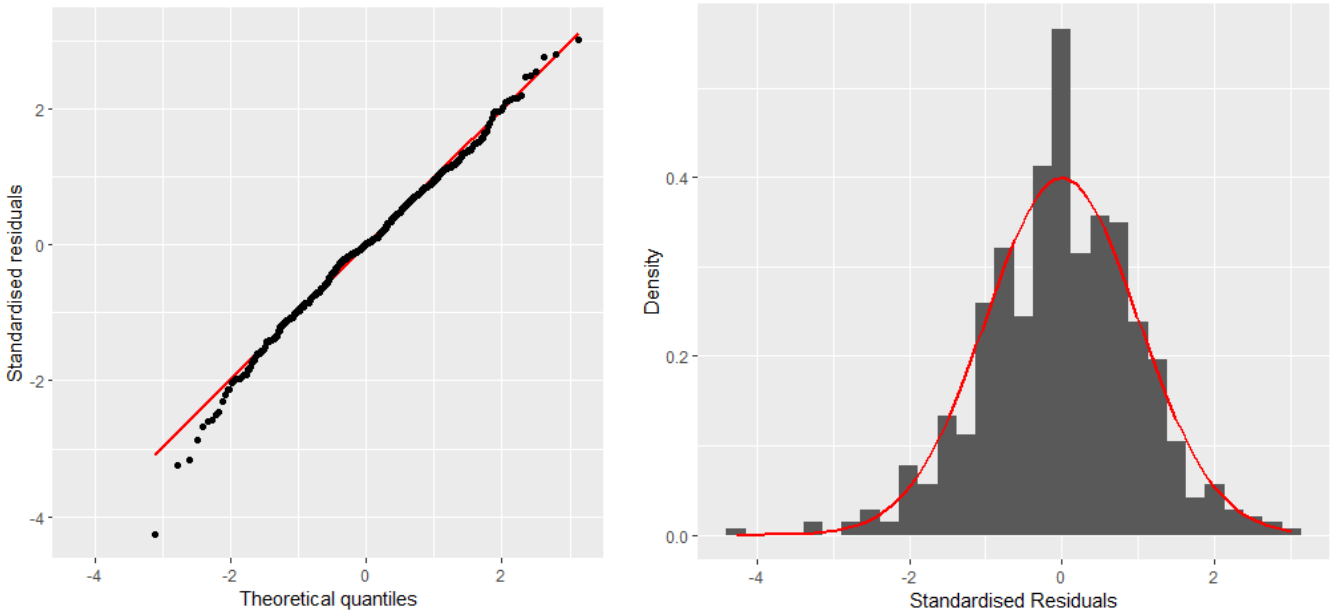

Figure 6 + 7: q-q and histogram plots to compare the std. residuals distribution to the expected $N(0,1)$ shown in red

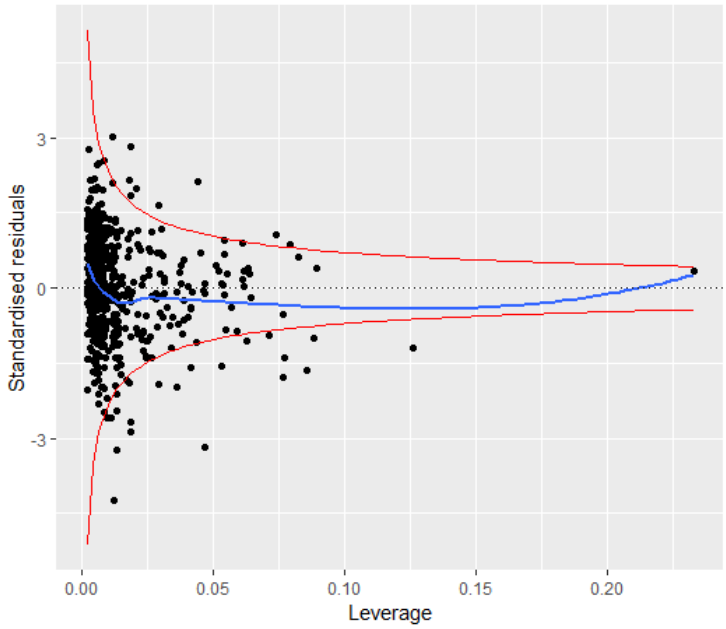

Figure 8: Residuals vs. Leverage plot showcasing the overly influential points based on Cooke's Distance, $level\ = \frac{4}{n}$

Datapoints with greater leverage have less freedom for variation that is causing a large residual and influencing the linear model coefficients. This is shown by the inner "red zone" between the red bounds in $Figure\ 8$ above. These red bounds show the limit for a residual as leverage increases to not be considered an overly influential datapoint using Cooke's Distance with

$level = \frac{4}{n}$. In this case, there does not seem to be any standout datapoints affecting the Linear Model negatively, one datapoint of interest is found at $leverage \approx 0.26$ which has the greatest potential to disrupt the model and has a residual on the bounds of being overly-influential based on Cooke's Distance, though this may not be the case with other measures of influence.

Conducting Multiple Testing for each HRR Index

Each model will have the form:

$$MAS \sim \beta_0 + \beta_2 Age + \beta_1 BMI + \beta_4 HRR$$

Repeated 4 times for each HRR measure: HRR_delta10, HRR_delta180, HRR10per and HRR180per.

Where the hypotheses conducted is:

$$H_0: \beta_4 = 0$$

$$H_1: \beta_4 \neq 0$$

First Method: Bonferroni correction can be used to preserve the 5% significance level by adjusting based on the number of tests conducted.

$$new\ sig.level = \frac{0.05}{4} = 0.0125$$

The coefficients and results for the HRR term from each corresponding model is as follows:

	Estimate	*Std. Error*	*t-value*	*p-value*
HRR_delta10	0.09834	0.03993	2.463	0.0157
HRR_delta180	0.09213	0.01656	5.563	2.70e-07
HRR10per	-0.1851	0.07344	-2.521	0.0135
HRR180per	-0.1621	0.02946	-5.503	3.48e-07

At the adjusted 1.25% sig. level, only the p-values for HRR_delta180 and HRR180per are significant:

$$2.7 * 10^{-7}, 3.48 * 10^{-7} < 0.0125 < 0.0135, 0.0157$$

Therefore, only these 2 corresponding models demonstrate a significant relationship between HRR_delta180 and MAS, and HRR180per and MAS respectively.

Method 2: Holm Method can be used to adjust each p-value directly and maintain the same 5% sig. level in the final evaluation by first re-ordering in increasing order of p-values and then applying the adjustment:

	Estimate	*p-value*	*Comparison Value*	*new p-value*
HRR_delta180	0.09213	2.70e-07	(4-1+1)*2.70e-07	1.08e-06
HRR180per	-0.1621	3.48e-07	(4-2+1)*3.48e-07	1.044e-06
HRR10per	-0.1851	0.0135	(4-3+1)*0.0135	0.027
HRR_delta10	0.09834	0.0157	(4-4+1)*0.0157	0.0157

The adjustment to each p-value is as follows:

$$new\ \ p-value = \min\{comparison\ value, 1\}$$

Now considering the new p-values at the 5% sig. level, all of the results are critically significant such that:

$$1.08 * 10^{-6}, 1.044 * 10^{-6}, 0.027, 0.0157 < 0.05$$

Therefore using Holm Method provides the contrasting result that all 4 models demonstrate a significant relationship between their corresponding HRR measure and MAS.

Section 2:

Finding the Log-Likelihood function and Estimates via Numerical Optimisation

Given the Beta Distribution:

$$f_Y(y;\alpha,\beta) = \frac{y^{\alpha-1}(1\ -y)^{\beta-1}}{B(\alpha,\beta)}\ ,\ \ 0\ \leq y\ \leq 1.$$

The Log-likelihood function can be derived as follows for some given sample y:

$$L(\alpha,\beta;y) = \prod_{i=1}^{n}\left(\frac{y^{\alpha-1}(1\ -y)^{\beta-1}}{B(\alpha,\beta)}\right)$$

$$= \prod_{i=1}^{n}(y^{\alpha-1}(1\ -y)^{\beta-1})\ \prod_{i=1}^{n}\left(\frac{1}{B(\alpha,\beta)}\right)$$

$$= \prod_{i=1}^{n} e^{(\alpha-1)\ln(y_i)}\ \prod_{i=1}^{n} e^{(\beta-1)\ln(1-y_i)} * \ B(\alpha,\beta)^{-n}$$

$$= \ \exp\left[\sum_{i=1}^{n}(\alpha-1)\ln\left(y_i\right)\ +\ \sum_{i=1}^{n}(\beta-1)\ln\left(1-y_i\right)\right] * \ B(\alpha,\beta)^{-n}$$

$$LogL(\alpha,\beta;y) = \sum_{i=1}^{n}((\alpha-1)\ln(y_i) + (\beta-1)\ln\left(1-y_i\right))\ -nlog\big(B(\alpha,\beta)\big)$$

$$= \sum_{i=1}^{n}((\alpha-1)\ln(y_i) + (\beta-1)\ln\left(1-y_i\right))\ -nlog\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right)$$

$$= \sum_{i=1}^{n}((\alpha-1)\ln(y_i) + (\beta-1)\ln\left(1-y_i\right))\ -nlog\big(\Gamma(\alpha)\big) - nlog\big(\Gamma(\beta)\big) + nlog\big(\Gamma(\alpha+\beta)\big)$$

The Maximum Likelihood Estimates can be calculated with the following code:

```
loglik = function(p, data) {
  a = p[1]
  b = p[2]
  n = nrow(data)

  #summation term:
  s = sum((a-1)*log(data$danceability) + (b-1)*log(1-data$danceability))

  #remaining terms:
  x = (- n*lgamma(a)-n*lgamma(b)
       + n*lgamma(a+b))

  return(x+s)
}

results = optim(c(10, 10), loglik, control=list(fnscale=-1), data=kendrick)

alpha_hat = results$par[1]
beta_hat = results$par[2]
```

Where:

$$\hat{\alpha} = 6.77\ (4sf.), \quad \hat{\beta} = 4.156\ (4sf.)$$

Assessing Reliability of Beta Model via Q-Q Plot

A Q-Q plot comparing theoretical quantiles to the sample quantiles given by the estimated beta distribution can be done with following code:

```
ggplot(kendrick, mapping = aes(sample=danceability)) +
  geom_abline(color="red", linewidth=1) +
  stat_qq(distribution=qbeta, dparams=list(alpha_hat, beta_hat)) +
  labs(x="Theoretical quantiles", y="Sample Quantiles") +
  xlim(c(0, 1)) +
  ylim(c(0, 1))
```

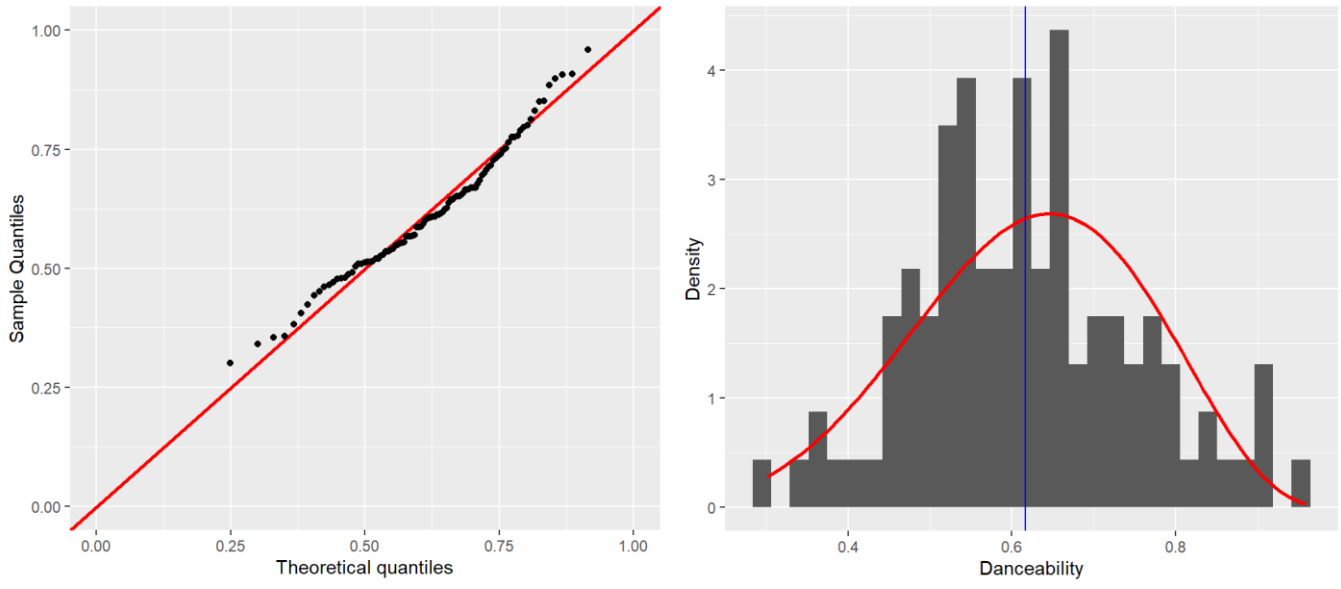

Figure 9 + 10: q-q and histogram plots to compare the danceability data to the estimated beta distribution (in red)

The Beta Distribution is a good fit for Kendrick Lamar's danceability data because the quantiles of the data fit to the estimated model very well as shown in $Figure\ 9$. There is no visually significant skew or mean shift between the Beta Model and sample data.

This is reinforced in $Figure\ 10$ where there is minimal difference between the mean of the danceability data (blue) and the Beta Model mean (red peak) and no obvious skew or outliers.

We can quantify the significance of this fit with the following method.

Using Chi-Squared to Determine Goodness of Fit

We can conduct the following Pearson's Chi-Squared Hypothesis Test:

$$H_0: danceability\ data\ is\ from\ B(\hat{\alpha}, \hat{\beta})$$

$$H_1: danceability\ data\ is\ not\ from\ B(\hat{\alpha}, \hat{\beta})$$

The results of the test (4sf.):

Bins	*(0,0.15]*	*(0.15,0.3]*	*(0.3,0.45]*	*(0.45,0.6]*	*(0.6,0.75]*	*(0.75,1]*
Observed	0	0	8	38	38	17
Expected	0.0225	1.455	11.17	30.64	38.54	19.17

The test statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 4.398\ (4sf.)$$

At the 5% sig. level:

$$DoF = 6 - 2 - 1 = 3$$

$$p - value \approx 0.221 > 0.05$$

Or:

$$critical\ value \approx 7.815 > \chi^2$$

Therefore we don't have enough evidence to reject the null hypothesis, so Kendrick Lamar's danceability data does follow the estimated Beta distribution. Similar to the results of the Q-Q plot, we can state that this Beta Model is a good estimation for the distribution of this data.

Done with the following code:

```
bins = c(0, 0.15, 0.3, 0.45, 0.6, 0.75, 1)
observed = table(cut(kendrick$danceability, bins))
cdf = pbeta(bins, alpha_hat, beta_hat)
expected = nrow(kendrick)*diff(cdf)

test_stat = sum(((observed-expected)^2)/expected)

dof = length(bins)-1 -2 -1
p_value = 1-pchisq(test_stat, dof)
critical = qchisq(0.95, 3)
```

Examining the Dispersion Ordering Between Drake and Kendrick Lamar's Music

The triangular area between 3 songs of an artist can be calculated by applying the following the function to 1000 sets of 3 randomly selected songs per artist:

$$Area = \frac{1}{2} |x11(x22 - x32) + x21(x32 - x12) + x31(x12 - x22)|$$

Where x1, x2 and x3 and the 3 random songs (rows from each artist's data frame) and the secondary index 1 refers to the speechiness score and 2 refers to the danceability score.

It can be applied with the following code:

```
iter = 1000
kendrick_areas = numeric(iter)
drake_areas = numeric(iter)

for(i in 1:iter) {
  kendrick_sample = kendrick[sample(nrow(kendrick), 3), ]
  drake_sample = drake[sample(nrow(drake), 3), ]

  kendrick_areas[i] = area(kendrick_sample[1, ],
                           kendrick_sample[2, ],
                           kendrick_sample[3, ])
  drake_areas[i] = area(drake_sample[1, ],
                        drake_sample[2, ],
                        drake_sample[3, ])
}
```

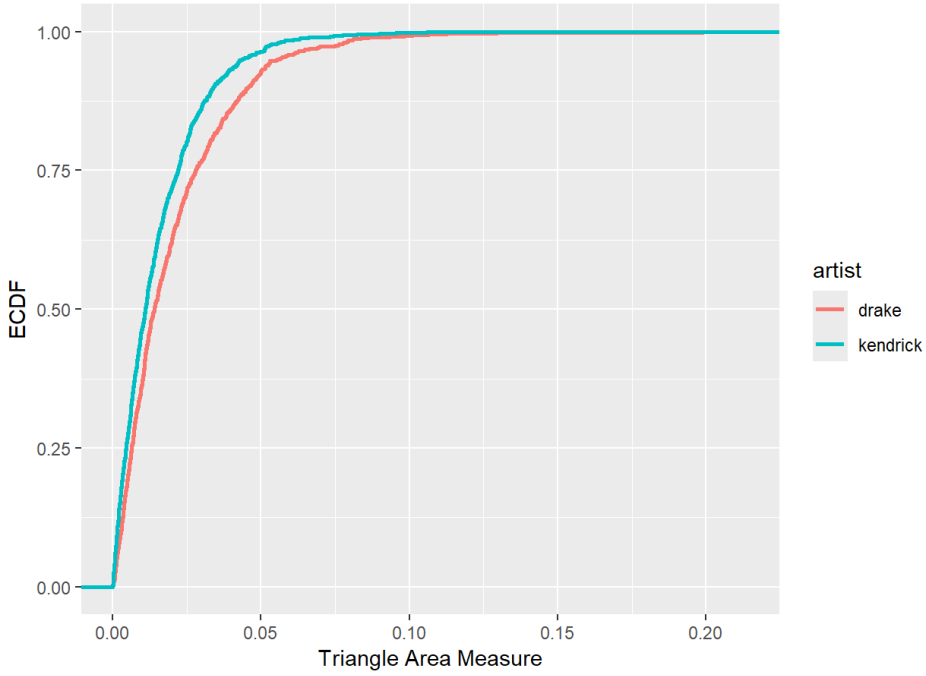

Figure 11: Empirical CDFs for each artists' triangular area measure

The Empirical CDF for the triangle area measure of Kendrick Lamar's music lies consistently above or equal to that of Drake's. This means that the area measures for Kendrick's music reach its maximum "sooner" (ie. at a lower value) than Drake's indicating that the area measures for Drake's music have a greater spread of values and have more datapoints at higher values.

To test whether Drake's area measure is greater than Kendrick's can be done using the following Kolmogorov-Smirnov Test Hypotheses:

$$H_0: \hat{F}_{Drake,1000}(z) = \hat{F}_{Kendrick,1000}(z)$$

$$H_1: \hat{F}_{Drake,1000}(z) > \hat{F}_{Kendrick,1000}(z)$$

With the resulting output:

$$test\ statistic\ D = 0.003, \qquad p - value = \ 0.991$$

From the following code:

```
ks.test(drake_areas, kendrick_areas, alternative="greater")
```

Given that at the 1% sig. level:

$$0.991 > 0.01$$

There is insufficient evidence to reject the null hypothesis so we can conclude that Kendrick Lamar's area data doesn't follow a statistically different continuous distribution than Drake's. Alternatively, we can say that the ECDF for Drake does not lie above that of Kendrick's, so this statistically support the earlier claim that Drake's triangle area measure has a greater spread based on the ECDFs plot in $Figure\ 11$.

In terms of dispersion ordering, this means we can say that Drake's triangle area measure stochastically dominates Kendrick Lamar's because:

$$\hat{F}_{Drake,1000}(z) \leq \hat{F}_{Kendrick,1000}(z), for\ all\ z\ and$$

$$\hat{F}_{Drake,1000}(z) < \hat{F}_{Kendrick,1000}(z), for\ some\ z$$

In other words, since Drake's Empirical CDF lies below or aligned with Kendrick's, despite not being statistically different, we can state for:

$$X = Drake's\ triangle\ area\ measures, Y = Kendrick's\ triangle\ area\ measures$$

That:

$$X \succeq_d Y \quad ie. X\ is\ more\ dispersive\ than\ Y$$