

ETC5521: Exploratory Data Analysis

Making comparisons between groups and strata

Lecturer: *Emi Tanaka*

Department of Econometrics and Business Statistics

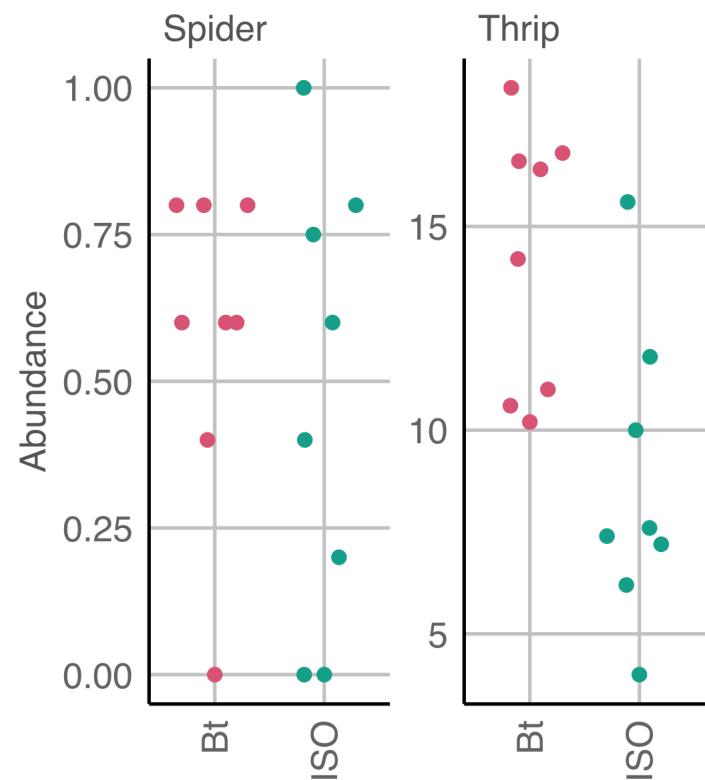
✉ ETC5521.Clayton-x@monash.edu

Week 7 - Session 2

Modelling and testing for comparisons

Revisiting Case study ① Pest resistance maize

data R



- The experiment compared abundance of spiders and thrips on *Bt* variety to the abundance of those on isogenic control variety.
- Would you say that the abundance of spiders and/or thrips are comparable between *Bt* variety and isogenic variety?

Two-sample parametric tests: *t*-test

- Assumes the two samples are independent and from the $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$, respectively.

$$H_0 : \mu_x - \mu_y = 0 \text{ vs. } H_1 : \mu_x - \mu_y \neq 0$$

$$T^* = \frac{\bar{X} - \bar{Y}}{SE(\bar{X} - \bar{Y})}.$$

- Assuming $\sigma_x^2 = \sigma_y^2$, then $T^* \sim t_{n_x+n_y-2}$.
- A $100(1 - \alpha)\%$ confidence interval for $\mu_x - \mu_y$ is given as (L, U) such that:

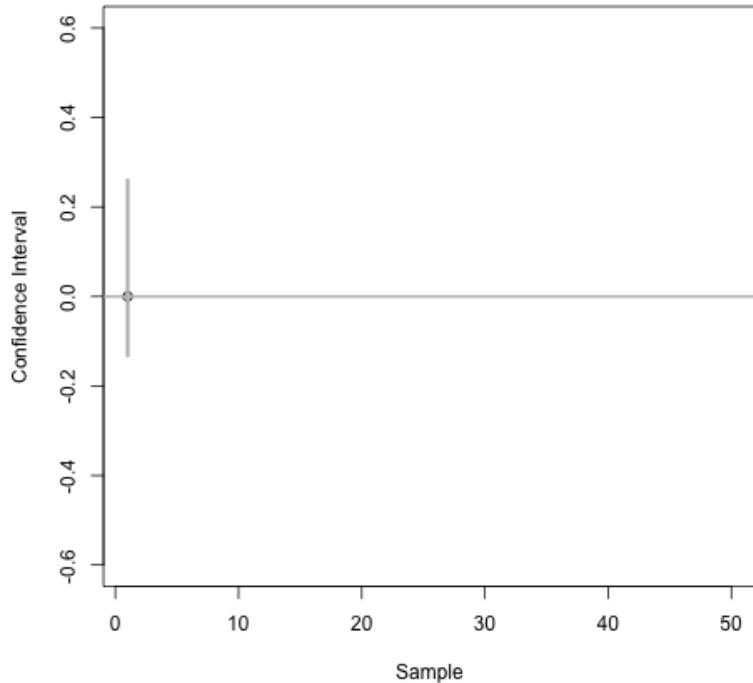
$$P(L < \mu_x - \mu_y < U) = 1 - \frac{\alpha}{2}.$$

- If $0 \in (L, U)$, consistent with H_0 .

```
with(gathmann.bt,
      t.test(thysan[gen=="ISO"],
             thysan[gen=="Bt"],
             alternative = "two.sided",
             var.equal = TRUE,
             conf.level = 0.95))

##
##      Two Sample t-test
##
## data: thysan[gen == "ISO"] and thysan[gen == "Bt"]
## t = -3.2182, df = 14, p-value = 0.006192
## alternative hypothesis: true difference in means is
## 95 percent confidence interval:
## -9.248813 -1.851187
## sample estimates:
## mean of x mean of y
##     8.725    14.275
```

Confidence interval for two sample difference



- In the right, a 95% confidence interval for population mean difference is constructed repeatedly, assuming population mean difference is Normally distributed, from 100 samples of the same population.
- The population mean is zero.
- Each confidence interval is calculated as

$$\bar{X} - \bar{Y} \pm t_{n-2,0.975} \times SE(\bar{X} - \bar{Y})$$

where $t_{n-2,0.975}$ is t^* such that

$$P(t_{n-2} < t^*) = 0.975.$$

Two sample non-parametric tests

Wilcoxon rank-sum test

- Suppose that X and Y are randomly selected values from two populations.

$$H_0 : P(X > Y) = P(X < Y)$$

vs.

$$H_1 : P(X > Y) \neq P(X < Y)$$

- All observations are ranked.
- Test statistic is based on the sum of the ranks of one group.

```
with(gathmann.bt,
      wilcox.test(thysan[gen=="ISO"],
                  thysan[gen=="Bt"],
                  alternative = "two.sided",
                  conf.int = TRUE,
                  conf.level = 0.95))

##
##      Wilcoxon rank sum exact test
##
## data: thysan[gen == "ISO"] and thysan[gen == "Bt"]
## W = 7, p-value = 0.006993
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   -9.4 -2.4
## sample estimates:
## difference in location
##                           -6.3
```

Equivalence of tests to testing model parameters

```
##      gen thysan aranei
## 1    Bt   16.6   0.80
## 2    Bt   16.4   0.80
## 3    Bt   11.0   0.60
## 4    Bt   16.8   0.40
## 5    Bt   10.6   0.60
## 6    Bt   18.4   0.80
## 7    Bt   14.2   0.00
## 8    Bt   10.2   0.60
## 9 ISO    6.2   0.75
## 10 ISO   10.0   0.20
## 11 ISO   11.8   1.00
## 12 ISO   15.6   0.80
## 13 ISO   7.6   0.00
## 14 ISO   7.4   0.00
## 15 ISO   7.2   0.60
## 16 ISO   4.0   0.40
```

$$\text{thysan}_i = \beta_0 + \beta_1 \mathbb{I}(\text{gen}_i = \text{ISO}) + e_i$$

where $e_i \sim NID(0, \sigma^2)$.

- The least squares estimate for $\hat{\beta}_1 = \bar{X} - \bar{Y}$.

```
lm(thysan ~ gen, data = gathmann.bt) %>%
  confint("genISO", level = 0.95)
```

```
##               2.5 %   97.5 %
## genISO -9.248813 -1.851187
```

- Notice that the above confidence interval is the same confidence interval from the t -test!

Revisiting Case study ④ Weight gain of calves Part 1/3

- 67 calves born in 1975 across 11 herds are fed of one of three diets with low, medium or high energy with their initial and final weights recorded.

 data R

Revisiting Case study 4 Weight gain of calves Part 2/3

- Modelling the response as weight gain with diet factor:

```
coef(lm((weight2 - weight1) ~ diet, data = urquhart.f
```

```
## (Intercept)      dietLow    dietMedium  
##  332.666667   -4.666667   -33.971014
```

- The herd is thought to be an important factor contributing to the response.
- Modelling the response as weight gain with diet and herd factor:

```
# herd needs to be factor not integer  
dat4 <- mutate(urquhart.feedlot, herd़f = factor(herd  
coef(lm((weight2 - weight1) ~ herd़f + diet, data = da  
  
## (Intercept)      herd़f9      herd़f16      herd़f19  
##  354.257353   -91.148529   -51.312039    7.410059  -6
```

- Last model is the same as modelling the final weight with the initial weight as a covariate with slope fixed to 1:

```
coef(lm(weight2 ~ offset(weight1) + herd़f + diet,  
       data = dat4))
```

```
## (Intercept)      herd़f9      herd़f16      herd़f19  
##  354.257353   -91.148529   -51.312039    7.410059  -6
```

- Estimating slope for initial weight from the data:

```
coef(lm(weight2 ~ weight1 + herd़f + diet,  
       data = dat4))
```

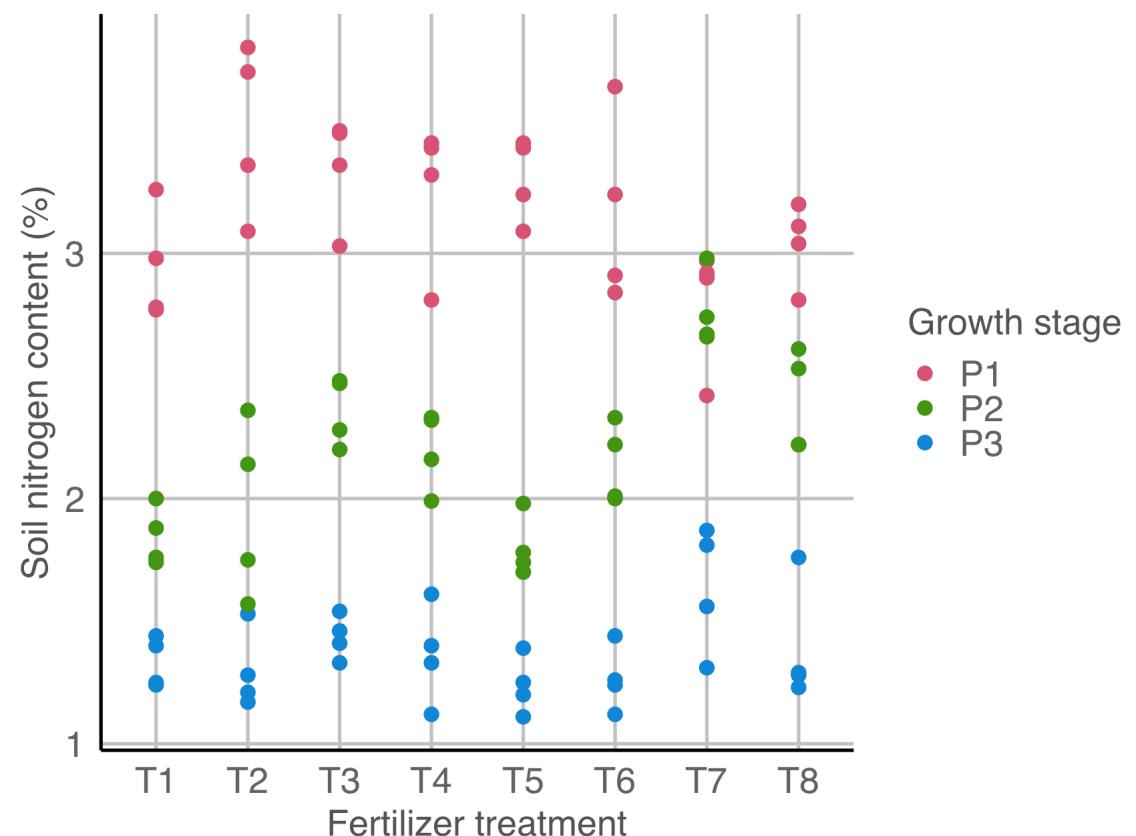
```
## (Intercept)      weight1      herd़f9      herd़f16  
##  200.440174    1.243238   -79.102111   -51.238137
```

Revisiting Case study ④ Weight gain of calves Part 3/3

```
dat4 <- lm(weight2 ~ weight1 + herdf + diet, data = dat4) %>% broom::augment()
ggplot(dat4, aes(.fitted, .resid)) +
  geom_point(data = select(dat4, -herdf), size = 2, color = "gray") +
  geom_point(size = 2, aes(color = herdf)) +
  geom_hline(yintercept = 0) +
  labs(x = "Fitted values", y = "Residual") +
  scale_color_discrete_qualitative() +
  facet_wrap(~herdf, nrow = 2) + guides(color = FALSE)
```

Case study 10 Soil nitrogen Part 1/3

data R



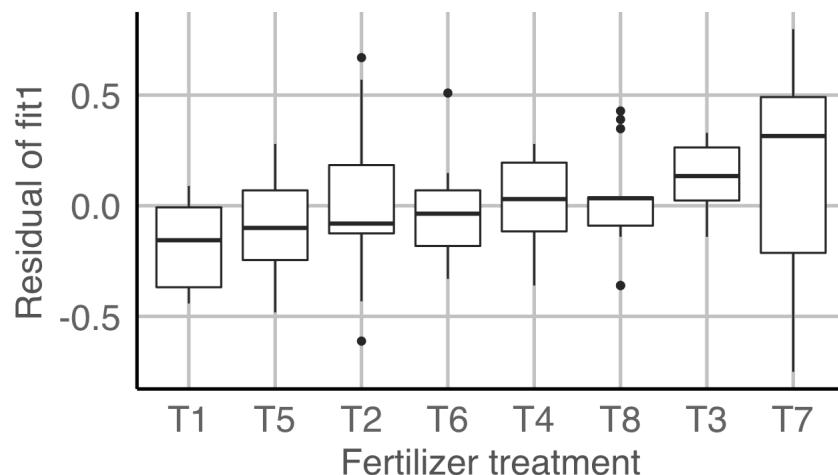
- Soil nitrogen content with 8 different fertilizer treatment is measured at 3 growth stage:
 - P1 = 15 days post transplanting
 - P2 = 40 days post transplanting
 - P3 = panicle initiation
- Clearly the growth stage affects the soil nitrogen content but this makes it hard to compare the fertilizer treatments.
- Let's model the nitrogen content as:

```
lm(nitro ~ stage + trt, data = gomez.nitrogen)
```

Case study 10 Soil nitrogen Part 2/3

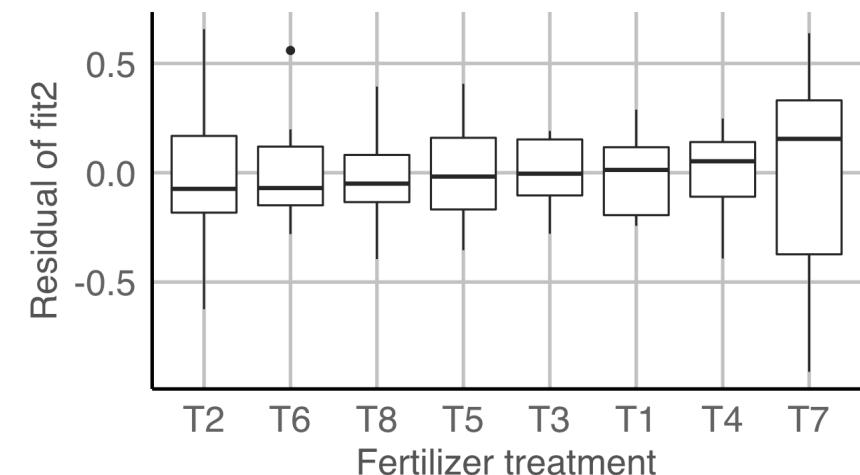
- Considering just the stage effect:

```
fit1 <- lm(nitro ~ stage, data = gomez.nitrogen)
fit1data <- broom::augment(fit1) %>%
  left_join(gomez.nitrogen, by=c("nitro", "stage")) %>%
  mutate(trt = fct_reorder(trt, .resid))
ggplot(fit1data, aes(trt, .resid)) +
  geom_boxplot() +
  labs(x = "Fertilizer treatment",
       y = "Residual of fit1")
```



- Here we expect no pattern:

```
fit2 <- lm(nitro ~ stage + trt,
            data = gomez.nitrogen)
fit2data <- broom::augment(fit2) %>%
  mutate(trt = fct_reorder(trt, .resid))
ggplot(fit2data, aes(trt, .resid)) +
  geom_boxplot() +
  labs(x = "Fertilizer treatment",
       y = "Residual of fit2")
```

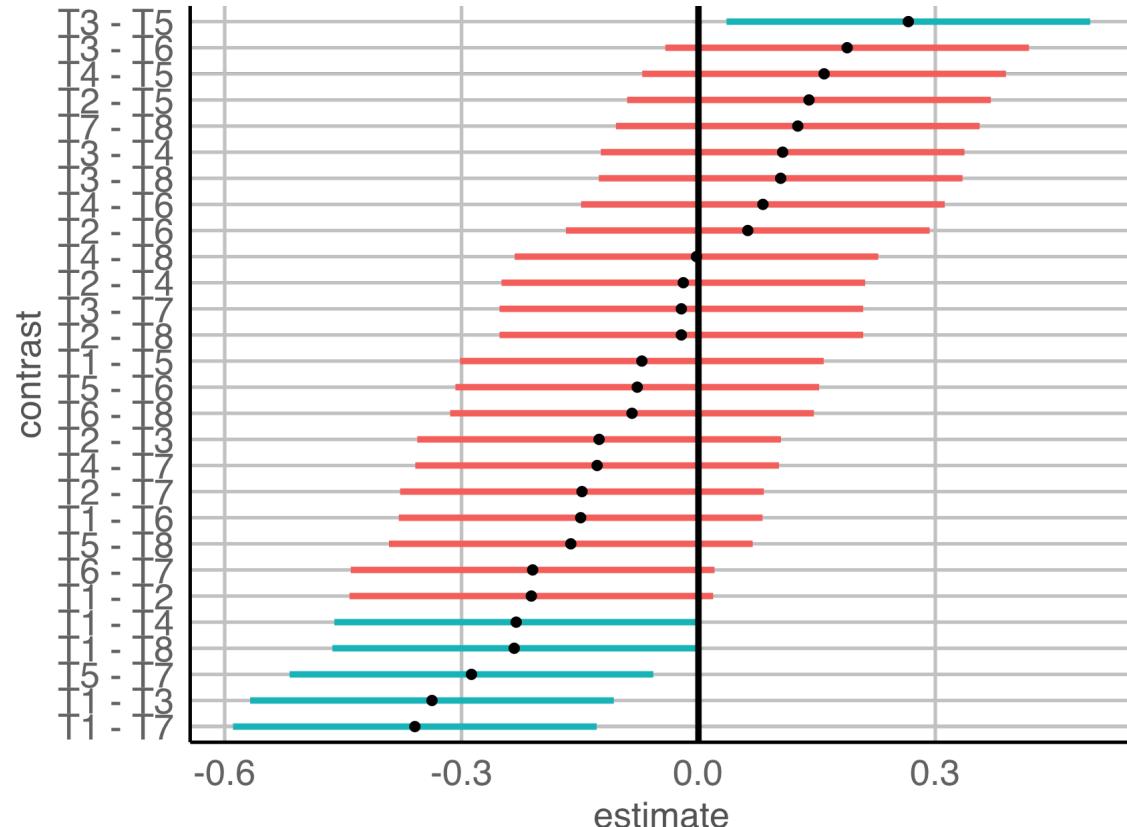


Case study 10 Soil nitrogen Part 3/3

```
library(emmeans)
confint(pairs(emmeans(fit2, "trt"), adjust="none"))

## contrast estimate    SE df lower.CL upper.CL
## T1 - T2   -0.2117 0.116 86  -0.4420  0.018654
## T1 - T3   -0.3375 0.116 86  -0.5678 -0.107180
## T1 - T4   -0.2308 0.116 86  -0.4612 -0.000513
## T1 - T5   -0.0717 0.116 86  -0.3020  0.158654
## T1 - T6   -0.1492 0.116 86  -0.3795  0.081154
## T1 - T7   -0.3592 0.116 86  -0.5895 -0.128846
## T1 - T8   -0.2333 0.116 86  -0.4637 -0.003013
## T2 - T3   -0.1258 0.116 86  -0.3562  0.104487
## T2 - T4   -0.0192 0.116 86  -0.2495  0.211154
## T2 - T5   0.1400 0.116 86  -0.0903  0.370320
## T2 - T6   0.0625 0.116 86  -0.1678  0.292820
## T2 - T7   -0.1475 0.116 86  -0.3778  0.082820
## T2 - T8   -0.0217 0.116 86  -0.2520  0.208654
```

scroll ↓



- From above, the 6 pairs of treatments: T3 & T5, T1 & T4, T1 & T8, T6 & T7, T1 & T3, T1 & T7 are significantly different.
- These confidence intervals are constructed *without taking any regard for others*.

Controlling the family-wise error rate

Unadjusted

- Each interval has been constructed using a procedure so that when the model is correct, the probability that the "correct" population contrast is covered is 0.95. . . individually.

$$\bar{X} - \bar{Y} \pm t_{n-t, 1-\alpha/2} \times SE(\bar{X} - \bar{Y})$$

where $\alpha = 0.05$ and t is the number of treatments.

- But, what is the probability that all intervals cover their corresponding true values simultaneously?

Bonferroni adjustment

- We can adjust the individual $100(1 - \alpha)\%$ confidence intervals so

$$\bar{X} - \bar{Y} \pm t_{n-t, 1-\alpha/(2m)} \times SE(\bar{X} - \bar{Y})$$

where m is the number of pairwise comparisons.

- So for 8 treatments, the number of pairwise comparisons is

```
choose(8, 2)
```

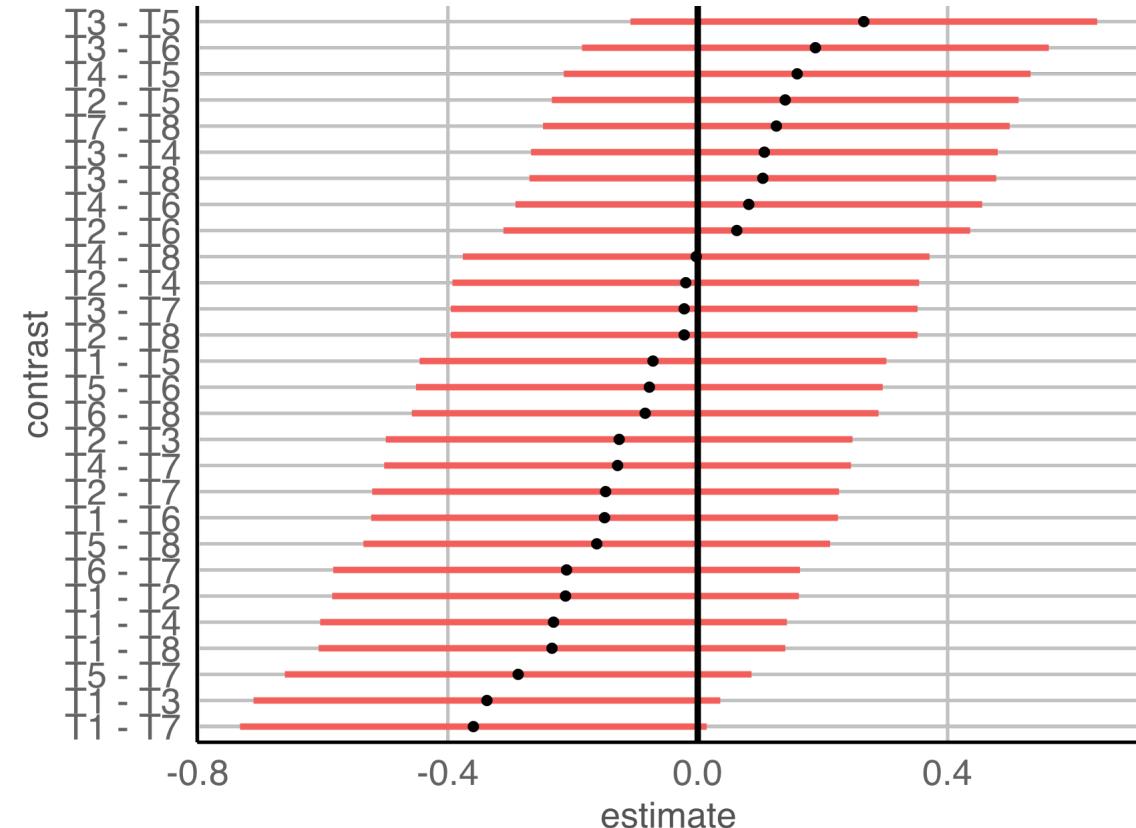
```
## [1] 28
```

Bonferroni adjusted confidence interval

```
confint(pairs(emmeans(fit2, "trt"),  
            adjust="bonferroni"))
```

```
##   contrast estimate    SE df lower.CL upper.CL  
## T1 - T2   -0.2117 0.116 86  -0.585  0.1619  
## T1 - T3   -0.3375 0.116 86  -0.711  0.0361  
## T1 - T4   -0.2308 0.116 86  -0.604  0.1427  
## T1 - T5   -0.0717 0.116 86  -0.445  0.3019  
## T1 - T6   -0.1492 0.116 86  -0.523  0.2244  
## T1 - T7   -0.3592 0.116 86  -0.733  0.0144  
## T1 - T8   -0.2333 0.116 86  -0.607  0.1402  
## T2 - T3   -0.1258 0.116 86  -0.499  0.2477  
## T2 - T4   -0.0192 0.116 86  -0.393  0.3544  
## T2 - T5   0.1400 0.116 86  -0.234  0.5136  
## T2 - T6   0.0625 0.116 86  -0.311  0.4361  
## T2 - T7   -0.1475 0.116 86  -0.521  0.2261  
## T2 - T8   -0.0217 0.116 86  -0.395  0.3519
```

scroll ↓

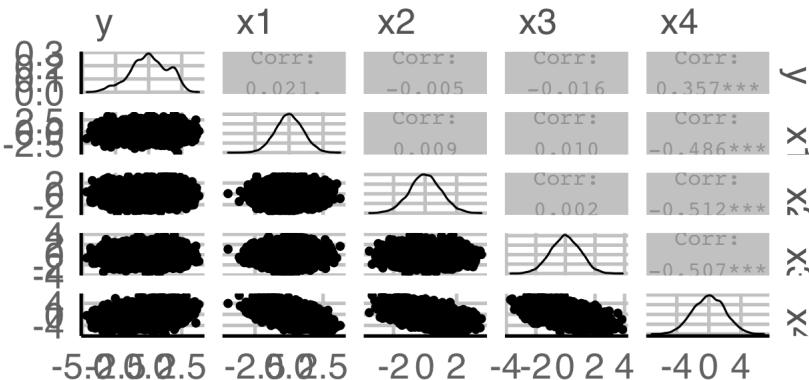


- Now none are significantly different.
- Note: Bonferroni adjustment is quite conservative.

Example 1 Mystery data Part 1/2

- Many inferences, e.g. using confidence intervals or p -values, are based on assumptions being met.
- From the model fit below can we suggest the following model?

$$\hat{Y} = -0.002 + 0.979x_1 + 0.998x_2 + 0.973x_3 + 0.995x_4$$

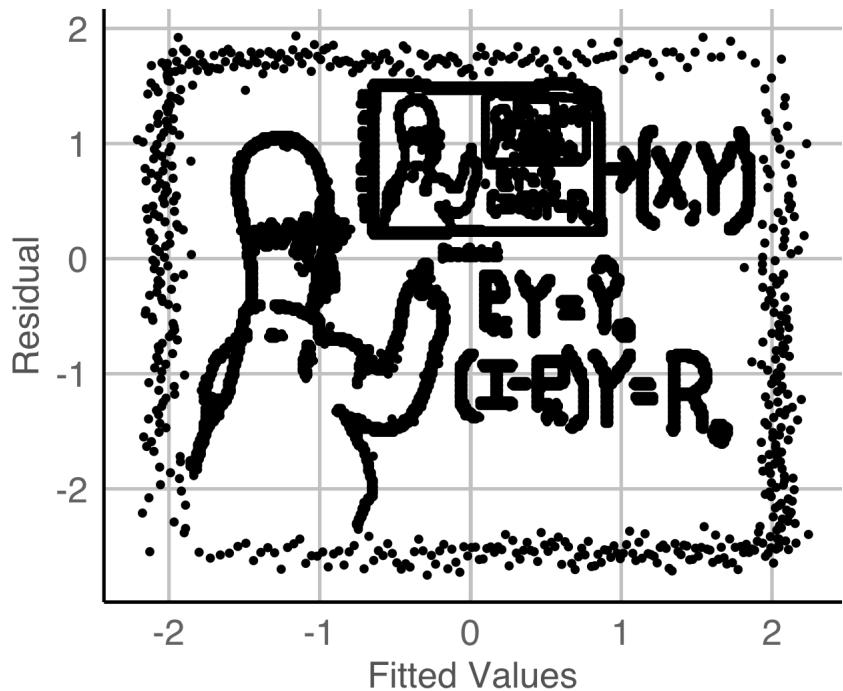


```
lm(y ~ x1 + x2 + x3 + x4, data=mystery_data) %>% broom::tidy()

## # A tibble: 5 x 5
##   term       estimate std.error statistic p.value
##   <chr>     <dbl>    <dbl>     <dbl>    <dbl>
## 1 (Intercept) -0.00204  0.0109   -0.187   0.852
## 2 x1          0.979    0.0151    64.8     0
## 3 x2          0.998    0.0155    64.4     0
## 4 x3          0.973    0.0154    63.1     0
## 5 x4          0.995    0.0109    91.6     0
```

Example 1 Mystery data Part 2/2

```
lm(y ~ x1 + x2 + x3 + x4, data=mystery_data) %>%  
  broom::augment() %>%  
  ggplot(aes(.fitted, .resid)) +  
  geom_point() +  
  labs(x="Fitted Values", y="Residual")
```



Moral of the story:
Don't forget to check model diagnostics.

That's it, for this lecture!



This work is licensed under a [Creative Commons
Attribution-ShareAlike 4.0 International License](#).

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

✉ ETC5521.Clayton-x@monash.edu