HW3 - Factors and friends

STUDENT NAME

02 March, 2024

As a reminder-

- Carefully read all questions. If there are errors/poorly worded questions, ask for clarification.

- Plan your approach and write (in comments) a step by step plan.

- Translate your plan into code, testing as you progress. Resolve errors or issues as you go.

- Write your (non-code) answers based on your results (after the code and your tables, please)

- Knit your rmd to pdf.

- Carefully review your pdf. Ensure all tables, figures, code, formulae, etc. fit on the page, are clean, clear, and easy to read, and are appropriate to the question asked; do not just dump everything in a tibble if the question only asks for a subset of it. E.g., if asked for the estimates for our βs, reduce the tibble you pass to kable to only the name and the value.

Again, be very sure to polish and refine your code and final document. Make 'beautiful' code that compiles to a beautiful PDF: Use sensible variable names, include concise yet meaningful comments in your code where appropriate, and constrain yourself to using tidyverse functions for your 'data carpentry', plotting, etc. This practice helps solidify what you've learned.

Please delete the note above and this sentence for your own assignment files

Problem 1

You are researching how GAD65 single nucleotide polymorphisms modify the enzyme's function with a goal of understanding a putative link to a broad spectrum of psychiatric conditions. You run an experiment in which you introduced mutations that result in single amino acid changes. To evaluate these changes, you measure enzyme activity in droplet microarray wells with a consistent concentration of the enzyme with a fluid substate containing both glutamate (GAD65's substrate) and pyridoxal phosphate drawn from a separate experiment. The substrate contains variable concentrations of glutamate (enzymatic substrate) and pyridoxal phosphate (enzyme cofactor).

1.A Load data and prepare data for analysis

```
# Reading in the data
raw_gad65_data <- readxl::read_excel("./data/HW3data_.xlsx")

## New names:
## * `` -> `...4`
## * `` -> `...6`
## * `` -> `...8`
# Turning categorical variables into factors
fixed_gad65_data <- mutate(raw_gad65_data,
```

```
across(c(WellID, `Plate ID`, `Enzyme Variant`),
as_factor))
```

1.B Considering *only* the enzyme variant, conduct an ANOVA analysis: (10 pt)

1.B.1 Write out reasonable scientific hypothesis in clear terms that this approach can test (2 pt)

Answer: There is a significant difference in the enzyme activity of GAD65 enzyme variants caused by single nucleotide polymorphisms.

This implies that the mutations introduced result in measurable changes in enzyme function, which could be linked to the enzyme's role in neurotransmitter regulation and, by extension, to psychiatric conditions. The observed differences in enzyme activity among variants suggest a potential functional impact of specific SNPs on GAD65, warranting further investigation into their role in the etiology of psychiatric disorders.**

1.B.2 Write out your statistical hypothesis (null hypothesis and alternate) and the formula you are fitting. (2 pt)
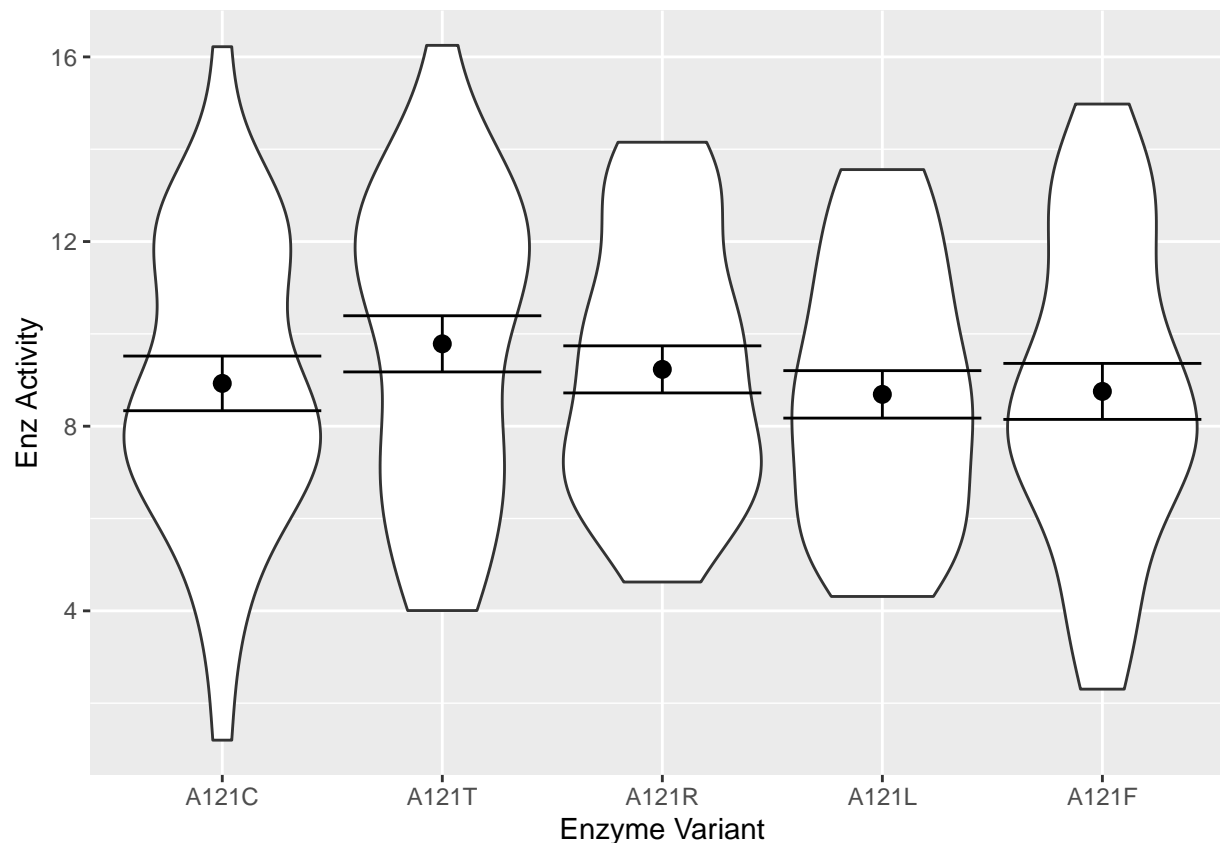
$$H_0 : ...$$

(beta also ok to use, using B for simplicity)

.

1.B.3 Make a figure which shows both the 'raw' data and summary relevant to this model/analysis (3 pt)

```
fixed_gad65_data |> ggplot(aes(x=`Enzyme Variant`, y=`Enz Activity`)) +
  geom_violin() +
  stat_summary(fun = mean) +
  stat_summary(geom = "errorbar", fun.data = mean_se)
```

```
## Warning: Removed 5 rows containing missing values (`geom_segment()`).
```

1.B.4 Fit your model; identify and interpret the results (3 pt (1pt each for: model, results and interpretation)). Use Anova from car for your ANOVA table.

```
variant_to_activity_model = lm(`Enz Activity` ~ `Enzyme Variant`, fixed_gad65_data)

tidy_variant_to_activity_model = tidy(variant_to_activity_model)
kable(tidy_variant_to_activity_model)
```

term	estimate	std.error	statistic	p.value
(Intercept)	8.92733	0.56801	15.71675	0.00000
Enzyme VariantA121T	0.85685	0.80329	1.06667	0.28789
Enzyme VariantA121R	0.30361	0.80329	0.37795	0.70602
Enzyme VariantA121L	-0.23785	0.80329	-0.29609	0.76758
Enzyme VariantA121F	-0.17366	0.80329	-0.21619	0.82915

```
car_variant_to_activity_model = car::Anova(variant_to_activity_model)
kable(car_variant_to_activity_model)
```

	Sum Sq	Df	F value	Pr(>F)
Enzyme Variant	24.027	4	0.6206	0.64853
Residuals	1403.482	145	NA	NA

text

1.C Considering only the enzyme variant and glutamate concentration, fit a GLM model that includes both: (15 pt)

1.C.1 Write out a reasonable scientific hypothesis in clear terms (2 pt)

text

1.C.2 Write out your statistical hypothesis and the formula you are fitting. (3 pt)

The hypotheses we care about are:

$$H_0 : ...$$

.

1.C.3 Make a figure which includes all relevant variables relevant to this model and *only* those variables (5 pt)

1.C.4 Fit your model; present the key components in one or more concise, well-formatted table(s), and interpret the results (5 pt)

text

1.D. Considering only the enzyme variant and pyridoxal phosphate concentration, fit a new GLM and interpret it: (15 pt)

1.D.1 Write out a reasonable scientific hypothesis in clear terms (2 pt)

text

1.D.2 Write out your statistical hypothesis and the formula you are fitting. (3 pt)

The hypothesis we care about is still:

$$H_0 :$$

$$\hat{Y} =$$

1.D.3 Make a figure which shows all relevant variables relevant to this model and *only* those variables (5 pt)

1.D.4 Fit your model; identify, present and interpret the results (5 pt)

text

1.E. Using your models from C and D, calculate activity after removing the estimated contribution of (1) glutamate and (2) pyridoxal phosphate concentrations (note: create 2 new variables). Recreate figure 1.B.3 for each of these with a separate facet for each (will require some data carpentry). (10 pt)

1.E.1 Describe this figure and provide an interpretation of each facet. (5 pt)

text

1.E.2 How does that interpretation relate to your interpretation of the statistical models you ran above?. (5 pt)

text

1.F. Considering the enzyme variant, glutamate and pyridoxal phosphate concentrations, conduct an ANCOVA/General Linear Model analysis: (40 pt)

1.F.1 Write out a reasonable scientific hypothesis in clear terms (3 pt)

text

1.F.2 Write out your statistical hypotheses and the formula you are fitting. (4 pt)

1.F.3 Fit your model; present and interpret the results (5)

text

1.F.4 Using your model from F, calculate activity after removing the estimated (model fitted) contribution of only glutamate and pyridoxal phosphate concentrations and both substrate and pyridoxal phosphate concentrations (note: 3 separate new adjusted-activity variables). Recreate figure 1.B.3 for each of these with a separate facet for each. Discuss (compare and contrast) to your figure from E. (16 pt)

text

1.F.5 Calculate the mean values for each factor level after correcting for only glutamate, only pyridoxal phosphate, and glutamate and pyridoxal phosphate using the model from F. Compare these values to the estimates from your model for each factor level. (12 pt)

text

1. Bonus! : Go back (revisit, do not duplicate here) to your plots above and use [theme] (https://ggplot2.tidyverse.org/reference/theme.html) (and any other ggplot functions you care to use) to polish them by adjusting theme and mapping characteristics (e.g., the fonts, font sizes, colors, color scales, etc.) of your plots. List each change, indicate what you changed, and clearly yet concisely explain the benefit of each change. (up to +5 pt; half a point per well-justified change)