Regression and ANOVA

Zack Treisman

Spring 2022

WESTERN
COLORADO UNIVERSITY

MATHEMATICS & COMPUTER
SCIENCE DEPARTMENT

Philosophy

Data analysis and statistics are tools used in modeling.

- A **model** is a proposed distribution of a variable or variables
- Separate any model into two parts: the **signal** and the **noise**.

A fundamental setting is a pair of variables, x and y. We know something about x, and would like to leverage this to learn something about y, to the extent that this is possible. We call x the **predictor** and y the **response**. Write

$$y = f(x) + \epsilon$$

where the model function $f(x)$ is what we call the **signal** and ϵ is the **noise**.

"All models are wrong but some are useful" - George Box

The usefulness of a model comes when the signal is not drowned out by the noise.

Regression

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**

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

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.

One Variable: Only a response y and no predictor x.

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

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

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

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

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

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

So minimizing MSE directs us to choose the mean.

Linear Regression

Suppose $f(x)$ is linear and ϵ is normally distributed.

Another way to say this is that if x and y are values of two numerical random variables X and Y, then

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

The parameters β_0 (the **intercept**), β_1 (the **slope**) and σ (the **standard deviation**) are estimated from observations (x_i, y_i) 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 calculus just like above,

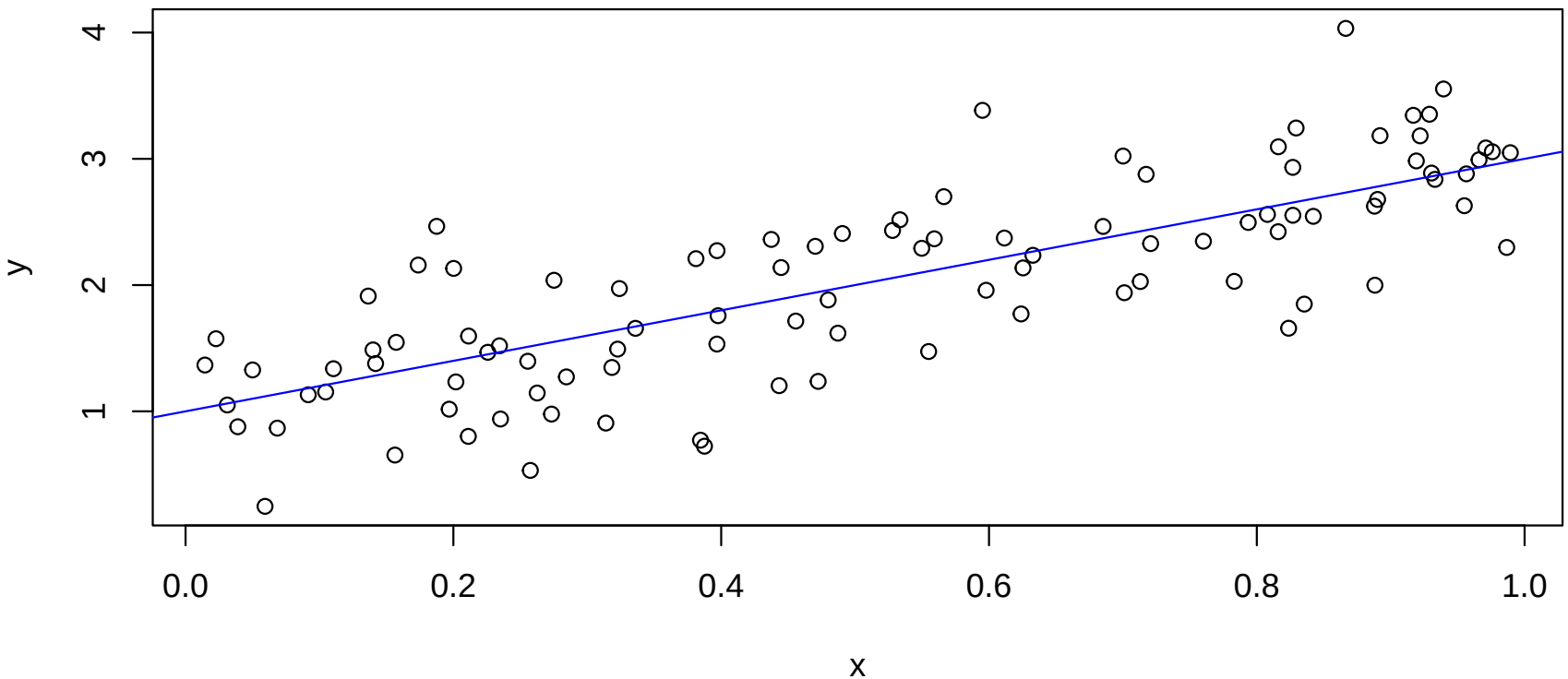
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

An estimate of σ is the standard deviation of the **residuals** $y_i - \hat{y}_i$.

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")
```

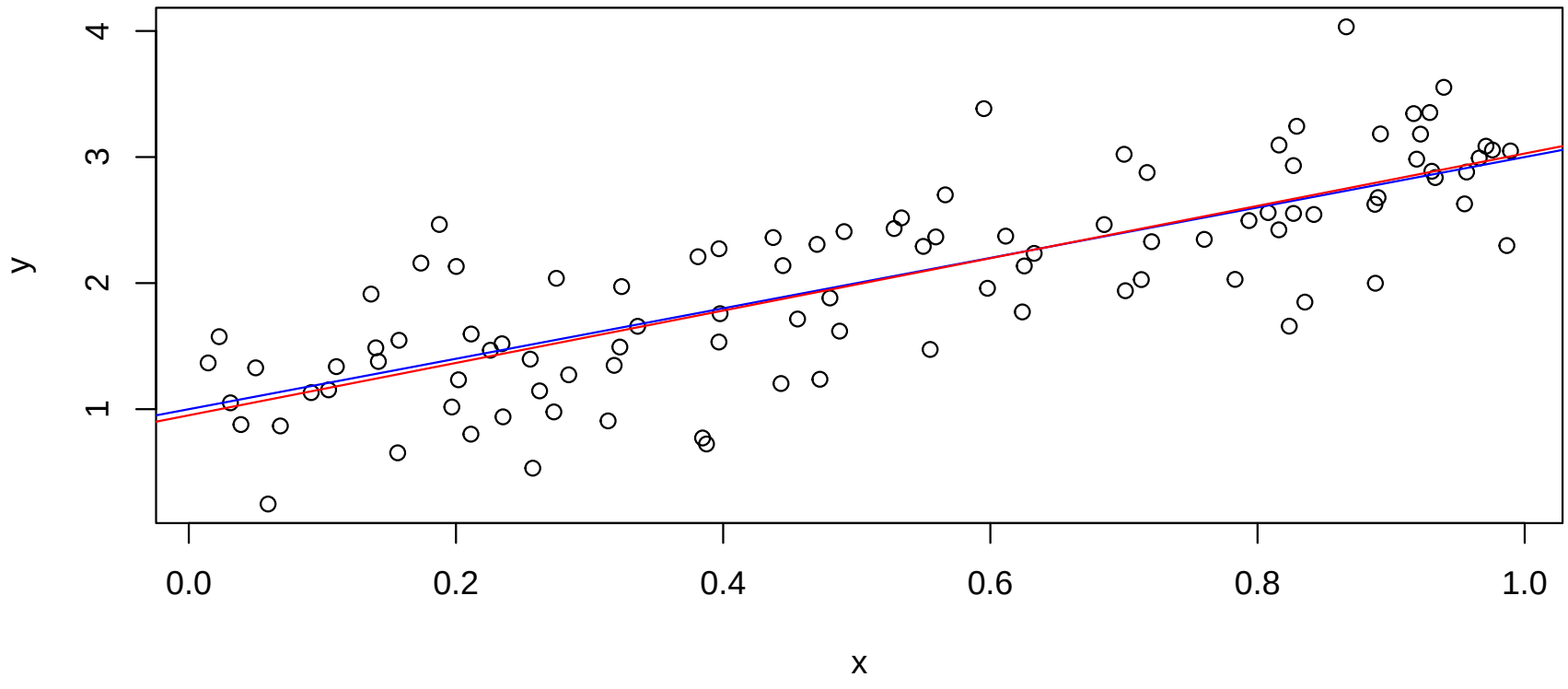


The `lm` function in R

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

```
## (Intercept)           x
##      0.9515      2.0765
```

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



Evaluating a linear model

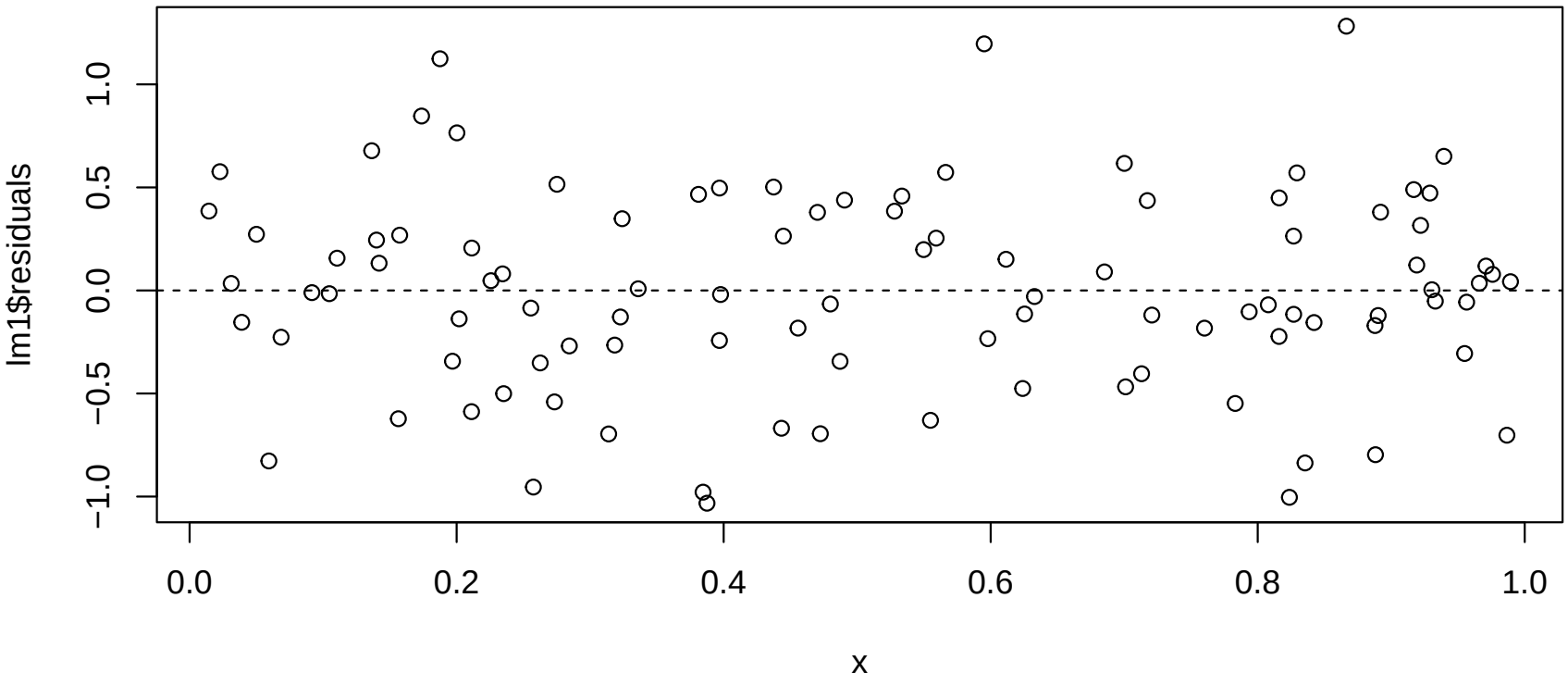
```
summary(lm1)
```

```
##
## 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

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

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

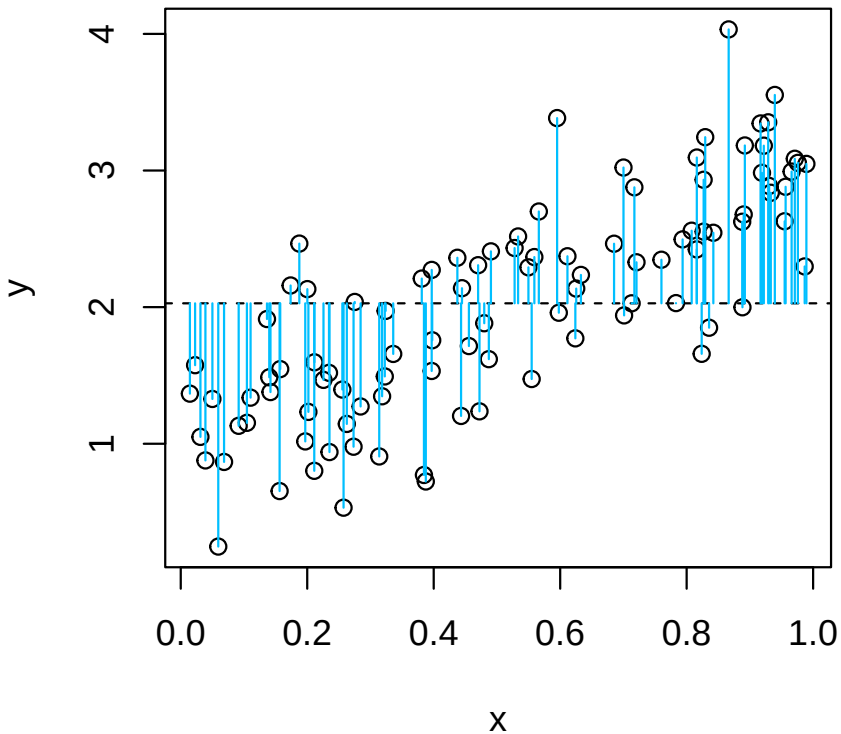


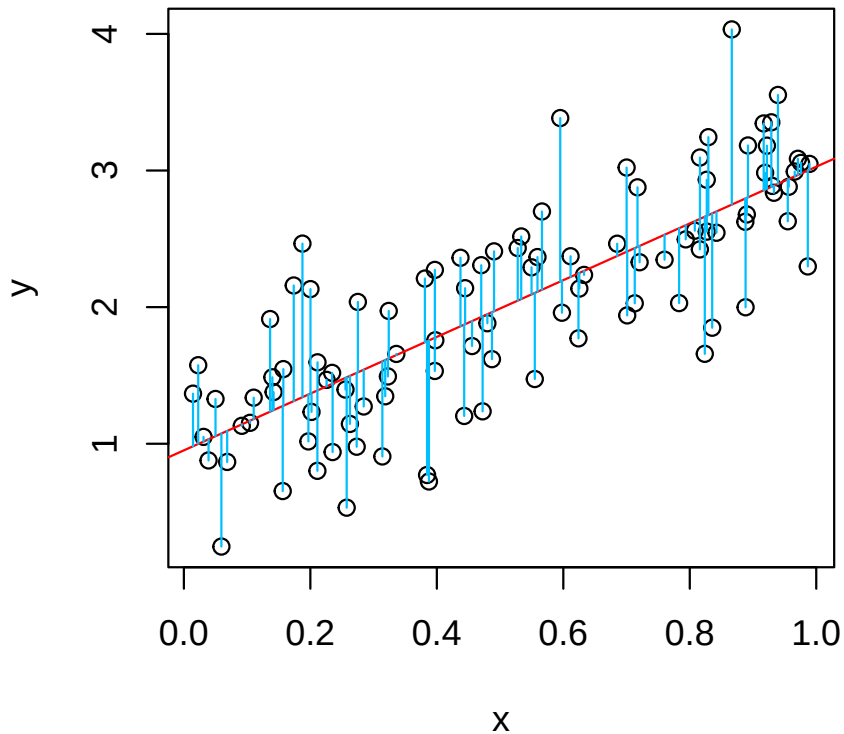
We'll discuss various ways to check homogeneity of variance. There's no one correct way to do it.

Overall accuracy of the model

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

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





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

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

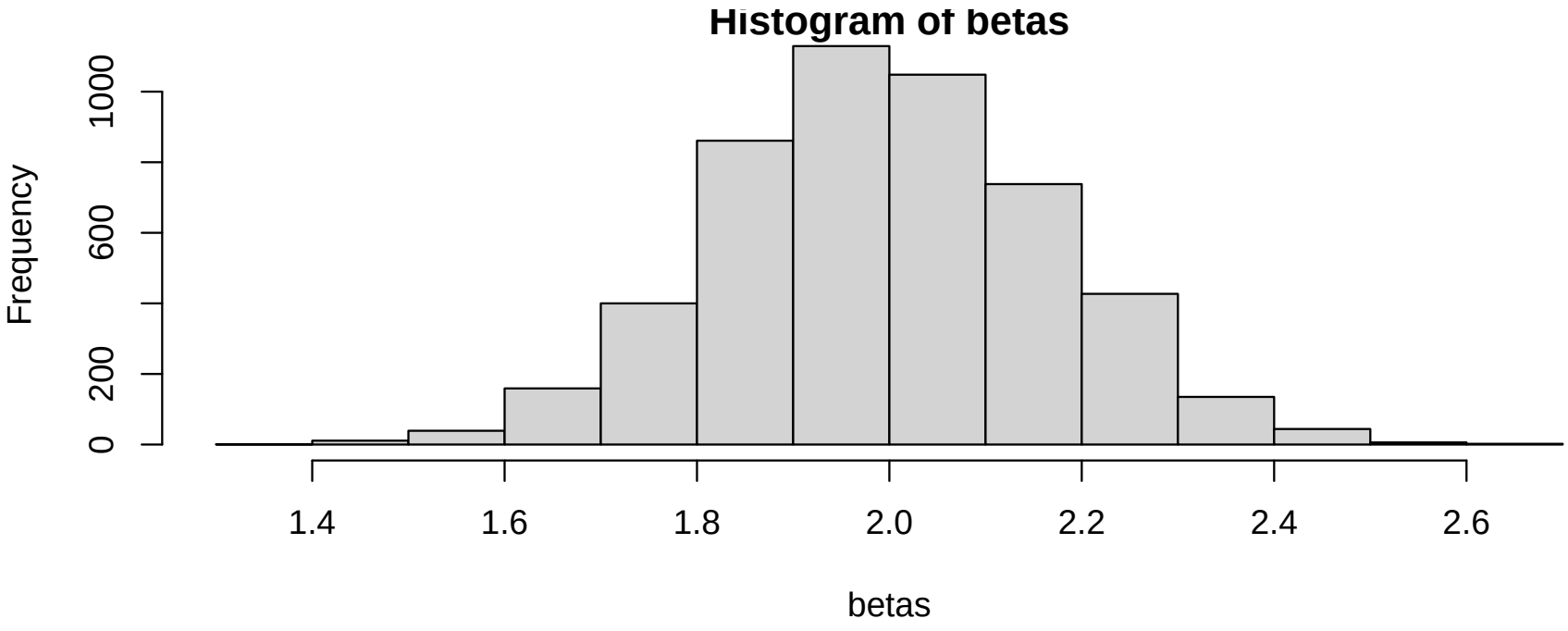
  contains the true value of β_1. (Which is 2.)
- In our case, the 95% confidence interval is

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

Simulate taking 5000 such samples and calculating 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

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.

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

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

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

- 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`

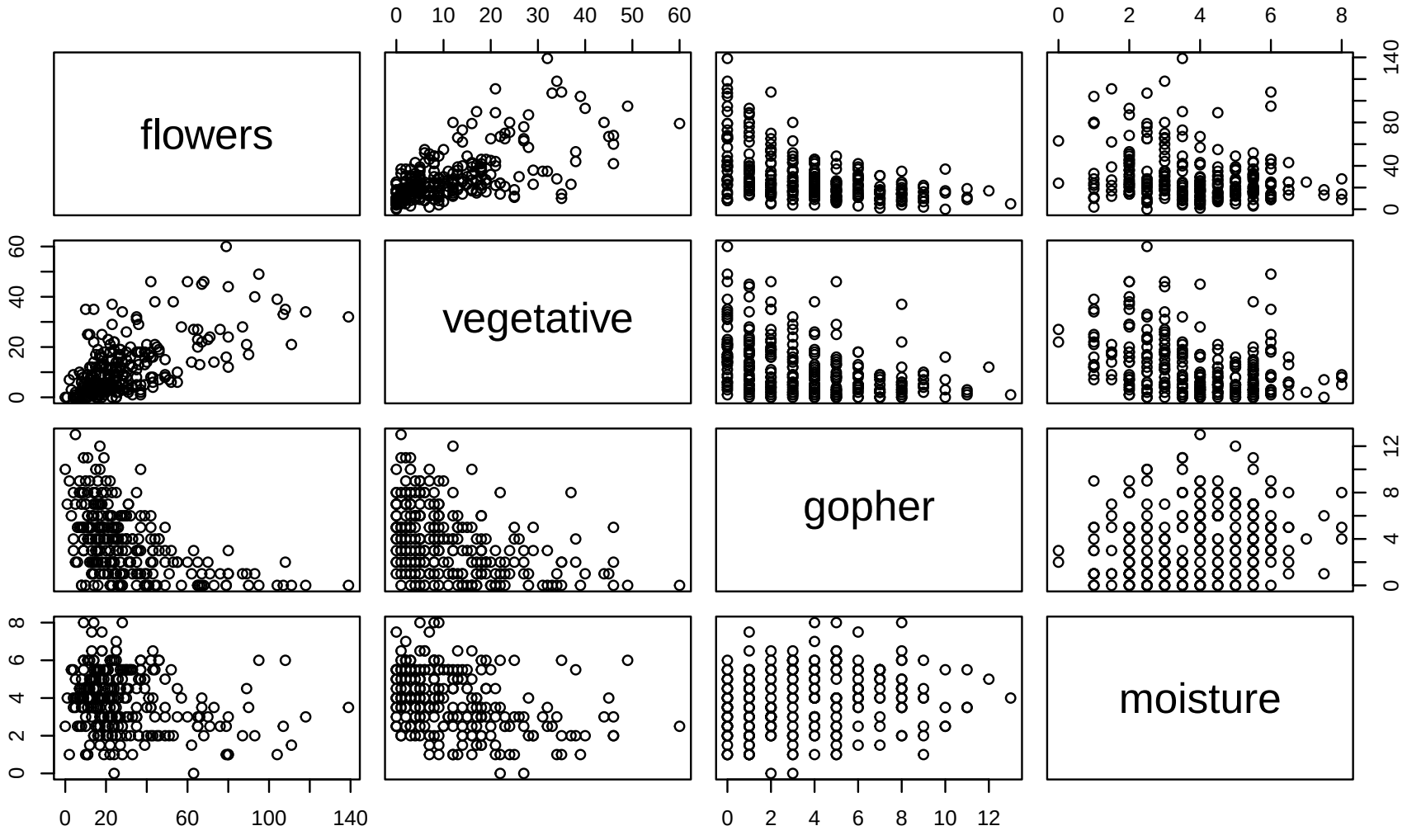
Lily data from Thomson et al. (1996).

```
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
## -52.51 -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")])
```



Confounders can have big effects

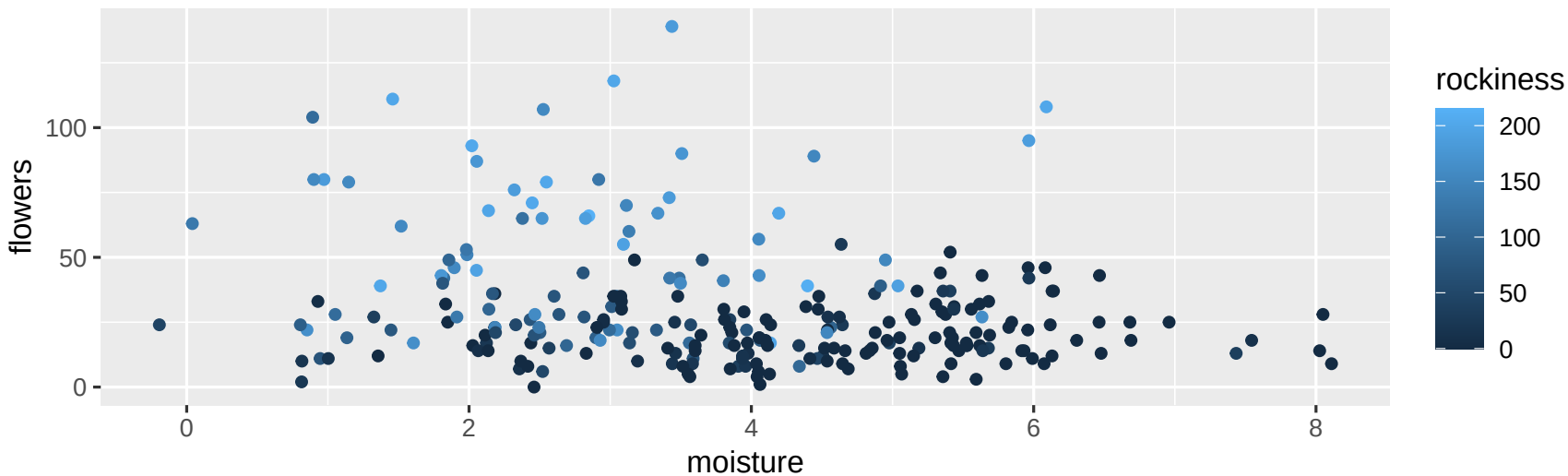
The Lily data contain an additional variable.

```
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)    8.5047    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 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

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

- 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

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

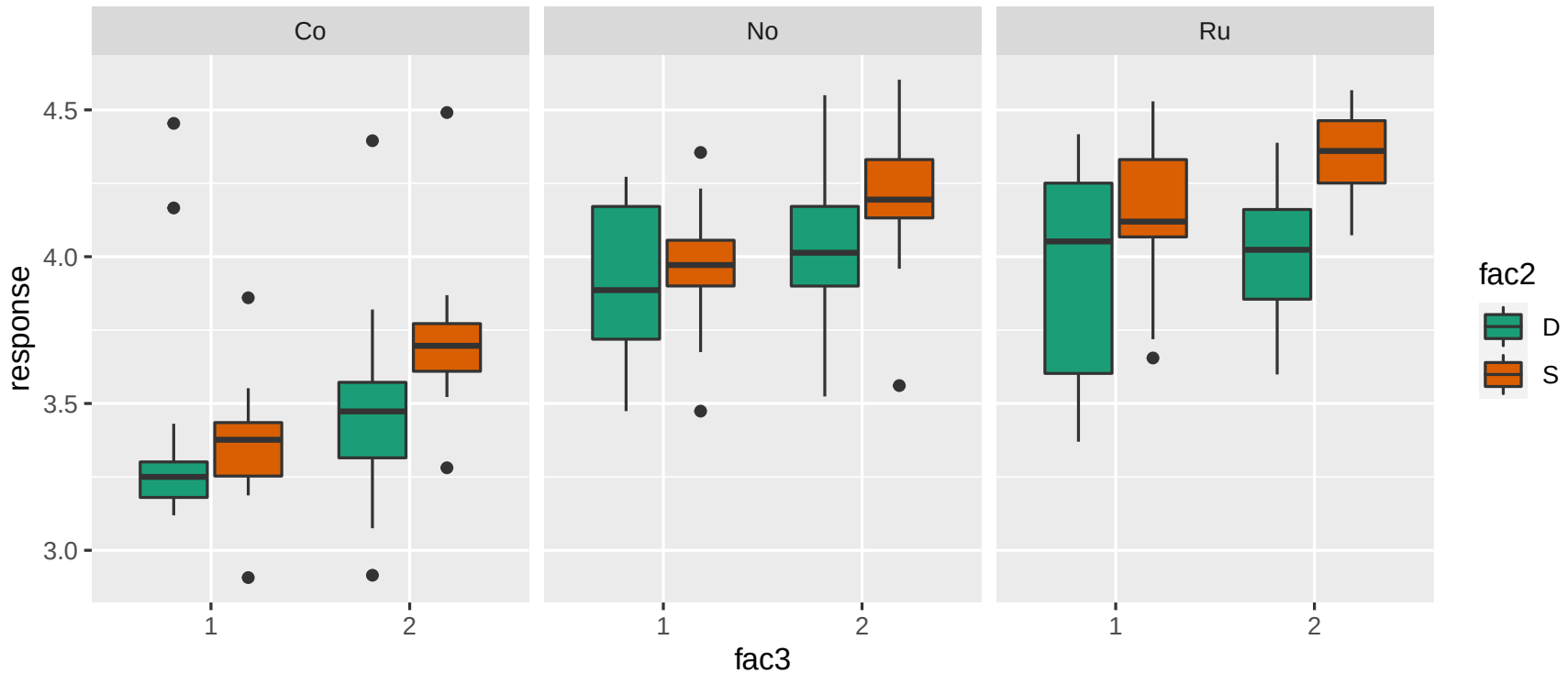
is what is classically called **analysis of variance** or ANOVA.

Multi-way ANOVA

Regression with multiple 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(fac3, response, fill = fac2))+
  geom_boxplot()+facet_wrap(~fac1)+
  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  18.43    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 the 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
```

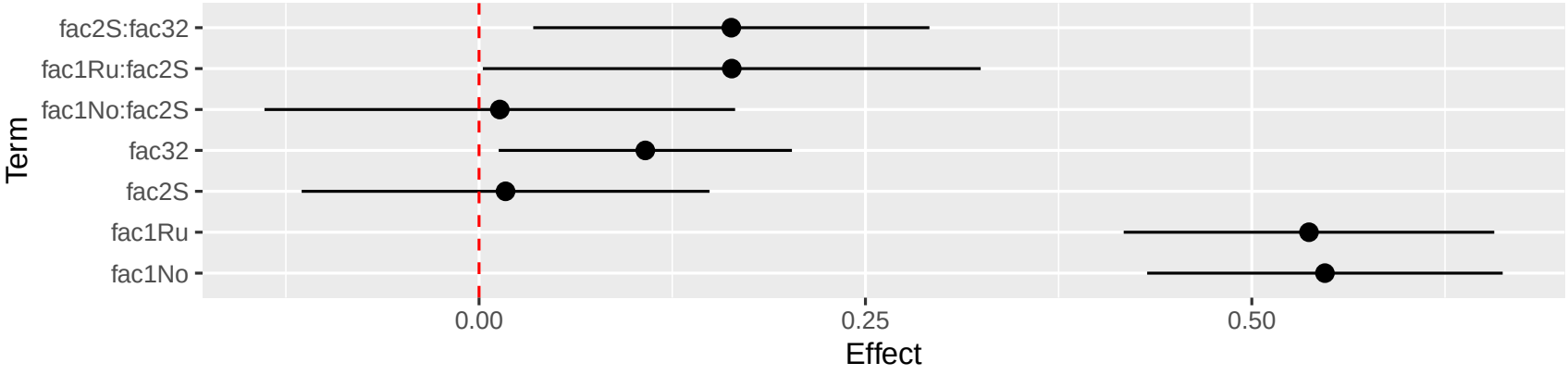
```
##                Estimate Std. Error t value   Pr(>|t|)
## (Intercept)    3.38324    0.05245 64.4995 1.024e-149
## fac1No         0.54730    0.05837  9.3760  6.583e-18
## fac1Ru         0.53699    0.06085  8.8249  2.755e-16
## fac2S          0.01715    0.06696  0.2561  7.981e-01
## fac32          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
```

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             18.08   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
```

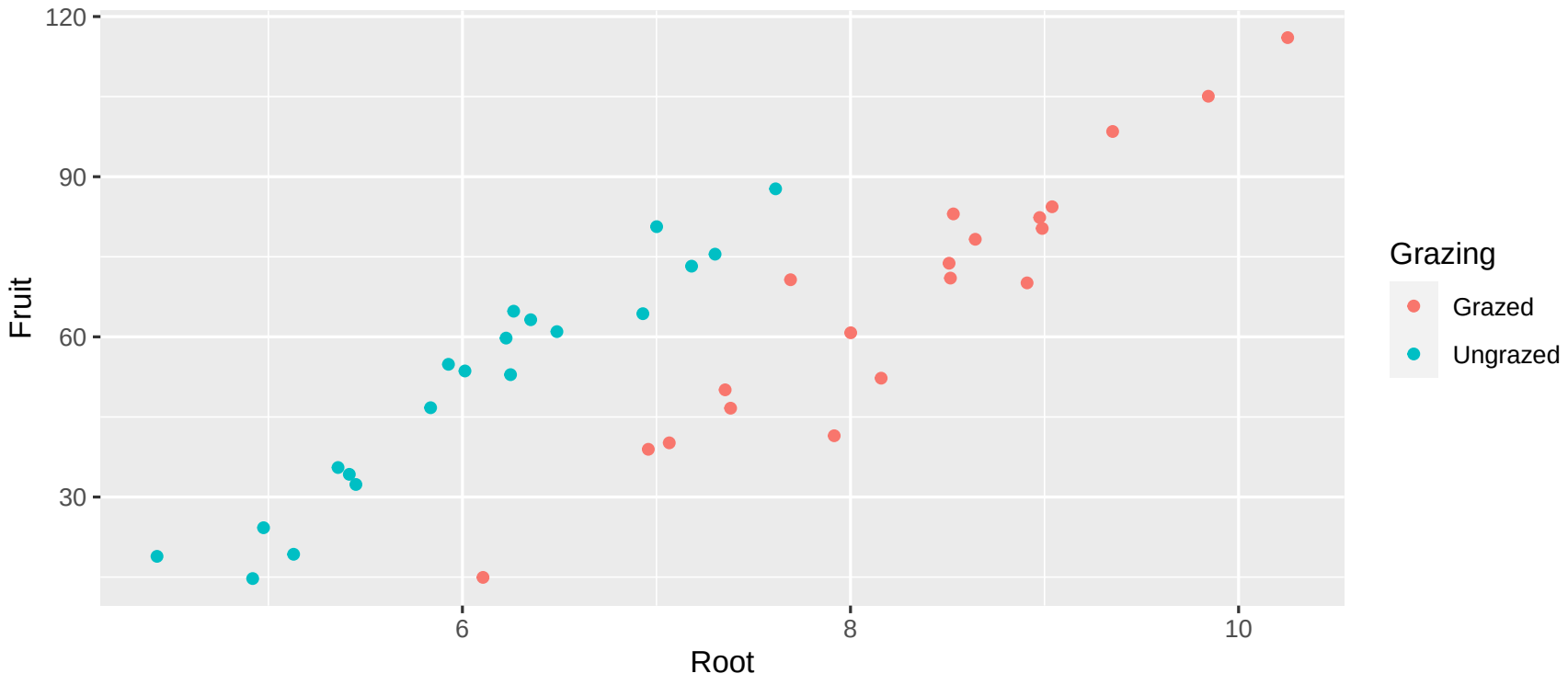
Results are similar to the type I analysis, but the p values for fac1:fac2 are on opposite sides of the bright line of 0.05.

Combining numerical and categorical predictors

Often we have both numerical and categorical predictors.

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   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
```

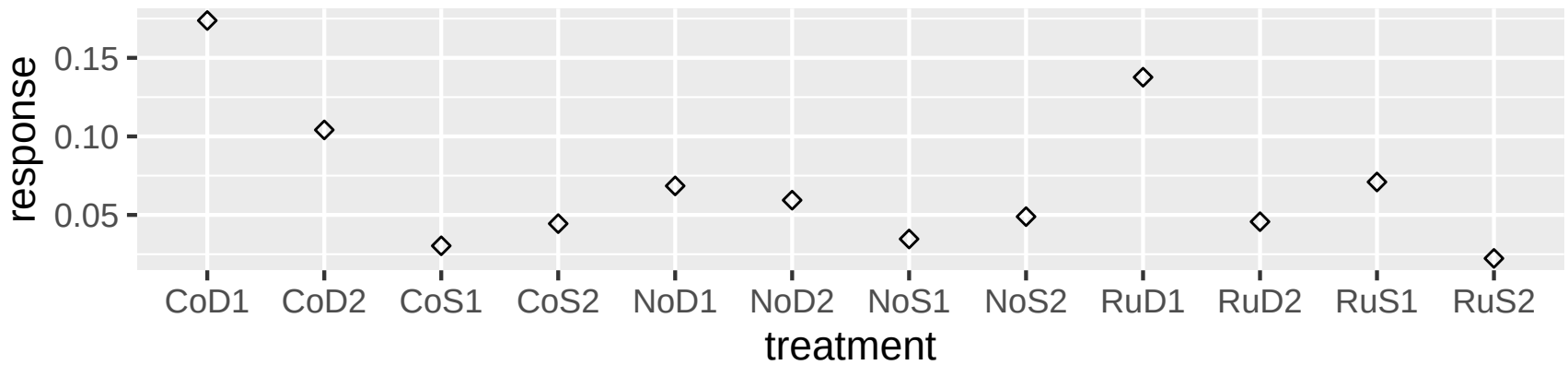
- `Root` and `Fruit` appear correlated.
- `Grazing` and `Fruit` appear correlated.
- The interaction between `Grazing` and `Root` does not appear to affect `Fruit`.

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."