

Linear Regression

Zack Treisman

Spring 2026

Linear Regression

- ▶ Model $\hat{y} = f(x)$
- ▶ Data (x_i, y_i)
- ▶ Choose parameters that minimize **mean squared error**

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

No predictor: $\hat{y} = \beta_0$.

Goal: Find \hat{y} to minimize $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2$.

Solution: Calculus $\implies \frac{-2}{n} \sum_{i=1}^n (y_i - \hat{y}) = 0$ and solve for \hat{y} .

$$\frac{-2}{n} \sum_{i=1}^n (y_i - \hat{y}) = 0$$

$$\frac{-2}{n} \left(\sum_{i=1}^n y_i - \sum_{i=1}^n \hat{y} \right) = 0$$

$$\frac{-1}{n} \sum_{i=1}^n y_i + \hat{y} = 0$$

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Minimizing $MSE \implies \hat{y} = \text{mean}(y_i)$.

Linear $f(x)$, normal errors

$$Y \sim N(\beta_0 + \beta_1 X, \sigma).$$

Parameters:

- ▶ β_0 (the **intercept**)
- ▶ β_1 (the **slope**)
- ▶ σ (the **standard deviation**)

Analytic solution:

Notation: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Calculus \implies

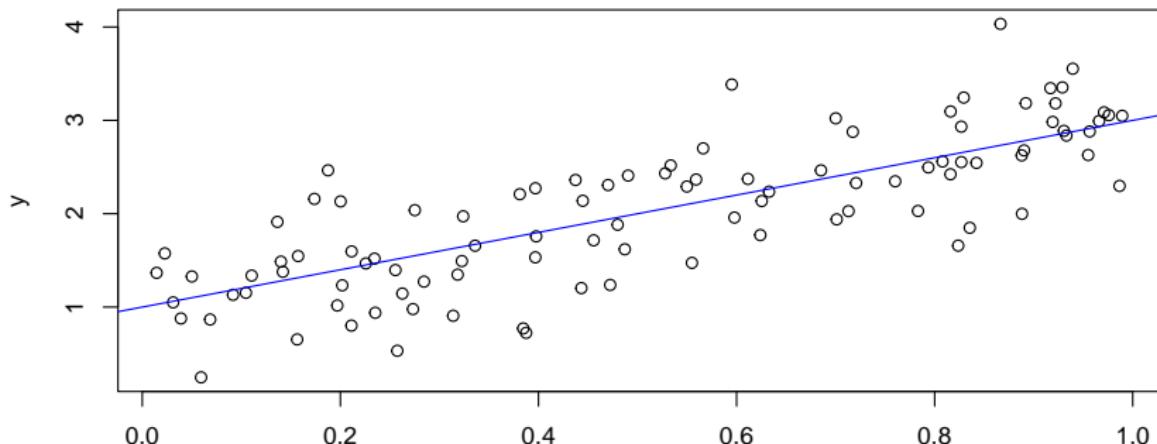
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

$$\hat{\sigma} = \text{st.dev}(y_i - \hat{y}_i)$$

Simulate an example

- ▶ $\beta_0 = 1, \beta_1 = 2, \sigma = 0.5$
- ▶ X is uniformly distributed
- ▶ Y is normally distributed about $1 + 2X$.

```
set.seed(5)
x <- runif(100)
y <- rnorm(100, mean = 1 + 2*x, sd = 0.5)
plot(y~x)
abline(1,2, col="blue")
```

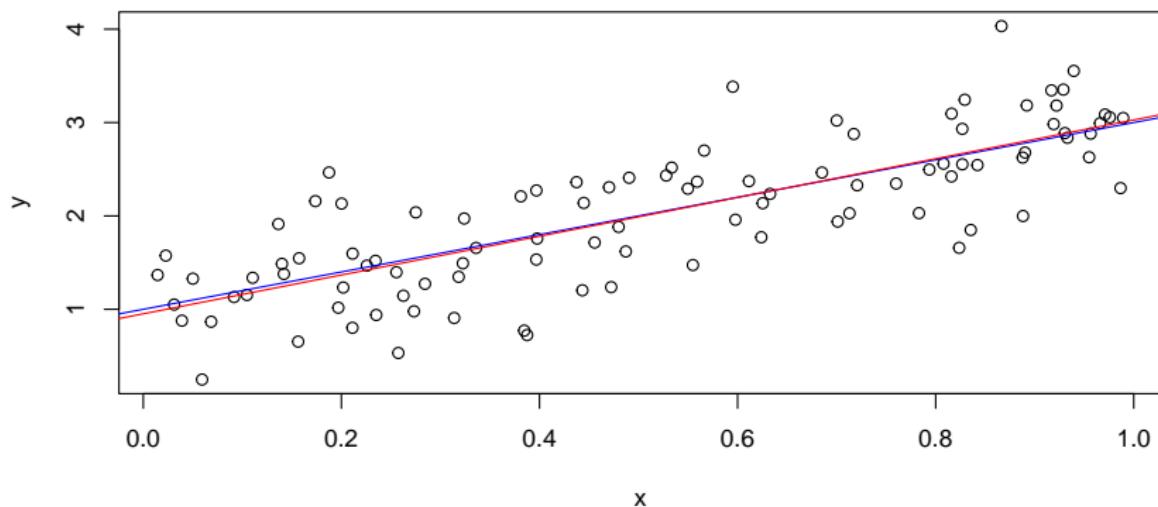


The lm function in R

```
lm1 <- lm(y~x)
lm1$coefficients

## (Intercept)          x
##      0.9515      2.0765

plot(y~x); abline(1,2, col="blue"); abline(lm1, col="red")
```



Evaluating a linear model

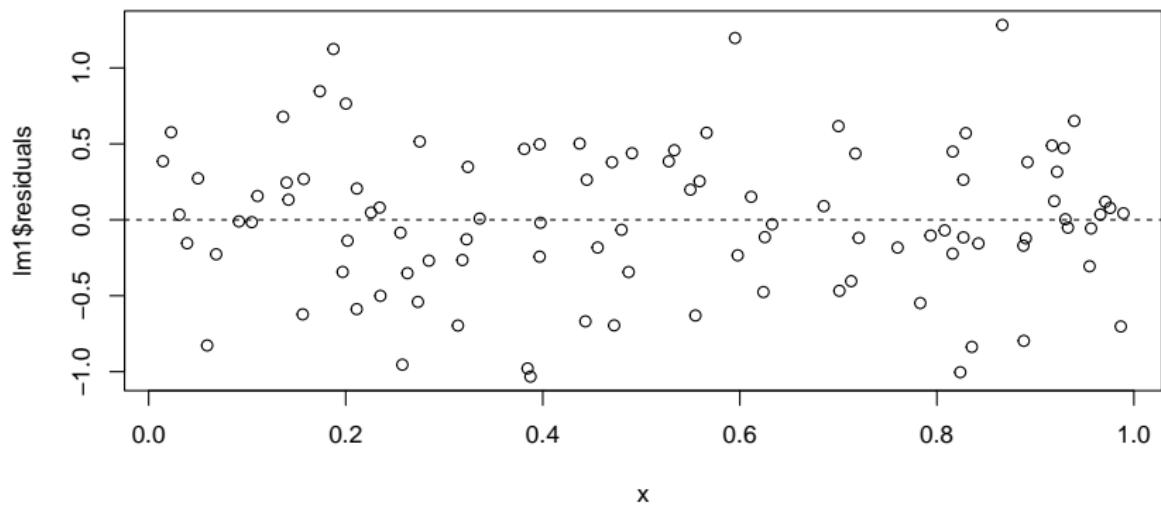
```
summary(lm1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0323 -0.2661 -0.0131  0.3558  1.2823
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.9515    0.0966   9.85  2.5e-16 ***
## x            2.0765    0.1610  12.89  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.485 on 98 degrees of freedom
## Multiple R-squared:  0.629, Adjusted R-squared:  0.625 
## F-statistic: 166 on 1 and 98 DF,  p-value: <2e-16
```

Check residuals

- ▶ Normally distributed around 0?
- ▶ Consistent variance?

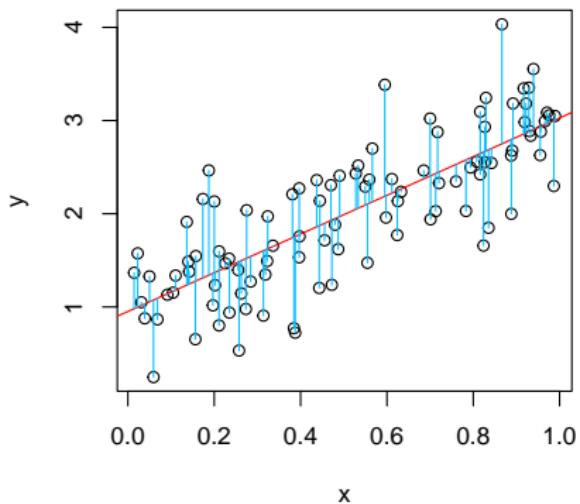
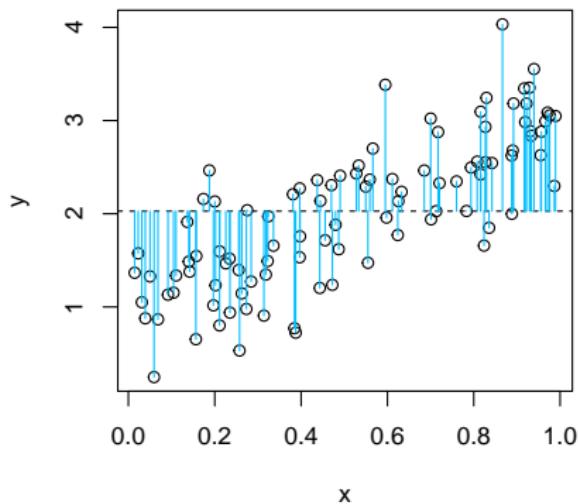
```
plot(lm1$residuals~x); abline(0,0, lty="dashed")
```



Overall accuracy

R^2 : how much of the variance in Y is described by the model.

$$R^2 = 1 - \frac{\sigma_{\text{residuals}}^2}{\sigma_y^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{(\text{blue on right})^2}{(\text{blue on left})^2}$$



Accuracy of the coefficient estimates

$SE(\hat{\beta}) \leftarrow$ how $\hat{\beta} \sim \beta$ varies from sample to sample.

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad SE(\hat{\beta}_0) = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- ▶ Standard errors → confidence intervals
- ▶ 95% chance that the interval

$$\hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1)$$

contains the true value of β_1 .

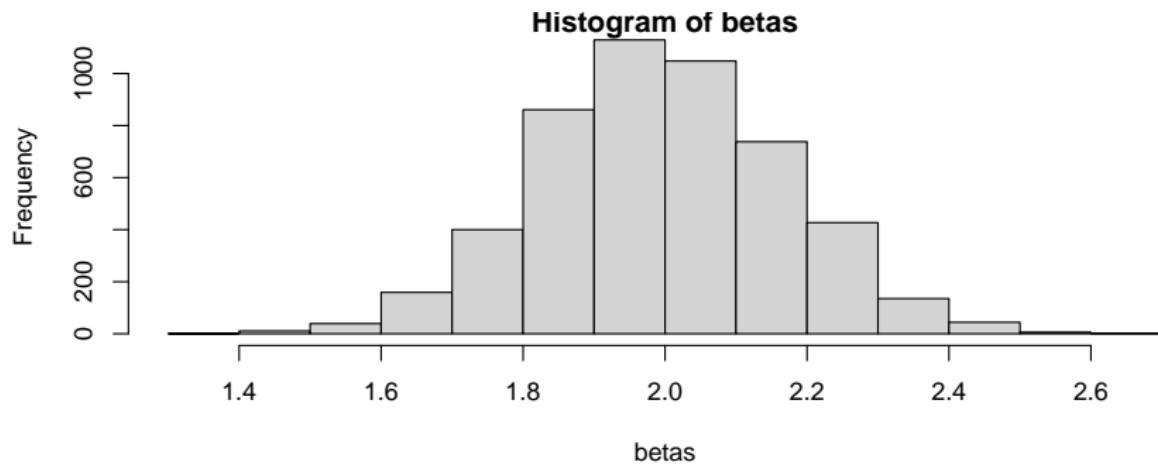
- ▶ $\beta_1 = 2$

$$2.076 \pm 1.96 \cdot 0.161 = (1.760, 2.392)$$

Simulate: $N = 5000$, calculate β_1

```
betas <- numeric(5000); b_captured <- logical(5000)
for(i in 1:5000){
  x <- runif(100); y <- rnorm(100, mean = 1 + 2*x, sd = 0.5)
  lmi <- lm(y~x); slope <- coef(summary(lmi))[2,]
  betas[i] <- slope[1]
  b_captured[i] <- (2>slope[1]-1.96*slope[2]) & (2<slope[1]+1.96*slope[2])
}
mean(b_captured); hist(betas)

## [1] 0.9526
```



Hypothesis testing

standard errors \implies hypothesis tests.

$$H_0 : \beta_1 = 0$$

(no relationship between X and Y)

- ▶ The t statistic for this test is

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}.$$

- ▶ Null hypothesis distribution: t_{n-2}

Multiple Linear Regression

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \sigma)$$

- ▶ β_j : average effect on Y of a one unit increase in X_j , holding all other predictors fixed.
- ▶ **balanced design** \leftrightarrow uncorrelated predictors
 - ▶ Each coefficient can be estimated and tested separately.
 - ▶ Interpreting β_j as above is possible.
- ▶ Correlated predictors cause problems:
 - ▶ coefficient variance \uparrow
 - ▶ Interpretability \downarrow

Accuracy revisited: R^2

Write $SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$ and $SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{tot} - SS_{res}}{SS_{tot}}.$$

- ▶ Problem:
 - ▶ Suppose X_j has *nothing* to do with Y .
 - ▶ $\hat{\beta}_j = 0$ is *very* unlikely - it comes from real data.
 - ▶ SS_{res} is smaller because X_j fit some noise.
 - ▶ R^2 is larger because of X_j .

Two alternatives to R^2

- ▶ Penalize models for having more predictors:

$$\text{Adjusted } R^2 = 1 - \frac{SS_{\text{res}}/(n-p-1)}{SS_{\text{tot}}/(n-1)}$$

- ▶ R^2 when $n \gg p$.
- ▶ Interpret as the proportion of the variance in the response explained by the model.
- ▶ F statistic.

$$F = \frac{(SS_{\text{tot}} - SS_{\text{res}})/p}{SS_{\text{res}}/(n-p-1)}$$

- ▶ Distributed as $F_{p,n-p-1}$ if $\beta_j = 0$ for all $j \neq 0$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_a : At least one of the β_j is non-zero.

Multiple regression with lm

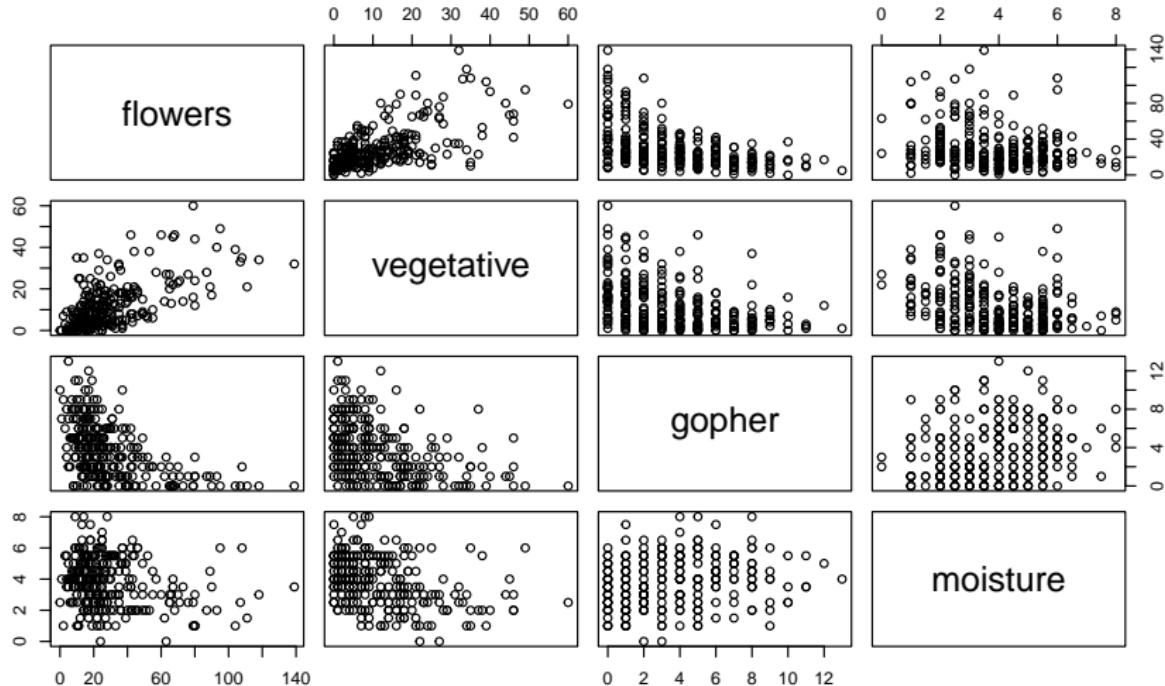
Lily data from Thomson et al. (1996).

```
lm2 <- lm(flowers~vegetative+gopher+moisture, data=Lily_sum) # data in emdbook
summary(lm2)
```

```
##
## Call:
## lm(formula = flowers ~ vegetative + gopher + moisture, data = Lily_sum)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -52.51 -10.45   -2.71    7.41   79.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.682     4.027    6.13  3.4e-09 ***
## vegetative   1.076     0.112    9.64  < 2e-16 ***
## gopher      -2.217     0.413   -5.37  1.8e-07 ***
## moisture     0.176     0.731    0.24     0.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.4 on 252 degrees of freedom
## Multiple R-squared:  0.452, Adjusted R-squared:  0.446
## F-statistic: 69.4 on 3 and 252 DF,  p-value: <2e-16
```

Pairs plot of the Lily data

```
pairs(Lily_sum[,c("flowers", "vegetative", "gopher", "moisture")])
```



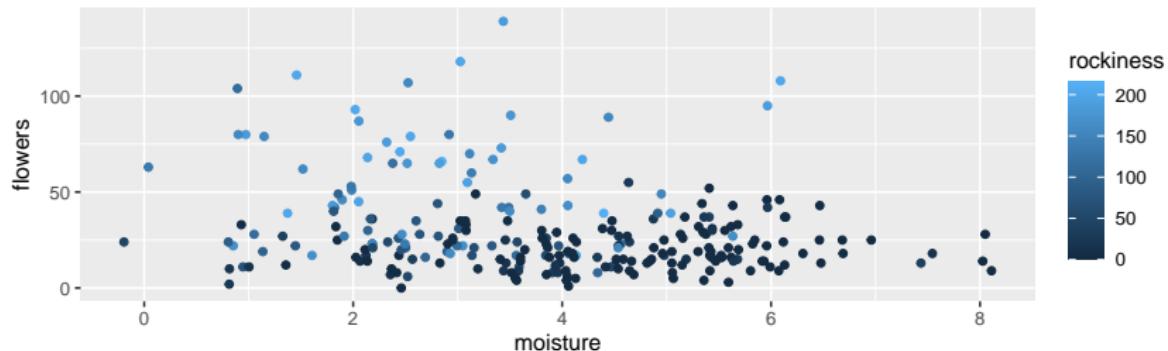
Confounding

The Lily data contain an additional variable.

```
lm3 <- lm(flowers~vegetative+gopher+moisture+rockiness, data=Lily_sum)
summary(lm3)$coefficients; summary(lm3)$adj.r.squared
```

```
##             Estimate Std. Error t value Pr(>|t|) 
## (Intercept) 8.5047    4.00178  2.125 3.454e-02
## vegetative   0.7344    0.10570  6.948 3.181e-11
## gopher      -1.0032    0.38888 -2.580 1.046e-02
## moisture     1.9940    0.67595  2.950 3.479e-03
## rockiness    0.1650    0.01903  8.669 5.500e-16
## [1] 0.5718
```

```
ggplot(Lily_sum, aes(moisture, flowers, color=rockiness))+
  geom_jitter(height = 0)
```



Interactions

To include an interaction term use * in the formula.

```
lm4 <- lm(flowers~vegetative+gopher+moisture*rockiness, data=Lily_sum)
summary(lm4)$coefficients; summary(lm4)$adj.r.squared
```

```
##                                     Estimate Std. Error t value Pr(>|t|) 
## (Intercept)                 4.51090   4.36902  1.032 3.028e-01
## vegetative                  0.73338   0.10491  6.990 2.485e-11
## gopher                     -1.07010   0.38717 -2.764 6.136e-03
## moisture                    2.98271   0.80818  3.691 2.746e-04
## rockiness                   0.24005   0.03909  6.140 3.215e-09
## moisture:rockiness          -0.02272   0.01036 -2.194 2.916e-02
## [1] 0.5782
```

The resulting model:

$$\widehat{\text{flw}} = 4.51 + 0.73\text{veg} - 1.07\text{gph} + 2.98\text{mst} + 0.24\text{rck} - 0.02(\text{mst} \times \text{rck})$$

Or, alternatively

$$\widehat{\text{flw}} = 4.51 + 0.73\text{veg} - 1.07\text{gph} + (2.98 - 0.02\text{rck})\text{mst} + 0.24\text{rck}$$

Categorical predictors

- ▶ Presence/ Absence
- ▶ Treatment levels: (low, medium, high)
- ▶ Species

Encoded for regression using **indicator variables**.

- ▶ Also called **dummy variables, one-hot encoding**

Example: X is a categorical variable with levels a, b, c . Arbitrarily choose a as the **reference level** and define

$$Z_b = \begin{cases} 0 & \text{if } X = a \text{ or } c \\ 1 & \text{if } X = b \end{cases} \quad Z_c = \begin{cases} 0 & \text{if } X = a \text{ or } b \\ 1 & \text{if } X = c \end{cases}$$

One way ANOVA

Continuing with the above example.

$$Y \sim N(\beta_0 + \beta_1 Z_b + \beta_2 Z_c, \sigma).$$

Regression gives a model.

- ▶ If $X = a$, the model predicts $\hat{y} = \hat{\beta}_0$.
- ▶ If $X = b$, the model predicts $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$.
- ▶ If $X = c$, the model predicts $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2$.

Using the F statistic to test

$$H_0 : \beta_1 = \beta_2 = 0, \quad H_a : \text{At least one of } \beta_1 \text{ or } \beta_2 \text{ is non-zero}$$

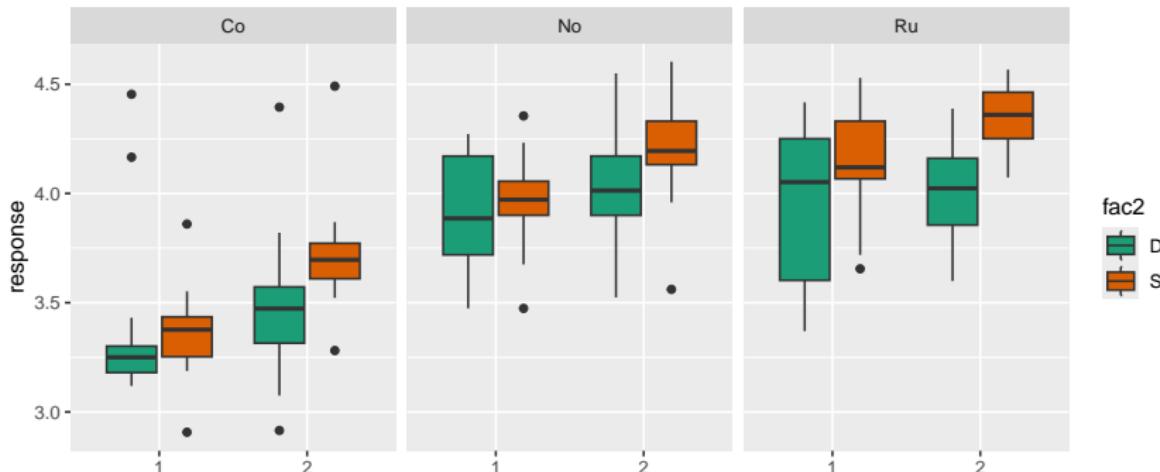
is called **analysis of variance** or ANOVA.

Multi-way ANOVA

Data from an experiment involving tadpoles.

- ▶ fac1 = experimental treatment
- ▶ fac2 = diet
- ▶ fac3 = tadpole genetics

```
tadpoles <- read.csv("data/tadpoles.csv")
tadpoles$fac3 <- as.factor(tadpoles$fac3) # It's coded as 1 or 2.
ggplot(tadpoles, aes(fac3, response, fill = fac2)) +
  geom_boxplot() + facet_wrap(~fac1) +
  scale_fill_brewer(palette = "Dark2")      # The default colors get boring.
```



A more general F statistic

F can compare nested models.

$$F = \frac{(SS_{old} - SS_{new})/df_{num}}{SS_{new}/df_{den}}.$$

- ▶ **numerator degrees of freedom** = number of new parameters
- ▶ **denominator degrees of freedom** = (number of data points)
- (total number of parameters in the extended model)

anova()

```
lm5 <- lm(response~fac1*fac2*fac3, data = tadpoles)
anova(lm5) # summary is not as useful as it analyzes indicator variables

## Analysis of Variance Table
##
## Response: response
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## fac1          2   18.43   9.22  151.79 < 2e-16 ***
## fac2          1    1.50   1.50   24.72 1.3e-06 ***
## fac3          1    2.28   2.28   37.50 4.0e-09 ***
## fac1:fac2     2    0.39   0.20    3.23   0.041 *
## fac1:fac3     2    0.08   0.04    0.69   0.503
## fac2:fac3     1    0.35   0.35    5.77   0.017 *
## fac1:fac2:fac3 2    0.07   0.03    0.57   0.565
## Residuals    227   13.78   0.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that fac1, fac2 and fac3 all have significant effects on the response, as do the interactions fac1:fac2 and fac2:fac3.

The model suggested by ANOVA

Now we can build a model using only those terms listed as significant.

```
lm5a <- lm(response~fac1+fac2+fac3+fac1:fac2+fac2:fac3, data = tadpoles)
summary(lm5a)$coefficients
```

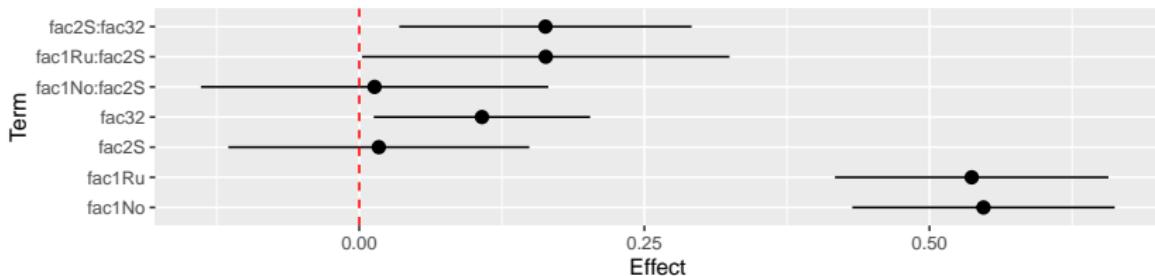
	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	3.38324	0.05245	64.4995	1.024e-149
## fac1No	0.54730	0.05837	9.3760	6.583e-18
## fac1Ru	0.53699	0.06085	8.8249	2.755e-16
## fac2S	0.01715	0.06696	0.2561	7.981e-01
## fac32	0.10758	0.04815	2.2343	2.642e-02
## fac1No:fac2S	0.01341	0.07728	0.1736	8.623e-01
## fac1Ru:fac2S	0.16350	0.08175	2.0001	4.666e-02
## fac2S:fac32	0.16323	0.06505	2.5094	1.278e-02

- ▶ The coefficients on fac2S and fac1No:fac2S are not significant, but are included because
 - ▶ include an interaction effect \implies include the corresponding main effects, and
 - ▶ include an effect from one level of a factor \implies include all levels.

Interpreting the result

Plot the coefficients with confidence intervals.

```
ests <- coef(lm5a)[-1] # The reference level is not of immediate interest.  
tad_model <- data.frame(var.labels=factor(names(ests)), levels=names(ests)), ests,  
    low95 = confint(lm5a)[-1,1], up95 = confint(lm5a)[-1,2])  
ggplot(tad_model, aes(var.labels, ests))+  
    geom_pointrange(aes(ymin=low95, ymax=up95))+  
    geom_hline(yintercept=0, linetype = "dashed", color = "red") +  
    labs(x = "Term", y = "Effect") + coord_flip()
```



The reference treatment is CoD1.

- ▶ Ru and No differ from Co but not each other.
- ▶ On its own, diet does not have a significant effect.
- ▶ Genetic factors 1 and 2 have different baseline response levels.
- ▶ Diet S in combination with Ru or genetic factor 2 has an effect.

Type I (Sequential) and Type II (Marginal) ANOVA

Type I anova **sequentially** adds each term in a list to a model.

Type II or marginal anova compares a model to the model including all possible other terms.

Often, type II is preferred.

- ▶ Why evaluate fac1 against the null model, fac2 against fac1, and fac3 against fac1 and fac2 if the order in which they are labelled is arbitrary?
- ▶ Type II anova is more robust to unequal group sizes.

Marginal ANOVA using the car package

`anova()` → sequential anova.

`Anova()` from package `car` → marginal anova.

`Anova(lm5)`

```
## Anova Table (Type II tests)
##
## Response: response
##           Sum Sq Df F value Pr(>F)
## fac1        18.08  2 148.86 < 2e-16 ***
## fac2         1.65  1   27.17 4.2e-07 ***
## fac3         2.23  1   36.72 5.6e-09 ***
## fac1:fac2    0.30  2     2.47   0.087 .
## fac1:fac3    0.05  2     0.45   0.637
## fac2:fac3    0.35  1     5.77   0.017 *
## fac1:fac2:fac3 0.07  2     0.57   0.565
## Residuals   13.78 227
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

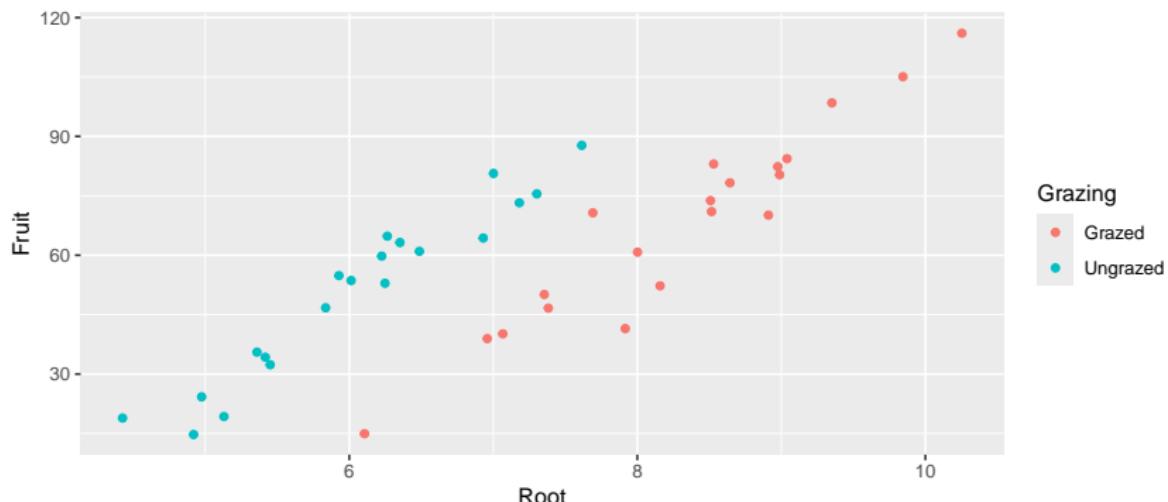
Results are similar to the type I analysis, but the p values for `fac1:fac2` are on opposite sides of the bright line of 0.05.

Combining numerical and categorical predictors

Often we have both numerical and categorical predictors.

Rabbits grazing on plants example from Crawley (2012, 538).

```
ipo <- read.csv('data/ipomopsis.csv')
ggplot(data=ipo, aes(x=Root, y=Fruit, color = Grazing))+
  geom_point()
```



ANCOVA

Does the categorical predictor *Grazing* affect the numerical response *Fruit*?

- ▶ The numerical variable *Root* is a confounder.
- ▶ This is classical **analysis of covariance** or ANCOVA.

```
lm6<- lm(Fruit~Root*Grazing, data=ipo)
anova(lm6) # sequential and marginal are identical in this case
```

```
## Analysis of Variance Table
##
## Response: Fruit
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Root          1 16795   16795   360.0 < 2e-16 ***
## Grazing       1  5264    5264   112.8 1.2e-12 ***
## Root:Grazing 1      5        5     0.1     0.75
## Residuals    36   1680      47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

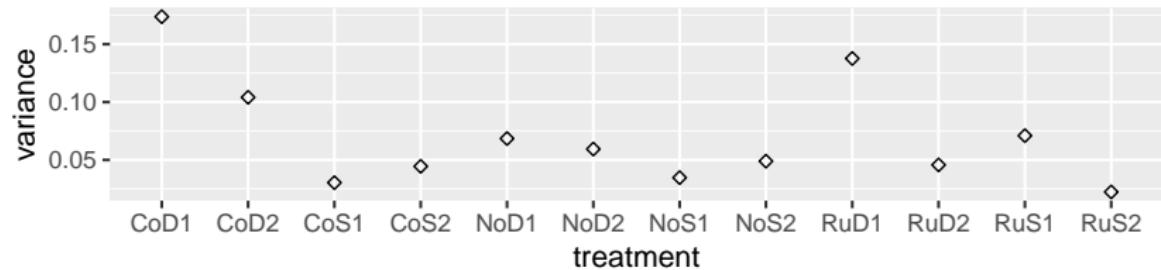
- ▶ *Root* → best fit lines have slope.
- ▶ *Grazing* → groups have different intercepts.
- ▶ *Root:Grazing* → slope is the same for both groups.

Assumption: Homogeneity of Variance

homoscedasticity = Constant residual variance

CoD1, CoD2 and RuD1 have higher variance. Is this a problem?

```
ggplot(tadpoles, aes(treatment, response)) +  
  stat_summary(fun=var, geom="point", shape = 23) + # Show variances.  
  labs(y="variance")
```



- ▶ CoD1 and CoD2: variance is due to outliers. Run the analysis without them - does the result change?
- ▶ RuD1: variance is large, but errors are symmetric. Maybe okay?

Assumption: Independence of observations

Pseudoreplication = dependent observations

- ▶ Artificially increases power.
- ▶ Common scenarios where it is encountered:
 - ▶ **Repeated measures:** Observe one individual multiple times.
 - ▶ **Block designs and Split plots:** Values of one variable are constant for grouped sets of observations.

We will discuss solutions to these issues later in the course.

References

- Crawley, Michael J. 2012. *The R Book*. 2nd ed. Wiley Publishing.
- Thomson, James D., George Weiblen, Barbara A. Thomson, Satie Alfaro, and Pierre Legendre. 1996. "Untangling Multiple Factors in Spatial Distributions: Lilies, Gophers, and Rocks." *Ecology* 77 (6): 1698–1715. <http://www.jstor.org/stable/2265776>.