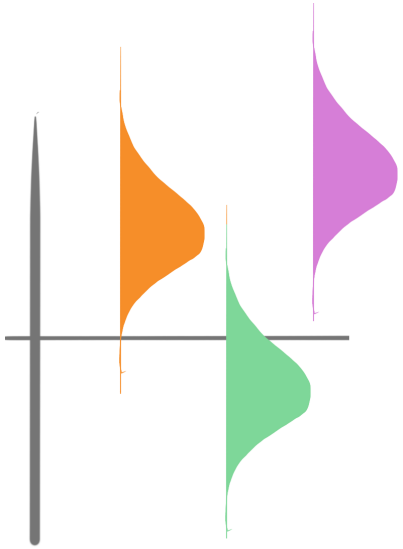
ANOVA



Statistical Experimental Design

Kris Sankaran | UW Madison | 21 September 2021

Feedback Response

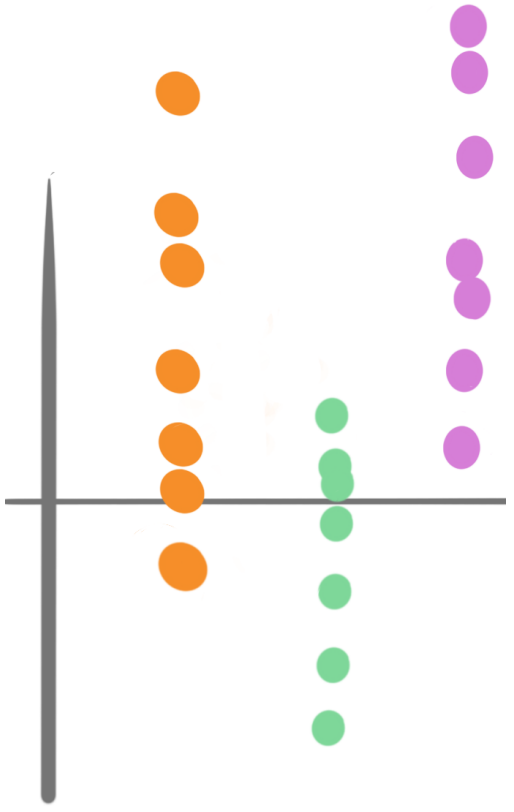
- Restructuring lectures: Separating conceptual and code elements
- R Code: Only including essential R code
  - Have started an "R Cheatsheet"
- Pace / Volume: I will try to enunciate clearly, but please do raise your hand if I should slow down or repeat!
- R Review Session: 9/27 from 6:30 - 7:30pm
  - Hybrid: 1217 Medical Sciences Center and Zoom
- We will drop your lowest HW score

Today

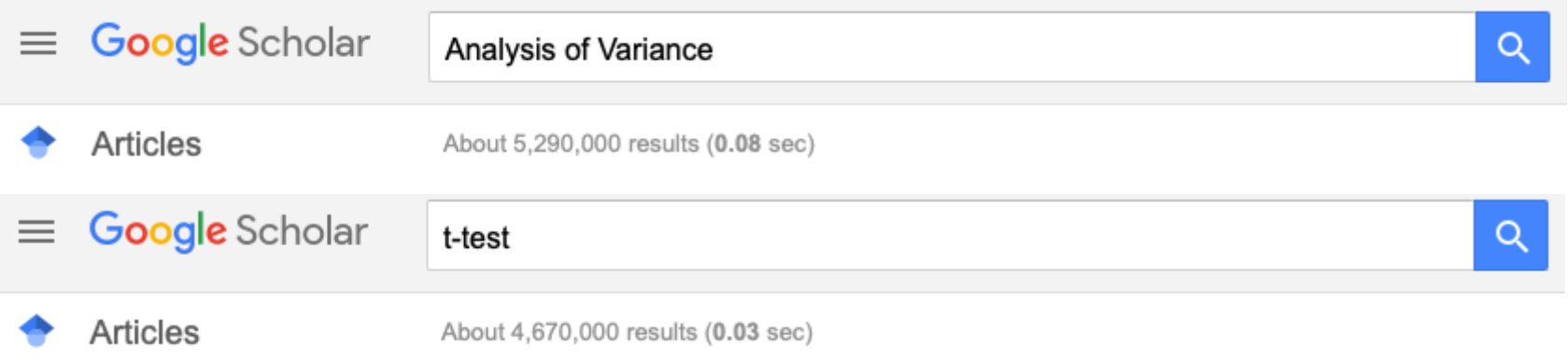
- Book Sections: 3.1 - 3.4
- Online Notes: Week 3 [1] and [2]

Motivation

- ANOVA helps gauge the effects of ≥ 3 different treatments on a continuous response
  - How does the etch rate of a tool depend on its power?
  - How do different foods affect blood sugar?
  - How do several job training programs compare?
- It is an extension of two sample testing when there are 3 or more levels



Motivation


Google Scholar
Analysis of Variance
Articles
About 5,290,000 results (0.08 sec)
Google Scholar
t-test
Articles
About 4,670,000 results (0.03 sec)

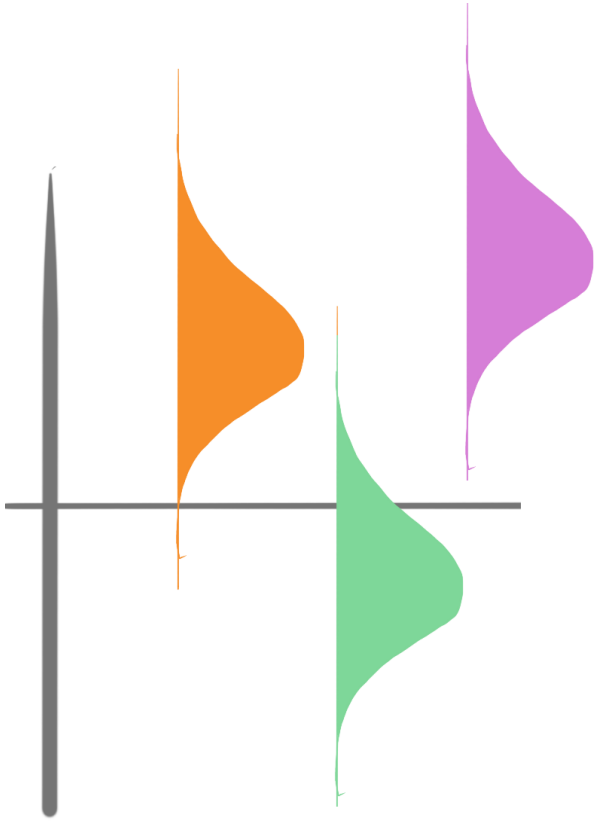
Model

ANOVA assumes,

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where $i = 1, \ldots, a$ and $j = 1, \ldots, n$ and the errors $\epsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$ are independent.

- i indexes different groups
- j indexes the samples within groups
- $N = na$ is the total number of samples



Hypothesis Testing

> Is there at least one group that differs from the rest?

Formally,

$$
\begin{aligned}
H_0 &: \tau_1 = \cdots = \tau_a = 0 \\
H_1 &: \tau_i \neq 0 \text{ for at least one } i.
\end{aligned}
$$

Q: What would be the version of the picture on the previous slide, if the null hypothesis were true?

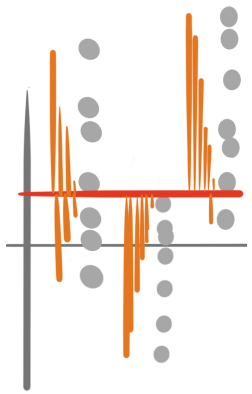
ANOVA Identity

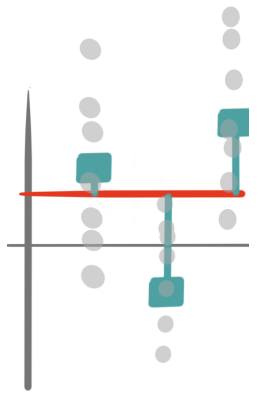
Before designing a test statistic, it helps to observe this identity,

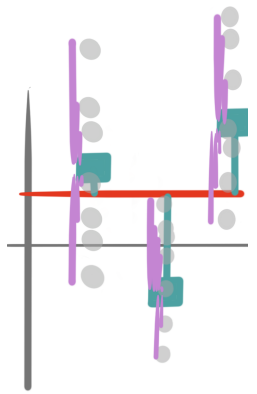
$$\sum_{ij} \left(y_{ij} - \bar{y}\right)^2 = n \sum_{i} \left(\bar{y}_i - \bar{y}\right)^2 + \sum_{i,j} \left(y_{ij} - \bar{y}_i\right)^2$$

which is usually abbreviated as

$$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{E}}.$$







Test Statistic

- If any of the groups are different from the global mean, then we expect $SS_{\text{treatment}}$ to be large
- How large is large enough?
- Consider the statistic,

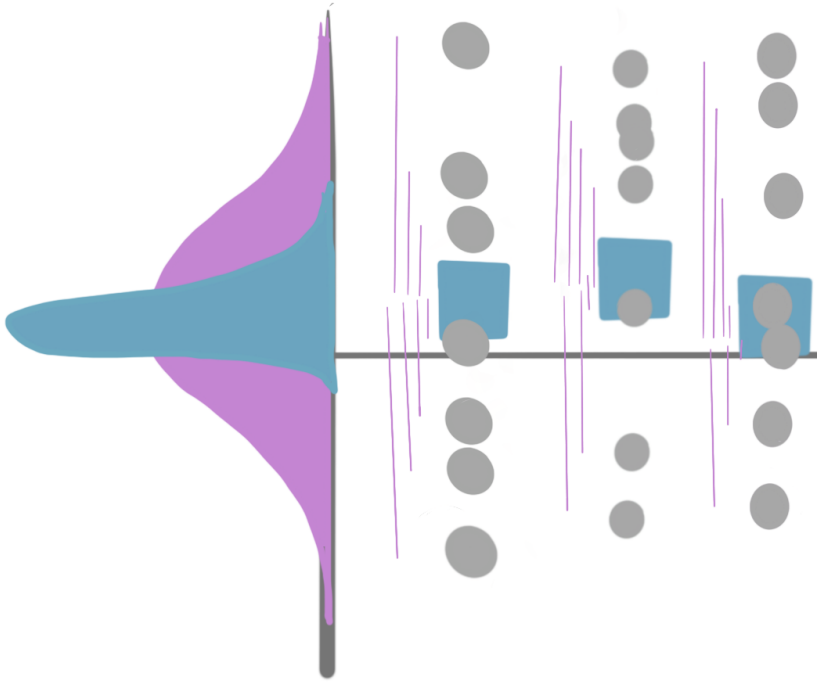
$$\frac{MS_{\text{treatment}}}{MS_E}$$

where we define,

$$MS_{\text{treatment}} := \frac{1}{a-1} SS_{\text{treatment}}$$
$$MS_E := \frac{1}{N-a} SS_E$$

Test Statistic

- It is not obvious, but reference distribution for this statistic is $F\left(a-1, N-a\right)$.
- If this statistic is at a large quantile of that distribution, we conclude $\tau_i \neq 0$ for at least one i



Model Checking

We assumed,

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

with i.i.d. errors $\epsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$

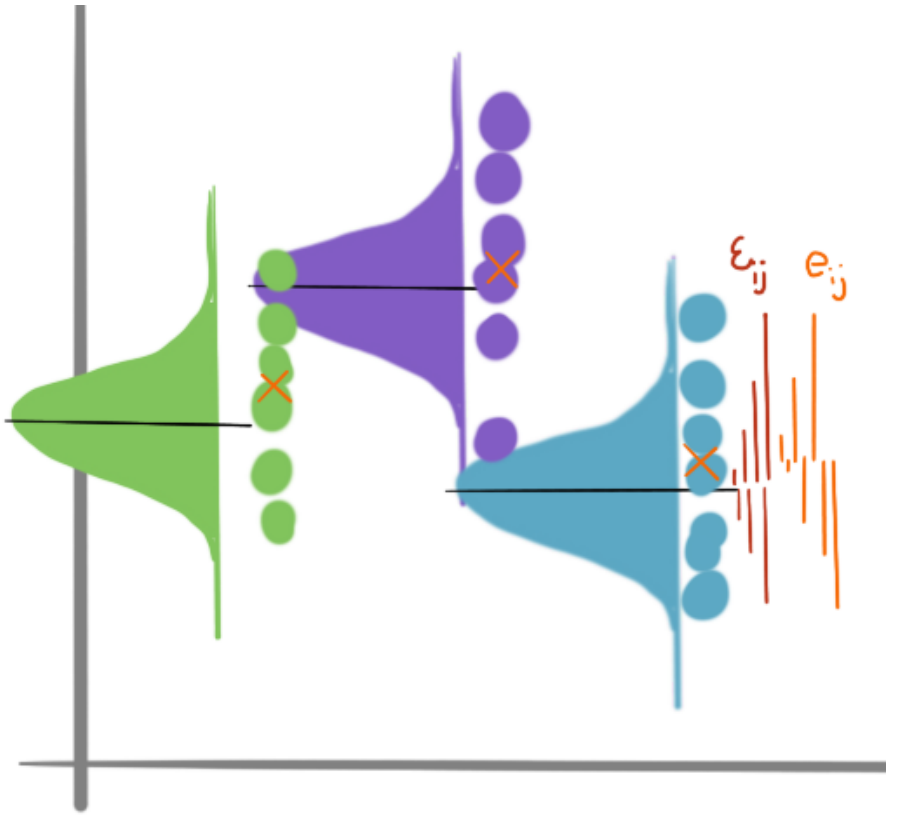
What could fail?

- The errors might not be normally distributed
- The variance might not be the same in each group
- The errors might not be independent
- There might be systematic variations besides the group deviations τ_i.

Residuals

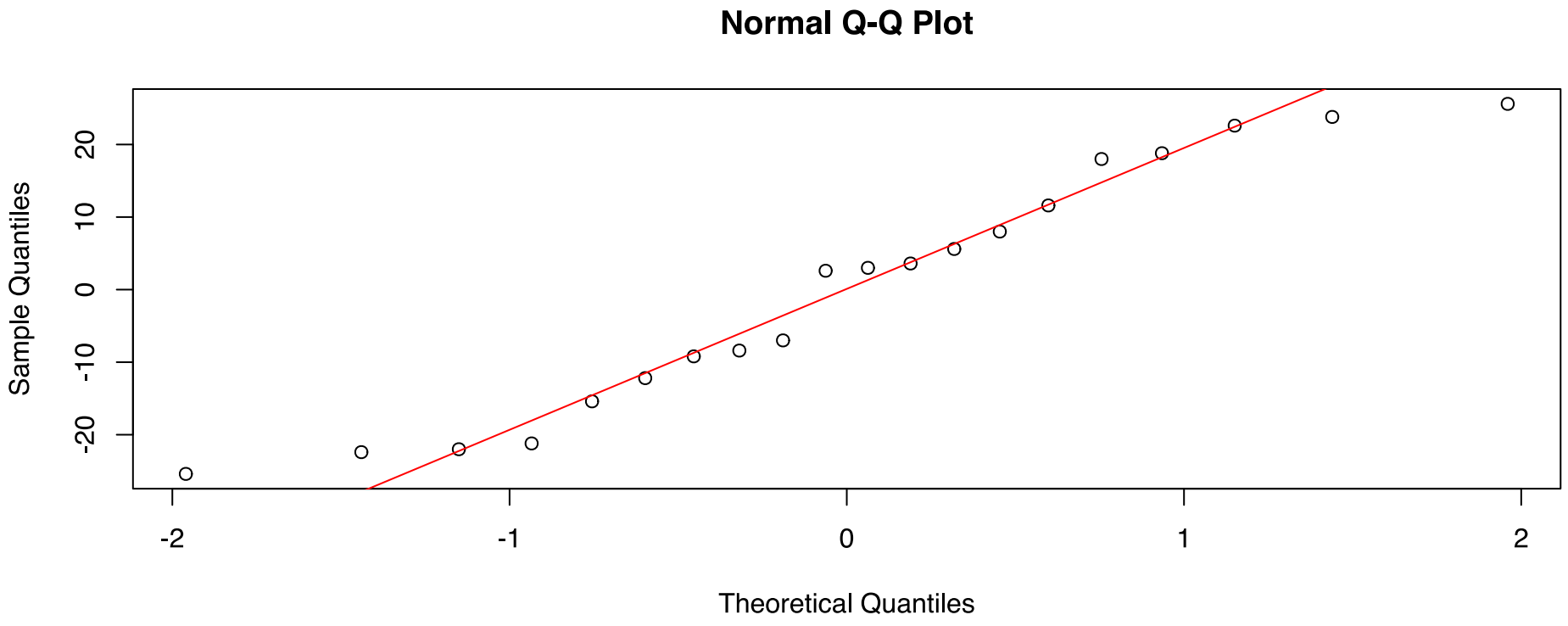
Residuals are our best guess of what the true random error ϵ_{ij}, and so are useful for model checking,

$$
\begin{aligned}
e_{ij} &= y_{ij} - \hat{y}_{ij} \\
&= y_{ij} - (\hat{\mu} + \hat{\tau}_i)
\end{aligned}
$$



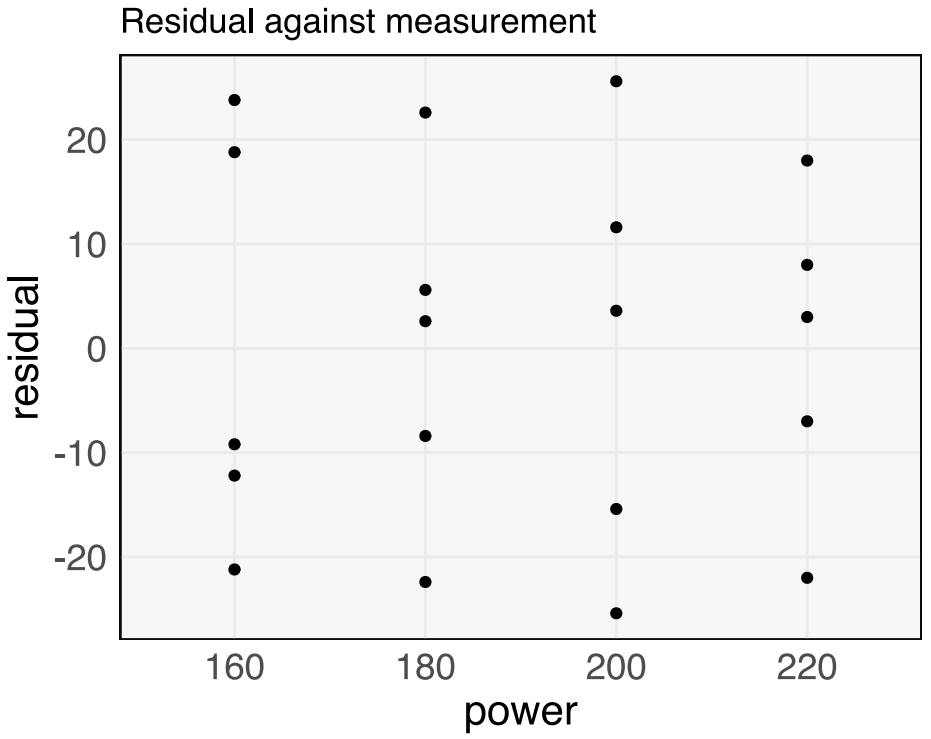
Normality assumption?

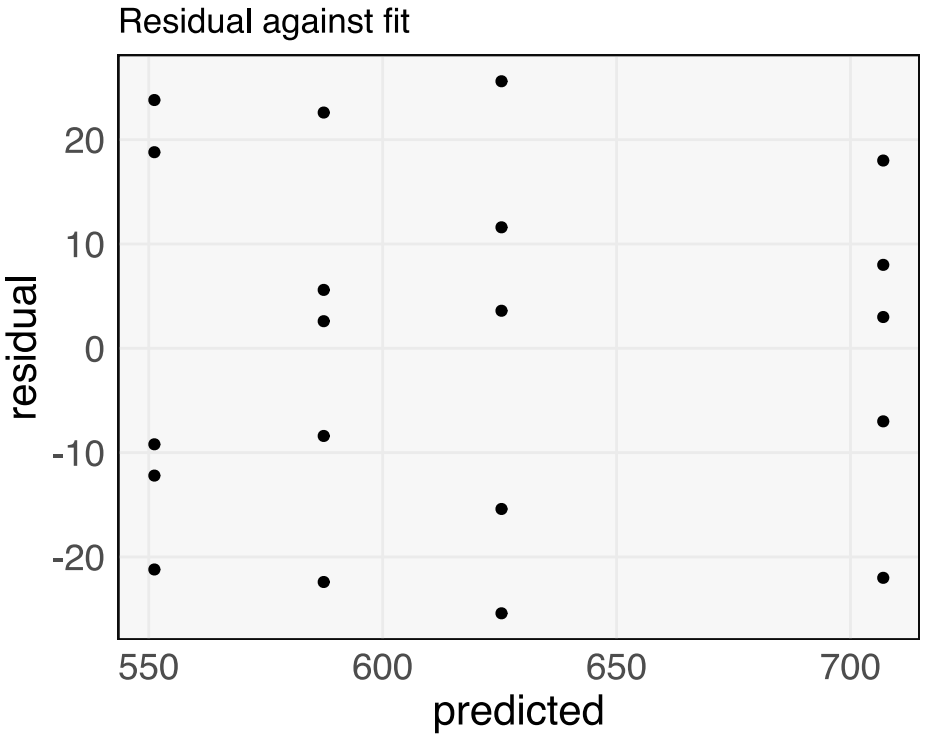
We can't check normality of ϵ_{ij} directly, but we can check normality of the residuals e_{ij}.



Equal variance across groups?

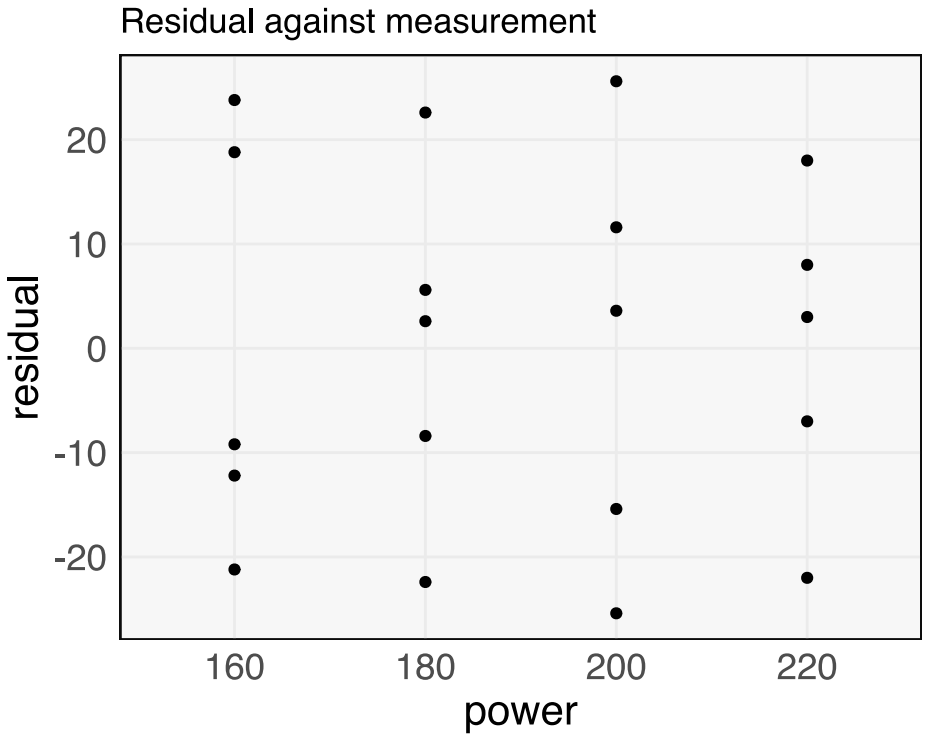
- Plot residuals against measured variables
- Plot residuals against fitted means

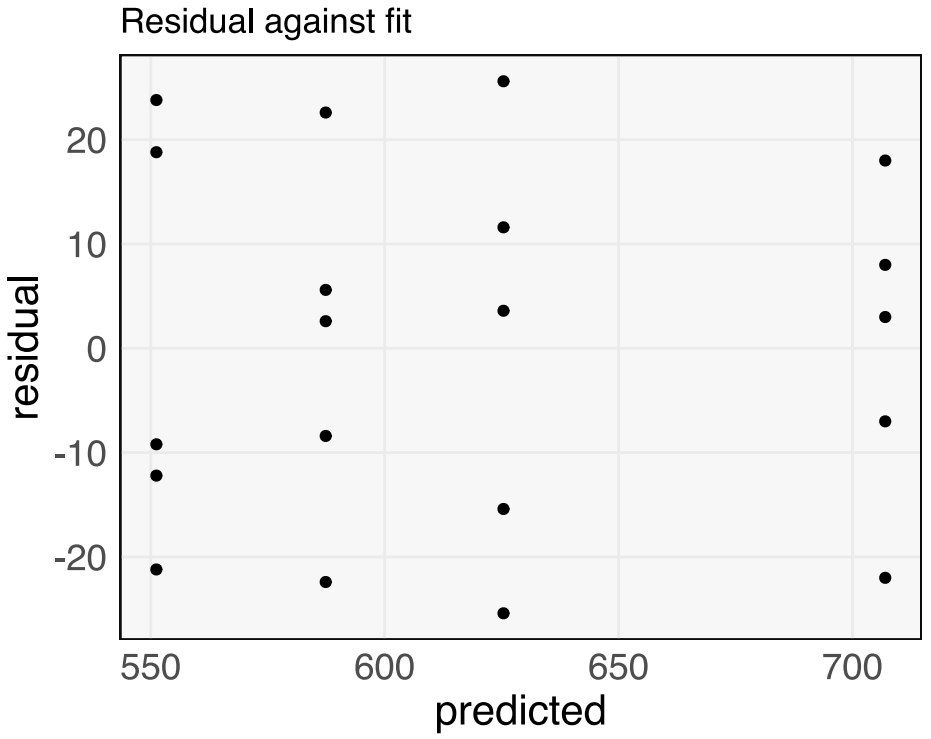




Unmeasured variation?

- Plot residuals against measured variables
- Plot residuals against fitted means

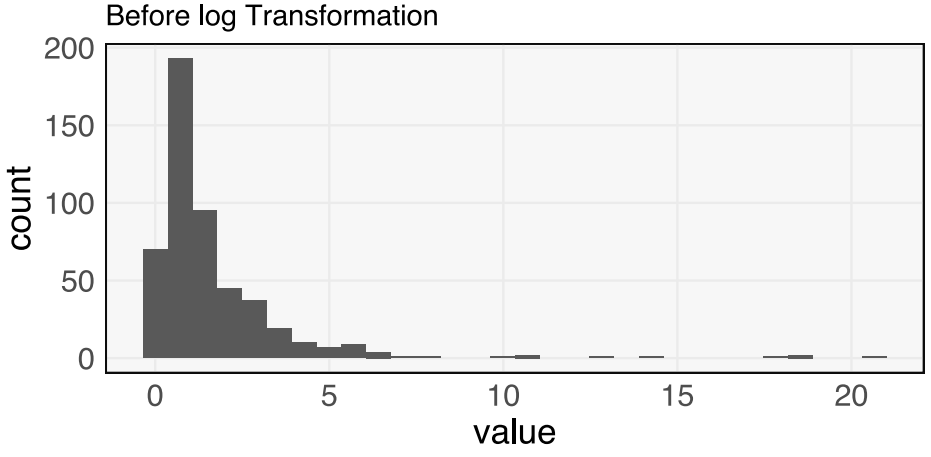


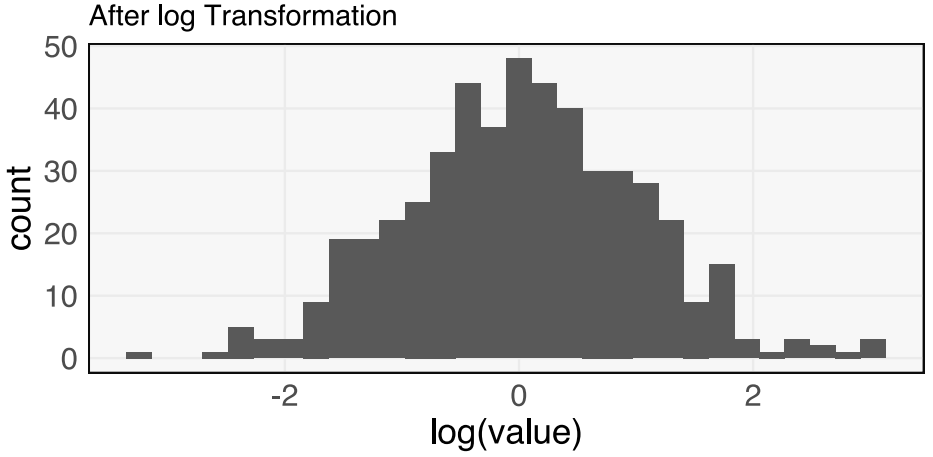


What if a check fails?

Transform the response variable to make it more bell-shaped

- Skewed, nonnegative data: Use $\log(x)$ or $\log(1+x)$
- Count data: $\sqrt{x}, \sqrt{1+x}$
- Proportions: $\arcsin\left(\sqrt{x}\right)$





Code Implementation

Etch Rate Dataset

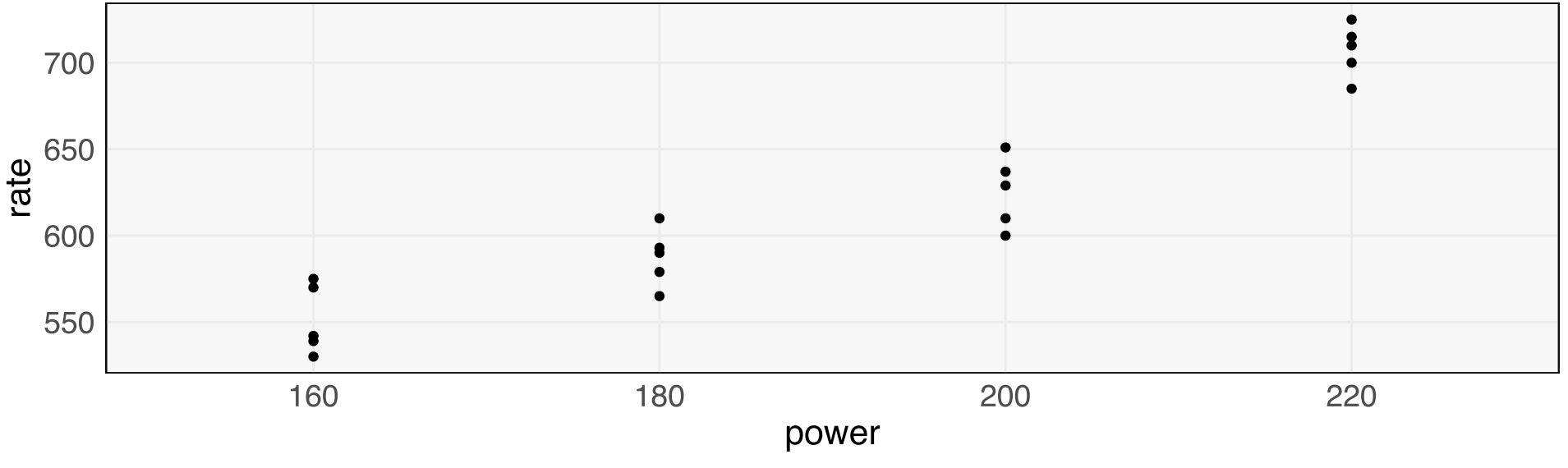
```r
library(readr)
etch_rate <- read_csv("https://uwmadison.box.com/shared/static/vw3ldbgvgn7rupt4tz3ditl1mpupw
etch_rate$power <- as.factor(etch_rate$power) # want to think of power as distinct groups
head(etch_rate, 10)
```

```
## # A tibble: 10 × 3
##    power replicate  rate
##    <fct> <chr>     <dbl>
##  1 160   Ob1         575
##  2 160   Ob2         542
##  3 160   Ob3         530
##  4 160   Ob4         539
##  5 160   Ob5         570
##  6 180   Ob1         565
##  7 180   Ob2         593
##  8 180   Ob3         590
##  9 180   Ob4         579
## 10 180   Ob5         610
```

Plot the Data

- `ggplot()` expects a data.frame with the whole dataset
- `geom_point()` asks which columns to use for the x and y axis.

```
library(ggplot2)
ggplot(etch_rate) +
  geom_point(aes(power, rate))
```



ANOVA Hypothesis Test

- Which element in the table corresponds to $SS_{\text{treatment}}$?
- Which element in the table corresponds to $MS_{\text{treatment}}$?

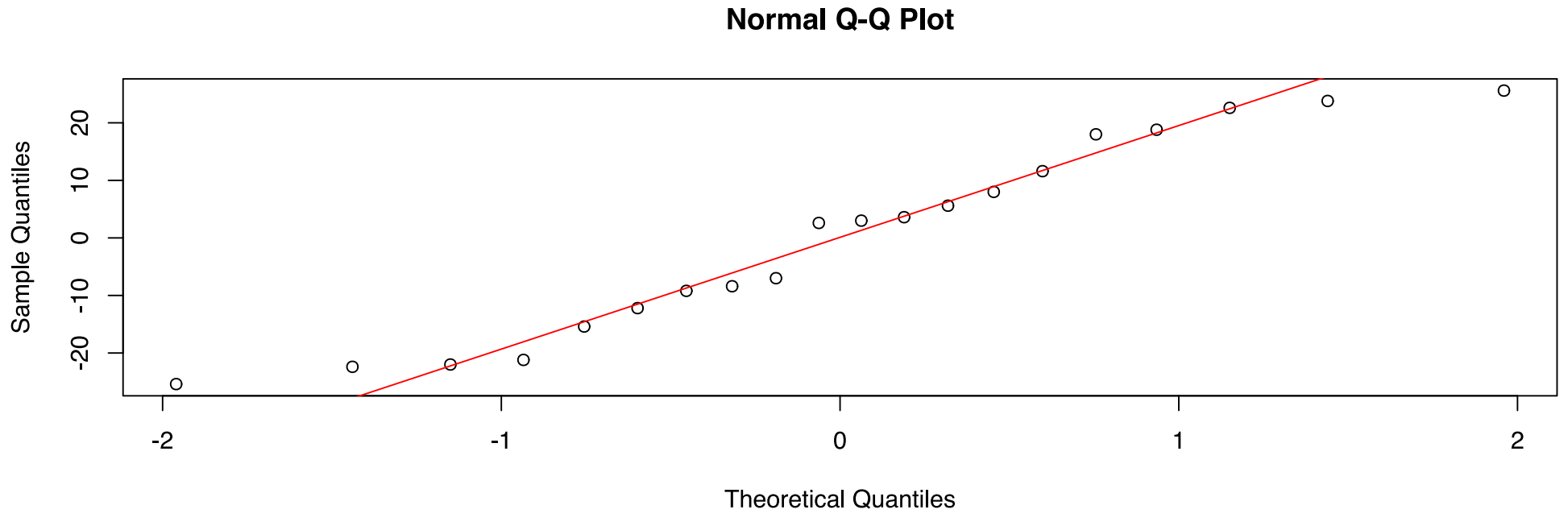
```r
fit <- lm(rate ~ power, data = etch_rate)
summary(aov(fit))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## power        3  66871   22290    66.8 2.88e-09 ***
## Residuals   16   5339     334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check normality of residuals

- Use the `qqnorm` and `qqline` pair to make a QQ Plot

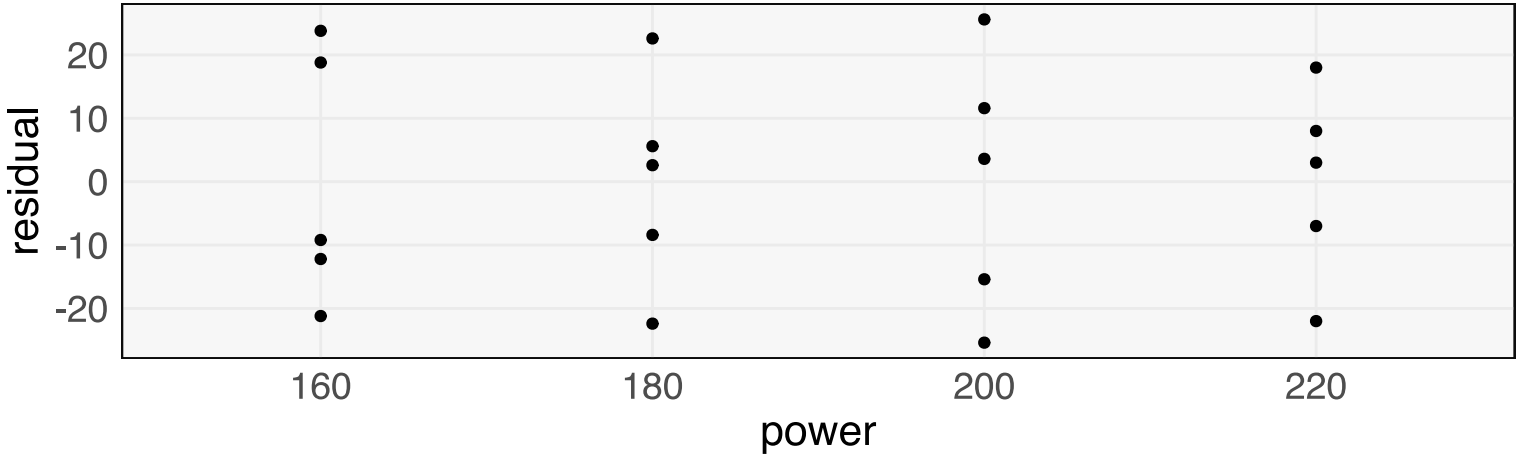
```
qqnorm(resid(fit))
qqline(resid(fit), col = "red")
```



Plot residuals against factor

- First add the residual to the `data.frame`
- Then make the y-axis the residual

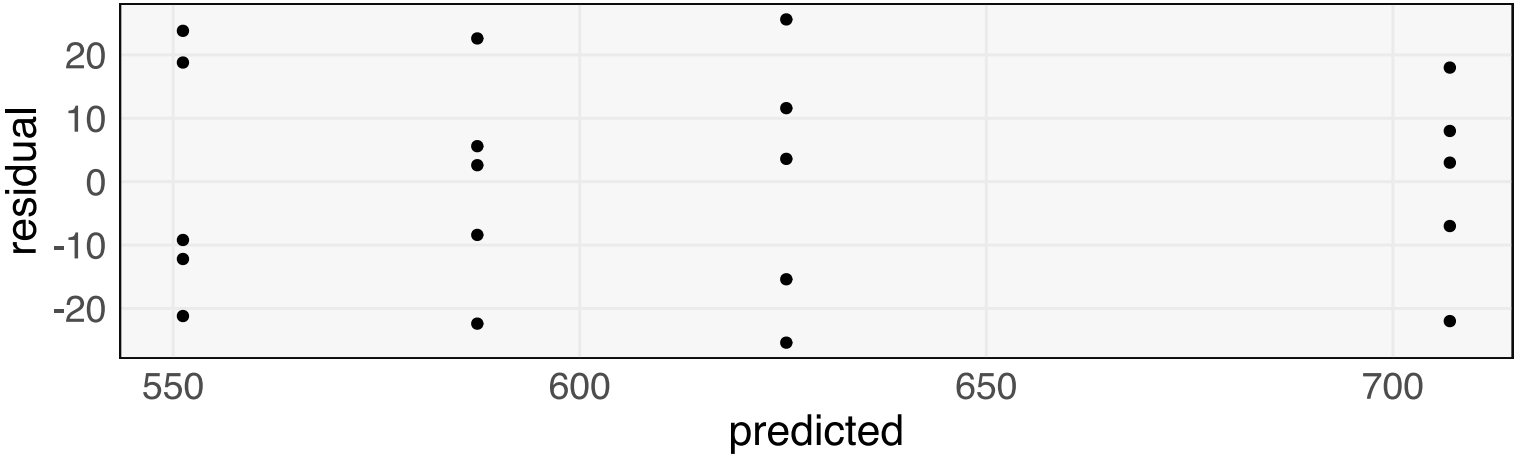
```
etch_rate$residual <- resid(fit)
ggplot(etch_rate) +
  geom_point(aes(power, residual))
```



Plot residuals against predicted value

- First add the fitted value to the `data.frame`
- Then make the x-axis the fitted value

```
etch_rate$predicted <- predict(fit)
ggplot(etch_rate) +
  geom_point(aes(predicted, residual))
```



Hint: Reshaping data

What if the etch dataset had been sent to us like this?

```
## # A tibble: 4 × 6
##   Power   Ob1   Ob2   Ob3   Ob4   Ob5
##   <fct> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 160     575   542   530   539   570
## 2 180     565   593   590   579   610
## 3 200     600   651   610   637   629
## 4 220     725   700   715   685   710
```

The `lm` function expects the outcome of interest to be in a single column.

Hint: Reshaping data

To reorganize the data into an acceptable form, we can use the `pivot_longer` function from the `tidyr` package.

```
library(tidyr)
etch <- pivot_longer(etch, -Power, names_to = "replicate", values_to = "etch_rate")
head(etch)
```

```
## # A tibble: 6 × 3
##   Power replicate etch_rate
##   <fct> <chr>         <dbl>
## 1 160   Ob1             575
## 2 160   Ob2             542
## 3 160   Ob3             530
## 4 160   Ob4             539
## 5 160   Ob5             570
## 6 180   Ob1             565
```

Exercise

This walks through problem 3.29 in the book.

A semiconductor manufacturer has developed three different methods for reducing particle counts on wafers. All three methods are tested on five different wafers and the after treatment particle count is obtained.

```r
particles <- data.frame(
  method = c("1", "2", "3"),
  rep1 = c(31, 62, 53),
  rep2 = c(10, 40, 27),
  rep3 = c(21, 24, 120),
  rep4 = c(4, 30, 97),
  rep5 = c(1, 35, 68)
)

particles
```

```
##   method rep1 rep2 rep3 rep4 rep5
## 1      1   31   10   21    4    1
## 2      2   62   40   24   30   35
## 3      3   53   27  120   97   68
```

Exercise

(1) Use `pivot_longer(particles, -method, ...)` to reshape the data so that the outcome is in a single column.

(2) Use `lm` and `aov` to make an ANOVA table. Does the method detect any difference in the means across the three groups?

(3) Plot the residuals against the group. Are the assumptions satisfied?

(4) Apply a transformation to the response, fit an ANOVA model, and recheck normality of the residuals with a QQ plot. Stop when you are satisfied that the assumptions are met.

Solution

(1) `pivot_longer` can reshape the data.

```
library(tidyr)
particles <- pivot_longer(particles, -method, names_to = "replicate")
head(particles)
```

```
## # A tibble: 6 × 3
##   method replicate value
##   <chr>  <chr>     <dbl>
## 1 1      rep1         31
## 2 1      rep2         10
## 3 1      rep3         21
## 4 1      rep4          4
## 5 1      rep5          1
## 6 2      rep1         62
```

Solution

(2) The $MS_{\text{treatment}}$ is much larger than MS_E (4481.9 vs. 566.3), and the F test indicates a difference between methods at the 0.01 level.

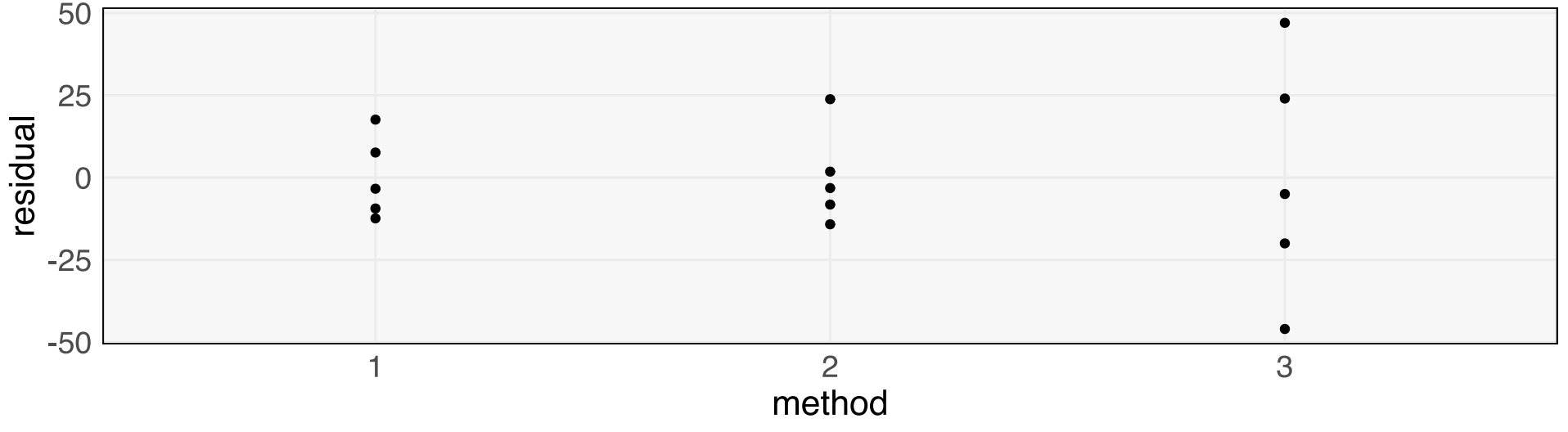
```
fit <- lm(value ~ method, data = particles)
summary(aov(fit))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## method       2   8964    4482   7.914 0.00643 **
## Residuals   12   6796     566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution

(3) There appears to be very nonconstant variance across groups.

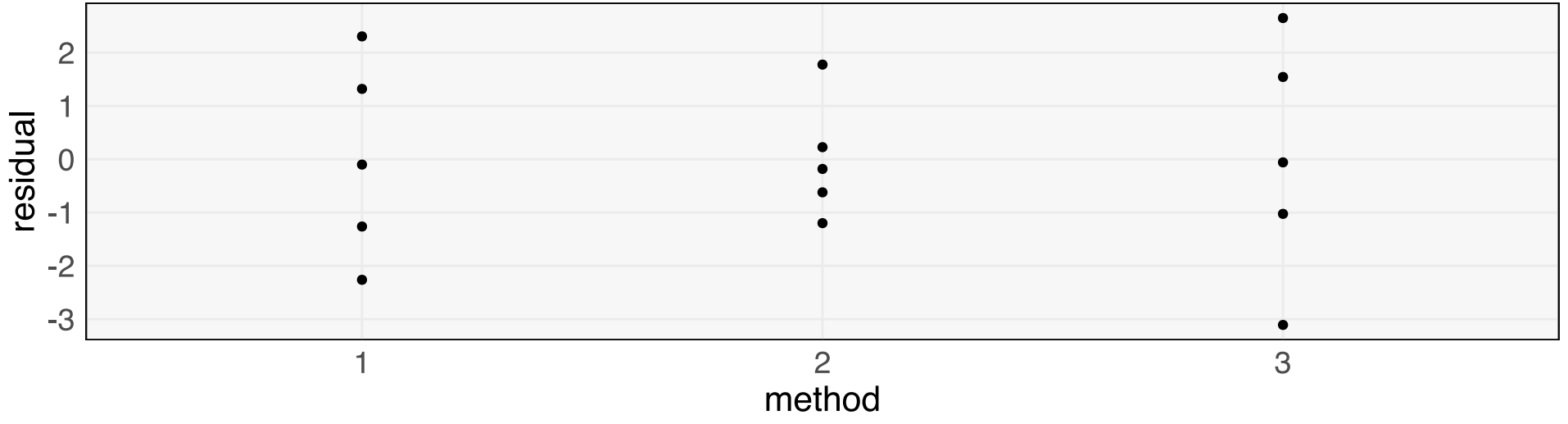
```
particles$residual <- resid(fit)
ggplot(particles) +
  geom_point(aes(method, residual))
```



Solution

(4) If we use a $\sqrt{x}$ transformation, the difference between groups seems to be somewhat reduced, though not completely removed.

```
fit <- lm(sqrt(value) ~ method, data = particles)
particles$residual <- resid(fit)
ggplot(particles) +
  geom_point(aes(method, residual))
```



Solution

(4) Nonetheless, there still appears to be a significant difference between the groups' means.

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## method       2  63.90   31.95    9.84 0.00295 **
## Residuals   12  38.96    3.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```