

STAT0006: ICA 1

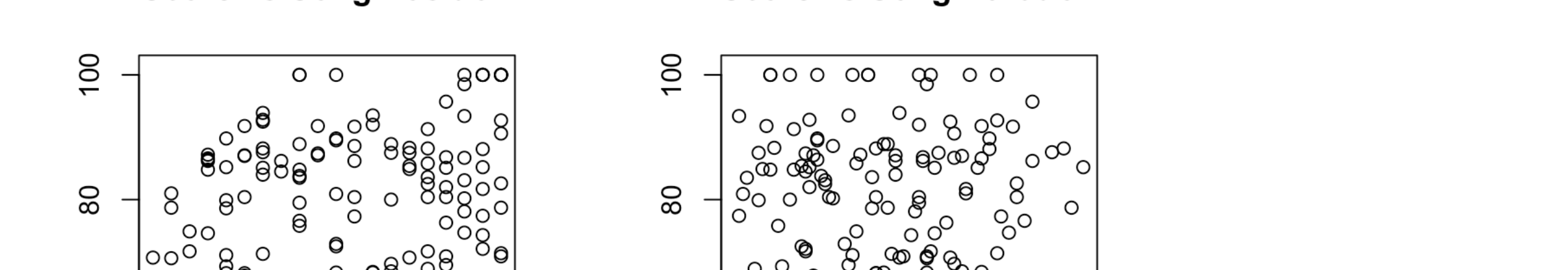
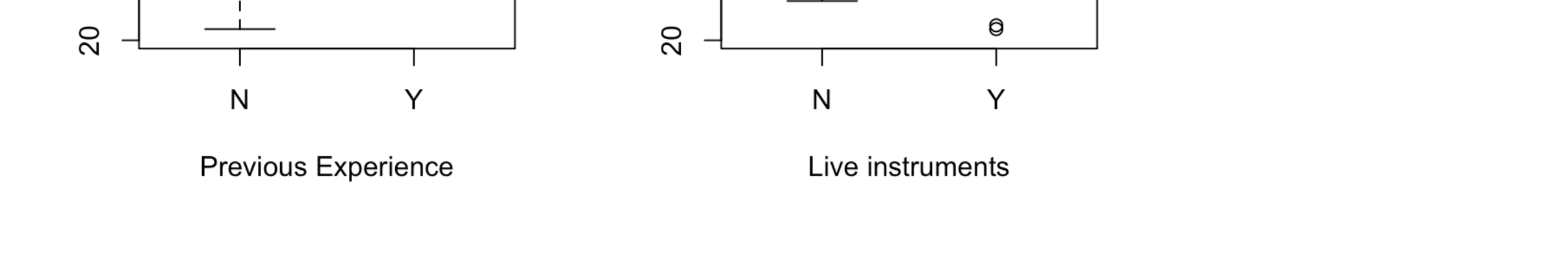
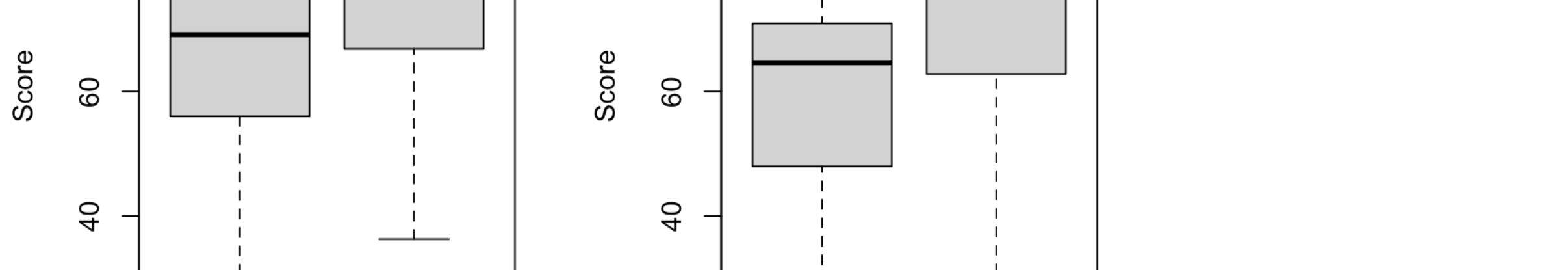
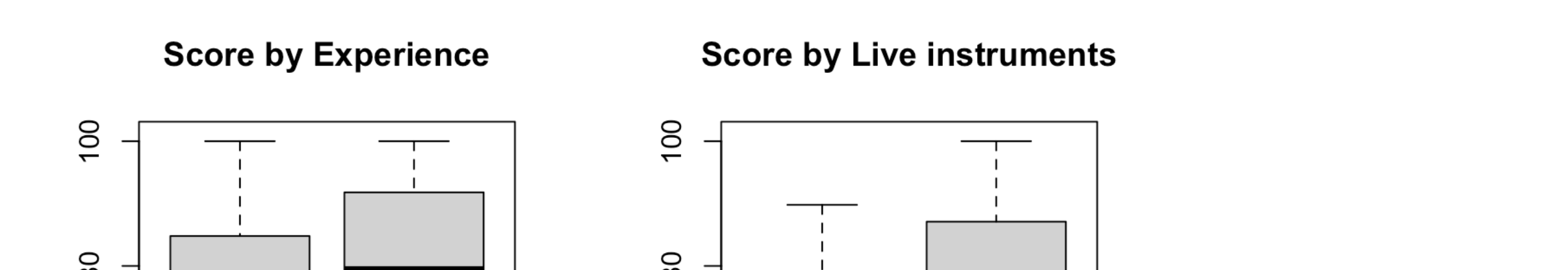
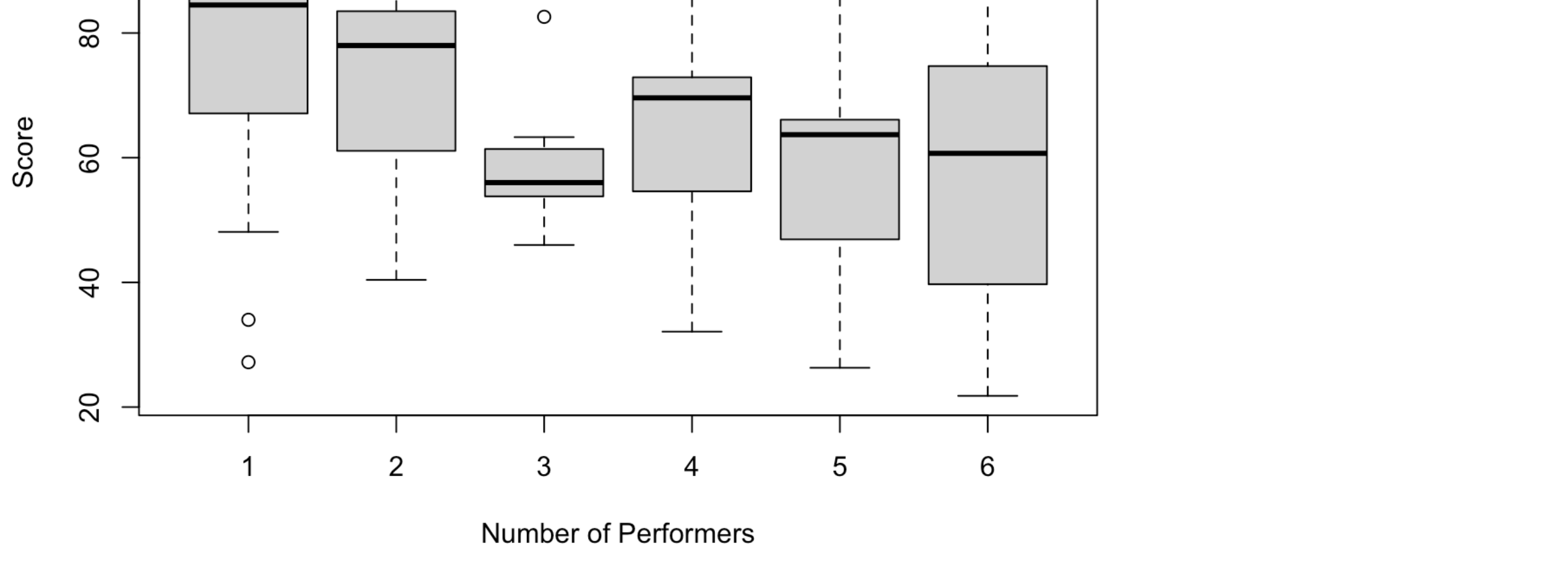
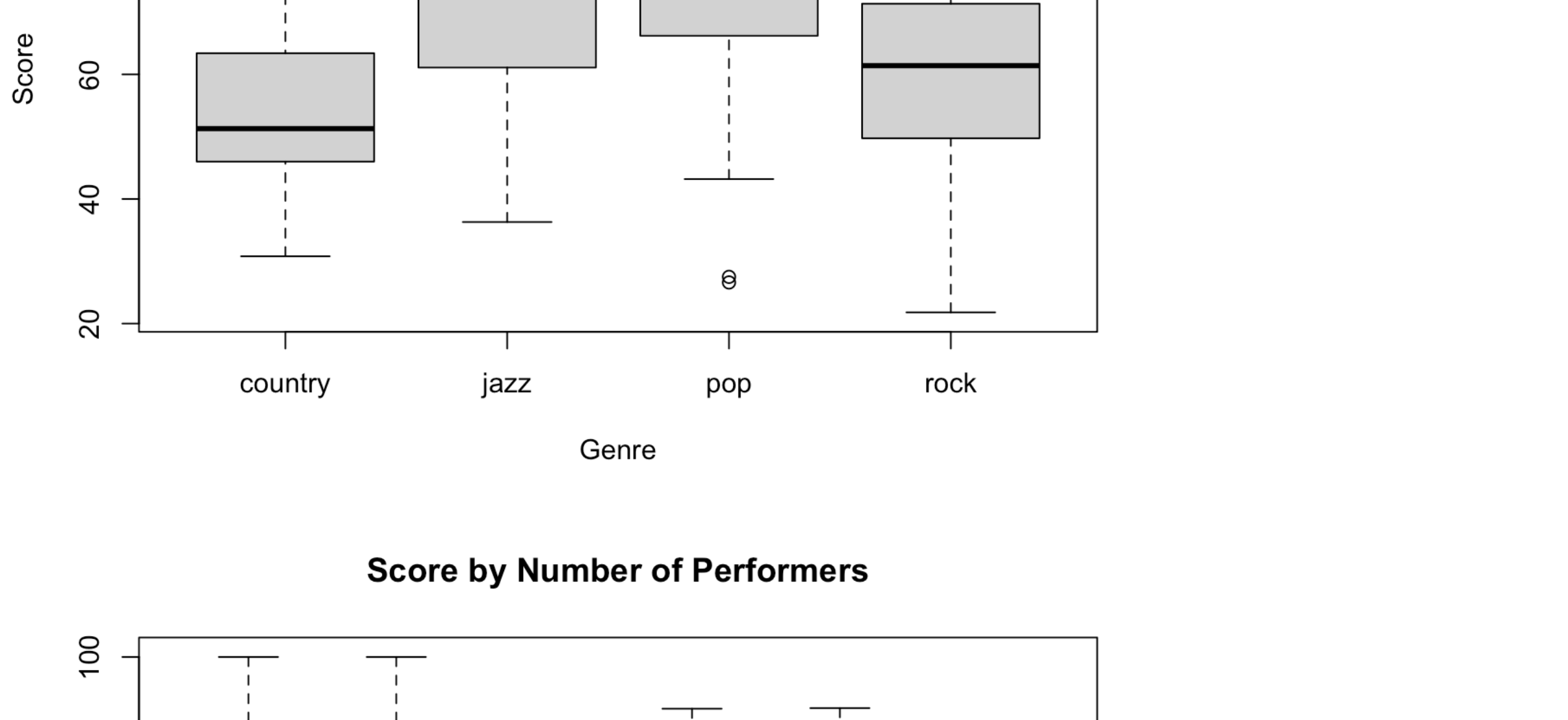
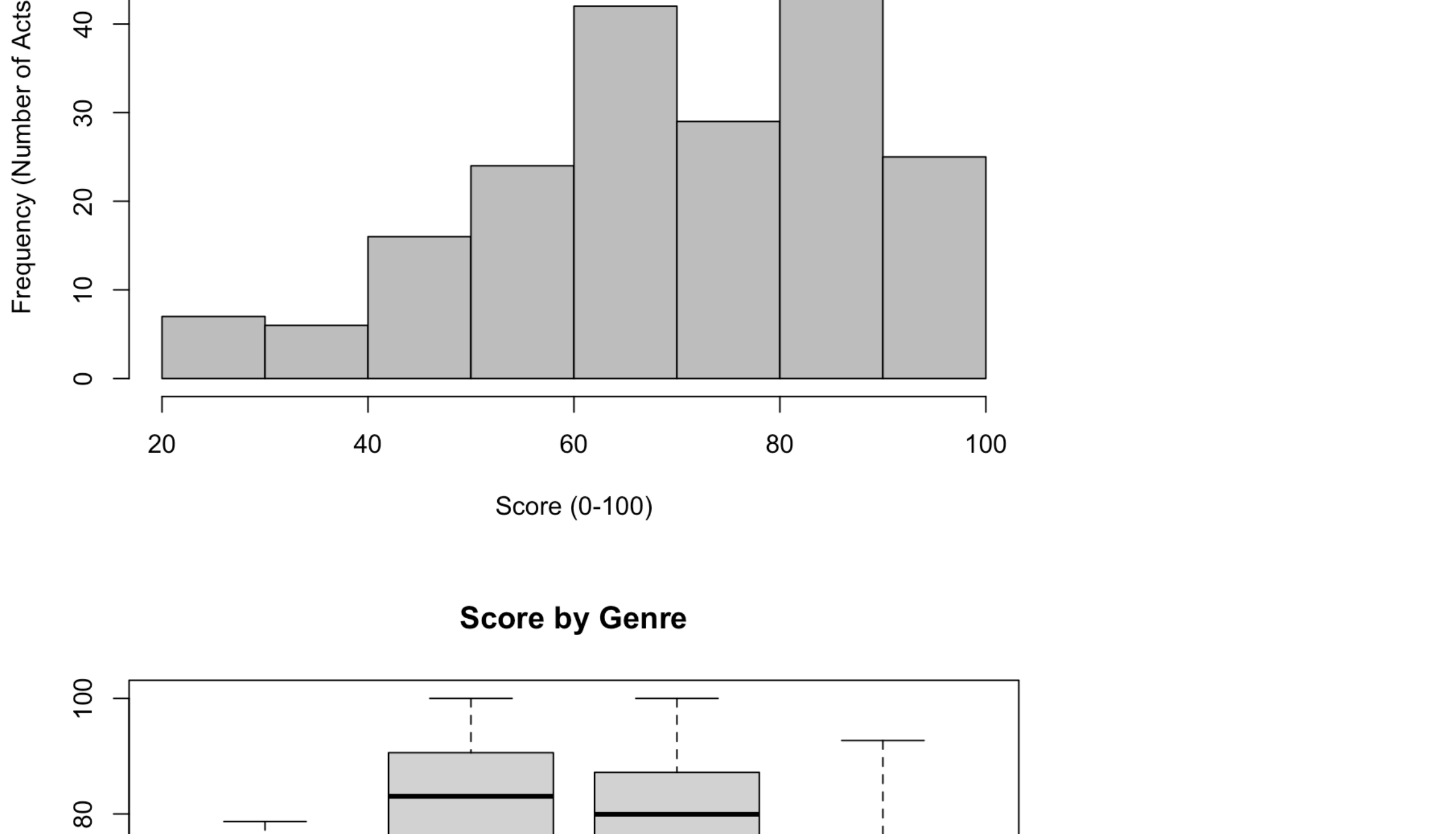
Student number: 24001114

Submission deadline: 12 noon on 18th November 2025 (DAP)

15/11/2025

Question 1

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	21.80	59.65	71.00	70.71	85.90	100.00



An examination of the song contest dataset highlights several impactful factors linked to success.

First, the overall average score for performances is about 70.7, but this score isn't distributed evenly; most entries tend to cluster near the higher end, which points to a negative skew in the results. Put simply, the majority of acts perform reasonably well, with relatively fewer low scores.

Music genre stands out as a major driver of success. Entries in Jazz and Pop consistently outperformed other styles, with Jazz acts averaging 77.0 points and Pop acts close behind at 76.4. By contrast, Rock entries averaged 59.7, and Country only 54.5, showing these styles may not be favoured strongly with the judges.

Another important factor is contest history. Experienced contestants (acts who have entered before) held a marked advantage, with an average score of 77.9, which is a nearly nine point lead over first-time entrants, who averaged 68.8. This suggests returning contestants benefit from practice or maybe greater recognition.

Instrumentation also played an important role. Acts featuring live instruments had an average score of 74.0, significantly higher than those without live instrumentation, who received just 60.7 on average. Live music appears to earn higher scores, possibly because judges associate it with greater musical skill or a stronger sense of audience connection.

Group size patterns show solo performers are most successful, scoring an impressive average of 78.3. Duos follow at 72.8, while three-person groups and larger receive much lower scores, usually between 56.8 and 66.8. So, smaller acts seem to have a small edge.

When examining running order (the performance sequence), there is only a slight benefit for acts performing later in the program. Points are somewhat higher for those in later slots, but the scatter is wide enough that no solid advantage is established. On the other hand, song length shows no meaningful relationship with scoring; whether a song ran longer or shorter, its duration didn't systematically impact marks.

In summary, the data strongly indicates that an ideal act for a high score would be a solo or duo performing Jazz or Pop, using live instruments, and coming in with prior contest experience. Factors like order of performance and song length are much less influential and shouldn't be a primary concern for entrants hoping to maximize their score.

Word count for Q1: (380 / 500).

Question 2

Model 1

```
##
## Call:
## lm(formula = score ~ genre, data = songs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -49.818  -10.361   3.248  11.115  33.011
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.535     3.950   13.806 < 2e-16 ***
## genrejazz    22.451     4.944    4.541 9.77e-06 ***
## genrepop     21.883     4.255    5.143 6.54e-07 ***
## generock      5.154     4.610    1.118  0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.29 on 196 degrees of freedom
## Multiple R-squared:  0.2215, Adjusted R-squared:  0.2096
## F-statistic: 18.59 on 3 and 196 DF,  p-value: 1.187e-10
```

Since genre is a categorical variable in Model 1, the model is comparing the average scores for each genre rather than fitting a numeric trend.

The linearity assumption primarily concerns numeric predictors, so it isn't relevant here (there's no underlying linear relationship to achieve through transformation). The model estimates group means for genres like jazz, country, pop, and rock based on score data. Any transformation to meet linearity is unnecessary, especially given the only mild skew observed in the EDA. Instead, to evaluate model assumptions, it's more useful to look at the residuals from the fit.

Word count for Q2: (96 / 150).

Question 3

Model 2

```
##
## Call:
## lm(formula = score ~ experience + genre + live + performers +
##      order + duration, data = songs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0252  -7.4698  -0.6262   4.8958  23.2237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.73893     4.83552   8.218 3.08e-14 ***
## experienceY   4.99943     1.61234    3.101 0.00222 **
## genrejazz     24.75710     2.80411    8.829 6.76e-16 ***
## genrepop      25.48813     2.42076    10.529 < 2e-16 ***
## generock       8.48759     2.62312    3.236 0.00143 **
## liveY         16.59626     1.52760    10.864 < 2e-16 ***
## performers   -4.98244     0.37973  -13.121 < 2e-16 ***
## order         1.22231     0.10913    11.201 < 2e-16 ***
## duration     -0.01535     0.03141   -0.489  0.62557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.194 on 191 degrees of freedom
## Multiple R-squared:  0.7582, Adjusted R-squared:  0.7481
## F-statistic: 74.88 on 8 and 191 DF,  p-value: < 2.2e-16
```

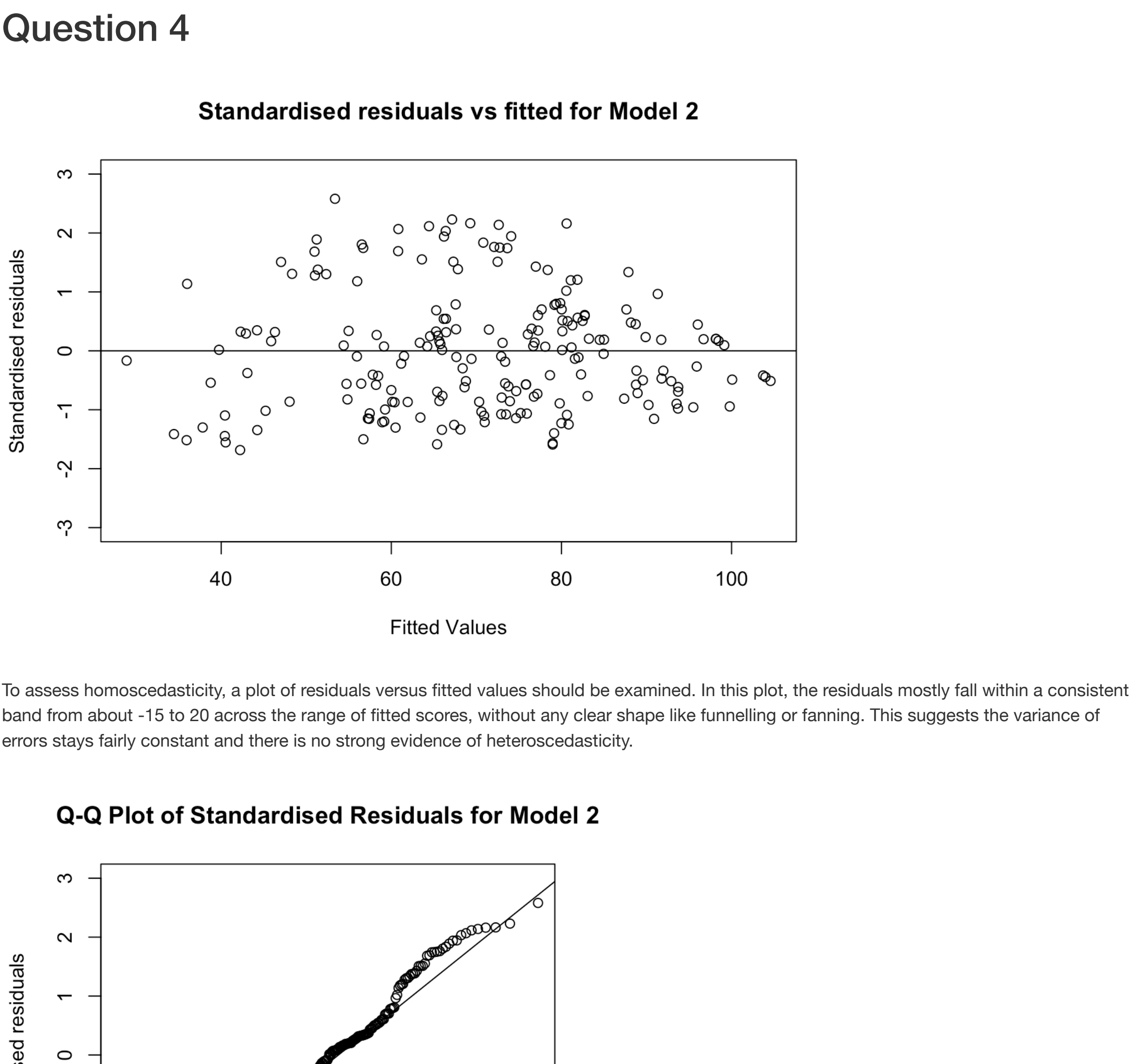
The Model 2 summary shows: Intercept = 39.74, generock = 8.49, and duration = -0.015. The intercept represents the average predicted score for a baseline act: a country band with no live instruments, no prior experience, 0 performers, 0 order, and 0 duration. This value is just a reference point rather than an actual scenario.

The generock coefficient of 8.49 indicates that, holding all other variables constant, a rock act is predicted to score 8.49 points higher than a country act. Duration's coefficient of -0.015 means that for each additional second of performance, the predicted score decreases by 0.015 points, assuming all other factors remain unchanged. This effect, however, is not statistically significant.

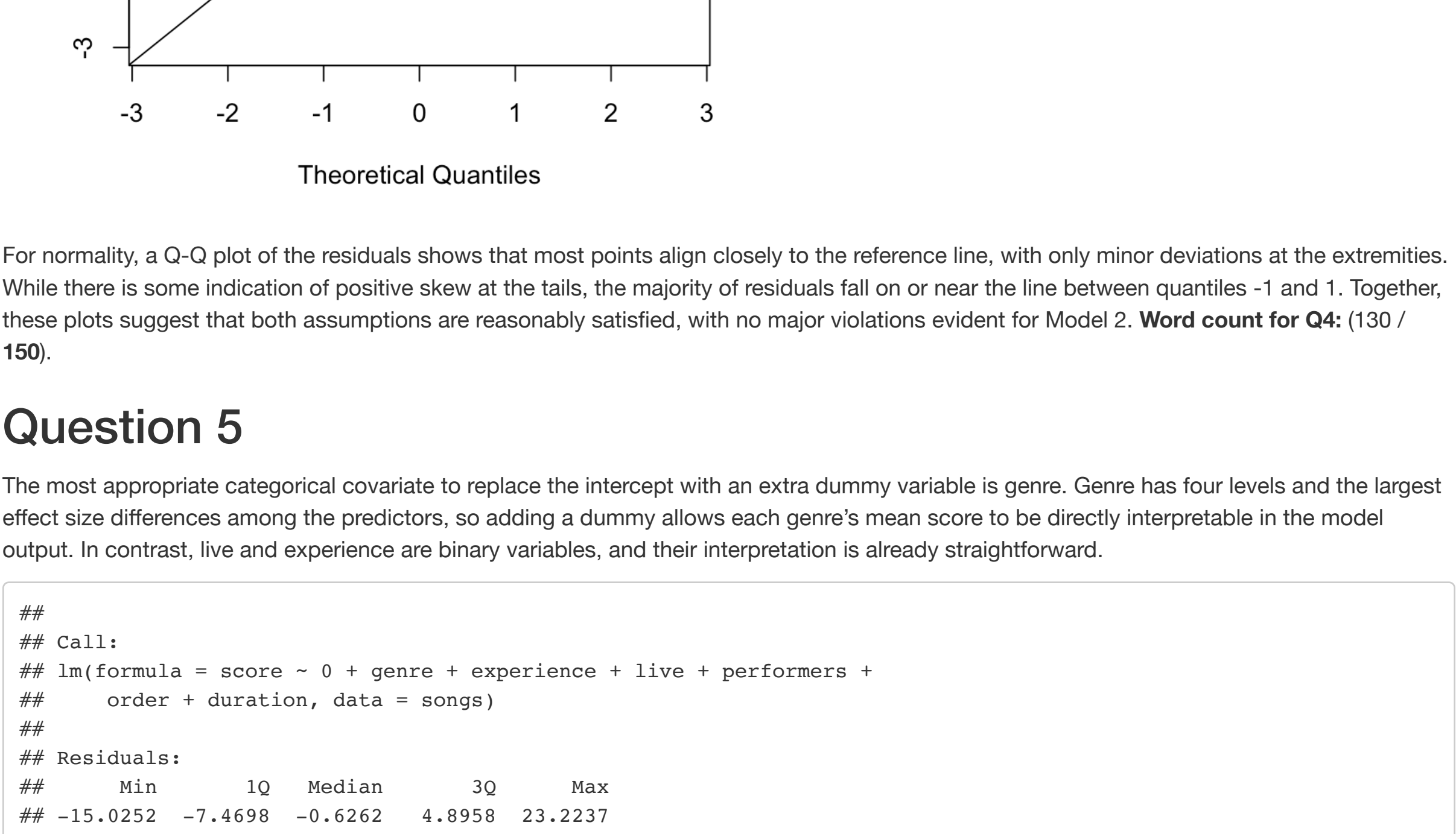
If duration were recorded in minutes instead of seconds, its coefficient would be multiplied by 60, resulting in -0.921 (since one minute equals sixty seconds). All other regression coefficients would remain unchanged, as the scaling only affects duration. So, the predicted change would represent points per minute rather than points per second.

Word count for Q3: (165 / 250).

Question 4



To assess homoscedasticity, a plot of the residuals versus fitted scores, without any clear shape like funneling or fanning. In this plot, the residuals mostly fall within a consistent band from about -15 to 20 across the range of predicted values, suggesting the variance of errors stays fairly constant and there is no strong evidence of heteroscedasticity.



For normality, a Q-Q plot of the residuals shows that most points align closely to the reference line, with only minor deviations at the extremities. While there is some indication of positive skew at the tails, the majority of residuals fall on or near the line between quantiles -1 and 1. Together, these plots suggest that both assumptions are reasonably satisfied, with no major violations evident for Model 2. **Word count for Q4:** (130 / 150).

Question 5

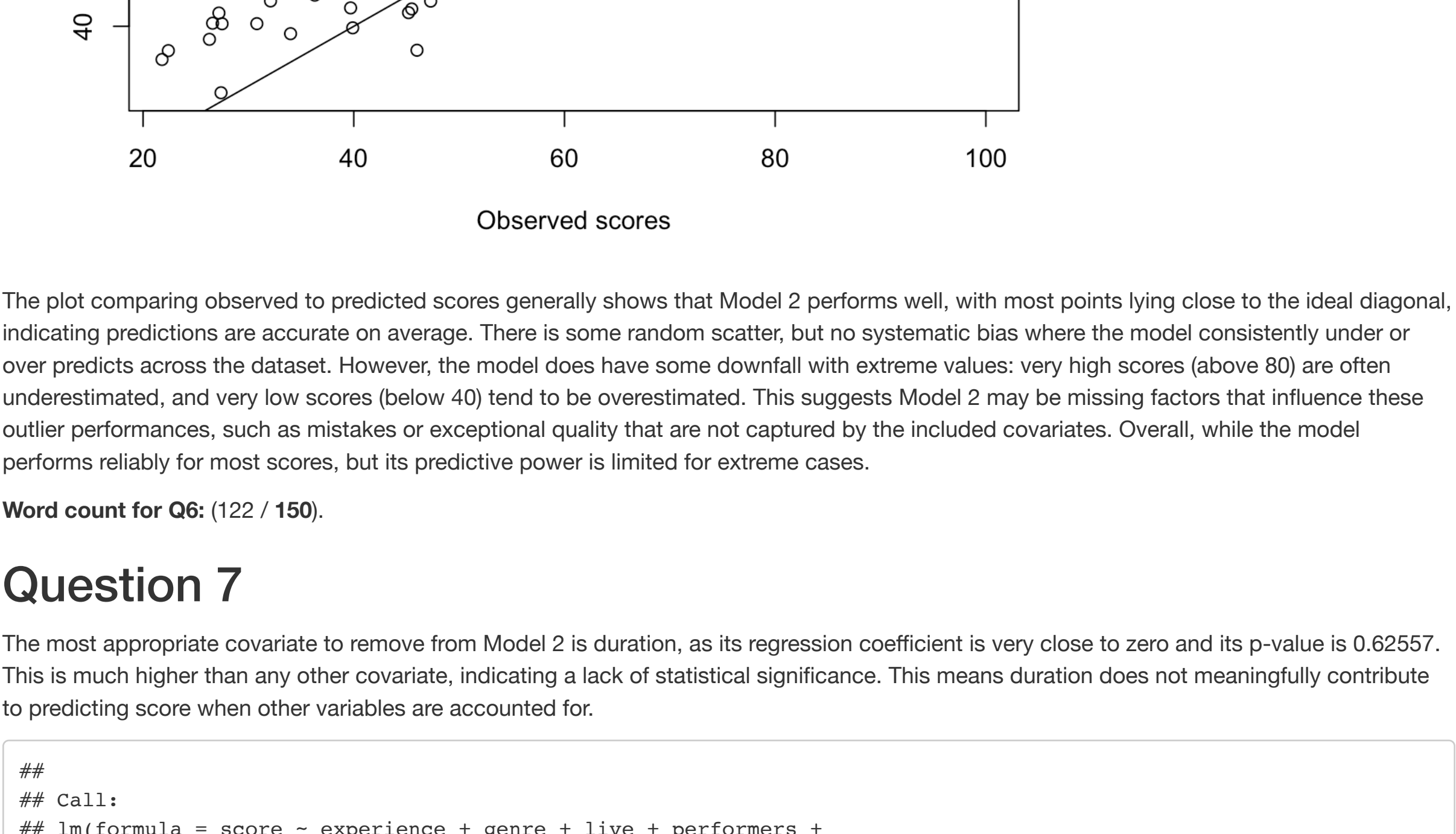
The most appropriate categorical covariate to replace the intercept with an extra dummy variable is genre. Genre has four levels and the largest effect size differences among the predictors, so adding a dummy allows each genre's mean score to be directly interpretable in the model output. In contrast, live and experience are binary variables, and their interpretation is already straightforward.

```
##
## Call:
## lm(formula = score ~ 0 + genre + experience + live + performers +
##      order + duration, data = songs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0252  -7.4698  -0.6262   4.8958  23.2237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## genrecountry  39.73893     4.83552   8.218 3.08e-14 ***
## genrejazz     64.49603     4.74597  13.590 < 2e-16 ***
## genrepop      65.22705     4.55687  14.314 < 2e-16 ***
## generock      48.22652     4.69596  10.270 < 2e-16 ***
## experienceY   4.99943     1.61234    3.101 0.00222 **
## liveY         16.59626     1.52760    10.864 < 2e-16 ***
## performers   -4.98244     0.37973  -13.121 < 2e-16 ***
## order         1.22231     0.10913    11.201 < 2e-16 ***
## duration     -0.01535     0.03141   -0.489  0.62557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.194 on 191 degrees of freedom
## Multiple R-squared:  0.9849, Adjusted R-squared:  0.9842
## F-statistic: 1381 on 9 and 191 DF,  p-value: < 2.2e-16
```

When fitting Model 3, the underlying statistical model remains unchanged apart from how coefficients are labelled. The fitted values and residuals do not change, only the interpretation of parameters shifts. Therefore, the assumptions of homoscedasticity and normality for the residuals are not improved in Model 3 compared to Model 2; both models fit the data identically, so model assumptions are unaffected by the reparameterization.

Word count for Q5: (124 / 200).

Question 6



The plot comparing observed to predicted scores generally shows that Model 2 performs well, with most points lying close to the ideal diagonal, indicating predictions are accurate on average. There is some random scatter, but no systematic bias where the model consistently under or over predicts across the dataset. However, the model does have some downfall with extreme values; very high scores (above 80) are often underestimated, and very low scores (below 40) tend to be overestimated. This suggests Model 2 may be missing factors that influence these outlier performances, such as mistakes or exceptional quality that are not captured by the included covariates. Overall, while the model performs reliably for most scores, but its predictive power is limited for extreme cases.

Word count for Q6: (122 / 150).

Question 7

The most appropriate covariate to remove from Model 2 is duration, as its regression coefficient is very close to zero and its p-value is 0.62557. This is much higher than any other covariate, indicating a lack of statistical significance. This means duration does not meaningfully contribute to predicting score when other variables are accounted for.

```
##
## Call:
## lm(formula = score ~ experience + genre + live + performers +
##      order, data = songs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2838  -7.4742  -0.3031   4.8975  22.6513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.8473     2.8931  13.082 < 2e-16 ***
## experienceY   5.0144     1.6089    3.117 0.00211 **
## genrejazz     24.6984     2.7960    8.834 6.38e-16 ***
## genrepop      25.3886     2.4074    10.546 < 2e-16 ***
## generock       8.3847     2.6095    3.213 0.00154 **
## liveY         16.6219     1.5237    10.909 < 2e-16 ***
## performers   -4.9658     0.3775  -13.156 < 2e-16 ***
## order         1.2176     0.1085    11.224 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.176 on 192 degrees of freedom
## Multiple R-squared:  0.7579, Adjusted R-squared:  0.7491
## F-statistic: 85.88 on 7 and 192 DF,  p-value: < 2.2e-16
```

No F-test is needed in this situation. When removing a single covariate, the t-test for that coefficient (already provided in Model 2's summary output) is equivalent to a partial F-test, and given the very high p-value, there is no evidence against its removal. So, simplifying to Model 4 by excluding duration is justified for, without the need for further testing.

Word count for Q7: (115 / 150).

Question 8

Using the model to advise a classical band is not advisable, since the dataset on which the model is based does not include 'classical' as a genre, meaning any prediction for this category would use unreliable extrapolation. The effect of genre is a primary driver of scores in the model, so without relevant data, the results cannot be meaningfully interpreted or trusted for classical performances.

However, insights regarding other covariates; such as live instrumentation, number of performers, and experience, may still offer broad guidance. The model suggests live performances and bands with fewer members tend to score higher, and performing later in the order may also help. Bu, these trends are observed only for the existing genres and may not apply to classical music. For example, classical music often benefits from more instruments, so minimizing performers may not lead to a higher score. Overall, while general performance factors could inform their approach, the lack of genre-specific data means the model is unsuitable to be used.

Word count for Q8: (165 / 200).

Question 9

One potential improvement is to relax the assumption of additivity by explicitly adding interaction terms to the model. Interaction terms allow for the effects of one variable, such as live performance, to depend on another, like genre. For example, live performances may contribute more positively to rock than jazz scores, which a model with only additive terms fails to capture. Including these enables the model to reflect more nuanced patterns in the data, rather than assuming each covariate operates independently across all levels of other covariates.

A second way to improve the model is to accommodate non-linear relationships by introducing polynomial or quadratic terms. The EDA revealed that the link between order and score was not strictly linear; beginning entries often had lower scores while later ones scored higher, which does not fit a simple linear pattern. By allowing the effect of order on score to vary in a non-linear way, the model can better account for such curved relationships.

Word count for Q9: (162/ 250).

Statement about use of generative AI tools

I used AI to help proofread my answers. All the code and text is done by me.