Regression and ANOVA

Zack Treisman

Spring 2021

Philosophy

Regression generally refers to a family of techniques where a model $\hat{y} = f(x)$ is fit to data (x_i, y_i) such that the **mean squared error**

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

is as small as possible given *a priori* assumptions on the form of $f(x)$, such as being a linear function $f(x) = \beta_0 + \beta_1 x$.

Regression is the starting point for linear regression, ANOVA and several other classical techniques.

Linear Regression

Linear regression posits a relationship between two numerical random variables X and Y of the form

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2).$$

The parameters β_0 (the **intercept**), β_1 (the **slope**) and σ^2 (the **variance**) are estimated from observations of X and Y. Specifically, writing $\bar{x} = \sum_{i=1}^{n} x_i$, and $\bar{y} = \sum_{i=1}^{n} y_i$ for the sample means and using a little calculus,
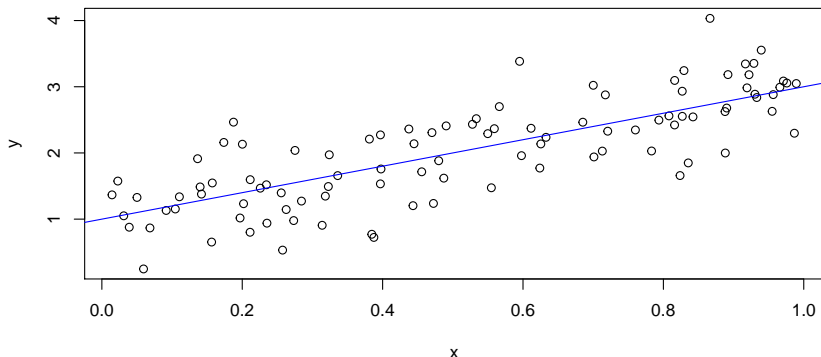
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The variance of the **residuals** $y_i - \hat{y}$ gives the estimate of σ^2.

The data linear regression expects

Estimates from these formulas are expected to be accurate when X is uniformly distributed and Y is normally distributed about a linear function of X.

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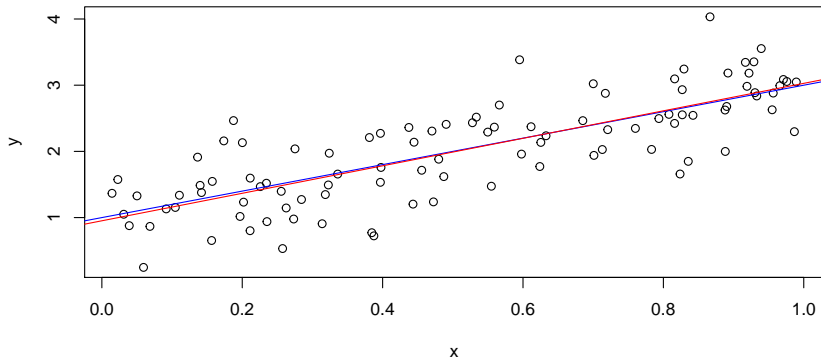
The `lm` function in R

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

```
## (Intercept)           x
##   0.9515095   2.0764806
```

```
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
## -1.03231 -0.26608 -0.01308  0.35581  1.28230
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95151    0.09657   9.853 2.52e-16 ***
## x            2.07648    0.16104  12.895  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4853 on 98 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.6254
## F-statistic: 166.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Check residuals

Residuals appear normally distributed around 0 with a consistent variance independent of X.

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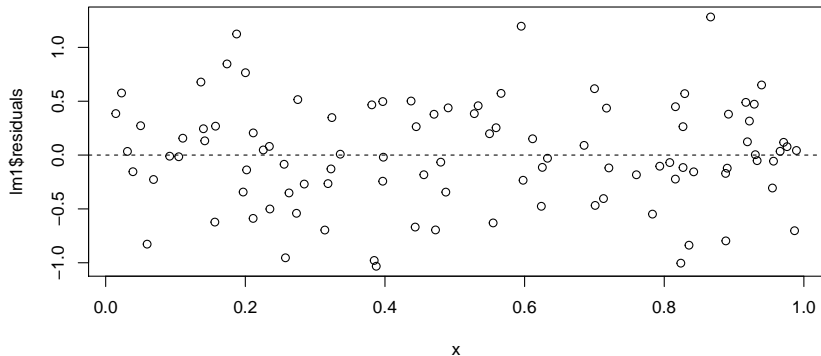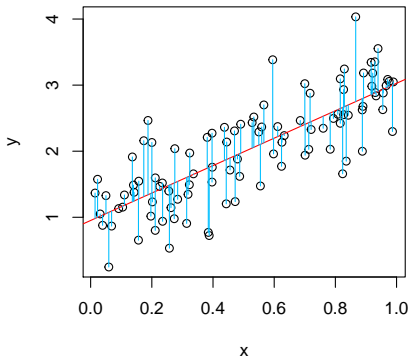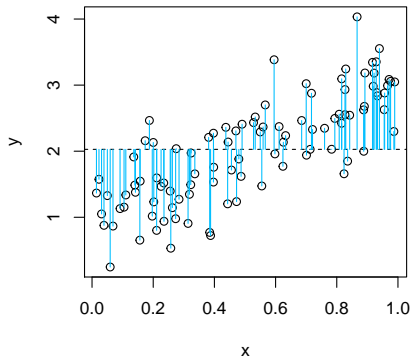

We'll discuss various ways to check homogeneity of variance.

There's no one correct way to do it.

Overall accuracy of the model

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

$$R^2 = 1 - \frac{\sigma^2_{\text{residuals}}}{\sigma^2_y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

We'll talk about Adjusted R^2 and the F-statistic shortly.

Accuracy of the coefficient estimates

The standard error of an estimator reflects how it varies under repeated sampling.

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

▶ Standard errors allow us to compute confidence intervals for these parameters.

▶ For samples such as ours, there is a 95% chance that the interval

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

contains the true value of β_1. (Which is 2.)

▶ In our case, the 95% confidence interval is

$$2.076 \pm 2 \cdot 0.161 = (1.754, 2.399)$$

Simulate taking 5000 such samples and calculating β_1

```r
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
```


Histogram of betas
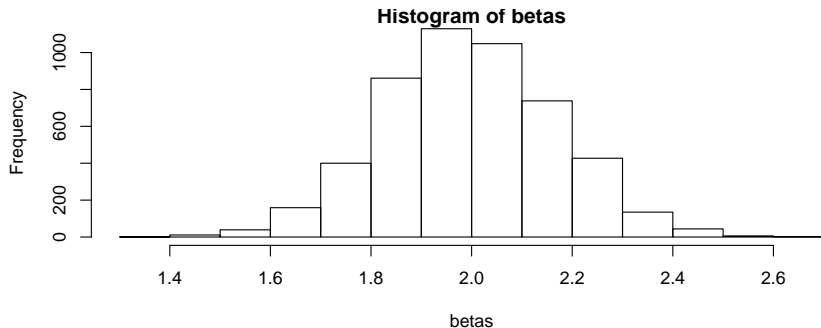
Hypothesis testing

Having the standard errors for the estimated parameters also allows us to do hypothesis tests.

▶ Generally we are not interested in testing the intercept.
▶ Testing the null hypothesis

$$H_0 : \beta_1 = 0$$

is equivalent to testing for no relationship between X and Y.
▶ The t statistic for this test is

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

▶ The null hypothesis distribution is a t distribution with $n - 2$ degrees of freedom.

Multiple Linear Regression

The theory is similar if there are multiple predictor variables.

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \sigma^2)$$

- The parameter β_j is the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.
- Ideally the predictors are uncorrelated — this is called a **balanced design**.
  - Each coefficient can be estimated and tested separately.
  - Interpreting β_j as above is possible.
- Correlations among predictors cause problems:
  - The variance of all coefficients tends to increase.
  - Interpretations become hazardous.

Overall accuracy of the model revisited

R^2 is defined exactly as for one variable linear models.

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

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

▶ An immediate problem is that even a predictor X_j that has *nothing* to do with Y is going to give *some* reduction in R^2, because $\hat{\beta}_j$ will not be *exactly* zero.

Two alternatives to R^2

- Adjust R^2 to penalize models for having more predictors:

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

  - This is still (loosely) interpretable as the proportion of the variance in the response explained by the model.
  - Nearly identical to R^2 when $n >> p$.
- Another alternative is the F statistic.

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

  - This will be distributed as $F_{p,n-p-1}$ if **all** of the β_j (except possibly β_0) are zero, so can be used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a : \text{At least one of the } \beta_j \text{ is non-zero.}$$

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
## -52.512 -10.447  -2.705   7.407  79.276
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.6817     4.0273   6.129 3.40e-09 ***
## vegetative    1.0758     0.1116   9.640  < 2e-16 ***
## gopher       -2.2171     0.4127  -5.372 1.77e-07 ***
## moisture      0.1761     0.7311   0.241     0.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.38 on 252 degrees of freedom
## Multiple R-squared:  0.4523,	Adjusted R-squared:  0.4458
## F-statistic: 69.38 on 3 and 252 DF,  p-value: < 2.2e-16
```

Pairs plot of the Lily data

```
pairs(Lily_sum[,c("flowers", "vegetative", "gopher", "moisture")])
```
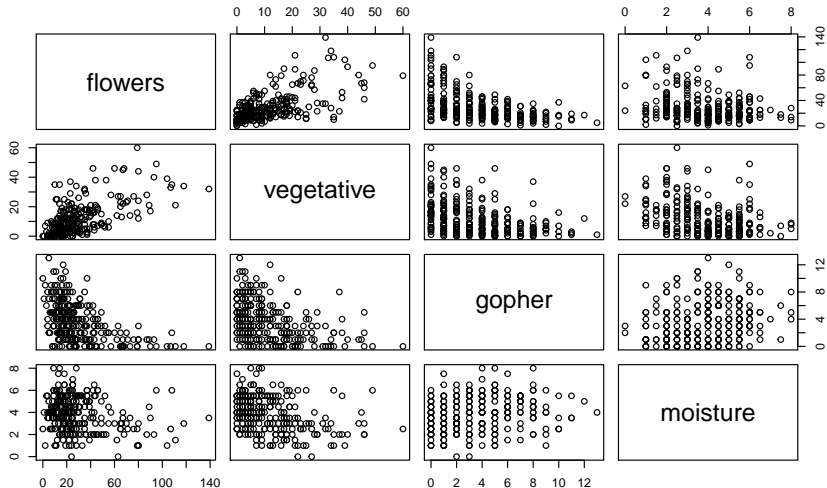
Confounders can have big effects

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)  8.5047487  4.0017825  2.125240 3.454361e-02
## vegetative   0.7343963  0.1057047  6.947620 3.181124e-11
## gopher      -1.0032202  0.3888850 -2.579735 1.045726e-02
## moisture     1.9940077  0.6759522  2.949924 3.478788e-03
## rockiness    0.1649542  0.0190285  8.668796 5.500007e-16
```

```
## [1] 0.5718083
```

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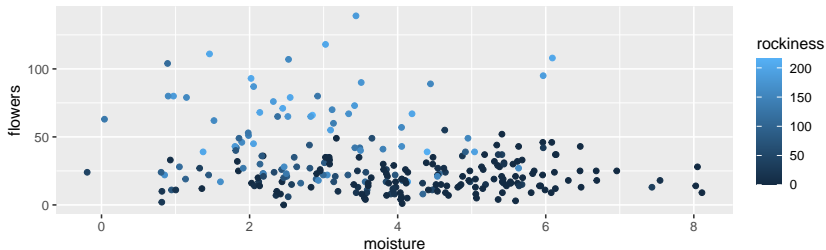
Interactions

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)        4.51089921 4.36901681  1.032475 3.028476e-01
## vegetative         0.73337897 0.10491177  6.990435 2.484626e-11
## gopher            -1.07010284 0.38716595 -2.763938 6.135862e-03
## moisture           2.98270503 0.80817645  3.690661 2.745527e-04
## rockiness          0.24005485 0.03909441  6.140389 3.215459e-09
## moisture:rockiness -0.02271932 0.01035527 -2.193987 2.915712e-02
```

```
## [1] 0.5782167
```

The resulting model has a term for the product of the interacting variables.

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

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

It is common for some or all of the predictor variables in a regression to be categorical.

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

Categorical variables are encoded for regression using **indicator (dummy) variables**.

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

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

One way ANOVA

Continuing with the above example. Given data, regression will estimate the parameters for a model

$$Y \sim N(\beta_0 + \beta_1 Z_b + \beta_2 Z_c, \sigma^2).$$

- ▶ 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 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2$.

The hypothesis test using the F statistic as above 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 what is classically called **analysis of variance** or ANOVA.

Multi-way ANOVA

Regression with multilple categorical predictors is called **multi-way** or **multi-factor anova**.

Tadpole data acquired from Weiss (2012). Note unequal variances.

```
tadpoles <- read.csv("data/tadpoles.csv")
tadpoles$fac3 <- as.factor(tadpoles$fac3) # It's coded as 1 or 2.
ggplot(tadpoles, aes(treatment, response, fill = fac1, color = fac2))+
  geom_boxplot()+
  stat_summary(fun = mean, geom = "point", size = 3)+ # Show means.
  scale_fill_brewer(palette = "Dark2")      # The default colors get boring.
```
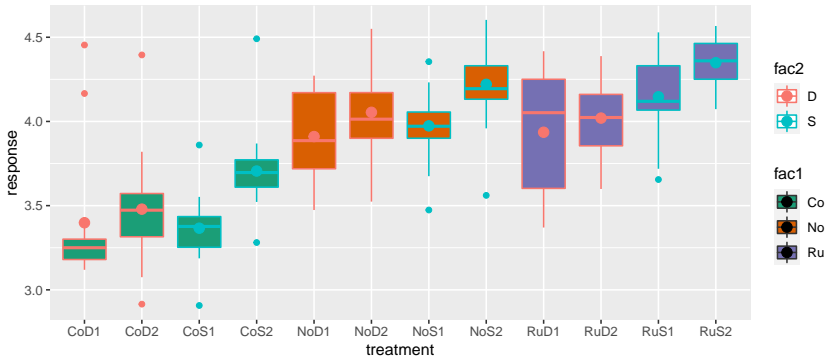
A more general F statistic

Earlier we used F to compare a model to the **null model**, with no predictors.

A more general F statistic can compare any two models where one is an extension of the other by adding predictors. To quantify the advantage of a new model obtained by adding variables to an existing model, compute

$$F = \frac{(SS_{old} - SS_{new})/(\text{number of new parameters})}{SS_{new}/(\text{number of data points less parameters})}.$$

► The number of new parameters, called the **numerator degrees of freedom** is the count of additional indicator variables in the extended model.

► The **denominator degrees of freedom** is the number of data points minus the total number of parameters in the extended model.

ANOVA tables

The anova command calculates F statistics to compare models.

```
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 18.4339  9.2169 151.7899 < 2.2e-16 ***
## fac2            1  1.5013  1.5013  24.7238 1.304e-06 ***
## fac3            1  2.2771  2.2771  37.5007 3.984e-09 ***
## fac1:fac2       2  0.3926  0.1963   3.2328   0.04127 *
## fac1:fac3       2  0.0838  0.0419   0.6900   0.50263
## fac2:fac3       1  0.3503  0.3503   5.7693   0.01711 *
## fac1:fac2:fac3  2  0.0695  0.0347   0.5723   0.56505
## Residuals     227 13.7838  0.0607
## ---
## 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 the ANOVA

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

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

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 3.38323831 0.05245374 64.4994720 1.024057e-149
## fac1No      0.54729873 0.05837246  9.3759750  6.582616e-18
## fac1Ru      0.53698650 0.06084917  8.8248782  2.755469e-16
## fac2S       0.01714866 0.06696214  0.2560949  7.981054e-01
## fac32       0.10757945 0.04814860  2.2343214  2.641948e-02
## fac1No:fac2S 0.01341375 0.07727544  0.1735836  8.623448e-01
## fac1Ru:fac2S 0.16350360 0.08174704  2.0001164  4.665860e-02
## fac2S:fac32  0.16322994 0.06504617  2.5094473  1.277781e-02
```

Note that the coefficients on `fac2S` and `fac1No:fac2S` are not significant, it is only in its interactions with `fac1Ru` and `fac32` that the diet factor appears to have an effect. We will still include these terms in the model, because if we include an interaction effect, we also include the corresponding main effects, and if we include an effect from one level of a factor, we 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()
```
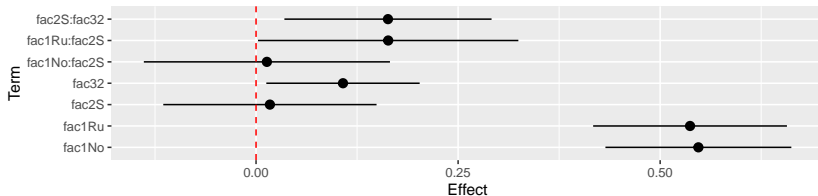

The reference treatment is `CoD1`.

- ▶ `Ru` and `No` differ from `Co` but not each other.
- ▶ On its own, Diet does not have a significant effect.
- ▶ Sibships 1 and 2 have different mitotic levels.
- ▶ Shrimp in combination with `Ru` or sibship 2 has an effect.

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

Type I anova **sequential**ly adds each term in a list to a model containing the terms before it on that list.

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

Often, type II is preferred. For example, 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?

Additionally, type II anova is more robust to deviations from the assumption of equal group sizes, resulting in unbalanced designs.

Marginal ANOVA using the `car` package

The base R command `anova` does sequential anova. Marginal anova is done using the `Anova` command in the `car` package.

```
Anova(lm5)
```

```
## Anova Table (Type II tests)
##
## Response: response
##                 Sum Sq  Df  F value    Pr(>F)
## fac1            18.0783   2 148.8618 < 2.2e-16 ***
## fac2             1.6498   1  27.1702 4.184e-07 ***
## fac3             2.2300   1  36.7243 5.609e-09 ***
## fac1:fac2        0.2997   2   2.4681   0.08702 .
## fac1:fac3        0.0550   2   0.4525   0.63659
## fac2:fac3        0.3503   1   5.7693   0.01711 *
## fac1:fac2:fac3   0.0695   2   0.5723   0.56505
## Residuals       13.7838 227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
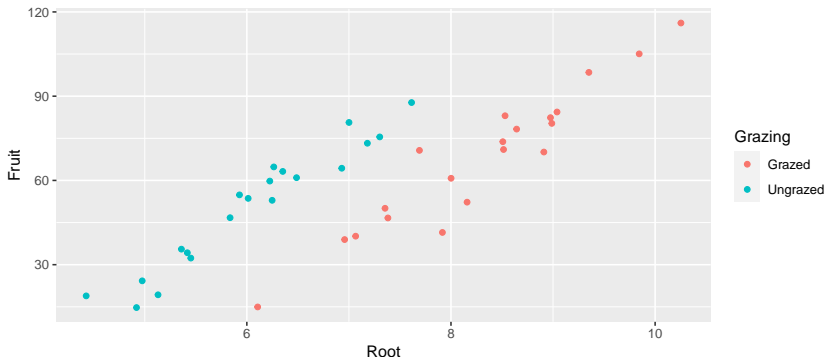
Results are similar to the type I analysis, but the p values for `fac1:fac2` are on opposite sides of the bright line of 0.05.

Combining numerical and categorical predictors

Often we have both numerical and categorical predictors.

Seed production 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` effect the numerical response `Fruit`? The numerical variable `Root` is a confounder. This is classical **analysis of covariance** or ANCOVA. Once again, it's just regression.

```
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 16795.0  16795.0 359.9681 < 2.2e-16 ***
## Grazing       1  5264.4   5264.4 112.8316 1.209e-12 ***
## Root:Grazing  1     4.8      4.8   0.1031      0.75
## Residuals    36  1679.6     46.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

▶ `Root` and `Fruit` appear correlated.
▶ `Grazing` and `Fruit` appear correlated.
▶ The interaction between `Grazing` and `Root` does not appear to affect `Fruit`.

Homogeneity of Variance

Regression assumes that the variance in the residuals is constant for all values of the predictors, or **homoscedasticity**. For real data, we must asses the variance in the residuals against each predictor.

In the tadpole data, the CoD1, CoD2 and RuD1 treatment groups have higher variances than the others. Is this a problem?

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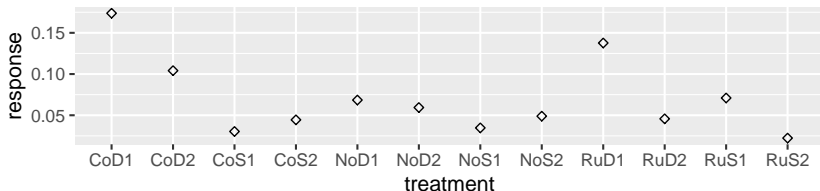

- ▶ For the CoD1 and CoD2 groups, the variance is due to outliers. Run the analysis without them - does the result change?
- ▶ RuD1 might be a problem, but it is only one group, and while the variance is large, at least there isn't much skew.

Independence of observations

Regression assumes that observations are independent, but this is often violated for real data.

Pseudoreplication is the technical term for data that includes dependent observations. It has the effect of artificially increasing the power of statistical tests. There are two very common scenarios where it is encountered:

- ▶ **Repeated measures**: Observe the same 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.

Weiss, Jack. 2012. "Ecology 563."