R Programming For Natural Resource Professionals

Week 12-13

t-tests and ANOVA in R

The week and a half ahead…

Learning objectives for this week (and a half)

1. Perform and interpret regression
    1. T-test
    2. ANOVA
    3. Linear regression
2. Assess model assumptions (i.e., model validation)
3. Model comparison

Learning objectives for this week (and a half)

- This is not a stats class therefore presentation of statistical underpinnings of methods will be minimal
- We will focus on statistical knowledge needed to understand code and interpret results
- **Consult your advisor or a statistician about the details of your analysis**

Learning objectives for this week (and a half)

Goals for this course

Understand how to write code and interpret results for numerous common statistical tests

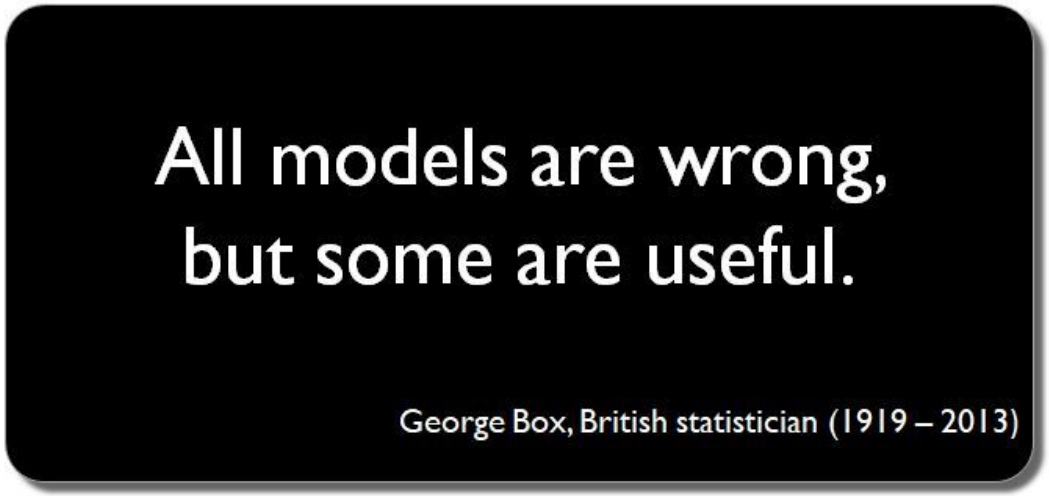
Statistical modeling

What is a model?

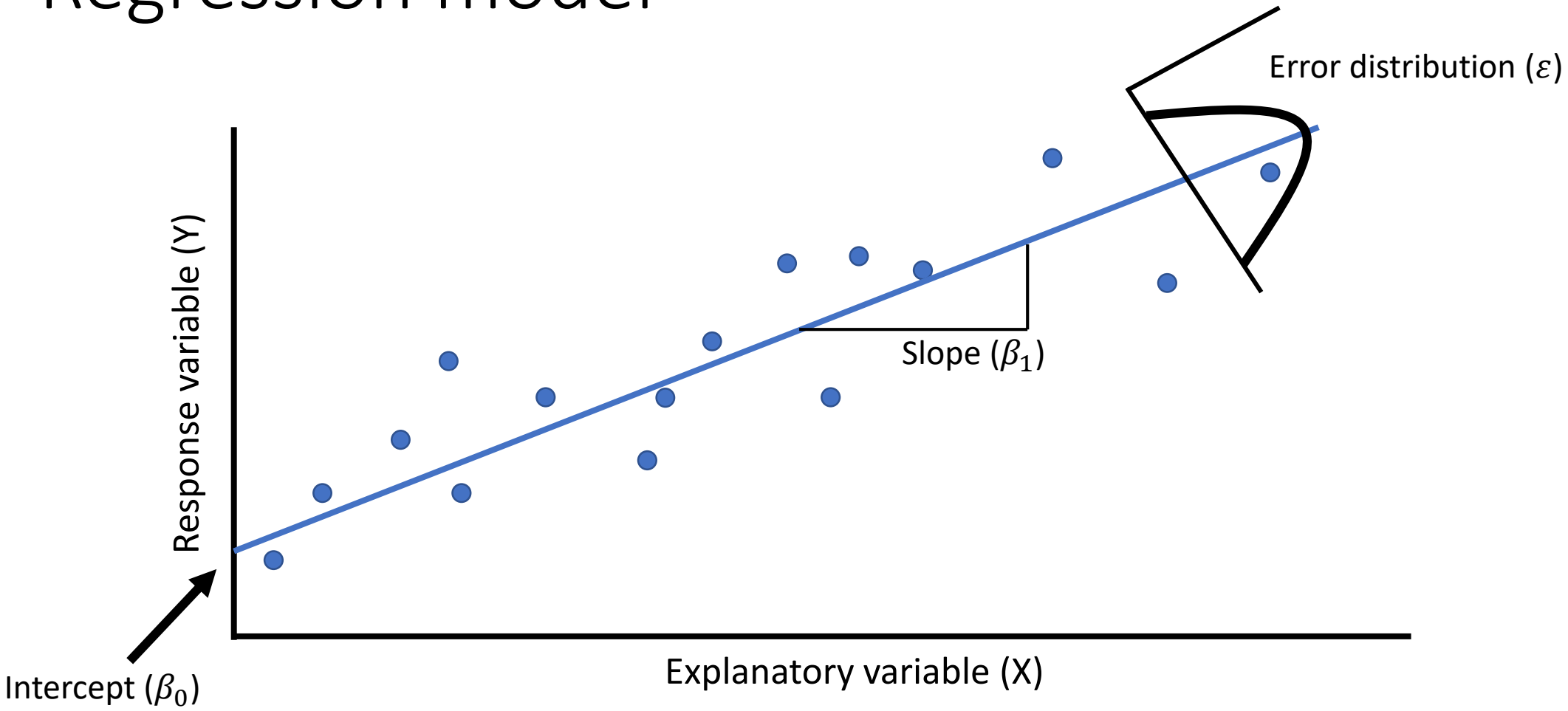
- Mathematical representation of observed data
- Allows relationships between variables to be identified
- Enables predictions

Statistical modeling

"Modeling is an art, as well as a science and, is directed toward finding a good approximating model … as the basis for statistical inference"
– Burnham & Anderson



Regression model



$$Y = \beta_0 + \beta_1 + \varepsilon$$

Goals of constructing models

1. Parameter estimation: What parameter values best fit the data?
2. Inference: How certain are the estimates the model produces?
3. Adequacy: Is the model the right choice?

Which model to use?

1. T-test: Continuous response variable with a 2-category predictor variable.
2. ANOVA: Continuous response variable with one or two 2+ category predictor variables.
3. Regression: Continuous response variable, 1 or more continuous predictor variables.
4. ANCOVA: Continuous response variable, mix of categorical and continuous predictor variables.

T-test

Common use: Predictor variable is two categories

Types of tests:

- One-sample t-test
- Independent two-sample t-test
- Paired t-test

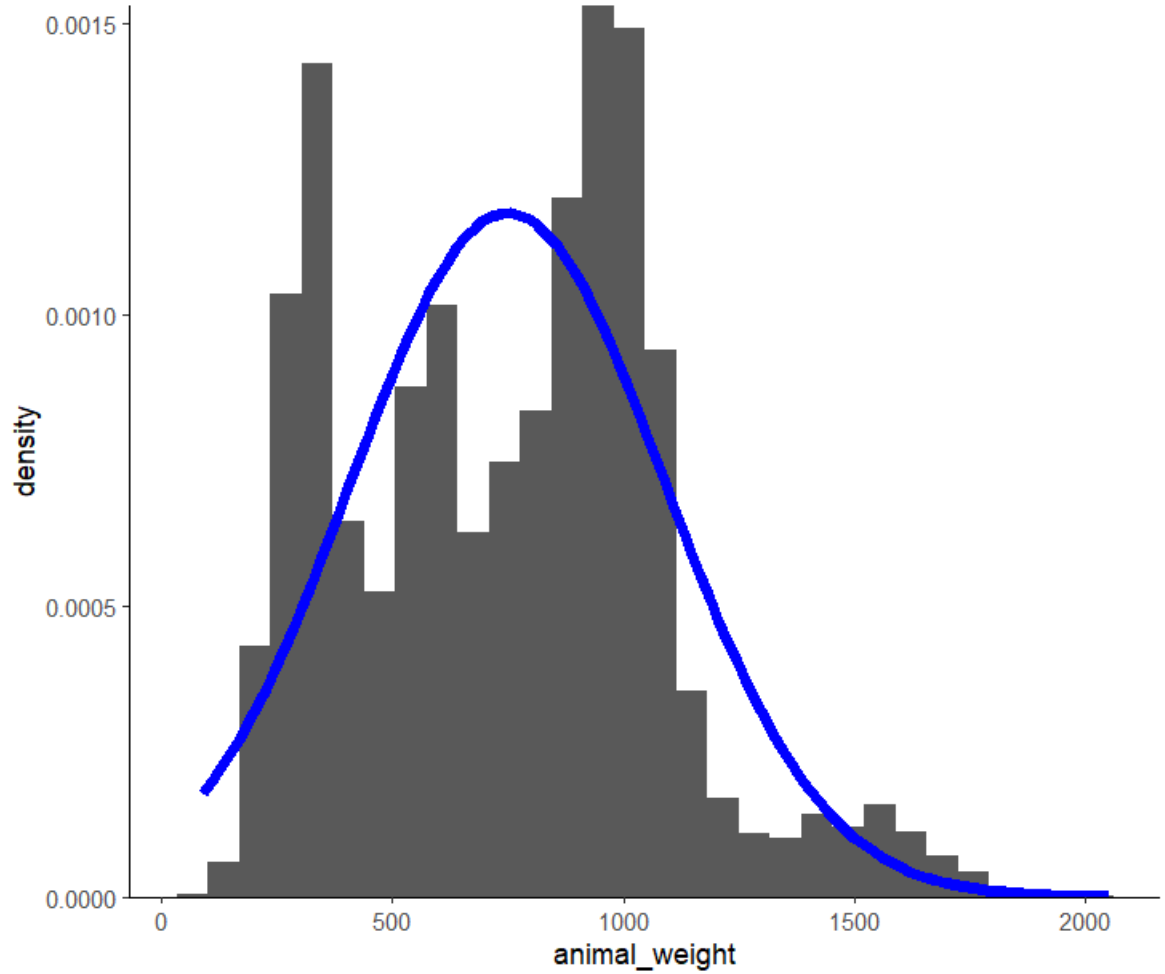
T-test assumptions

1. **Normality:** Model residuals are approximately normally distributed.
2. **Homogeneity of variances:** Both samples have approximately the same variance.
3. **Random sampling:** Both samples were obtained using a random sampling method.
4. **Independence:** The observations in one sample are independent of the observations in the other sample.

Testing normality: Visual approach

Create a histogram

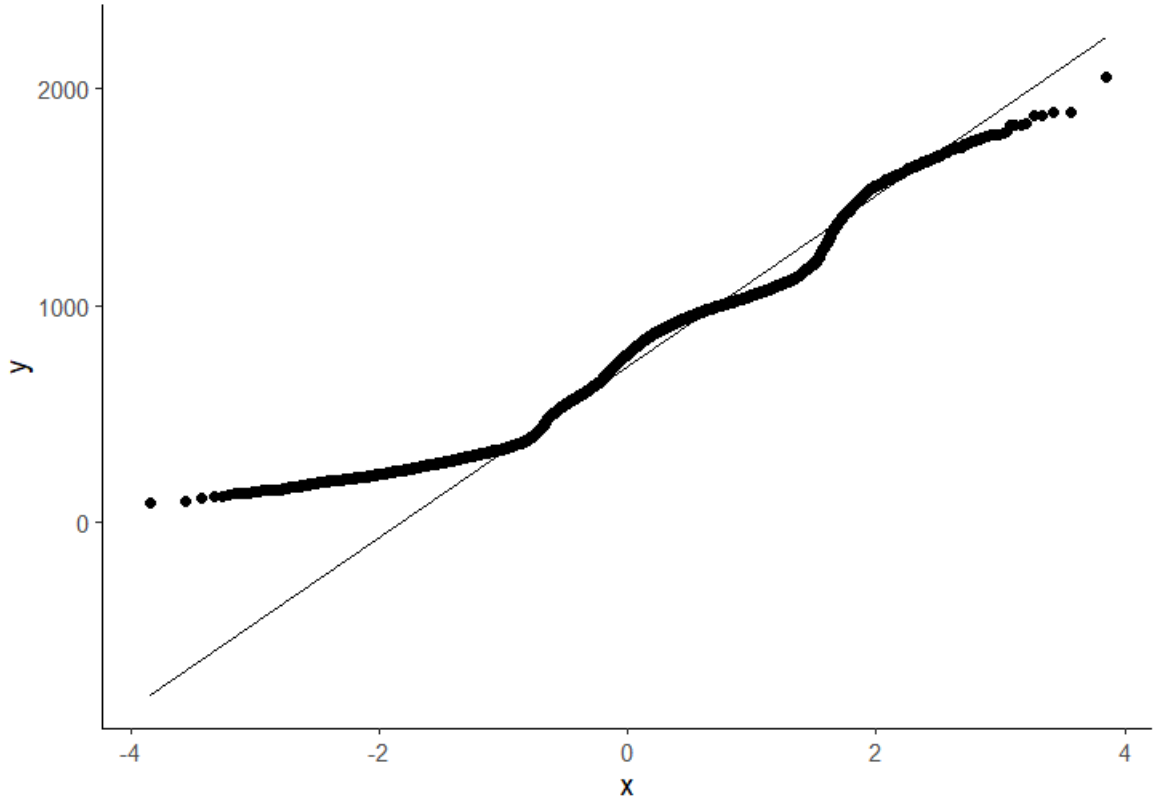
- Normal distribution shown in blue
- No clear threshold for “normal or not”
- Qualitative approach



Testing normality: Visual approach

Create a QQ plot

- Quantile-Quantile plot
  - Quantiles of 1 dataset against the other
  - 1 is observed (y) and 1 is theoretical (x)
- Normality follows the 1:1 line
- No clear threshold for “normal or not”
- Qualitative approach



Testing normality: Statistical approach

Shapiro-Wilk Test via rstatix::shapiro_test

Dat %>% shapiro_test(var1, var2, var3…)

Null hypothesis: Data are normally distributed

- p-value > 0.05 indicates normal distribution

<u>If non-normality is observed?</u>

Switch to non-parametric Welch's t-test using t.test() argument var.equal = FALSE

T-test assumptions

Homogeneity of Variances: Samples have approximately the same variance.

Statistical assessment methods

- F-test (ratio of variances)
  - var.test()
- Bartlett's test
  - bartlett.test()
- Levene's test
  - rstatix::levene_test()
  - Tidyverse-friendly approach

<u>If variances are heterogeneous?</u>

Switch to non-parametric Welch's t-test using t.test() argument var.equal = FALSE

One sample t-test

Common use: Predictor variable is two categories

Is the mean of the sample different from an expected value?

```
t.test(sample, mu = ref) %>%
    tidy()
```

Sample mean

t-statistic

Degrees of Freedom

Sample mean lower & upper CI

Alternative hypothesis (two sided, upper, lower)

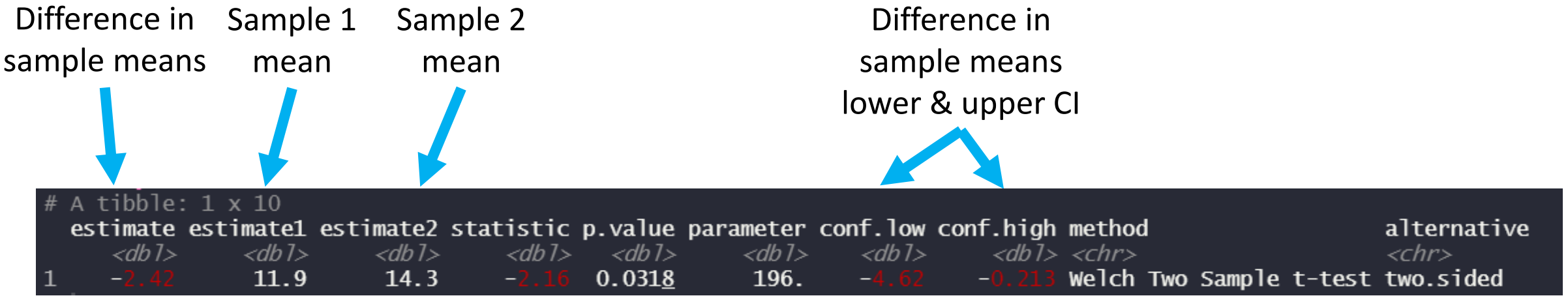
```
# A tibble: 1 x 8
  estimate statistic     p.value parameter conf.low conf.high method          alternative
     <dbl>     <dbl>       <dbl>     <dbl>    <dbl>     <dbl> <chr>           <chr>
1     14.9     -6.34 0.000000176        39     14.2      15.6 One Sample t-test two.sided
```

Two sample t-test

Common use: Predictor variable is two categories

Are the means of two independent predictor variables different?

```
t.test(response ~ predictor , data = dat, var.equal = TRUE) %>%
    tidy()
```



Paired t-test

Common use: Predictor variable is two non-independent categories

Do the means of two non-independent variables different?

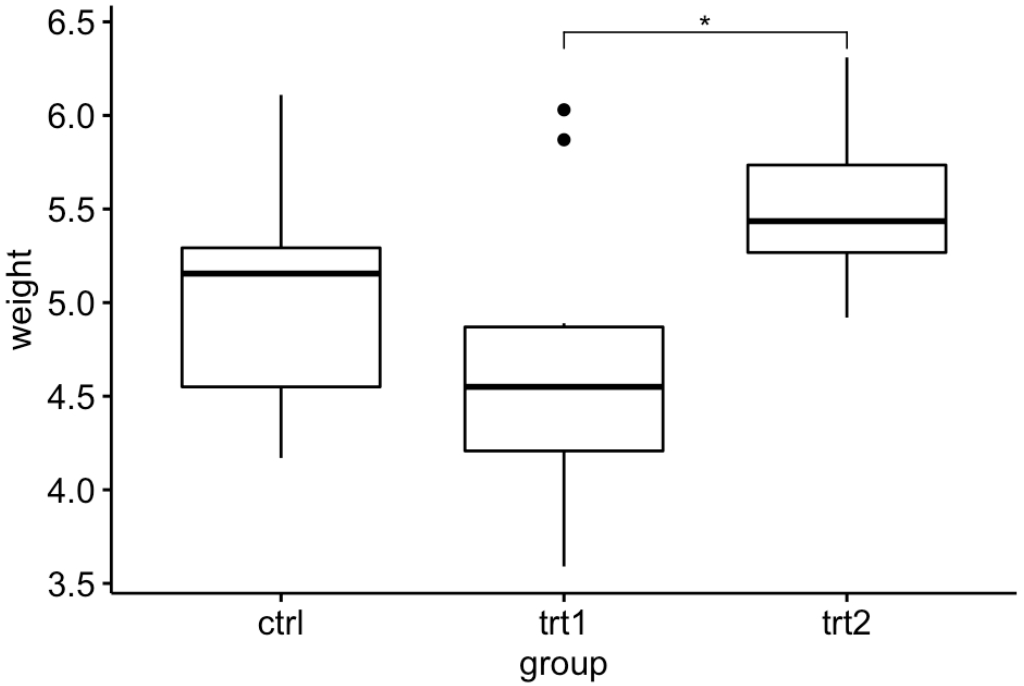
```
t.test(response ~ predictor , data = dat, paired = TRUE) %>%
    tidy()
```

```
# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high method      alternative
     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <chr>       <chr>
1    -1.22     -1.84  0.0680       164    -2.54    0.0918 Paired t-test two.sided
```

ANOVA

Common use: Explanatory variable is more than two categories

Do the means of more than two independent samples differ?



ANOVA assumptions

1. Normality: Model residuals are approximately normally distributed.

2. Homogeneity of variances: Both samples have approximately the same variance.

3. Random sampling: Samples were obtained using a random sampling method.

4. Independence: The observations in one sample are independent of the observations in the other sample.

Violated ANOVA assumptions

One-way ANOVA: Kruskal-Wallis ANOVA

kruskal.test(response ~ predictor, data = dat)

Two-way ANOVA: variable transformations

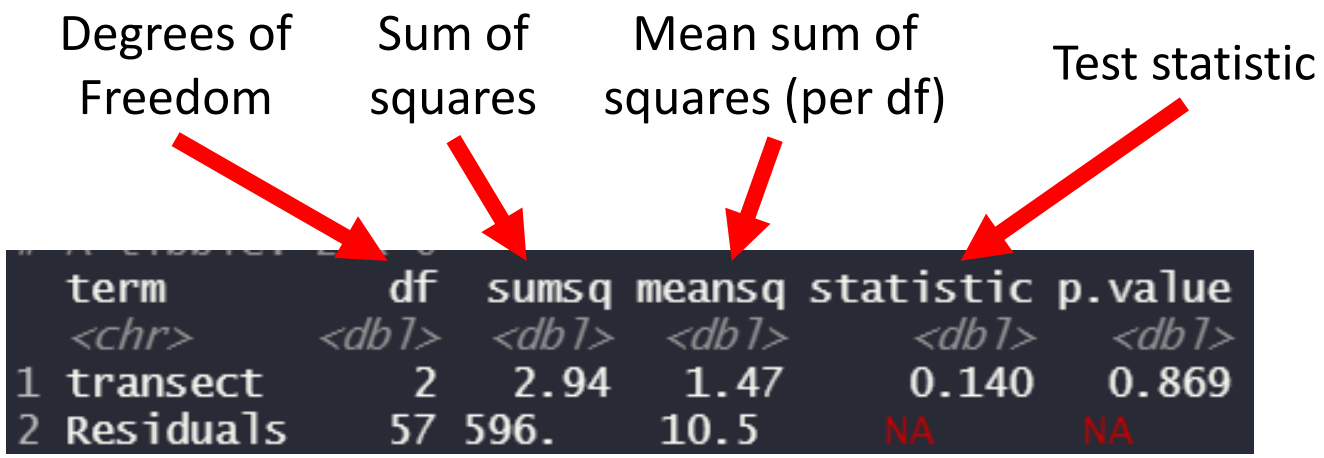
[Applications are beyond the scope of this course]

log(var), sqrt(var), etc.

One-way ANOVA

Common use: Determine whether differences exist between the means of three or more independent (unrelated) samples.

```
aov(response ~ predictor, data = dat) %>%
  tidy()
```



term	df	sumsq	meansq	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 transect	2	2.94	1.47	0.140	0.869
2 Residuals	57	596.	10.5	NA	NA

One-way ANOVA: Alternative methods

```
lm(response ~ predictor, data = dat) %>%
  summary()
```

```
lm(response ~ predictor, data = dat) %>%
  anova() %>%
  tidy()
```

```
Call:
lm(formula = leaf1area ~ transect, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9911 -2.5478  0.1162  2.4588  7.7210

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8788     0.7232  13.660   <2e-16 ***
transectR2   -0.5129     1.0228  -0.501    0.618
transectR6   -0.4078     1.0228  -0.399    0.692
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.234 on 57 degrees of freedom
Multiple R-squared:  0.0049,    Adjusted R-squared:  -0.03002
F-statistic: 0.1403 on 2 and 57 DF,  p-value: 0.8694
```

term	df	sumsq	meansq	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 transect	2	2.94	1.47	0.140	0.869
2 Residuals	57	596.	10.5	NA	NA

t-test tests whether coefficients are significantly different from zero

Assessment of model fit

Two-way ANOVA

Used to determine the effect of two categorical predictor variables on a continuous response variable

```
aov(response ~ predictor1 + predictor2, data = dat) %>%
    tidy()
```

	term	df	sumsq	meansq	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	basin	2	0.00323	0.00162	15.3	0.0000132
2	habitat	1	0.000746	0.000746	7.08	0.0114
3	Residuals	38	0.00401	0.000105	NA	NA

Two-way ANOVA: Post-hoc analysis

Post hoc: Latin phrase meaning “after this.” Used to describe follow-up analyses.

Tukey Honest Significant Differences: Evaluates combinations of categories within variables (additive model) or within and among variables (interactive model)

```
model <- aov(response ~ predictor1 + predictor2, data = dat)
TukeyHSD(model)
```