STAT 204 Exam Study Guide & Practice Problems

Due Wednesday, November 18, 2026, 11:59 pm (Pacific)

Fall 2026

2026-11-04

Table of contents

1 Overview **2**

2 What to Study **2**

2.1 Key Topics . . . 2

2.1.1 EDA and PCS . . . 2

2.1.2 Linear Regression . . . 3

2.1.3 ANOVA . . . 3

2.1.4 Logistic Regression . . . 3

2.2 Formulas to know . . . 3

3 Practice Problems **3**

3.1 Section 1: Exploratory Data Analysis . . . 3

3.1.1 Practice Problem 1.1: Danceability by Genre . . . 4

3.1.2 Practice Problem 1.2: Energy and Loudness . . . 5

3.2 Section 2: Linear Regression . . . 6

3.2.1 Practice Problem 2.1: Cocoa Percentage and Chocolate Ratings . . . 6

3.2.2 Practice Problem 2.2: Model Diagnostics . . . 7

3.3 Section 3: Analysis of Variance (ANOVA) . . . 8

3.3.1 Practice Problem 3.1: Genre, Mode, and Danceability . . . 8

3.3.2 Practice Problem 3.2: ANOVA Table Interpretation . . . 9

3.4 Section 4: Comprehensive Practice Problems . . . 11

3.4.1 Practice Problem 4.1: Predicting Energy from Loudness . . . 11

3.4.2 Practice Problem 4.2: Ramen Ratings by Style and Country . . . 13

3.5 Section 5: Logistic Regression . . . 15

3.5.1 Practice Problem 5.1: Age and Summiting Everest . . . 15

3.5.2 Practice Problem 5.2: Adding Supplemental Oxygen . . . 16

1 Overview

This study guide will help you prepare for the STAT 204 exam covering:

1. **Exploratory Data Analysis (EDA)**: interpreting visualizations and checking findings with **PCS** (predictability, computability, stability)
2. **Linear Regression**: interpreting model output and diagnostics
3. **Analysis of Variance (ANOVA)**: two-way ANOVA with interactions
4. **Logistic Regression**: interpreting coefficients as odds, and predicting probabilities

Exam Format:

- **60 minutes**, in person, with no AI tools or other electronic devices (other than a calculator)
- The exam has the same structure as the practice below (EDA, linear regression, ANOVA, logistic regression) but is **shorter**, and it uses a **different dataset**
- Most questions ask you to **interpret R output** in the context of the data. All code is base R with the native pipe `|>`
- Closed book. You may bring **one formula sheet**: the paper provided by the instructor, hand-written on **one side only**, with **formulas only**. No interpretations, worked examples, or solved problems. Formula sheets are collected and graded with the exam, and adding anything other than formulas is considered academic misconduct.
- Calculator permitted

AI and studying: AI use is encouraged for studying (see the syllabus AI Policy), but solve each problem yourself first. Where AI would fail here: it can give confident but incorrect worked solutions, misread an interpretation, or explain a diagnostic plot it cannot see. Check every AI answer against the class notes. The exam is closed-book with no AI, so practice without it. Do not upload course materials (including this guide) to AI platforms.

All datasets below come from TidyTuesday and are read directly from GitHub, so you can run every chunk yourself.

```
library(tidyverse)
theme_set(theme_minimal(base_size = 11))
```

2 What to Study

2.1 Key Topics

2.1.1 EDA and PCS

- Interpreting boxplots (median, quartiles, outliers, variability)
- Interpreting scatterplots (direction, form, strength, unusual observations)
- **Predictability:** does a finding reemerge in held-out data, and does it agree with domain knowledge?
- **Computability:** can someone else rerun it? (raw data read by code, judgment calls commented in code, `set.seed()` before anything random)

- **Stability:** does a finding survive reasonable changes to judgment calls (cleaning, subsetting) and to the data (bootstrap)?

2.1.2 Linear Regression

- Writing fitted equations and interpreting slope and intercept in context
- Hypothesis tests for coefficients
- Predictions, and what the residual standard error says about their accuracy
- R^2 and what it does (and does not) tell you
- Checking assumptions (linearity, constant variance, normality, independence) from diagnostic plots, and remedies

2.1.3 ANOVA

- Two-way ANOVA model equations
- Predicted values from treatment-coded (`lm`) output: which coefficients to add
- Interpreting main-effect and interaction coefficients, and what the reference levels are
- Interpreting ANOVA tables (F, p-values) and interaction plots

2.1.4 Logistic Regression

- Log-odds (logit) and probability forms
- Interpreting coefficients through odds ratios, e^{β}, including for a c-unit change
- Converting log-odds $\rightarrow$ odds $\rightarrow$ probability
- Significance of predictors

2.2 Formulas to know

These are the kinds of formulas worth putting on your formula sheet. Your sheet must contain formulas only.

- **Simple linear regression:** $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$; $\hat{Y} = b_0 + b_1 X$
- **Two-way ANOVA:** $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$
- **Logistic regression:** $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots, \quad p = \dfrac{e^{\eta}}{1+e^{\eta}}$ with $\eta = \beta_0 + \beta_1 X_1 + \cdots$, odds ratio for a c-unit change $= e^{c\beta_1}$

3 Practice Problems

3.1 Section 1: Exploratory Data Analysis

The first two problems use **Spotify songs** (TidyTuesday, January 21, 2020): audio features for songs from playlists of six genres. `danceability` and `energy` run from 0 to 1, `loudness` is in

decibels (0 dB is the loudest), and `mode` is major or minor. Songs that appeared in several playlists were listed more than once in the raw file.

```
spotify_raw <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
nrow(spotify_raw)
```

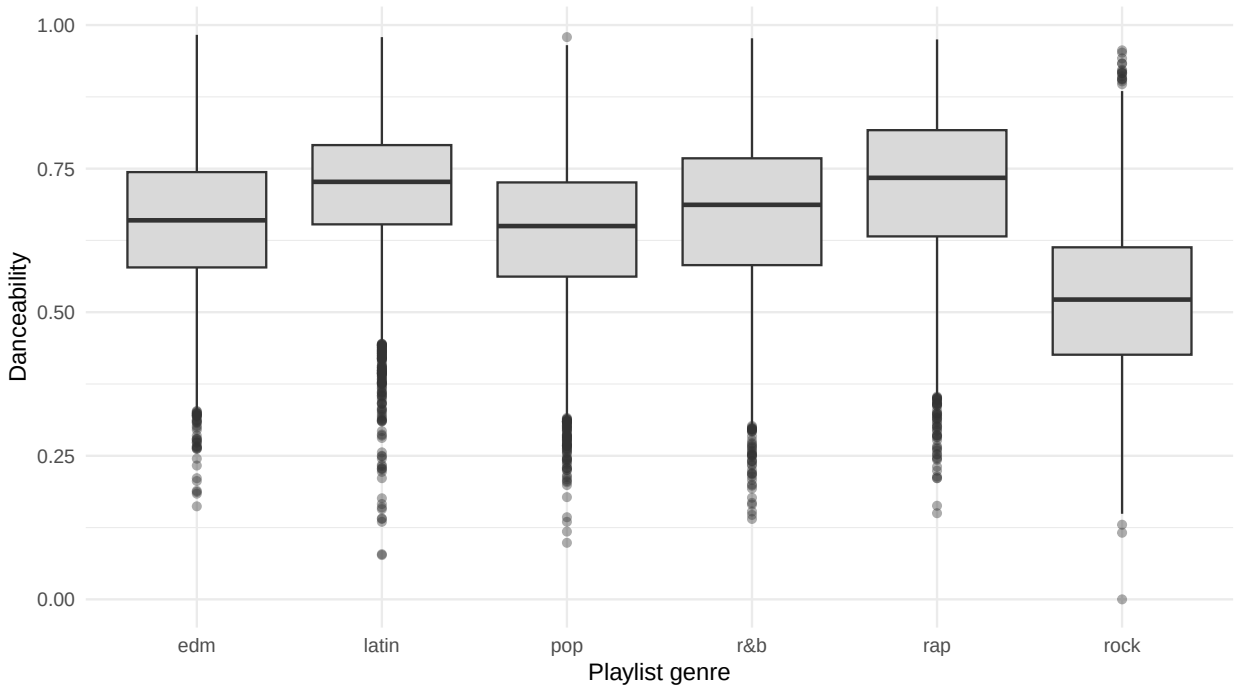
```
[1] 32833
```

```
# JUDGMENT CALL: keep one row per song (track_id); the raw file lists
# a song once for every playlist it appears in.
spotify <- spotify_raw |>
  distinct(track_id, .keep_all = TRUE) |>
  mutate(playlist_genre = factor(playlist_genre),
         mode = factor(mode, levels = 0:1, labels = c("minor", "major")))
nrow(spotify)
```

```
[1] 28356
```

3.1.1 Practice Problem 1.1: Danceability by Genre

```
ggplot(spotify, aes(x = playlist_genre, y = danceability)) +
  geom_boxplot(fill = "grey85", outlier.alpha = 0.4) +
  labs(x = "Playlist genre", y = "Danceability")
```



An analyst reports: *"Rap and latin are the most danceable genres, and rock is the least."* They checked it like this:

```
set.seed(204)
spotify <- spotify |>
  mutate(half = sample(rep(c("explore", "hold out"), length.out = n())))

# median danceability by genre in each half
spotify |>
  group_by(playlist_genre, half) |>
  summarise(median = round(median(danceability), 3), .groups = "drop") |>
  pivot_wider(names_from = half, values_from = median)
```

```
# A tibble: 6 x 3
  playlist_genre explore `hold out`
  <fct>            <dbl>      <dbl>
1 edm              0.659      0.661
2 latin            0.725      0.728
3 pop              0.647      0.652
4 r&b              0.685      0.689
5 rap              0.734      0.734
6 rock             0.518      0.524
```

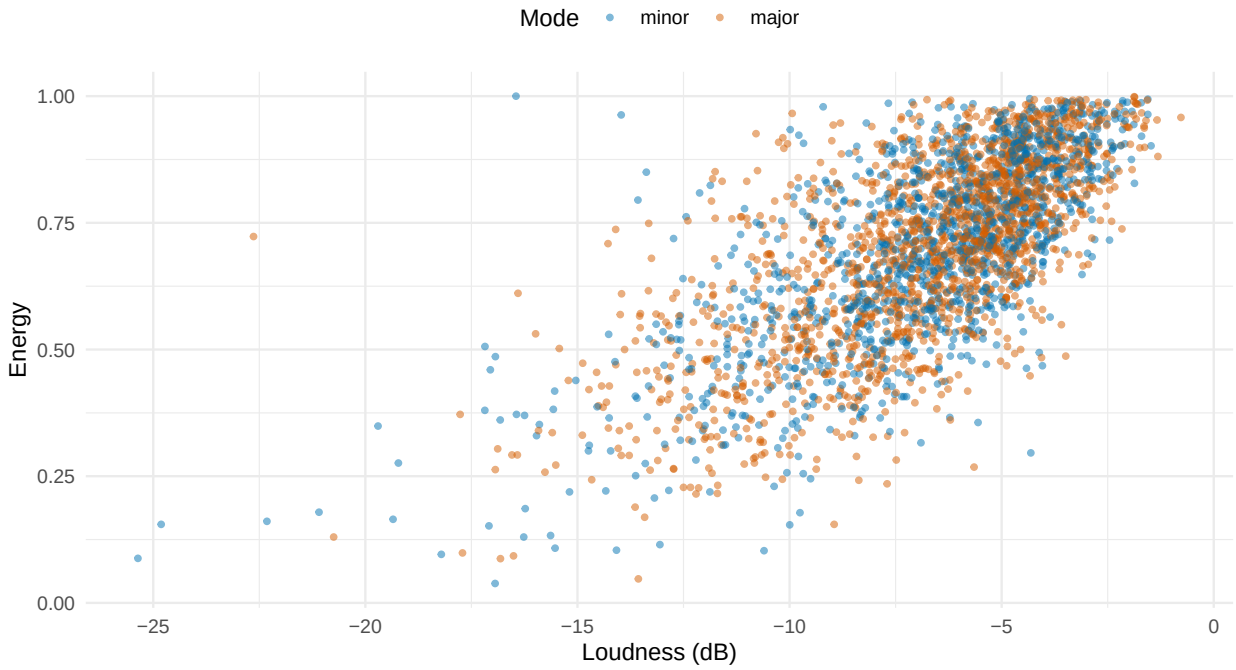
Questions:

(a) Which genre has the highest median danceability? The lowest?
(b) Which genre shows the most variability? Which has the most low outliers?
(c) Which PCS principle does the table above check? Does the finding pass?
(d) The analyst removed duplicate songs before plotting. Is that a stability, predictability, or computability issue? Describe how you would check whether this choice matters.

3.1.2 Practice Problem 1.2: Energy and Loudness

```
set.seed(204)
# JUDGMENT CALL: plot a random 3,000 songs so the points are readable
sub <- slice_sample(spotify, n = 3000)

ggplot(sub, aes(x = loudness, y = energy, color = mode)) +
  geom_point(alpha = 0.5, size = 1) +
  scale_color_manual(values = c("minor" = "#0072B2", "major" = "#D55E00")) +
  labs(x = "Loudness (dB)", y = "Energy", color = "Mode") +
  theme(legend.position = "top")
```



Questions:

(a) Describe the relationship between loudness and energy (direction, form, strength, unusual points).

(b) Does mode (major vs. minor) appear to be associated with energy, once loudness is taken into account?

(c) The plot shows only 3,000 songs. Which line of code makes this figure *computable*? Why would the figure change from run to run without it, and how would you check that the pattern is *stable* to which 3,000 songs were drawn?

3.2 Section 2: Linear Regression

3.2.1 Practice Problem 2.1: Cocoa Percentage and Chocolate Ratings

Chocolate bar ratings (TidyTuesday, January 18, 2022): expert ratings (1 to 4, in steps of 0.25) of 2,530 dark chocolate bars, along with their cocoa percentage.

```
chocolate <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/dat
  mutate(cocoa = parse_number(cocoa_percent))

M1 <- lm(rating ~ cocoa, data = chocolate)
summary(M1)
```

```
Call:
lm(formula = rating ~ cocoa, data = chocolate)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21541 -0.23867  0.03459  0.28459  0.99393

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.02953    0.11209  35.949  < 2e-16 ***
cocoa       -0.01163    0.00156  -7.456 1.22e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

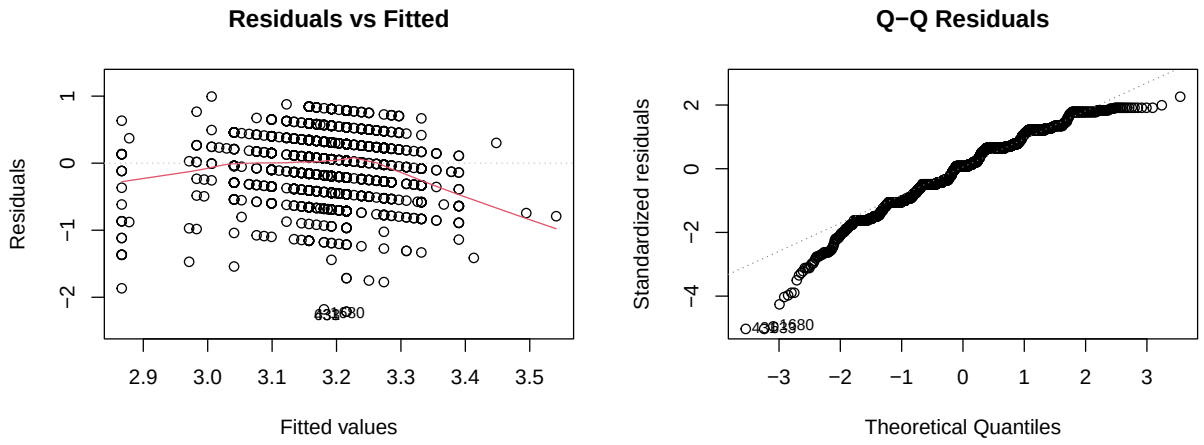
Residual standard error: 0.4406 on 2528 degrees of freedom
Multiple R-squared:  0.02152,   Adjusted R-squared:  0.02113
F-statistic: 55.59 on 1 and 2528 DF,  p-value: 1.218e-13
```

Questions:

(a) Write the fitted regression equation.
(b) Interpret the slope in context, both per 1 percentage point and per 10 percentage points of cocoa.
(c) Predict the rating of a 70% bar. How far off is such a prediction typically?
(d) Write the null and alternative hypotheses for the slope and state your conclusion using $\alpha = 0.05$.
(e) Interpret R^2. Is cocoa percentage a statistically significant predictor? Is it a *useful* one? Explain the difference.

3.2.2 Practice Problem 2.2: Model Diagnostics

```r
par(mfrow = c(1, 2))
plot(M1, which = 1, main = "Residuals vs Fitted", caption = "")
plot(M1, which = 2, main = "Q-Q Residuals", caption = "")
```




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

Questions:

(a) Why do the residuals form diagonal stripes? (Hint: what values can `rating` take?) Is this a violation?
(b) Does the linearity assumption appear satisfied? Look at the smoother at both ends.
(c) Does the normality assumption appear satisfied? Describe the shape of the departure.
(d) Many bars come from the same company. Which assumption does that threaten, and what would you do about it?

3.3 Section 3: Analysis of Variance (ANOVA)

3.3.1 Practice Problem 3.1: Genre, Mode, and Danceability

Using the Spotify data, an analyst keeps three genres and fits a two-way ANOVA with interaction:

```
songs <- spotify |>
  filter(playlist_genre %in% c("pop", "rap", "rock")) |>
  mutate(playlist_genre = droplevels(playlist_genre))

table(songs$playlist_genre, songs$mode)
```

```
       minor major
  pop   2094  3038
  rap   2578  2823
  rock  1323  2982
```

```r
M2 <- lm(danceability ~ playlist_genre * mode, data = songs)
summary(M2)
```

```
Call:
lm(formula = danceability ~ playlist_genre * mode, data = songs)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56901 -0.08499  0.01201  0.09499  0.43476

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.648886   0.002948 220.129  < 2e-16 ***
playlist_genrerap            0.063798   0.003968  16.077  < 2e-16 ***
playlist_genrerock          -0.136508   0.004737 -28.815  < 2e-16 ***
modemajor                   -0.018899   0.003831  -4.933 8.19e-07 ***
playlist_genrerap:modemajor  0.025226   0.005309   4.752 2.03e-06 ***
playlist_genrerock:modemajor 0.027765   0.005876   4.725 2.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1349 on 14832 degrees of freedom
Multiple R-squared:  0.2589,    Adjusted R-squared:  0.2586
F-statistic:  1036 on 5 and 14832 DF,  p-value: < 2.2e-16
```

Questions:

(a) Write the two-way ANOVA model equation. Define all terms.
(b) What is the predicted danceability of a **pop** song in a **minor** key?
(c) What is the predicted danceability of a **rock** song in a **major** key?
(d) Interpret the `modemajor` coefficient. Which genre does it apply to?

3.3.2 Practice Problem 3.2: ANOVA Table Interpretation

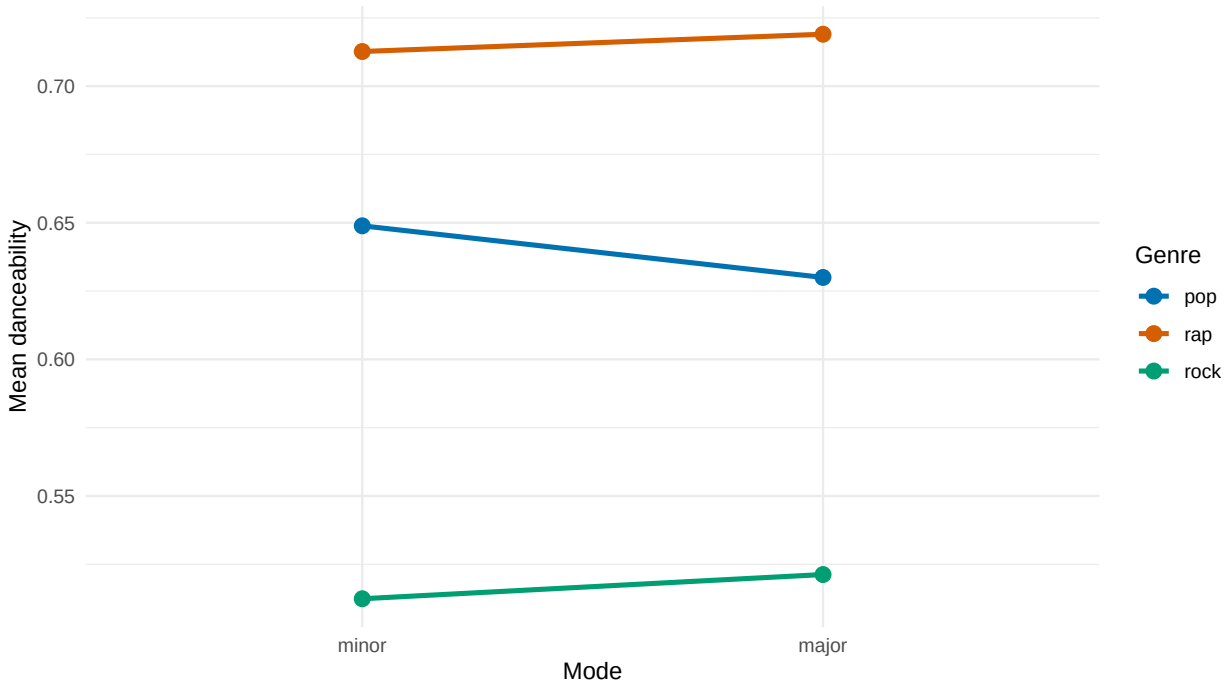
```r
anova(M2)
```

```
Analysis of Variance Table

Response: danceability
                      Df  Sum Sq Mean Sq   F value    Pr(>F)
playlist_genre         2  93.705  46.853 2574.9999 < 2.2e-16 ***
```

```
mode                    1   0.013   0.013    0.7206     0.396
playlist_genre:mode     2   0.556   0.278   15.2681 2.377e-07 ***
Residuals           14832 269.872   0.018
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
songs |>
  group_by(playlist_genre, mode) |>
  summarise(mean_dance = mean(danceability), .groups = "drop") |>
  ggplot(aes(x = mode, y = mean_dance, color = playlist_genre, group = playlist_genre)) +
  geom_line(linewidth = 1.1) +
  geom_point(size = 3) +
  scale_color_manual(values = c("#0072B2", "#D55E00", "#009E73")) +
  labs(x = "Mode", y = "Mean danceability", color = "Genre")
```



Questions:

(a) Write the hypotheses for the interaction test. What is your conclusion using $\alpha = 0.05$?

(b) The main effect of mode is *not* significant, but the interaction is. Using the interaction plot, explain how both can be true.

(c) The interaction is highly significant, but the lines in the plot differ by only about 0.02–0.03 on a 0–1 scale. Why can such a small difference be "significant" here? Is it practically important?

3.4 Section 4: Comprehensive Practice Problems

3.4.1 Practice Problem 4.1: Predicting Energy from Loudness

```r
M3 <- lm(energy ~ loudness, data = spotify)
summary(M3)
```

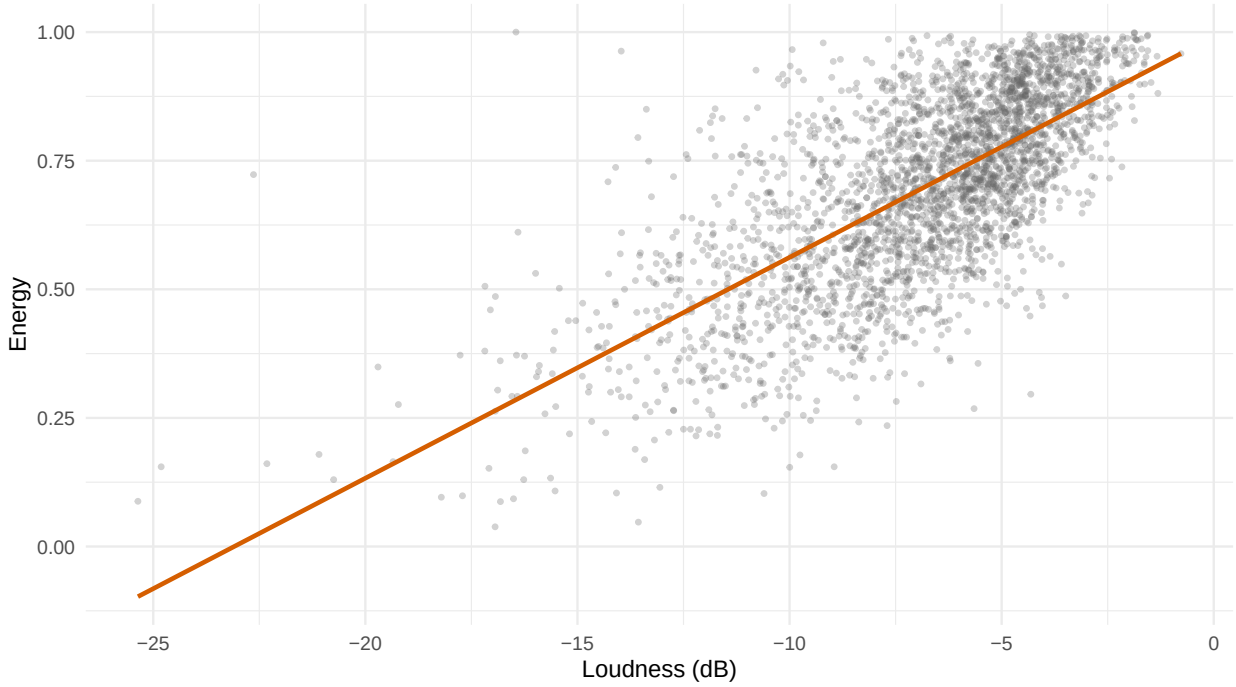
```
Call:
lm(formula = energy ~ loudness, data = spotify)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54304 -0.08864  0.00586  0.09173  1.36044

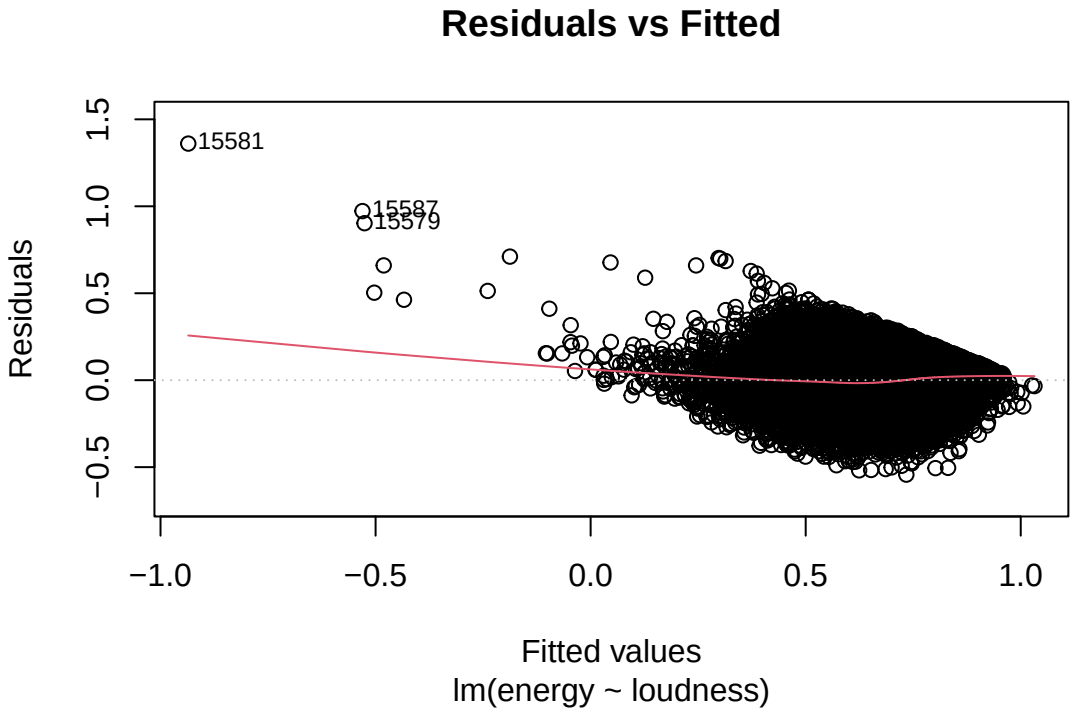
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9794588  0.0019587   500.0   <2e-16 ***
loudness    0.0412267  0.0002625   157.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1342 on 28354 degrees of freedom
Multiple R-squared:  0.4653,    Adjusted R-squared:  0.4653
F-statistic: 2.468e+04 on 1 and 28354 DF,  p-value: < 2.2e-16
```

```r
ggplot(sub, aes(x = loudness, y = energy)) +
  geom_point(alpha = 0.3, size = 0.8, color = "grey40") +
  geom_smooth(method = "lm", se = FALSE, color = "#D55E00") +
  labs(x = "Loudness (dB)", y = "Energy")
```




```r
plot(M3, which = 1, main = "Residuals vs Fitted", caption = "")
```



Questions:

(a) Write the fitted equation and interpret both coefficients. Is the intercept meaningful here?

(b) Predict the energy of a song at -6 dB and of a song at -40 dB. Is either prediction a problem?
(c) What fraction of the variation in energy is explained by loudness?
(d) The residuals vs. fitted plot has a sharp diagonal upper edge. What feature of the response causes it, and which assumption does it affect?

3.4.2 Practice Problem 4.2: Ramen Ratings by Style and Country

Ramen ratings (TidyTuesday, June 4, 2019): star ratings (0 to 5) from The Ramen Rater. We keep three styles and three countries:

```
ramen <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2
  filter(style %in% c("Pack", "Bowl", "Cup"),
         country %in% c("Japan", "South Korea", "United States"),
         !is.na(stars)) |>
  mutate(style = factor(style, levels = c("Bowl", "Cup", "Pack")),
         country = factor(country))

M4 <- lm(stars ~ style * country, data = ramen)
summary(M4)
```

```
Call:
lm(formula = stars ~ style * country, data = ramen)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0074 -0.3705  0.0995  0.6615  1.6295

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.00743    0.07063  56.736  < 2e-16 ***
styleCup                      -0.26993    0.11924  -2.264 0.023781 *
stylePack                     -0.16897    0.09860  -1.714 0.086863 .
countrySouth Korea            -0.17761    0.12692  -1.399 0.161971
countryUnited States          -0.49524    0.12745  -3.886 0.000108 ***
styleCup:countrySouth Korea   -0.03766    0.21412  -0.176 0.860403
stylePack:countrySouth Korea   0.23963    0.15893   1.508 0.131891
styleCup:countryUnited States  0.12827    0.19095   0.672 0.501874
stylePack:countryUnited States 0.29022    0.16497   1.759 0.078804 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

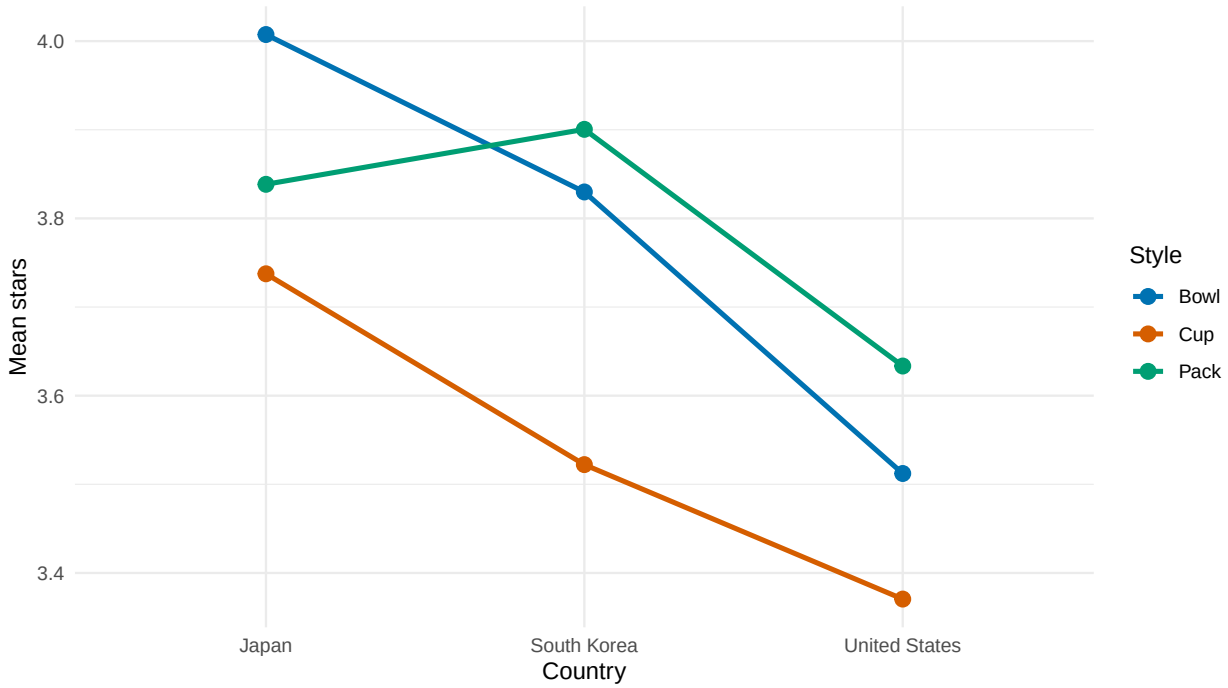
Residual standard error: 0.9607 on 1122 degrees of freedom
Multiple R-squared:  0.0369,    Adjusted R-squared:  0.03004
F-statistic: 5.374 on 8 and 1122 DF,  p-value: 1.189e-06
```

```r
anova(M4)
```

```
Analysis of Variance Table

Response: stars
                Df  Sum Sq Mean Sq F value    Pr(>F)
style            2   13.00  6.4985  7.0409 0.0009145 ***
country          2   22.18 11.0915 12.0173 6.856e-06 ***
style:country    4    4.50  1.1251  1.2190 0.3009325
Residuals     1122 1035.57  0.9230
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ramen |>
  group_by(country, style) |>
  summarise(mean_stars = mean(stars), .groups = "drop") |>
  ggplot(aes(x = country, y = mean_stars, color = style, group = style)) +
  geom_line(linewidth = 1.1) +
  geom_point(size = 3) +
  scale_color_manual(values = c("#0072B2", "#D55E00", "#009E73")) +
  labs(x = "Country", y = "Mean stars", color = "Style")
```



Questions:

(a) What is the predicted rating for a **Bowl** from **Japan**?
(b) What is the predicted rating for a **Pack** from the **United States**?

(c) Is there evidence of an interaction between style and country? Which main effects are significant, and what do they mean in context?

3.5 Section 5: Logistic Regression

3.5.1 Practice Problem 5.1: Age and Summiting Everest

Himalayan expeditions (TidyTuesday, September 22, 2020): we use members of Everest expeditions since 2000, excluding hired staff. `success` is TRUE if the person reached the summit.

```
everest <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/
  # JUDGMENT CALL: Everest only, since 2000, excluding hired staff
  filter(peak_name == "Everest", year >= 2000, !hired, !is.na(age))
nrow(everest)
```

```
[1] 9139
```

```
M5 <- glm(success ~ age, data = everest, family = binomial)
summary(M5)
```

```
Call:
glm(formula = success ~ age, family = binomial, data = everest)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.697636   0.082725   8.433   <2e-16 ***
age         -0.020760   0.001984 -10.462   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12624  on 9138  degrees of freedom
Residual deviance: 12512  on 9137  degrees of freedom
AIC: 12516

Number of Fisher Scoring iterations: 4
```

```
exp(coef(M5))
```

```
(Intercept)         age
  2.0089985   0.9794541
```

Questions:

(a) Write the model in log-odds (logit) form.
(b) Interpret the age coefficient in terms of odds, per year and per 10 years of age.
(c) Predict the probability that a 40-year-old climber summits.
(d) Is age a significant predictor at $\alpha = 0.05$?

3.5.2 Practice Problem 5.2: Adding Supplemental Oxygen

```r
M6 <- glm(success ~ age + oxygen_used, data = everest, family = binomial)
summary(M6)
```

```
Call:
glm(formula = success ~ age + oxygen_used, family = binomial,
    data = everest)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.986736   0.161078 -18.542  < 2e-16 ***
age             -0.019934   0.002934  -6.794 1.09e-11 ***
oxygen_usedTRUE  4.958756   0.116946  42.402  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12623.8  on 9138  degrees of freedom
Residual deviance:  6714.5  on 9136  degrees of freedom
AIC: 6720.5

Number of Fisher Scoring iterations: 6
```

```r
exp(coef(M6))
```

```
    (Intercept)             age oxygen_usedTRUE
     0.05045185      0.98026383    142.41651438
```

Questions:

(a) What are the predicted log-odds of summiting for a 40-year-old who does **not** use oxygen?
(b) What is the predicted probability of summiting for a 40-year-old who **does** use oxygen?

(c) Interpret the `oxygen_usedTRUE` odds ratio in context. Can we conclude that oxygen *causes* success? Name one alternative explanation.

(d) The age coefficient barely changes between M5 and M6. What does that suggest about the relationship between age and oxygen use?