Sample Work 2: Model Comparison and Linear Regression Analysis

By Yu Wu - Student Number: 23007515

10/11/2025

Background

This work was completed as part of the assessment for **STAT0006 Regression Modelling**, an undergraduate course offered by Department of Statistics at UCL. The project involves the application of simple and multiple linear regression models to simulated data, with a focus on model comparison, residual diagnostics, and predictive performance evaluation.

This report is assembled and restructured from individual assignments, which reflects core components of scientific computing: reproducible analysis, quantitative model selection, and statistical reasoning. The entire workflow was carried out using R and R Markdown, ensuring transparency and clarity in data processing and model evaluation.

The tasks include:

1. Fitting simple and multiple linear regression models
2. Assessing goodness-of-fit using residual analysis
3. Comparing models via adjusted R squared and ANOVA
4. Identifying potential issues such as multicollinearity and nonlinearity

The final grade for this work has not yet been released. Through this project, I deepened my understanding of:

1. How to implement and interpret linear regression models in R
2. The practical considerations involved in model selection and evaluation
3. The importance of graphical diagnostics in validating assumptions
4. Communicating statistical results clearly through code and commentary

Introduction to the Data

Suppose that an annual song contest has been running for the past 50 years. The organisers have put together a dataset containing information about a random selection of entries during this time. They would like your help to understand the factors that influence the scores received by different entries.

The dataset includes the following variables:

- `score` : the number of points achieved by the song (0 to 100)
- `experience` : whether the act had previously entered the competition (`Y` or `N`)
- `genre` : the type of music performed (`pop` , `country` , `rock` , or `jazz`)
- `performers` : number of performers on stage
- `live` : whether the performance included live instruments (`Y` or `N`)
- `order` : the position of the act in the running order (1 to 20)
- `duration` : the length of the performance (in seconds)

The dataset is contained in the file `songs.csv`.

Task 1: Exploratory Data Analysis

Summary:

- **Conduct an exploratory analysis of the dataset without fitting any models.**
- **Accessible to non-statistician.**
- **Focus on visualising distributions, relationships, and potential patterns in the data.**

Answer:

```
summary(songs)
```

```
##      score           genre             performers        live
##  Min.   : 21.80   Length:200         Min.   :1.000   Length:200
##  1st Qu.: 59.65   Class :character   1st Qu.:1.000   Class :character
##  Median : 71.00   Mode  :character   Median :2.000   Mode  :character
##  Mean   : 70.71                      Mean   :2.515
##  3rd Qu.: 85.90                      3rd Qu.:4.000
##  Max.   :100.00                      Max.   :6.000
##   experience            order          duration
##  Length:200         Min.   : 1.00   Min.   : 90.0
##  Class :character   1st Qu.: 5.00   1st Qu.:110.8
##  Mode  :character   Median :11.00   Median :127.0
##                     Mean   :10.37   Mean   :127.8
##                     3rd Qu.:16.00   3rd Qu.:145.0
##                     Max.   :20.00   Max.   :179.0
```

Interest lies in understanding which factors of song performance influence scores in an annual song contest. This data set has a total of 200 entries and contains six variables: three factors that are categorical and cannot compare by values directly (experience, genre and live) and three covariates that can make numeric comparison (performers, order and duration). There are no missing values in the variables used here.

The scores span a reasonably wide range, ranging from 21.8 to 100, with mean around 70.71 and standard deviation around 18.32.

We investigate the relationships between contest score and the 3 categorical covariates using the boxplots below. The scores have been separated according to previous experience or not, live instruments or not and different types of genres respectively in the 3 boxplots below.

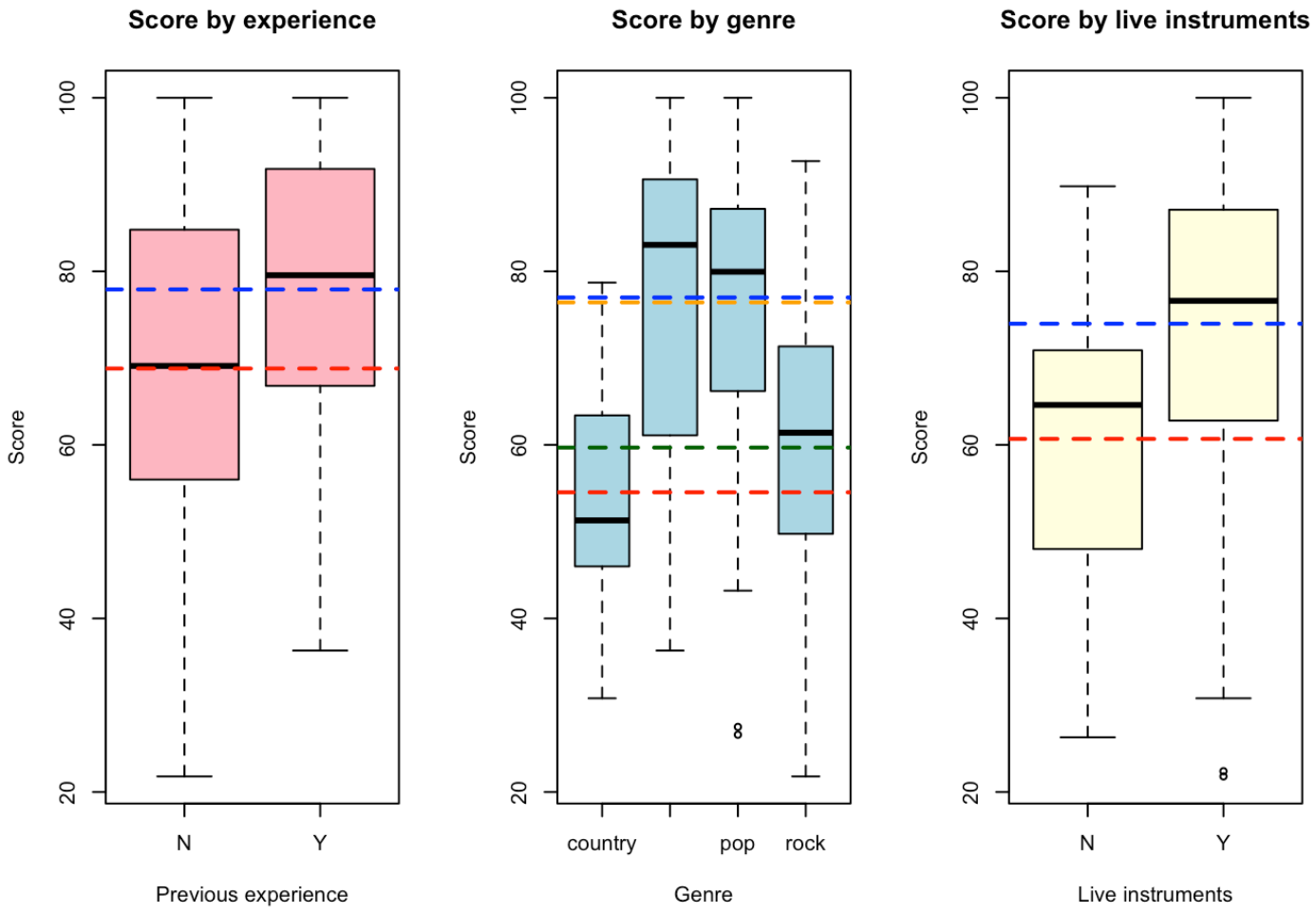
The first plot compares scores between performance with and without previous experience. The boxplot of experience suggests that acts with previous experience tend to achieve higher scores, with a mean of 77.91 compared to 68.8 for inexperienced performers. This suggests that experience would improve performance scores.

The second plot shows scores by different genres. Jazz and pop tend to have higher mean scores of 76.99 and 76.42 respectively, while country and rock are generally lower, with rock showing the largest spread ranging from 21.8 to 92.7. This indicates that genre could play an important role, with certain musical styles being more acknowledged to judges.

The final boxplot compares performances with and without live instruments. Those including live instruments (Y) appear to receive higher mean scores (73.96 compared with 60.69) Thus, having live elements could positively improve scores.

Key codes used are:

```
# The following code illustrates how to create boxplots with horizontal reference
lines for each category:
boxplot(score ~ covariate, data = songs,
        xlab = "Covariate Name",
        ylab = "Score",
        col = "Chosen Colour",
        main = "Score by Covariate Name")
abline(h = mean(songs$score[songs$experience == "Category 1"]), col = "red", lwd =
2, lty = 2)
abline(h = mean(songs$score[songs$experience == "Category 2"]), col = "blue", lwd
= 2, lty = 2)
```



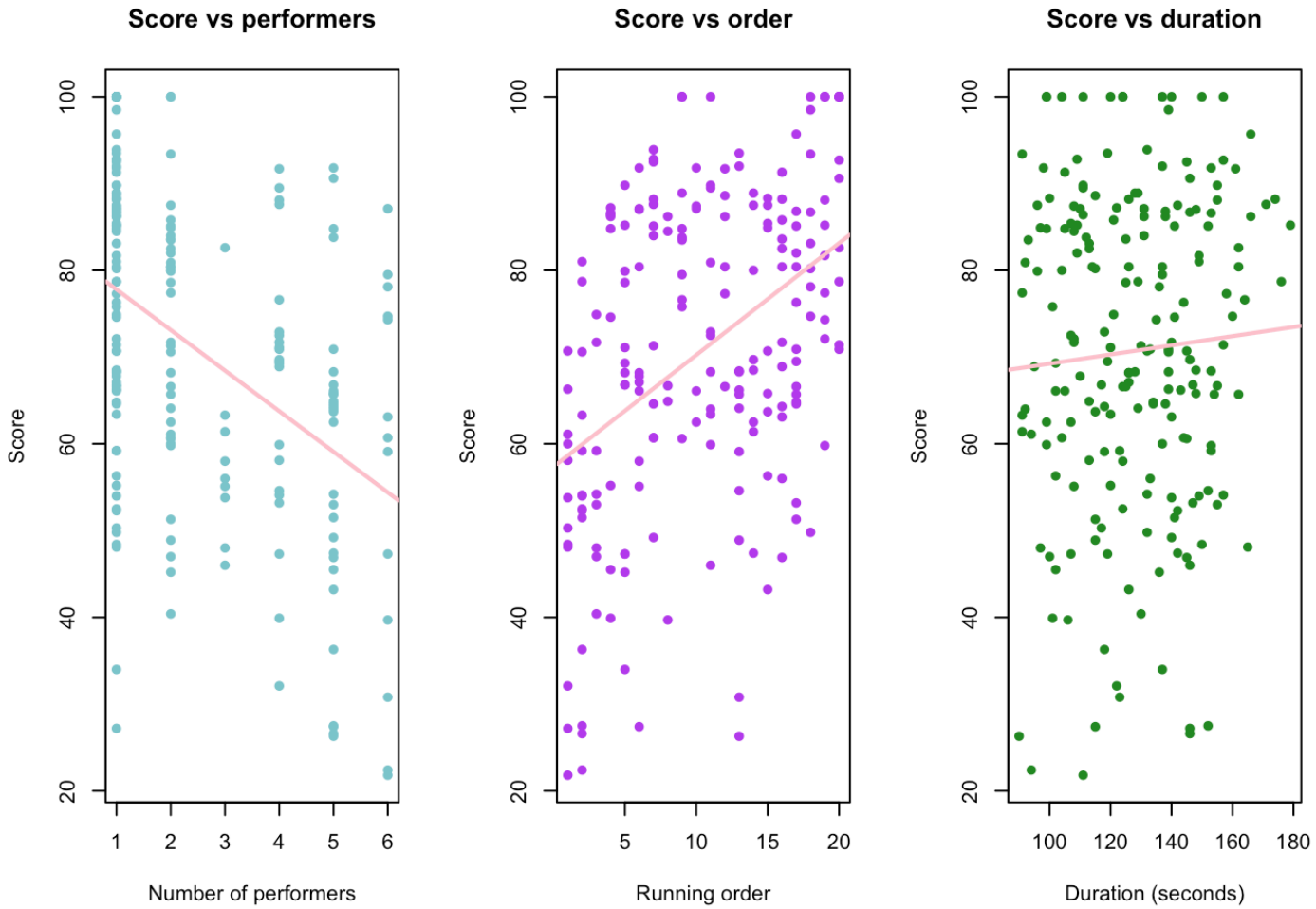
We then look at the numeric covariates using scatter plots of score against performers, order and duration respectively. The average score in the data set is 70.7. There appears to be a negative relationship between the number of performers and score (coefficient being -4.679 per additional performer), showing that smaller groups tend to score higher.

There appears to be a positive relationship between the running order and score, with later performers achieving relatively higher scores on average (coefficient being 1.289 per position).

There appears to be a weak positive relationship between the duration and score (coefficient being 0.053 per second), although the fitted line is nearly flat, suggesting that song length has minimal impact on the score.

Overall, these findings suggest that order and number of performers may have significant impact on score, while duration has minimal impact.

```
# Below is an illustration how to create scatter plots with a fitted regression li
ne for a numeric covariate:
plot(songs$covariate_name, songs$score,
     xlab = "Covariate Name",
     ylab = "Score",
     main = "Score vs covariate_name",
     col = "Chosen Colour",
     pch = 16)
abline(lm(score ~ covariate_name, data = songs), col = "Another Colour", lwd = 2)
```



In summary, our initial findings suggest that previous experience, musical genre, and use of live instruments may all be related to performance scores, while the number of performers, running order, and duration show weaker or no clear associations.

Word count for Q1: (insert word count here 473 / **500**).

Task 2: Fit Model 1 (Simple Linear Regression)

Background:

- **This task introduces a baseline model to evaluate how genre alone influences the contest scores.**
- **Simple linear regression allows us to isolate the effect of a single categorical predictor and assess the basic model assumptions.**

Summary:

- **Fit a linear regression model 1 using `genre` as the only explanatory variable for `score`**
- **Assess whether a transformation of `score` is necessary due to the assumption of linearity based on the model and EDA.**

Answer:

```
# Code for Model 1
model1 <- lm(score ~ genre, data = songs)
summary(model1)
```

```
##
## 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 ***
## genrerock      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
```

Model 1 is a linear regression examining how genre affects score. The R summary shows clear mean differences between genres, using country as the baseline. The intercept (54.5) gives the mean score for country songs. Compared with these, jazz songs average 22.5 points higher, pop 21.9 higher, and rock 5.2 higher. These differences agree with the EDA box plots.

Since genre is categorical, the model compares group means rather than fitting a continuous line, so linearity is irrelevant. The outcome (score, 0–100 scale) shows no strong skew or outliers, so no transformation is needed.

Word count for Q2: (insert word count here 95 / **150**).

Task 3: Fit Model 2 (Multiple Linear Regression)

Background:

- **Multiple linear regression allows us to control for multiple covariates simultaneously and obtain adjusted estimates of their associations with the outcome.**

Summary:

- **Fit a linear model 2 with all categorical (experience, genre, live) and numeric (performers, order, duration) predictors.**
- **Interpret key coefficients (eg. Intercept, Genrerock, Duration) and assess how unit changes (e.g., seconds to minutes) affect the the interpretation of regression coefficients.**

Answer:

```
# Code for Model 2
model2 <- lm(score ~ experience + genre + live + performers + order + duration,
             data = songs)
summary(model2)
```

```
##
## 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 ***
## genrerock    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
```

Model 2 is a multiple linear regression model for score considering all covariates. The R model summary is shown above. The intercept provides a baseline (reference) and is 39.7, representing the expected score for a country song performed by an inexperienced act (experience = N), without live instruments (live = N), with zero performers, order = 0 and duration = 0 seconds as indicated in the question.

The estimated coefficient for genrerock is 8.49, meaning that holding all other covariates constant, the expected number of rock scores is 8.49 points higher than country songs on average. Since the p-value is 0.00143, this difference is statistically significant, meaning that genre appears to be an important factor influencing score.

The estimated coefficient for duration is −0.015 which is negative, suggesting that for every additional second of performance, the expected score decreases very slightly by about 0.015 points, when other covariates are fixed. The p-value is 0.62557, so this term is not statistically significant, suggesting that song length has tiny effect on score.

If duration is measured in minutes instead of seconds, since 1 minute = 60 seconds, the estimated coefficient would be multiplied by 60. The new coefficient would therefore be approximately −0.921, representing the expected change in score for a 1-min increase in performance time. However, the intercept and all other regression coefficients would remain unchanged, because changing the units of one covariate does not affect the estimated effects of the others.

Word count for Q3: (insert word count here 241 / **250**).

Task 4: Model Assumption Diagnostics – Homoscedasticity & Normality

Background:

- **Linear regression relies on certain assumptions, including that residuals have constant variance (homoscedasticity) and follow a normal distribution.**
- **Violations of these assumptions can lead to biased estimates or invalid inference.**
- **In this task, I assess these assumptions for Model 2 using graphical diagnostics.**

Summary:

- **Produce one plot to assess homoscedasticity of errors, and one plot to assess normality of errors.**
- **Discuss whether any violations are evident based on the diagnostic plots.**

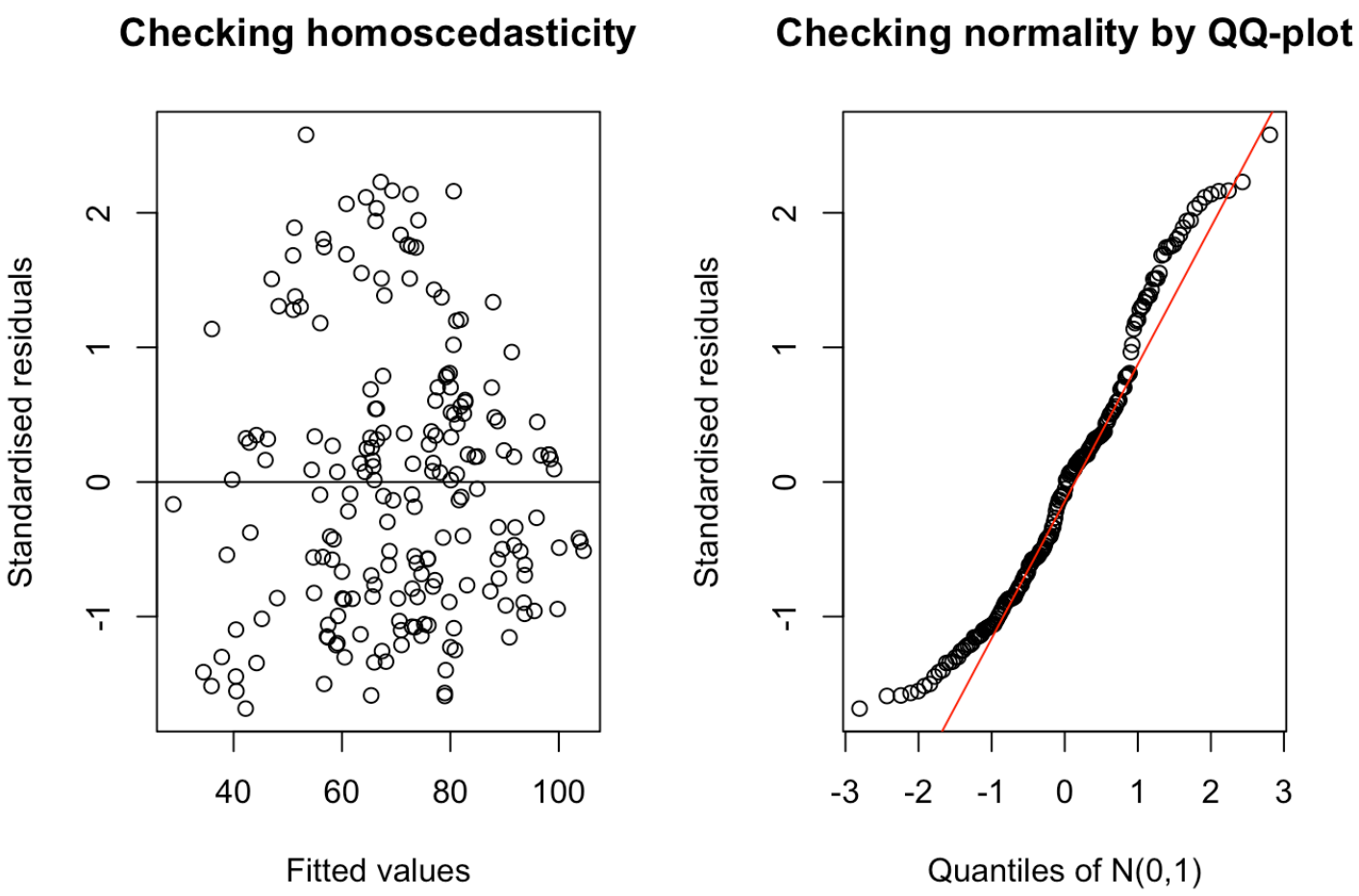
Answer:

```
# Compute standardised residuals and fitted values
model2_stdres <- rstandard(model2)
model2_fitted <- fitted(model2)

par(mfrow = c(1, 2))

# Check homoscedasticity
plot(model2_fitted, model2_stdres,
     xlab = "Fitted values",
     ylab = "Standardised residuals",
     main = "Checking homoscedasticity")
abline(a = 0, b = 0)

# Check normality
qqnorm(model2_stdres,
       main = "Checking normality by QQ-plot",
       ylab = "Standardised residuals",
       xlab = "Quantiles of N(0,1)")
qqline(model2_stdres, col = "red")
```



Homoscedasticity means the variance of errors remain the same. To be met, ideally, residuals should be randomly scattered around the zero line with a constant spread across all fitted values. The left plot of standardised residuals against fitted values shows no strong pattern, with points roughly scattered evenly

around zero horizontal line across the fitted range. This suggests that the assumption of homoscedasticity is reasonable. There is a small increase in spread for higher fitted values, but the change is minor and not enough to violate the assumption.

Normality means the error components are normally distributed. We use a QQ-plot to check this assumption. The right plot indicates that standardised residuals do not appear to be normally distributed, as evidenced by the deviations from the reference line in the plot, especially in the tails. Also, there is some bending away from the line. Therefore, the assumption of normality is violated.

Word count for Q4: (insert word count here 150 / **150**).

Task 5: Fit Model 3 (Modified Intercept with Dummy Variable)

Background:

- **In regression models with categorical predictors, the choice of reference category can influence interpretability and numerical stability.**
- **This task examines the effect of changing the intercept structure by including a dummy variable for the reference level.**

Summary:

- **Replace the intercept in Model 2 with a dummy variable for a reference category in one categorical variable.**
- **Fit Model 3 accordingly and compare error assumptions with Model 2.**
- **Model 3 code will depend on the chosen categorical variable, e.g. genre.**

Answer:

```
# Code for Model 3
model3 <- lm(score ~ -1 + as.factor(genre) + experience + live + performers + orde
r + duration,
             data = songs)
summary(model3)
```

```
##
## Call:
## lm(formula = score ~ -1 + as.factor(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|)
## as.factor(genre)country 39.73893    4.83552   8.218 3.08e-14 ***
## as.factor(genre)jazz    64.49603    4.74597  13.590  < 2e-16 ***
## as.factor(genre)pop     65.22705    4.55687  14.314  < 2e-16 ***
## as.factor(genre)rock    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
```

Genre is the most appropriate categorical covariate to have the extra dummy variable for. This variable has four distinct categories (country, jazz, pop, and rock) and does not have a natural reference category. By including a full set of dummy variables and removing the intercept, we can estimate an adjusted mean score for each genre directly. This makes interpretation clearer than in Model 2 where each coefficient represented a difference from the reference group (country). For other categorical variables such as experience and live, the binary nature of the variable already gives a clear baseline (Y/N), so adding a second dummy would be redundant.

Model 3 is mathematically equivalent to Model 2. It contains the same variables and therefore produces identical fitted values and residuals. Thus, the assessments of homoscedasticity and normality will not change. The diagnostic plots from Question 4 would look identical, meaning there is no improvement or deterioration in how well these assumptions are satisfied.

Word count for Q5: (insert word count here 158 / **200**).

Task 6: Prediction with Model 2

Background:

- **Comparing observed and predicted values is a key way to assess a model's overall fit and identify potential patterns in residuals.**
- **A strong model should closely match observed values without systematic bias.**

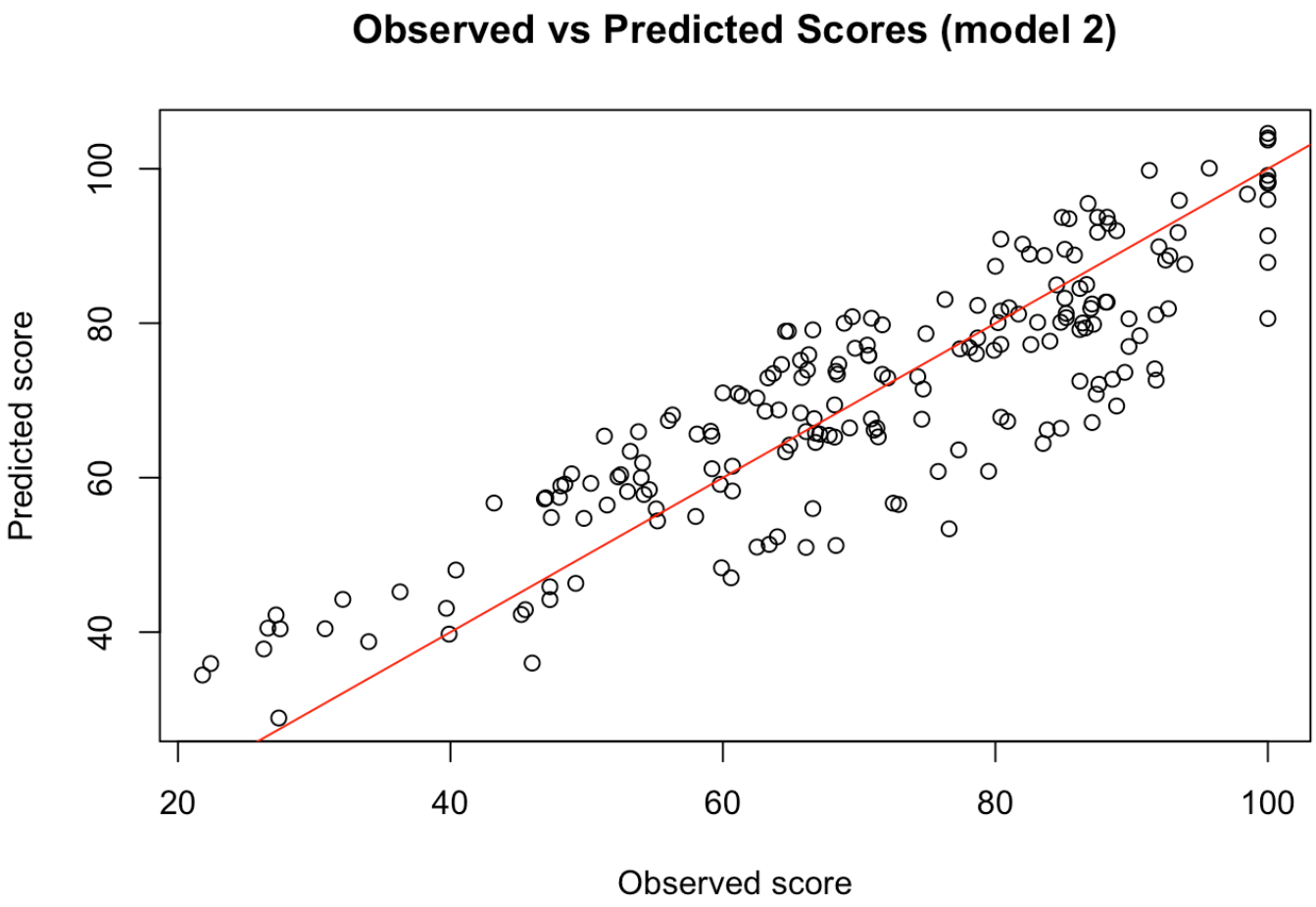
Summary:

- **Predict `score` for all observations using Model 2.**
- **Plot predicted vs. observed values and discuss model fit.**

Answer:

```
# Predicted values for each observation
fits <- fitted(model2)

# Create plot comparing observed vs predicted scores
plot(songs$score, fits,
     xlab = "Observed score",
     ylab = "Predicted score",
     main = "Observed vs Predicted Scores (model 2)")
abline(a = 0, b = 1, col = "red")   # 45-degree reference line
```



The predicted scores from Model 2 were obtained using the `fitted()` function in R, which returns the model's fitted values for each observation. The plot compares the observed scores (x-axis) with the predicted scores (y-axis). If the model fits well, the points should lie close to the red 45° line, indicating that predicted and observed values are similar.

The plot shows a strong positive relationship, and most points fall near the reference line, suggesting that Model 2 predicts the scores reasonably well. There is no clear systematic pattern of over- or under-prediction, although a few points at lower scores lie slightly below the line, indicating minor underestimation for low scores. Overall, the model does not raise serious concerns from this plot.

Word count for Q6: (insert word count here 121 / **150**).

Task 7: Simplify Model 2 to Model 4 by Removing One Covariate

Background:

- **Simplifying a regression model can improve interpretability and reduce overfitting, but should be done carefully using statistical evidence.**
- **The F-test is a standard tool to compare nested models and assess whether a removed covariate significantly contributes to model fit.**

Summary:

- **Propose one covariate to remove from Model 2 to create Model 4.**
- **Discuss whether an F-test is needed to compare the two models.**

Answer:

Duration is the most appropriate covariate to remove based on Model 2 output. Its estimated coefficient (–0.015) is very close to zero and has the largest p-value of 0.63, meaning we fail to reject H0: $\beta = 0$ of the t-test. This indicates that after controlling for the other covariates, duration does not provide additional impact on score. The EDA in Q1 also shows no clear relationship between duration and score, supporting its removal from the model.

F-test is needed if several coefficients were tested jointly. When removing only one covariate, the t-test for that coefficient in Model 2 is equivalent to the F-test for comparing Model 2 and Model 4 (since $F = t^2$ for one degree of freedom). Therefore, conducting a separate F-test is not required, as the t-test already indicates that removing duration will not significantly reduce model fit.

Word count for Q7: (insert word count here 143 / **150**).

Task 8: Application - Advice for Classical Band

Summary:

- **Use the model insights to give practical advice to a hypothetical classical band considering entering the contest.**
- **Assess whether using the model is appropriate for decision-making.**

Answer:

The model results and boxplots from the EDA both showed that acts with previous experience and live instruments scored more highly. Also, performances in jazz and pop genres tended to achieve higher scores on average. Therefore, if the classical band decides to enter, they might focus on improving performance quality through repetitively rehearsing and including live instrumental elements to enhance their stage presence.

However, since classical music was not included among the genres used to build this model, any prediction for this style could involve extrapolation outside the data range. The model cannot reliably estimate how classical pieces would be judged, because its coefficients are based only on country, jazz,

pop, and rock songs.

Thus, although the model provides useful insights about general factors affecting higher scores, using this model to predict results for a new genre classical music might not be a sensible idea.

Word count for Q8: (insert word count here 145 / **200**).

Task 9: Model Improvement Ideas

Summary:

- **Propose two improvements to the modelling approach that have not yet been considered.**
- **These can be related to model structure, variable treatment, or methodological enhancements.**

Answer:

One potential improvement would be to include interaction terms between key covariates. In the current model, all covariates are assumed to be independent, but in reality, these factors may interact. For instance, for different genres, there could be different optimal number of performers, as some genres might be performed better with more instruments, therefore more performers. Similarly, order might have different effect on different genres, as some genres might benefit more from being performed at a later time in the contest. The model can better adjust for these conditional effects by including those interaction terms.

Another way to improve the model is by including relevant external factors not currently measured in the data set. For example, the time of day or location of the contest might influence performance quality or judges' perception. For example, performing late in the day might be unfair as judges become bored and tired. Also, unfamiliar venues could bring uncertainty and influence sound quality. Including these would help explain additional variability in scores that cannot be captured by the current covariates.

Word count for Q9: (insert word count here 175 / **250**).

Statement about use of generative AI tools

I only use AI to check the grammar of my answers.