# HW3 - Factors and friends

Ryo Iwata

11 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 $\beta s$, 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

```r
# Reading in the data
raw_gad65_data <- readxl::read_excel("./data/HW3data_.xlsx", .name_repair = "minimal")
```

```r
fixed_gad65_data <- raw_gad65_data |>
  # Removing all empty nan only columns
  select(-where(is.logical)) |>
  # Turning categorical variables into factors
  mutate(across(c(WellID, `Plate ID`, `Enzyme Variant`), as_factor)) |>
  # Removing rows with empty data
  na.omit()
```

## 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: When considering only enzyme variants, there is a significant difference in the enzyme activity in at least one of the tested GAD65 enzyme variants compared to the other tested GAD65 enzyme variants.**

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

```
kable(distinct(fixed_gad65_data, `Enzyme Variant`))
```

| Enzyme Variant |
|---|
| A121C |
| A121T |
| A121R |
| A121L |
| A121F |

**The hypotheses we care about are:**

$$H_0 : \mu_{A121C} = \mu_{A121T} = \mu_{A121R} = \mu_{A121L} = \mu_{A121F}$$

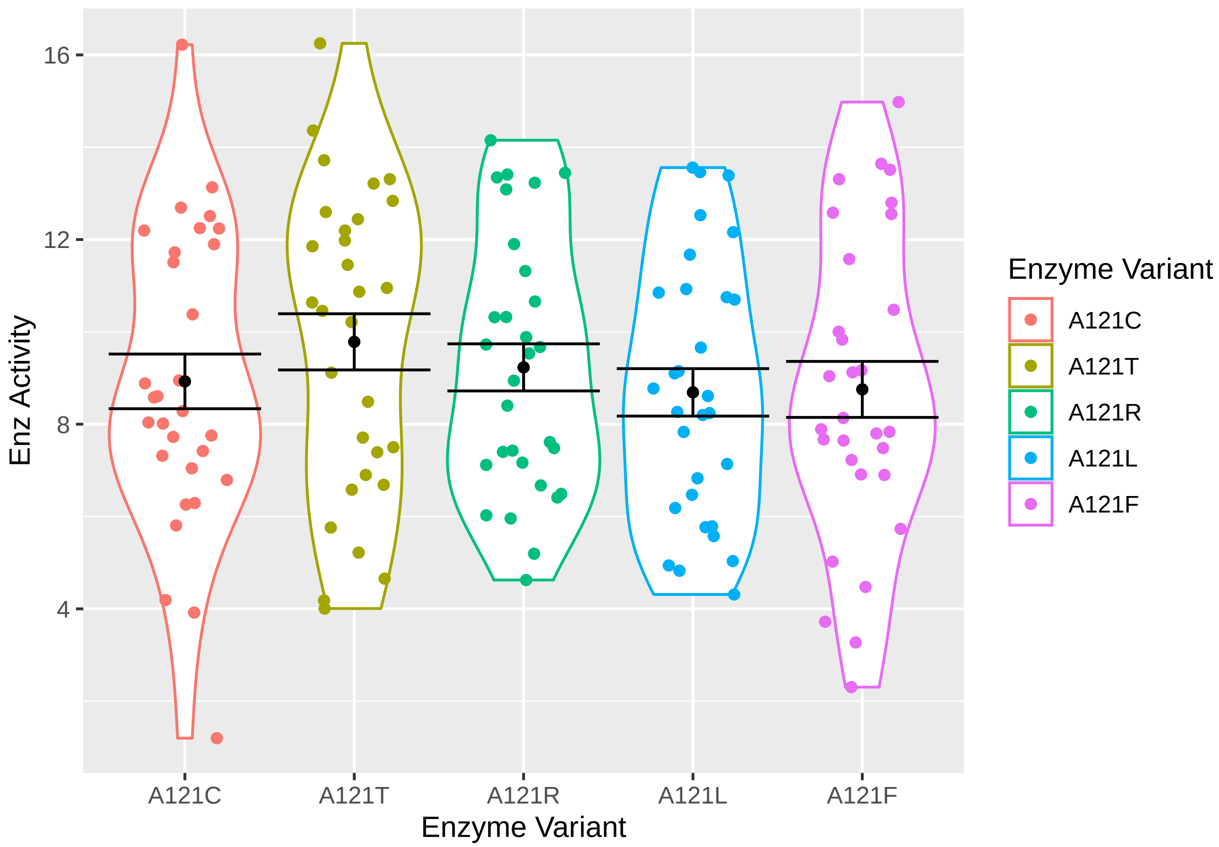
$$H_1 : \mu_{A121C} \neq \mu_{A121T} \neq \mu_{A121R} \neq \mu_{A121L} \neq \mu_{A121F}$$

**Formula that is being fitted** (beta also ok to use, using B for simplicity)

$$\hat{y} = \beta_0 + \beta_{A121T} x_{A121T?} + \beta_{A121R} x_{A121R?} + \beta_{A121L} x_{A121L?} + \beta_{A121F} x_{A121F?}$$

### 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`,
                               color = `Enzyme Variant`)) +
  geom_violin() +
  geom_jitter(width = 0.25) +
  stat_summary(fun = mean, color="black", geom = "point") +
  stat_summary(geom = "errorbar", fun.data = mean_se, color="black")
```



**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)

variant_to_activity_car = car::Anova(variant_to_activity_model)

kable(variant_to_activity_car)
```

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

```
variant_to_activity_pvalue <-
  pull(filter(variant_to_activity_car,
              row.names(variant_to_activity_car)=="`Enzyme Variant`"),
       "Pr(>F)")

variant_to_activity_fstat <-
  pull(filter(variant_to_activity_car,
              row.names(variant_to_activity_car)=="`Enzyme Variant`"),
       "F value")

variant_to_activity_Df <-
```

```r
  pull(filter(variant_to_activity_car,
              row.names(variant_to_activity_car)=="`Enzyme Variant`"),
       "Df")

print("F-statistic and associated p-value for all slope terms")
```

```
## [1] "F-statistic and associated p-value for all slope terms"
```

```r
print(paste("F(",
            variant_to_activity_Df,
            ",",
            df.residual(variant_to_activity_model),
            ") = ",
            variant_to_activity_fstat,
            ", p = ",
            variant_to_activity_pvalue))
```

```
## [1] "F( 4 , 145 ) =  0.620596876015268 , p =  0.648530942510027"
```

**Assuming significance with a p-value less than 0.05, we can not reject the null hypothesis that the mean enzyme activity of different enzyme variants are not significantly different from each other when considering only enzyme variant.**

## 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)

**Answer: There is a significant difference in the enzyme activity of at least one of the tested GAD65 enzyme variants compared to the other tested GAD65 enzyme variants after controlling for the effect of the concentrations of glutamate.**

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

**The hypotheses we care about are:**

$$H_0 : \mu_{A121C} = \mu_{A121T} = \mu_{A121R} = \mu_{A121L} = \mu_{A121F}$$
$$H_1 : \mu_{A121C} \neq \mu_{A121T} \neq \mu_{A121R} \neq \mu_{A121L} \neq \mu_{A121F}$$

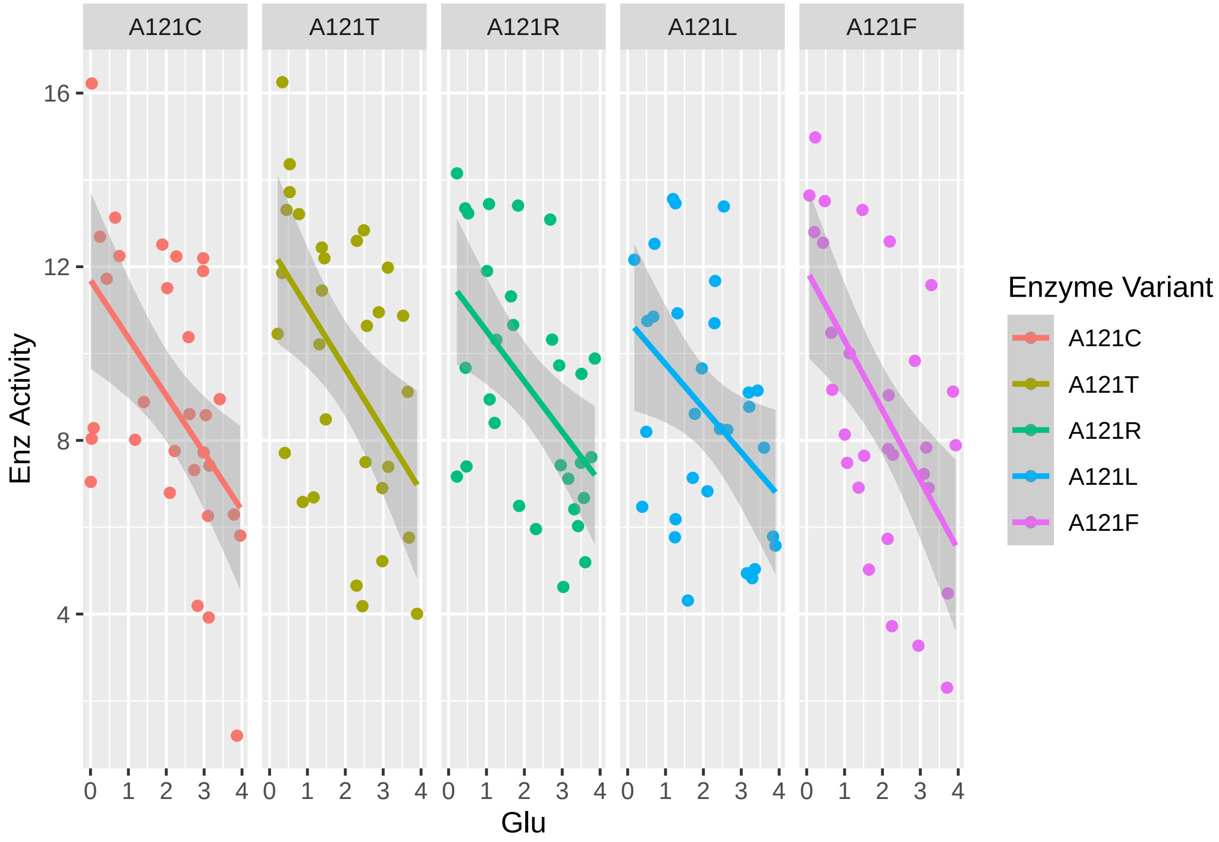
**Formula that is being fitted**

$$\hat{y} = \beta_0 + \beta_{A121T} x_{A121T?} + \beta_{A121R} x_{A121R?} + \beta_{A121L} x_{A121L?} + \beta_{A121F} x_{A121F?} + \beta_{glutamate} x_{glutamate}$$

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

```r
fixed_gad65_data |> ggplot(aes(x=Glu, y=`Enz Activity`, color=`Enzyme Variant`)) +
  geom_point() +
  facet_wrap(vars(`Enzyme Variant`), nrow=1) +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



**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)**

```
variant_and_glu_to_activity_model = lm(`Enz Activity` ~ `Enzyme Variant` + Glu, fixed_gad65_data)

variant_and_glu_to_activity_car = car::Anova(variant_and_glu_to_activity_model)
variant_and_glu_to_activity_tidy = tidy(variant_and_glu_to_activity_model)

kable(variant_and_glu_to_activity_car)
```

| | Sum Sq | Df | F value | Pr(>F) |
|---|---|---|---|---|
| Enzyme Variant | 19.915 | 4 | 0.69099 | 0.59933 |
| Glu | 365.930 | 1 | 50.78685 | 0.00000 |
| Residuals | 1037.552 | 144 | NA | NA |

```
variant_and_glu_pvalue_for_variant <- pull(filter(variant_and_glu_to_activity_car, row.names(variant_an

variant_and_glu_fstat_for_variant <- pull(filter(variant_and_glu_to_activity_car, row.names(variant_and
                       "F value")

variant_and_glu_df_for_variant <- pull(filter(variant_and_glu_to_activity_car, row.names(variant_and_gl
                       "Df")
```

```
print("F-statistic and associated p-value for all slope terms")
```

```
## [1] "F-statistic and associated p-value for all slope terms"
```

```
print(paste("F(", variant_and_glu_df_for_variant, ",", df.residual(variant_and_glu_to_activity_model),
            variant_and_glu_fstat_for_variant, ", p = ",
            variant_and_glu_pvalue_for_variant))
```

```
## [1] "F( 4 , 144 ) =  0.690987159431602 , p =  0.599334439526976"
```

```
variant_and_glu_pvalue_for_glu <- pull(filter(variant_and_glu_to_activity_car, row.names(variant_and_glu

variant_and_glu_fstat_for_variant <- pull(filter(variant_and_glu_to_activity_car, row.names(variant_and_
                       "F value")

variant_and_glu_df_for_variant <- pull(filter(variant_and_glu_to_activity_car, row.names(variant_and_glu
                       "Df")


print("F-statistic and associated p-value for all slope terms")
```

```
## [1] "F-statistic and associated p-value for all slope terms"
```

```
print(paste("F(", variant_and_glu_df_for_variant, ",", df.residual(variant_and_glu_to_activity_model),
            variant_and_glu_fstat_for_variant, ", p = ",
            variant_and_glu_pvalue_for_glu))
```

```
## [1] "F( 1 , 144 ) =  50.786845009417 , p =  4.56721100008259e-11"
```

**Assuming significance with a p-value less than 0.05, we can not reject the null hypothesis that the mean enzyme activity of different enzyme variants are not significantly different from each other. We can not reject the null hypothesis that the mean enzyme activity of different enzyme variants are not significantly different from each other. We can reject the null hypothesis that the Glumate concentration has a statistically significant effect on the enzyme activity.**

## 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)

**Answer: There is a significant difference in the enzyme activity in at least one of the tested GAD65 enzyme variants compared to the other tested GAD65 enzyme variants after controlling for the effect of the concentrations of pyridoxal phosphate.**

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

**The hypothesis we care about is still:**

$$H_0 : \mu_{A121C} = \mu_{A121T} = \mu_{A121R} = \mu_{A121L} = \mu_{A121F}$$

$$H_1 : \mu_{A121C} \neq \mu_{A121T} \neq \mu_{A121R} \neq \mu_{A121L} \neq \mu_{A121F}$$

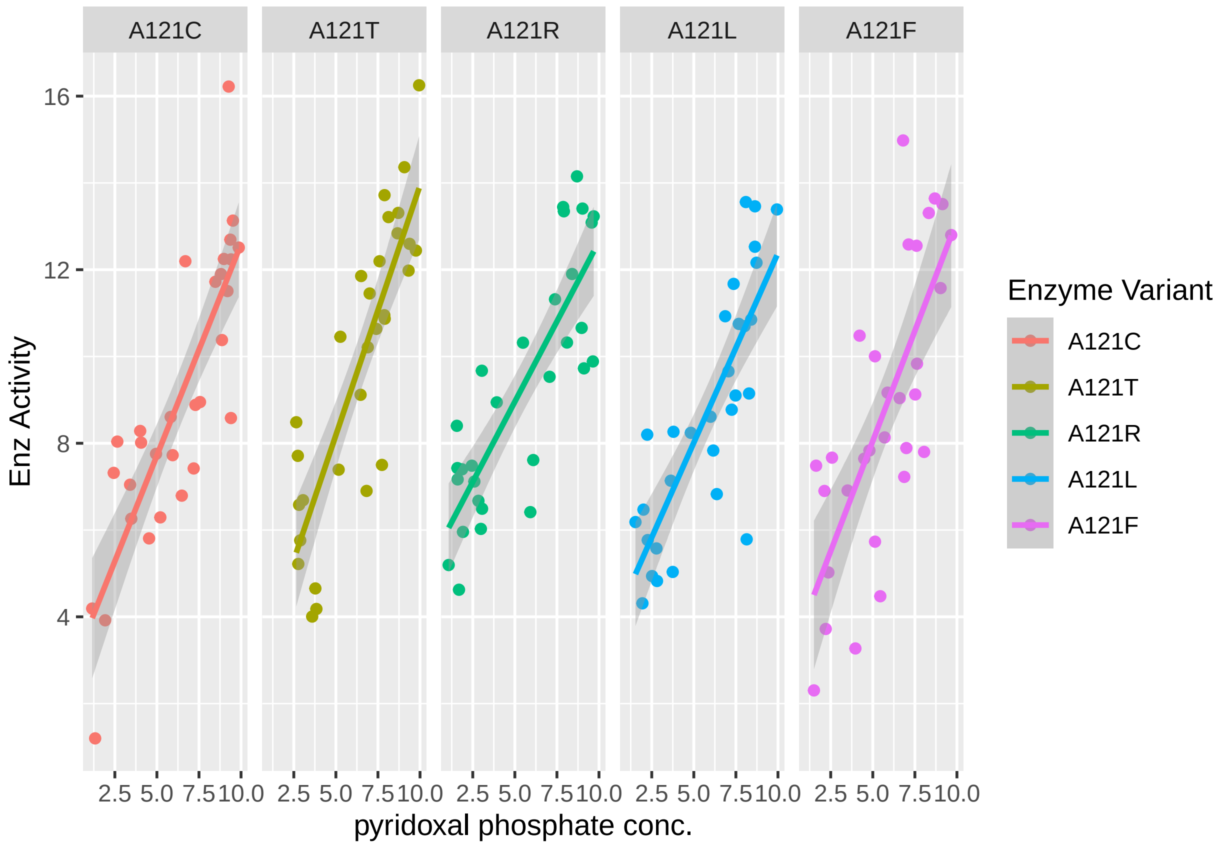
**Formula that is being fitted**

$$\hat{y} = \beta_0 + \beta_{A121T} x_{A121T?} + \beta_{A121R} x_{A121R?} + \beta_{A121L} x_{A121L?} + \beta_{A121F} x_{A121F?} + \beta_{phosphate} x_{phosphate}$$

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

```
fixed_gad65_data |> ggplot(aes(x=`pyridoxal phosphate conc.`, y=`Enz Activity`, color=`Enzyme Variant`)
  geom_point() +
  facet_wrap(vars(`Enzyme Variant`), nrow=1) +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



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

```
variant_and_PLP_to_activity_model = lm(`Enz Activity` ~ `Enzyme Variant` + `pyridoxal phosphate conc.`,

variant_and_PLP_to_activity_car = car::Anova(variant_and_PLP_to_activity_model)
variant_and_PLP_to_activity_tidy = tidy(variant_and_PLP_to_activity_model)

kable(variant_and_PLP_to_activity_car)
```

| | Sum Sq | Df | F value | Pr(>F) |
|---|---|---|---|---|
| Enzyme Variant | 23.609 | 4 | 1.7281 | 0.14695 |
| pyridoxal phosphate conc. | 911.653 | 1 | 266.9175 | 0.00000 |
| Residuals | 491.830 | 144 | NA | NA |

**text**

```
variant_and_PLP_pvalue_for_variant <- pull(filter(variant_and_PLP_to_activity_car, row.names(variant_and

variant_and_PLP_fstat_for_variant <- pull(filter(variant_and_PLP_to_activity_car, row.names(variant_and_
                       "F value")

variant_and_PLP_df_for_variant <- pull(filter(variant_and_PLP_to_activity_car, row.names(variant_and_PL
                       "Df")


print("F-statistic and associated p-value for all slope terms")
```

```
## [1] "F-statistic and associated p-value for all slope terms"
```

```
print(paste("F(", variant_and_PLP_df_for_variant, ",", df.residual(variant_and_PLP_to_activity_model),
            variant_and_PLP_fstat_for_variant, ", p = ",
            variant_and_PLP_pvalue_for_variant))
```

```
## [1] "F( 4 , 144 ) =  1.72807052414683 , p =  0.146953276991127"
```

```
variant_and_PLP_pvalue_for_glu <- pull(filter(variant_and_PLP_to_activity_car, row.names(variant_and_PL

variant_and_PLP_fstat_for_variant <- pull(filter(variant_and_PLP_to_activity_car, row.names(variant_and_
                       "F value")

variant_and_PLP_df_for_variant <- pull(filter(variant_and_PLP_to_activity_car, row.names(variant_and_PL
                       "Df")


print("F-statistic and associated p-value for all slope terms")
```

```
## [1] "F-statistic and associated p-value for all slope terms"
```

```
print(paste("F(", variant_and_PLP_df_for_variant, ",", df.residual(variant_and_PLP_to_activity_model),
            variant_and_PLP_fstat_for_variant, ", p = ",
            variant_and_PLP_pvalue_for_glu))
```

```
## [1] "F( 1 , 144 ) =  266.917537036642 , p =  1.33618287642415e-34"
```

**Assuming significance with a p-value less than 0.05, we can not reject the null hypothesis that the mean enzyme activity of different enzyme variants are not significantly different from each other. We can not reject the null hypothesis that the mean enzyme activity of different enzyme variants are not significantly different from each other. We can reject the null hypothesis that the pyridoxal phosphate concentration has a statistically significant effect on the enzyme activity.**

### 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)

```
variant_and_glu_to_activity_tidy
```

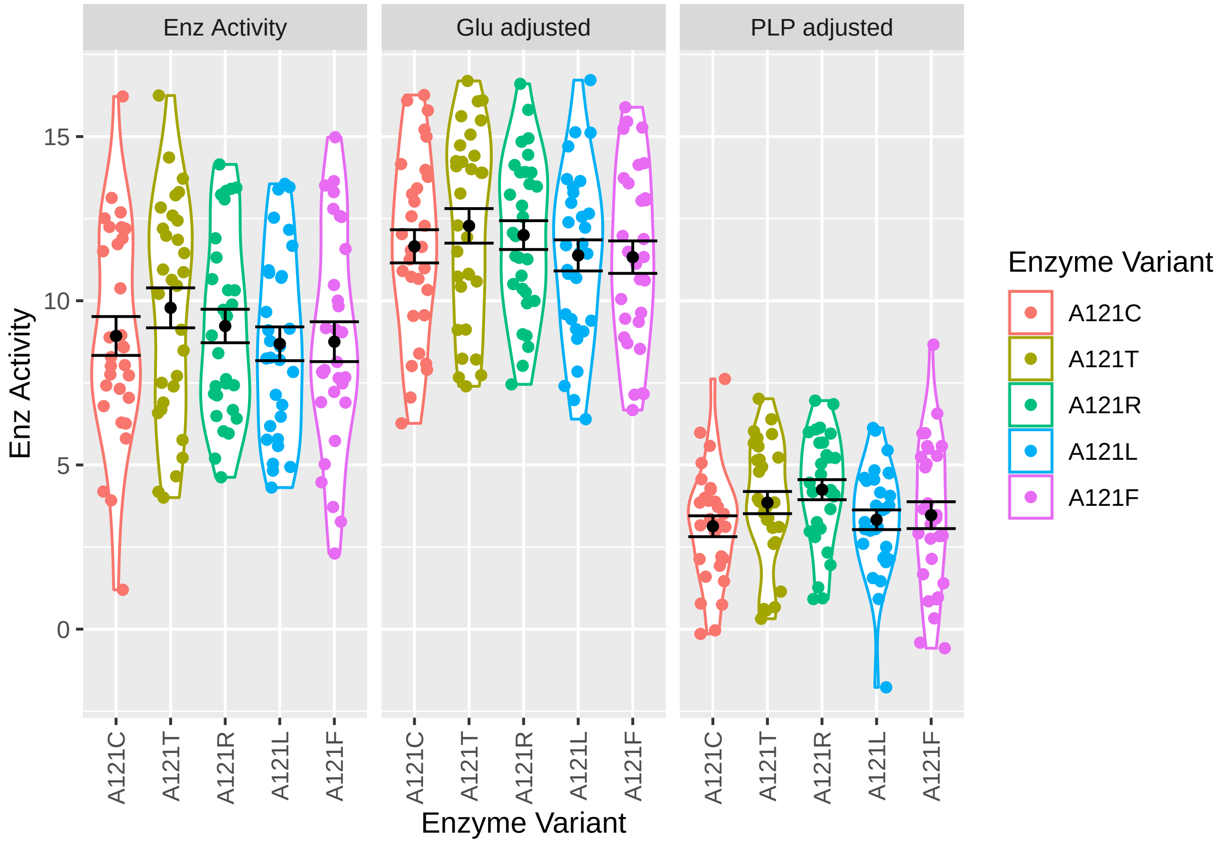
```
## # A tibble: 6 x 5
##   term                    estimate std.error statistic  p.value
##   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)               11.7       0.622     18.7   1.87e-40
## 2 `Enzyme Variant`A121T    0.623     0.694     0.898 3.71e- 1
## 3 `Enzyme Variant`A121R    0.342     0.693     0.494 6.22e- 1
## 4 `Enzyme Variant`A121L   -0.276     0.693    -0.399 6.91e- 1
## 5 `Enzyme Variant`A121F   -0.329     0.693    -0.474 6.36e- 1
## 6 Glu                     -1.31      0.184    -7.13  4.57e-11
```

```
glu_coef_for_variant_and_glu_to_activity <- pull(filter(variant_and_glu_to_activity_tidy, term == "Glu"

phosphate_coef_for_variant_and_glu_to_activity <- pull(filter(variant_and_PLP_to_activity_tidy, term ==

glu_and_PLP_adjusted_gad65_data <- fixed_gad65_data |>
  mutate(`Glu adjusted` = `Enz Activity` - Glu * glu_coef_for_variant_and_glu_to_activity) |>
  mutate(`PLP adjusted` = `Enz Activity` - `pyridoxal phosphate conc.`
 * phosphate_coef_for_variant_and_glu_to_activity) |>
  pivot_longer(cols=c(`Enz Activity`, `Glu adjusted`, `PLP adjusted`), values_to = "Enz Activity")
```

```
glu_and_PLP_adjusted_gad65_data |> ggplot(aes(x=`Enzyme Variant`,
                                y=`Enz Activity`,
                                color = `Enzyme Variant`)) +
  geom_violin() +
  geom_jitter(width = 0.25) +
  stat_summary(fun = mean, color="black", geom = "point") +
  stat_summary(geom = "errorbar", fun.data = mean_se, color="black") +
  facet_wrap(vars(name)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



### 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)

```
variant_ancova_model = lm(`Enz Activity` ~ `Enzyme Variant` + `Glu` + `pyridoxal phosphate conc.`, fixe

variant_ancova_car = car::Anova(variant_ancova_model)
variant_ancova_tidy = tidy(variant_ancova_model)


kable(variant_ancova_tidy)
```

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | 5.70177 | 0.43430 | 13.12869 | 0.00000 |
| Enzyme VariantA121T | 0.54703 | 0.35519 | 1.54010 | 0.12575 |
| Enzyme VariantA121R | 1.08556 | 0.35670 | 3.04338 | 0.00278 |
| Enzyme VariantA121L | 0.13605 | 0.35537 | 0.38285 | 0.70240 |
| Enzyme VariantA121F | 0.17882 | 0.35584 | 0.50254 | 0.61606 |
| Glu | -1.03185 | 0.09519 | -10.83998 | 0.00000 |
| pyridoxal phosphate conc. | 0.86103 | 0.04270 | 20.16322 | 0.00000 |

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)

```
variant_ancova_tidy
```

```
## # A tibble: 7 x 5
##   term                         estimate std.error statistic  p.value
##   <chr>                           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                     5.70      0.434     13.1   2.48e-26
## 2 `Enzyme Variant`A121T           0.547     0.355      1.54  1.26e- 1
## 3 `Enzyme Variant`A121R           1.09      0.357      3.04  2.78e- 3
## 4 `Enzyme Variant`A121L           0.136     0.355      0.383 7.02e- 1
## 5 `Enzyme Variant`A121F          0.179    0.356      0.503 6.16e- 1
## 6 Glu                           -1.03     0.0952   -10.8   2.34e-20
## 7 `pyridoxal phosphate conc.`    0.861    0.0427    20.2   1.21e-43
```

```
glu_coef_ancova <- pull(filter(variant_ancova_tidy, term == "Glu"), estimate)

phosphate_coef_ancova <- pull(filter(variant_ancova_tidy, term == "`pyridoxal phosphate conc.`"), estima
```

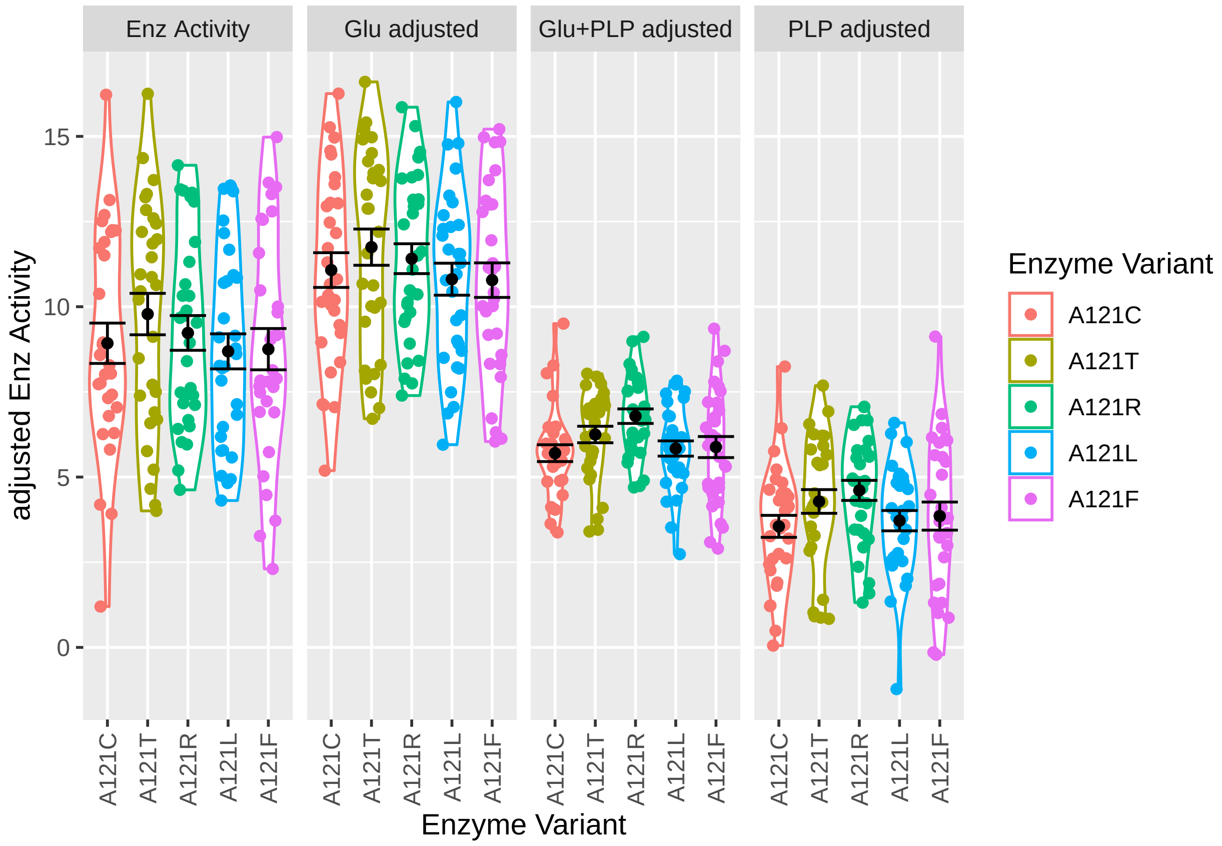
```
glu_and_PLP_adjusted_ancova_gad65_data <- augment(variant_ancova_model) |>
  mutate(`Glu adjusted` = `Enz Activity` - Glu * glu_coef_ancova) |>
  mutate(`PLP adjusted` = `Enz Activity` - `pyridoxal phosphate conc.`
 * phosphate_coef_ancova) |>
   mutate(`Glu+PLP adjusted` = `Enz Activity` - `pyridoxal phosphate conc.` * phosphate_coef_ancova - Gl


glu_and_PLP_adjusted_ancova_gad65_pivoted <- glu_and_PLP_adjusted_ancova_gad65_data |>
  pivot_longer(cols=c(`Enz Activity`, `Glu adjusted`, `PLP adjusted`, `Glu+PLP adjusted`), values_to =
```

```
glu_and_PLP_adjusted_ancova_gad65_data
```

```
## # A tibble: 150 x 13
##    `Enz Activity` `Enzyme Variant`   Glu pyridoxal phosphate c~1 .fitted  .resid
##              <dbl> <fct>            <dbl>                  <dbl>   <dbl>   <dbl>
##  1          13.1  A121C            0.651                    9.52   13.2  -0.0927
##  2           4.01 A121T            3.89                     3.59    5.32 -1.32
##  3           7.61 A121R            3.77                     6.09    8.15 -0.535
##  4           7.14 A121L            1.72                     3.63    7.19 -0.0546
##  5          12.6  A121F            0.432                    7.60   12.0   0.573
##  6           8.58 A121C            3.04                     9.40   10.7  -2.07
##  7          14.4  A121T            0.533                    9.07   13.5   0.852
##  8           6.49 A121R            1.86                     3.05    7.49 -1.00
##  9           8.26 A121L            2.44                     3.79    6.59  1.67
## 10          15.0  A121F            0.225                    6.80   11.5   3.47
## # i 140 more rows
## # i abbreviated name: 1: `pyridoxal phosphate conc.`
## # i 7 more variables: .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## #   .std.resid <dbl>, `Glu adjusted` <dbl>, `PLP adjusted` <dbl>,
## #   `Glu+PLP adjusted` <dbl>
```

```
glu_and_PLP_adjusted_ancova_gad65_pivoted |> ggplot(aes(x=`Enzyme Variant`,
                               y=`adjusted Enz Activity`,
                               color = `Enzyme Variant`)) +
  geom_violin() +
  geom_jitter(width = 0.25) +
  stat_summary(fun = mean, color="black", geom = "point") +
  stat_summary(geom = "errorbar", fun.data = mean_se, color="black") +
  facet_wrap(vars(name), nrow=1) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



**Glutamate goes down and phosphate goes up**

**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)**

```
glu_and_PLP_adjusted_ancova_gad65_data |> summarise(`Model Estimates` = mean(.fitted), `Glu adjusted` =
  kable()
```

| Enzyme Variant | Model Estimates | Glu adjusted | PLP adjusted | Glu+PLP adjusted |
|---|---:|---:|---:|---:|
| A121C | 8.9273 | 11.076 | 3.5534 | 5.7018 |
| A121T | 9.7842 | 11.749 | 4.2844 | 6.2488 |
| A121R | 9.2309 | 11.410 | 4.6084 | 6.7873 |
| A121L | 8.6895 | 10.808 | 3.7197 | 5.8378 |
| A121F | 8.7537 | 10.780 | 3.8543 | 5.8806 |

**Glutamate goes down and phosphate goes up**

**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)**

1. Rotated the tick values for the variants so they are vertical and not overlapping with `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))`
2. facet_wrap(vars(name), nrow=1)

## TODO

- add titles to plots