STAR 513: Final Exam

Spring 2025

Yvette Uwineza

Instructions:

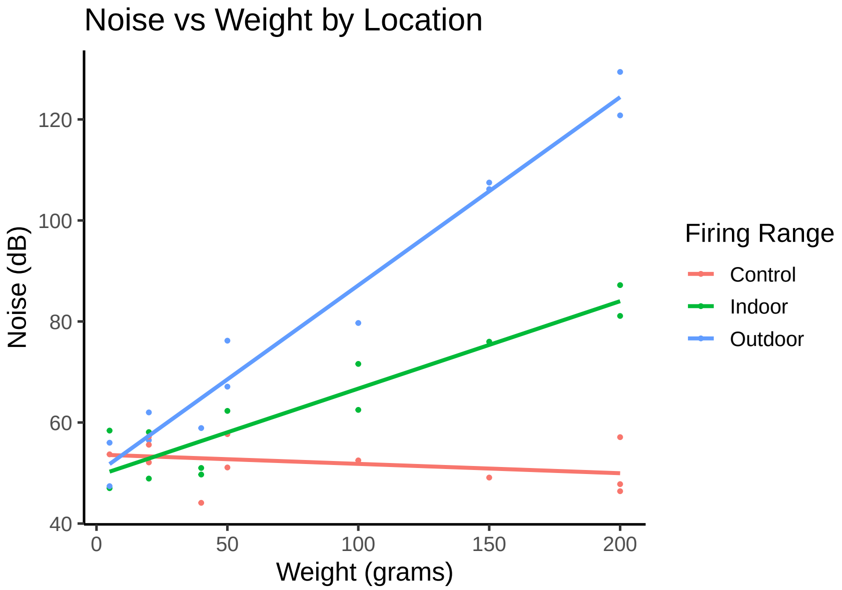
- This exam is due by Tues 5/13 at midnight.
- **Students are required to work independently on the exam.** Do NOT discuss the exam with anyone else (including other students).
- You may use the textbook, class notes, examples, HW solutions posted in the current Canvas course. You may use any other publicly available (print or online) statistics references or resources that you find helpful. Use of homework "helper" websites (ex: Chegg, NoteHall, etc) is NOT allowed. Use of chatbots (ex: ChatGPT) is NOT allowed.
- Knit frequently to avoid last minute problems. You may add or delete code chunks as needed. **It is the student's responsibility to check the knitted document (for correctness and completeness) before submitting.**
- For any questions that require calculations, you should provide R code for full credit.
- For some questions, there may be more than one possible answer, analysis or graph that could be used for full credit. **Choose one approach**, making a reasonable choice and justifying if needed.
- Given this is the final exam, you should present your best work. I will deduct points for things like printing full data to knitted document, unreadable tables, unclear, excess or unnecessary output, etc.
- Use $\alpha = 0.05$ and/or 95% confidence where needed.
- All questions are worth 4 points except where noted. Maximum score is 92.
- **I believe all students can do well on this exam. Please don't cheat!!!**

Firing Range (Q1 - Q8)

A study was done to examine noise exposure at police firing ranges. The primary question was whether the response variable **Noise** (measured in dB) differed based on the shot **Weight** (measured in grams) and **Location** ("Indoor" firing range, "Outdoor" firing range, sound proof "Control" box). Information was recorded for a total of n=36 shots across all 3 Locations.
This data is available from Canvas as `FiringRange.csv`.

Q1

Create a scatterplot of Noise (y) vs Weight (x) color coded by Location. Overlay separate regression lines for each Location. Your plot should include axis labels with units (where applicable).



Q2

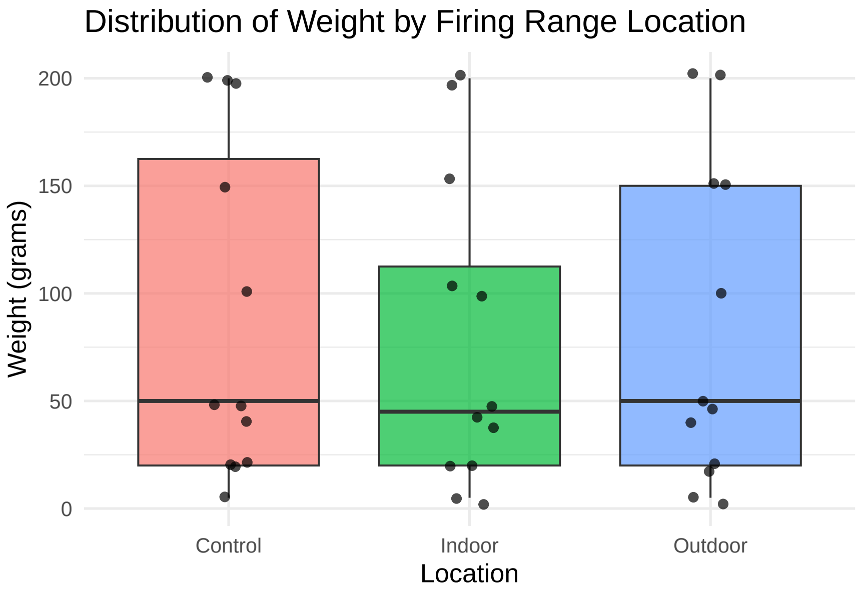
Based on your plot from the previous question, do you see visual evidence of an **interaction** between Location and Weight? Your discussion should be brief, but still be specific.

Yes, the scatterplot with separate regression lines for each location provides evidence of an interaction between location and weight. The slopes of the regression lines are different between indoor and outdoor ranges. This non parallel pattern of lines suggest a potential interaction effect.

Q3

People sometimes confuse "interaction" and "correlation". Create a summary plot to look at the association ("correlation") between Weight (y) and Location (x). Do you see visual evidence of an **association** between Location and Weight?

There isn't a strong evidence of association between Location and weight. There is little visual evidence based on the plots below.



Q4

Fit an appropriate model using Noise (y) as the response. Include Location, Weight and interaction as predictors. Show the coefficients table AND a Type 3 ANOVA table.

```
## # A tibble: 6 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      51.0         1.25      40.9   7.18e-28
## 2 Weight            0.176       0.0114    15.4   8.98e-16
## 3 Location1         2.67        1.77       1.51  1.43e- 1
## 4 Location2        -1.59        1.75      -0.906 3.72e- 1
## 5 Weight:Location1 -0.194       0.0158   -12.3   2.99e-13
## 6 Weight:Location2 -0.00262     0.0165    -0.159 8.75e- 1
```

```
## Anova Table (Type III tests)
##
## Response: Noise
##                 Sum Sq Df   F value    Pr(>F)
## (Intercept)      40223  1 1671.9001 < 2.2e-16 ***
## Weight            5692  1  236.5904 8.977e-16 ***
## Location            55  2    1.1494    0.3304
## Weight:Location   4893  2  101.6878 4.325e-14 ***
## Residuals          722 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q5

What Location (Indoor, Outdoor or Control) is being used as the "reference" category? Why is this level used?

The reference category is Control.

R orders factor levels alphabetically, unless we change it.

Q6 (6 pts)

For each Location, provide the estimated intercept and slope (corresponding to Weight). For this question you should use the interaction model from Q4 to provide numeric values for each estimate.

Control:
Estimated Intercept = 50.992
Estimated Slope = 0.176

Indoor:
Estimated Intercept = 2.668
Estimated Slope = -1.588

Outdoor:
Estimated Intercept = -0.194
Estimated Slope = -0.003

Q7

Briefly discuss how you can (informally) check your estimated intercepts OR slopes using the graph from Q1. Your discussion should be brief, but still be specific.

---

we can use the regression lines we fit in the scatterplot from question 1. The slopes can be checked by looking at the steepness of the lines and whether they are upward or downward sloping. The intercepts moves the line up or down.

Q8

Should we consider dropping the interaction from this model? Briefly discuss.

Should we consider dropping either of the main effects (Location, Weight) from this model? Briefly discuss.

Interaction: No, we should not drop the interaction. We sass wht the pvalue from the type 3 anova corresponding to the interaction effect was <0.001. This means that the relationship between weight and noise differs by location

Main Effects: we should not drop weight since it is highly significant with a pvalue of <0.001. We should keep location, because even though it is not highly significant, the principle of hierachy tells us that when an interaction is included, the corresponding main effects my also be retained.

MS Activity (Q9 - Q17)

Researchers are interested in identifying variables that are associated with physical activity in people with multiple sclerosis (MS). A total of n = 34 subjects were included in this study. The response variable (y) is Total.Activity (minutes/day), measured via Actigraph accelerometer. Several potential predictor variables are considered:

- Walk.Speed: Walking speed (m/s)
- Peg.Test: Time to complete peg test requiring manual dexterity (seconds)
- Chair.Rise: Time to rise from a seated position (seconds)
- TUG: "timed up and go" to rise from a chair, walk a set distance and return (seconds)
- LA.TotStr: Strength of the "less affected" leg (N/kg)
- MA.TotStr: Strength of the "more affected" leg (N/kg)

This data is available from Canvas as `MSActivity.csv`.

Q9

Calculate pairwise (Pearson) correlations between all variables. This should be presented as a 7 × 7 matrix, rounded to 2 decimal places.
Based on your result, what predictor has the strongest correlation with Total.Activity (positive or negative)? Mention at least one other notable finding based on correlations.

Response

```
##                Total.Activity Walk.Speed Peg.Test Chair.Rise   TUG LA.TotStr
## Total.Activity           1.00       0.55    -0.40      -0.58 -0.45      0.51
## Walk.Speed               0.55       1.00    -0.62      -0.80 -0.84      0.50
## Peg.Test                -0.40      -0.62     1.00       0.64  0.72     -0.18
## Chair.Rise              -0.58      -0.80     0.64       1.00  0.91     -0.50
## TUG                     -0.45      -0.84     0.72       0.91  1.00     -0.41
## LA.TotStr                0.51       0.50    -0.18      -0.50 -0.41      1.00
## MA.TotStr                0.49       0.60    -0.31      -0.61 -0.55      0.90
##                MA.TotStr
## Total.Activity      0.49
## Walk.Speed          0.60
## Peg.Test           -0.31
## Chair.Rise         -0.61
## TUG                -0.55
## LA.TotStr           0.90
## MA.TotStr           1.00
```

Q10 (2 pts)

Fit the full model, including all 6 predictors. Do NOT show the summary() output! Rather, show the (multiple) R2 value.

R2 =

```
## [1] 0.499107
```

Q11

Using the full model from the previous question, provide the variance inflation factors. Which predictor has the largest vif value? Briefly discuss why this predictor has high vif (even if it does not meet the “rule of thumb”).

Response

```
## Walk.Speed   Peg.Test Chair.Rise        TUG  LA.TotStr  MA.TotStr
##   3.874243   2.178964   6.851501   9.123130   5.791174   6.680443
```

The predictor with the highest value is TUG (Timed Up and Go test), it is likely to be highly correlated with other mobility and strength measures which assess related aspects of physical function. The rule of them is 10 showing that this vif has moderated multicollinearity among predictors.

Q12 (6 pts)

Use **AIC all subsets** selection to choose a model. Consider only additive models (no interactions). For the selected model show the (multiple) R2 value and the coefficients table.

```
#Q12


MS_full_additive <- lm(Total.Activity ~ Walk.Speed + Peg.Test + Chair.Rise + TUG + LA.TotStr + MA.TotSti


options(na.action = "na.fail")
all_models <- dredge(MS_full_additive, rank = "AIC")


best_model <- get.models(all_models, 1)[[1]]


summary(best_model)$r.squared
```

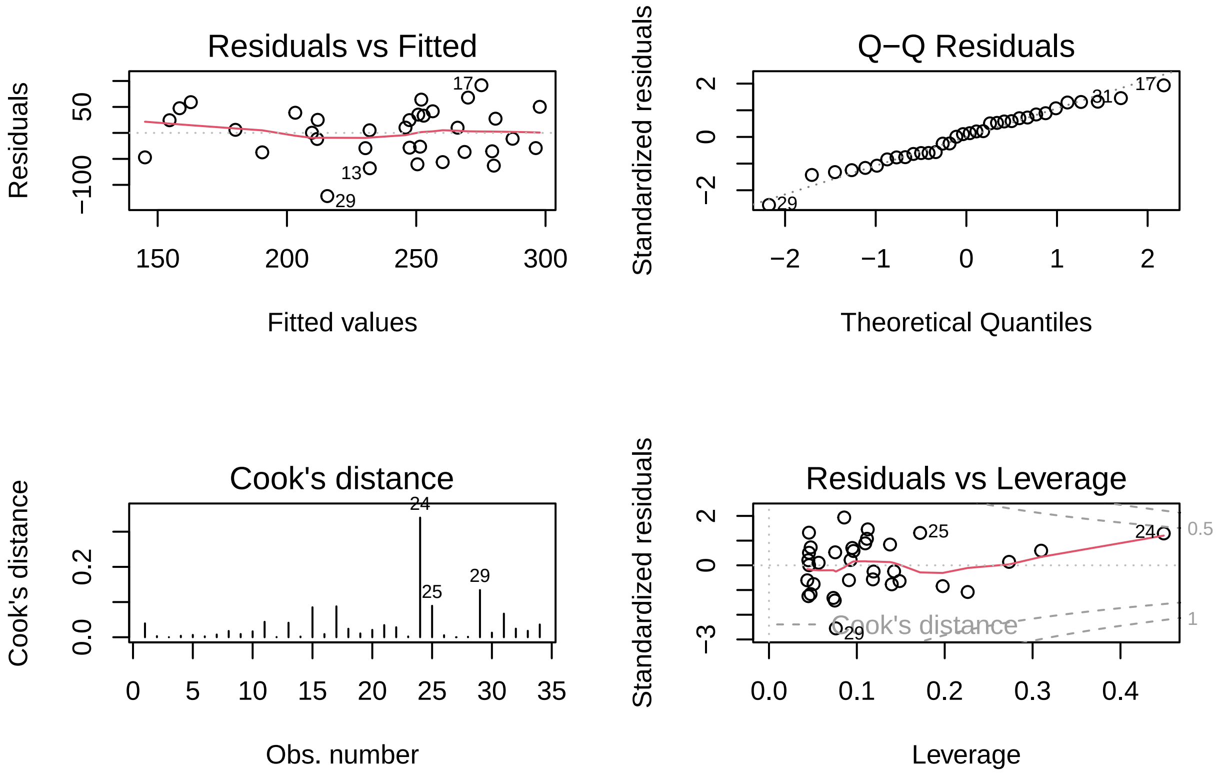
```
## [1] 0.4400385
```

```
tidy(best_model)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   148.        88.5       1.67 0.104
## 2 Chair.Rise     -9.06       3.28     -2.76 0.00979
## 3 TUG            12.3        5.85      2.10 0.0442
## 4 Walk.Speed     69.8       39.3       1.77 0.0861
```

Q13

Using your selected model from the previous question, provide diagnostic plots: (1) Residuals vs Fitted, (2) QQplot of Residuals, (4) Cook's Distance and (5) Residuals vs Leverage.



Q14 (8 pts)

State the four assumptions of linear regression (in words) and briefly discuss whether each assumption is satisfied for your selected model. Where applicable, your discussion should mention evidence from a specific plot (graph) from the previous question.

Assumption 1: Independence: in this study, each row represents a different participant. So, the observations are independent.

Assumption 2: Equal Variance: Looking athe the residual vs Fitted plot, we have some spread variability at higher fitted values, but no clear funnel shape. So, the equal variance is satisfied.

Assumption 3: Linearity: We don’t have strong curvature as we look at the residuals vs fitted plot. So, the linear assumption is sort of satisfied.

Assumption 4: Normality of Residuals: the Q-Q plot shows that the residuals follows the theoreticall ine, so it is satisfied as well.

Q15

Using your selected model, provide a detailed interpretation of the estimated coefficient corresponding to **Chair.Rise** in context of this study. Your interpretation should include appropriate units (where applicable) and the numeric value for the estimated coefficient.

Holding other variables constant, an increase in one second in Chair Rise time leads an estimated decrease of 9.06 units in total activity.

Q16

Using your selected model, do we have statistical evidence of a linear association between Total.Activity and **Chair.Rise**? Provide a brief response in context, including direction and appropriate statistical evidence (ex: test statistic and p-value or confidence interval.)

Yes, we have evidence of linear association between chair rise time and total activity. The association is negative, the t stat is -2.76 and the p_value is 0.0098 < 0.05

Q17

Consider your selected model from Q12, could this model have been selected by **backwards elimination using p-values** (with alpha = 0.05)? Your discussion should be brief, but still be specific.

Using Backward elimination, only variables with a pvalue less that 0.05 would stay. So, this model would not have been selected by backward elimination using p-values. Walk.speed has a pvalue of 0.086, and it would have been dropped.

Fasting (Q18 - Q22)

A study was done to examine the effect of fasting (not eating) on brain function.

Investigators considered two Age groups: Younger (20-30 years old, n = 15) and Older (50-60 years old, n = 15). Hence, there were a total of n = 30 volunteer participants. Investigators suspect that the effect of fasting (on BDNF) may be different between the Age groups.

All subjects had blood samples taken at three Time points: 0 hours (immediately after a light breakfast), 12 hours (fasting period, with no food consumed since breakfast) and 24 hours (after normal eating resumed).

BDNF (Brain-Derived Neurotrophic Factor) was measured in ng/ml from each of the blood samples. Higher values of BDNF correspond to better brain function.

Hence the data would include the following variables:

- Subject: 1-30
- Age: Older or Younger
- Time: 0, 12, or 24 hours
- BDNF: measured in ng/ml

Notes:

- No data is provided! You have to think about the analysis based on the description above.
- This study is similar to things we have seen in the course. BUT there is not an exact matching example in the notes.

Q18 (2 pts)

What is the total number of rows in the data?

Each subject gets measured 3 Time points: 30 subjects x 3 (Time points per subject) = 90 rows.

Q19

Explain why a mixed model (ex: lmer) is appropriate here. Your response should specifically address what assumption of a linear model (ex: lm) is NOT satisfied for this data.

A mixed model is appropriated here because the same subjects are measured at multiple time points, creating non-independent observations. The usual lm assumes that ll observation are independent which is violated in our case where we have repeated models. A mixed model addresses this by including a random effects for subject, which adjusts for the correlation among measurements from the same individual.

Q20

What is the response variable? What are the fixed effect(s)? What are the random effect(s)? No need to justify.

Response: BDNF (ng/ml)

Fixed effect(s): Age, Time, Age x Time interaction

Random effect(s): Subject

Q21

Would you include a Age:Time interaction in your model? Briefly justify your response.

Yes, I would include the interaction because the study tell us that the effect of fasting on BDNF levels may differ between Age groups.

Q22

In the default data format, Time (0, 12, 24) is numeric. But of course, we can convert Time to a factor (categorical predictor). Discuss the impact of treating Time as numeric vs as factor. For example, how would this change the interpretation of parameters?

If time is considered as numeric, the model assumes a linear relationship between Time and BDNF. The parameter estimate for Time would be interpreted as the expected change in BDNF per one hour increase. On the other hand, if time is treated as a factor, the model won't take into consideration change across Time points. Each level time is compared to a reference group.

Appendix

```r
#Retain this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
library(tidyverse)
library(broom)
library(car)
library(emmeans)
library(MuMIn)
#install.packages("MuMIn")

# loading the dataset

FiringRange <- read_csv("~/star_513/Final_Exam/FiringRange.csv")

#glimpse(FiringRange)

MSActivity <- read_csv("~/star_513/Final_Exam/MSActivity.csv")

#glimpse(MSActivity)
```

```
#Q1

ggplot(FiringRange, aes(x = Weight, y = Noise, color = Location)) +
  geom_point(size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(group = Location)) +
  labs(
    x = "Weight (grams)",
    y = "Noise (dB)",
    title = "Noise vs Weight by Location",
    color = "Firing Range "
  ) +
  theme_classic(base_size = 14)

#Q3


ggplot(FiringRange, aes(x = Location, y = Weight, fill = Location)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.1, size = 2, alpha = 0.7) +
  labs(
    x = "Location",
    y = "Weight (grams)",
    title = "Distribution of Weight by Firing Range Location"
  ) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")

#Q4


options(contrasts = c("contr.sum", "contr.poly"))


Firing_Model2 <- lm(Noise ~ Weight * Location, data = FiringRange)


tidy(Firing_Model2)


Anova(Firing_Model2, type = 3)

#Q9

cor_matrix <- MSActivity %>%
  cor() %>%
  round(2)
cor_matrix

#Q10

MS_full_model <- lm(Total.Activity ~ Walk.Speed + Peg.Test + Chair.Rise + TUG + LA.TotStr + MA.TotStr,
```

```r
summary(MS_full_model)$r.squared


#Q11

vif_values <- vif(MS_full_model)
vif_values
#Q12


MS_full_additive <- lm(Total.Activity ~ Walk.Speed + Peg.Test + Chair.Rise + TUG + LA.TotStr + MA.TotSt


options(na.action = "na.fail")
all_models <- dredge(MS_full_additive, rank = "AIC")


best_model <- get.models(all_models, 1)[[1]]


summary(best_model)$r.squared

tidy(best_model)

#Q13


par(mfrow = c(2, 2))


plot(best_model, which = c(1, 2, 4, 5))
```