Session 5 - Linear Regression and ANOVA

Brian Leroux

Wednesday, February 13, 2019

Outline

The focus for this lecture will be on the use and interpretation of linear regression models. Next week we will examine more of the theory behind regression including how the computations are done.

1. Linear Regression
2. Regression and ANOVA with Two or More Factors

1. Linear Regression

- The linear regression model
- Interpretation of parameters in regression models
- Relationship between linear regression, t-tests and ANOVA
- Relationship between regression and correlation
- Correlation vs causation
- All models are wrong
- Dangers of extrapolation

Example: Dose-Response in the Tooth-Growth Experiment

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Data: ToothGrowth (built-in R dataset)

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

The variables are

len Tooth (odontoblast) length

supp Supplement type (VC or OJ)

dose Dose (mg/day)

A 2-factor experiment

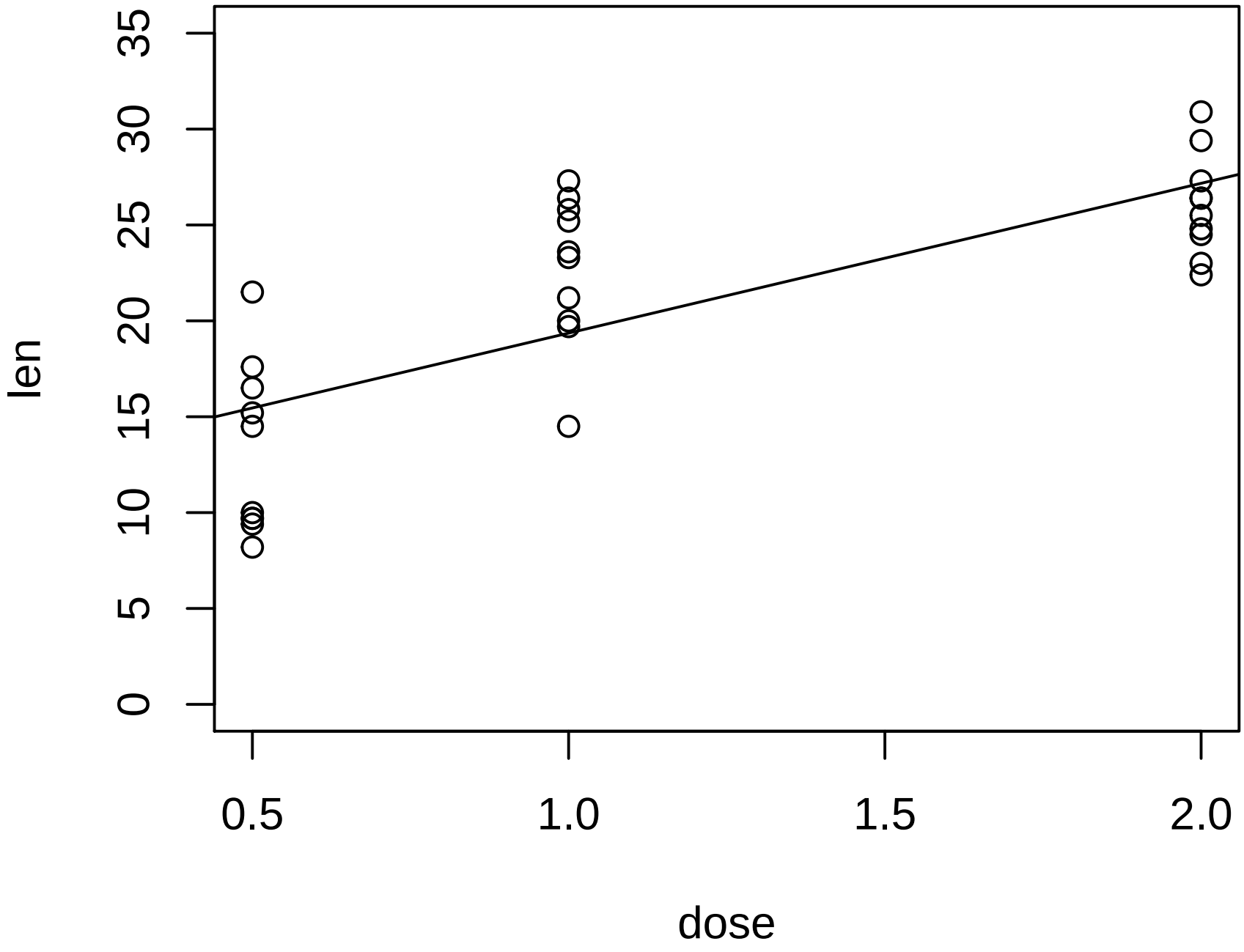
In the tooth-growth experiment (data 'ToothGrowth'), there are 2 experimental factors:

1. Delivery method: orange juice (OJ) or ascorbic acid (VC)
2. Dose of vitamin C: 0.5, 1, or 2 mg/day All 6 combinations of the 2 factors were used:

```
    dose
supp 0.5  1  2
  OJ  10 10 10
  VC  10 10 10
```

For now we will consider only the dose factor and we will conduct separate analyses for each delivery method.

Length vs dose for the OJ group



The Fitted Regression Line

The fitted line was obtained using *least-squares* (LS) regression. In R this is done with the 'lm' function. (We will see how this is done next week.)

```
lm(len ~ dose,data=oj)
```

```
Call:
lm(formula = len ~ dose, data = oj)

Coefficients:
(Intercept)         dose
     11.550        7.811
```

The equation of the fitted line is

$$\text{len} = 11.550 + 7.811 \times \text{dose}$$

The Linear Regression Model

Linear regression is based on an underlying statistical model

$$E[Y|X = x] = \alpha + \beta x,$$

where

Y is the *response* variable (also called "outcome", "dependent variable")

X is the *predictor* variable (also called "explanatory" or "independent" variable)

α is the *intercept*

β is the *regression coefficient* for X

This model says that the conditional expectation of Y given X is a linear fuction of X. It is sometimes written as $E[Y] = \alpha + \beta X$.

Another form of the model that is used sometimes is:

$$Y = \alpha + \beta x + \epsilon,$$

where ϵ represents the "error", assumed to have mean 0.

Interpretation of the parameters

The interpretation of α is the mean of Y given $X = 0$, i.e., $\mathrm{E}(Y|X = 0) = \alpha + \beta \times 0 = \alpha$. This is the point where the regression line crosses the y-axis.

The interpretation of β is the average *difference* in the mean of Y per unit *difference* in X.

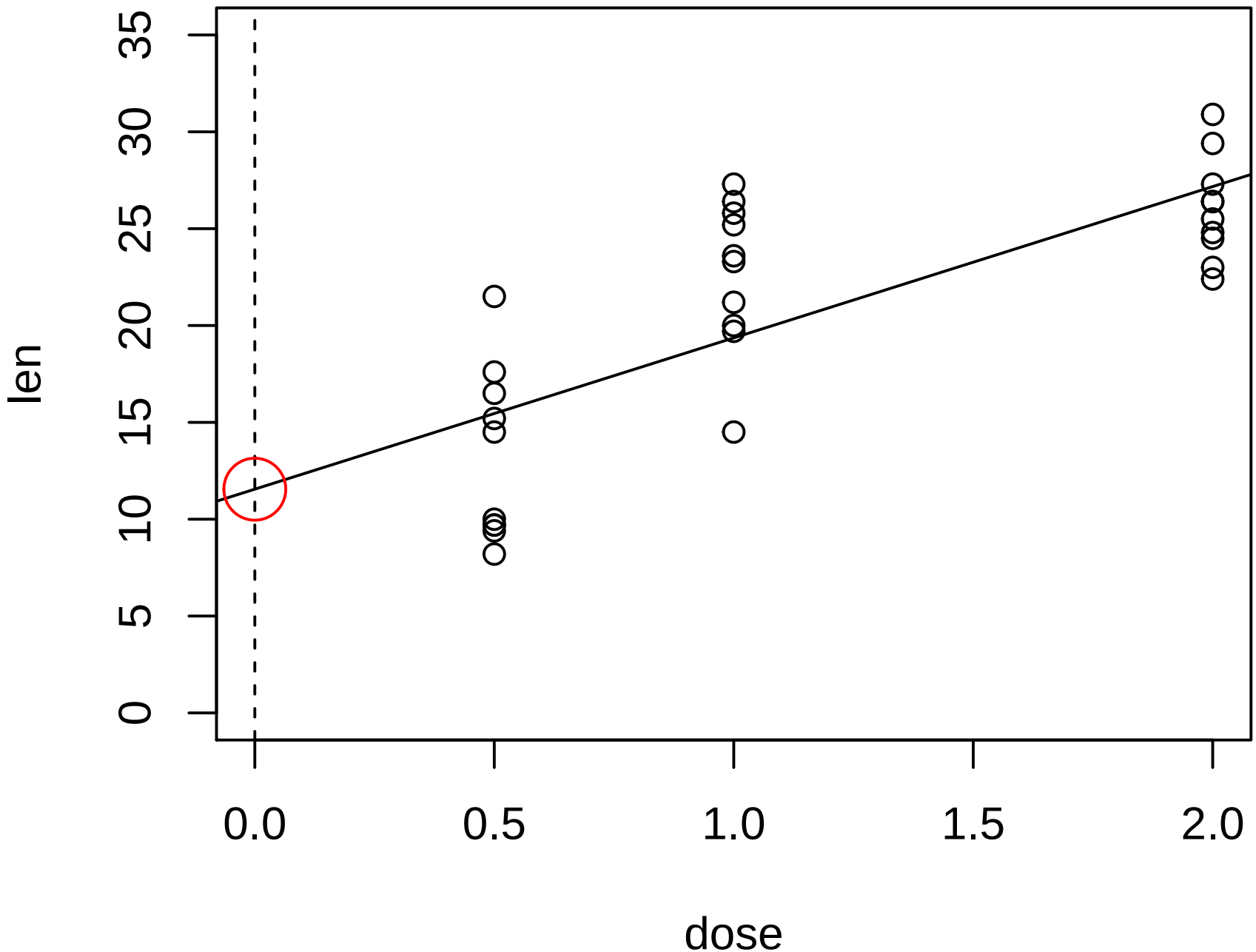
Sometimes this is expressed as the average difference in Y corresponding to a 1-unit difference in X, i.e.,

$$\mathrm{E}(Y|X = x + 1) - \mathrm{E}(Y|X = x) = \alpha + \beta(x + 1) - (\alpha + \beta x) = \beta.$$

For a given data set, the fitted regression model is written as $\mathrm{E}(Y) = \hat{\alpha} + \hat{\beta}X$, where $\hat{\alpha}$ is the point where the fitted regression line crosses the y-axis and $\hat{\beta}$ is the slope of the fitted regression line.

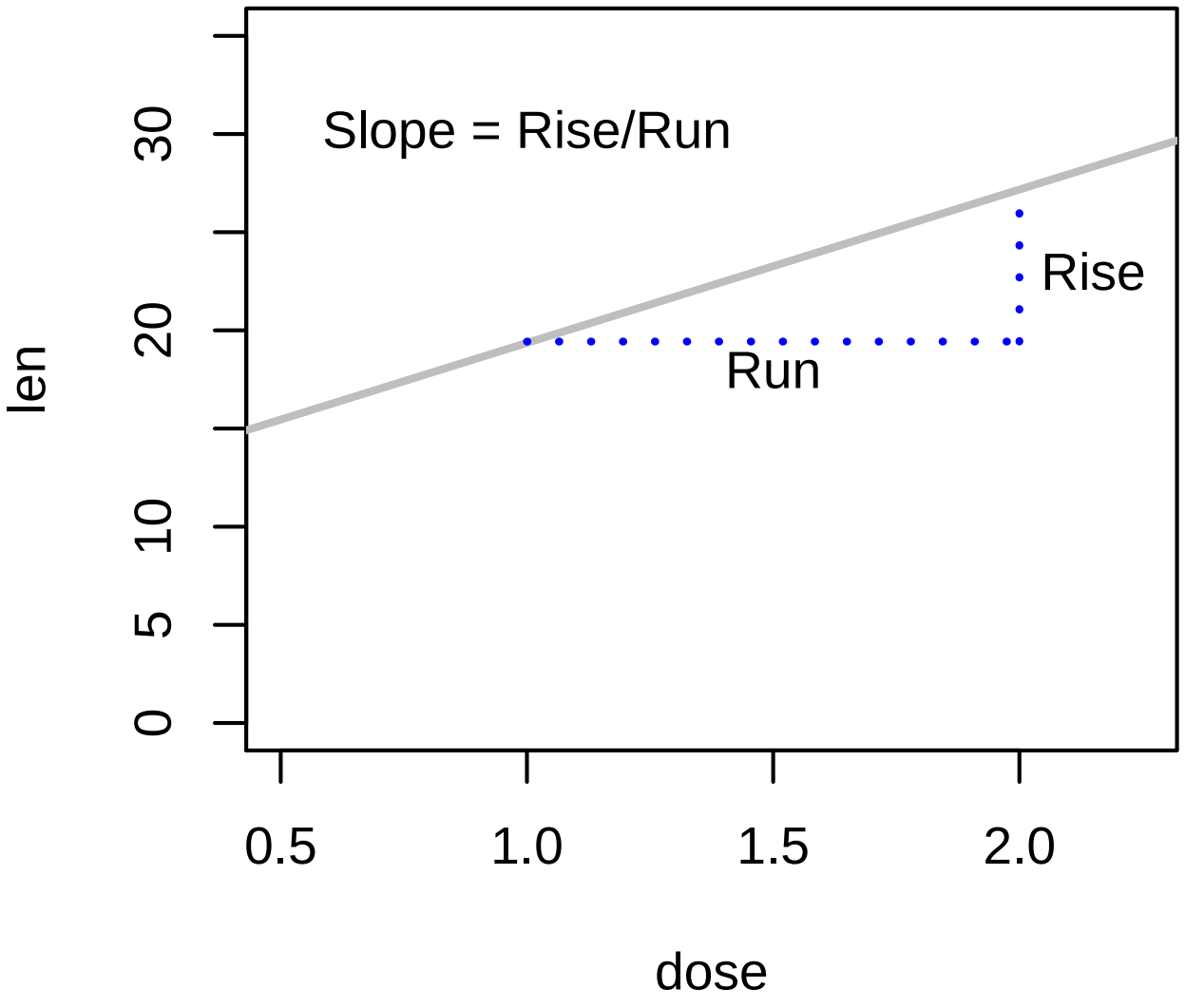
Interpretation of the estimate of the intercept

The estimated intercept $\hat{\alpha}$ is an estimate of the mean response for $X = 0$.



Interpretation of the estimate of the regression coefficient

$\hat{\beta}$ is the slope of the regression line, i.e., "rise over run"



Interpretation of the parameters for the OJ data

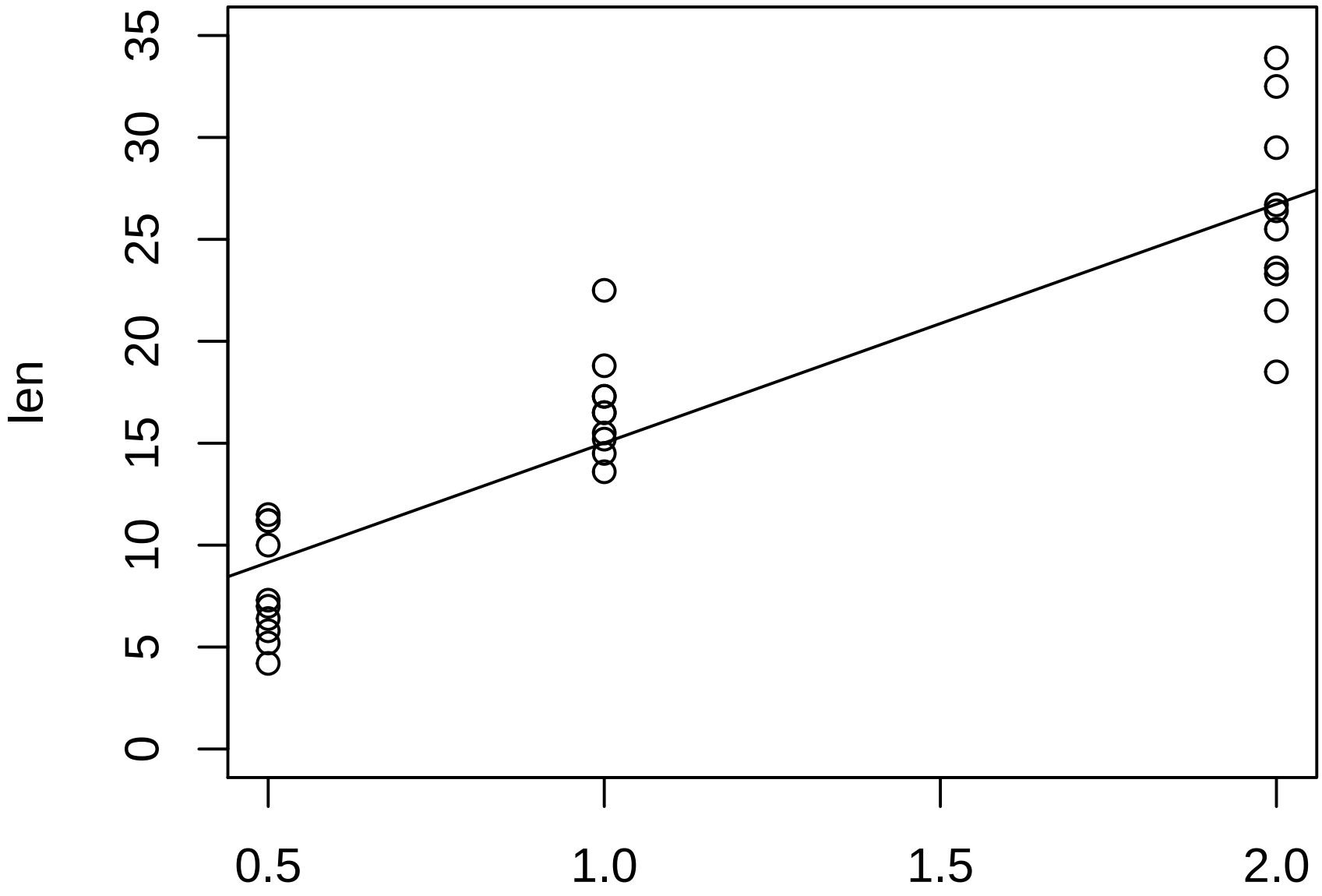
$\hat{\alpha} = 11.55$ is the estimated mean tooth length if the vitamin C dose is set to 0.

$\hat{\beta} = 7.881$ is the estimated average difference in tooth length per unit difference in vitamin C dose.

Linear regression for the "VC" group

Fitted regression line:

$$len = 3.295 + 11.716$$



Testing hypotheses about regression coefficients

In regression, typically the null hypothesis of interest is $H_0 : \beta = 0$, because this represents the hypothesis that the mean response is the same for any value of the predictor variable, i.e., there is no effect of X on Y.

The test statistic is calculated using the formula Z =Estimate/SE(Estimate). This is sometimes called a "t-statistic" because it has the same general form as the t-test statistic. Statistical significance is determined using a t-distribution with $n - 2$ degrees of freedom. (We will see this in more detail, including how to calculate the SE, next week.)

The subtraction of 2 represents the two parameters estimated (α and β) to fit the regression line.

For the OJ group, we get $Z = 7.8114/1.3017 = 6.0011$. The p-value is the 2-sided tail probability for the t_{28} distribution.

```
2*(1-pt(6.0011,df=28))
```

```
[1] 1.824655e-06
```

OJ group:

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 11.550000   1.721951 6.707508 2.788784e-07
dose         7.811429   1.301673 6.001070 1.824801e-06
```

VC group:

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  3.29500   1.427060  2.308943 2.854201e-02
dose        11.71571   1.078756 10.860392 1.509369e-11
```

Note: p-values this small are typically reported as $p < 0.001$ or $p < 0.0001$.

The relationship between regression and ANOVA

We can also apply ANOVA to test the significance of dose. The difference is that with ANOVA we are fitting a separate mean value to each of the three dose groups, rather than assuming a straight-line relationship.

```
summary(aov(len ~ factor(dose), data=oj))
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(dose)  2  885.3   442.6   31.44 8.89e-08 ***
Residuals    27  380.1    14.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis being tested by the F-statistic in this case is $H_0 : \mu_1 = \mu_2 = \mu_3$, where μ_1, μ_2, and μ_3 represent the mean response in each of the three dose groups. Again, there is strong evidence against the null hypothesis. But note that the p-values for ANOVA and linear regression are different.

Regression versus ANOVA

For ANOVA, dose is treated as a 'factor' (i.e., categorical) variable, whereas for regression dose is treated as a quantitative variable. However, the **null hypotheses** for ANOVA and regression are equivalent: they both imply that the mean response does not depend on dose. It is just expressed differently for regression ($\beta = 0$) than for ANOVA ($H_0 : \mu_1 = \mu_2 = \mu_3$).

For the OJ data, the conclusions are similar for ANOVA and regression: both methods provide evidence against the null hypothesis. Thus, with either method we would conclude that there is an effect of dose on tooth length.

ANOVA and regression will not always agree in this way. Note that in this example, the p-values are not the same for the two methods (approximately 10^{-6} for regression versus 10^{-8} for ANOVA).

Regression or ANOVA?

In practice, the choice of method will be based on what type of **alternative** hypotheses we are interested in detecting.

If we believe there is an *approximate* linear dose-response relationship between X and Y then regression is more appropriate because it will tend to have higher power to detect this relationship than ANOVA. Note that the assumption of linearity does not have to hold exactly (it rarely if ever does) in order to apply linear regression.

However, if we have no reason to suspect the relationship to be close to linear (e.g., if we expect a non-monotone relationship), then ANOVA is the best approach.

Regression versus ANOVA for comparing 2 groups

There is one special case in which regression and ANOVA give the exact same results: the comparison of two groups. As an illustration, consider the comparison of the dose-1 and dose-2 groups in the OJ data.

```
summary(lm(len ~ dose, data=oj,subset=(dose >= 1)))$coef
```

```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept)    19.34   2.363519 8.182714 1.774084e-07
dose            3.36   1.494821 2.247761 3.736280e-02
```

```
summary(aov(len ~ factor(dose), data=oj, subset=(dose >= 1)))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
factor(dose)  1  56.45   56.45   5.052 0.0374 *
Residuals    18 201.10   11.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results are the same (except for differences in rounding).

Regression versus the equal-variance t-test for comparing 2 groups

With only two values of the dose variable, linearity does not constrain the model in any way (hence regression and ANOVA are the same). In this case, the equal-variance t-test would also give the same result.

```
summary(lm(len ~ dose, data=oj, subset=(dose >= 1)))$coef
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept)    19.34   2.363519 8.182714 1.774084e-07
dose            3.36   1.494821 2.247761 3.736280e-02
```

```
t.test(oj$len[oj$dose==1],oj$len[oj$dose==2],var.equal=T)
```

```
    Two Sample t-test

data:  oj$len[oj$dose == 1] and oj$len[oj$dose == 2]
t = -2.2478, df = 18, p-value = 0.03736
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.5005017 -0.2194983
sample estimates:
mean of x mean of y
    22.70     26.06
```

The p-values from regression and the equal-variance t-test are the same. Also, the estimated coefficient for dose (3.36) is equal to the difference between the group means (26.06 - 22.70).

Comparison of regression, ANOVA and the equal-variance t-test

For comparison of 2 group means:

> *The equal-variance two-sample t-test, ANOVA, and linear regression are equivalent.*

For comparison of 3 or more group means:

> *ANOVA and linear regression can give different results*

Other regression analysis output

```
summary(lm(len ~ dose, data=oj))
```

```
Call:
lm(formula = len ~ dose, data = oj)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2557 -3.7979 -0.0643  3.3521  7.9386

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.550      1.722   6.708 2.79e-07 ***
dose           7.811      1.302   6.001 1.82e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.446 on 28 degrees of freedom
Multiple R-squared:  0.5626,    Adjusted R-squared:  0.547
F-statistic: 36.01 on 1 and 28 DF,  p-value: 1.825e-06
```

Residuals, Residual Standard Error, R-squared, and the F-test

Residuals are used for checking assumptions (next week).

"Residual standard error" is an estimate of the SD of the data points around the regression line (SD of the error term in the regression model).

"Multiple R-squared" is a measure of how much variation in the response is explained by the model. (For this model, it is just the square of the correlation.)

"Adjusted R-squared" is an adjusted value of R-squared based on the number of parameters in the model. It is used in model selection (other model selection methods we will see include AIC, BIC, and cross-validation).

The F-test for this model is equivalent to the t-statistic for β (note that $F = t^2$ and the p-values agree).

Correlation

The Pearson correlation coefficient is a measure of the strength of *linear* association between two variables. The formula is

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{(}Y))}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$$

It is a measure between -1 and 1 which is interpreted as follows:

$r = 0$: there is no *linear* association between X and Y

$r > 0$: there is a positive linear association between X and Y

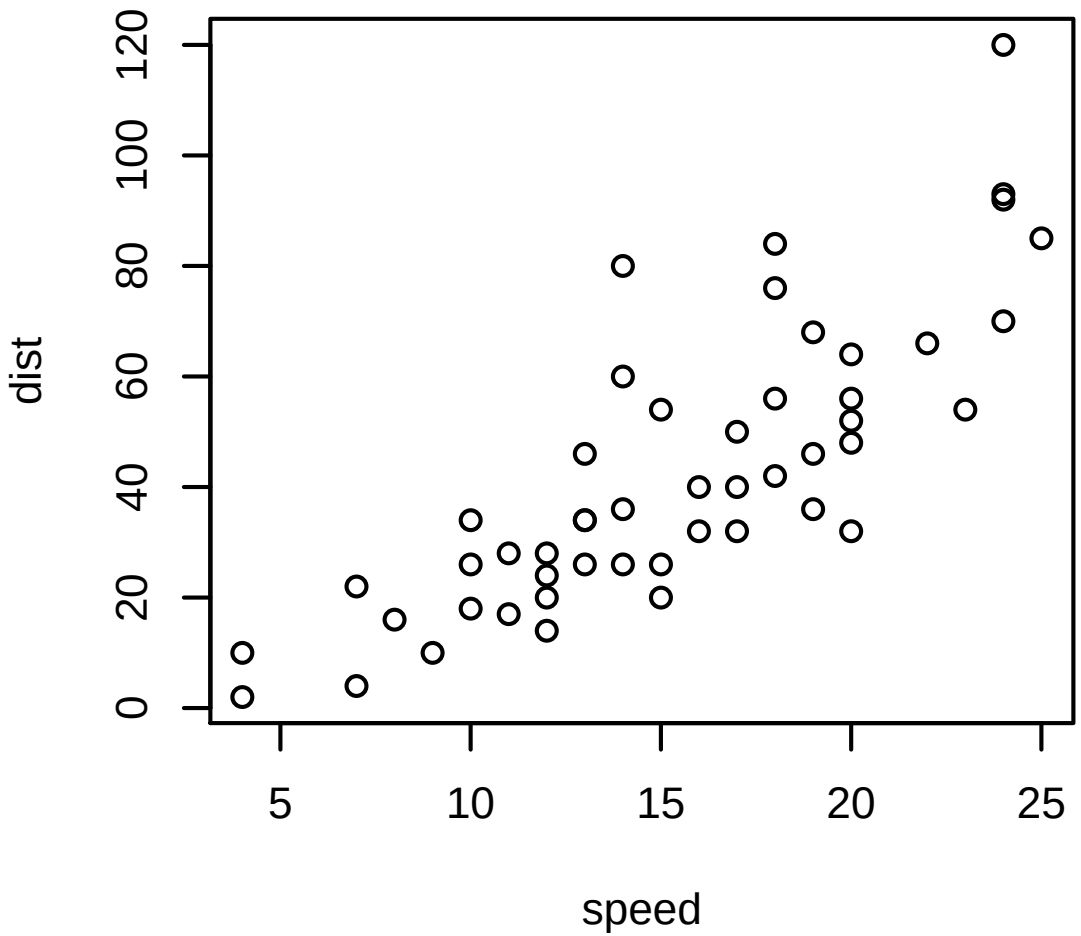
$r < 0$: there is a negative linear association between X and Y

Correlation versus regression

Linear regression and correlation are closely related. The Pearson correlation coefficient is a measure of how well the data points in a scatterplot follow a straight line. The least-squares regression line is the line in question.

Example: speed vs stopping distance of cars (R dataset 'cars').

```
plot(dist ~ speed, data=cars)
```



Regression versus correlation for the 'cars' data

```
cor(cars$dist,cars$speed)
```

```
[1] 0.8068949
```

```
summary(lm(dist ~ speed,data=cars))
```

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

The correlation between distance and speed is $r = 0.807$.

The square of the correlation is equal to the R^2 value from the regression: $r^2 = 0.807^2 = 0.651 = R^2$.

For simple linear regression models, the R-squared is just the square of the Pearson correlation coefficient. For models with more than 1 predictor R-squared has an interpretation in terms of correlation between observed and fitted values and also as a percentage of variance explained by the model (we will come back to this in the context of prediction).

Equivalence of hypothesis testing for correlation and regression

Testing a null hypothesis of 0 correlation is equivalent to testing the null hypothesis of 0 for the linear regression coefficient.

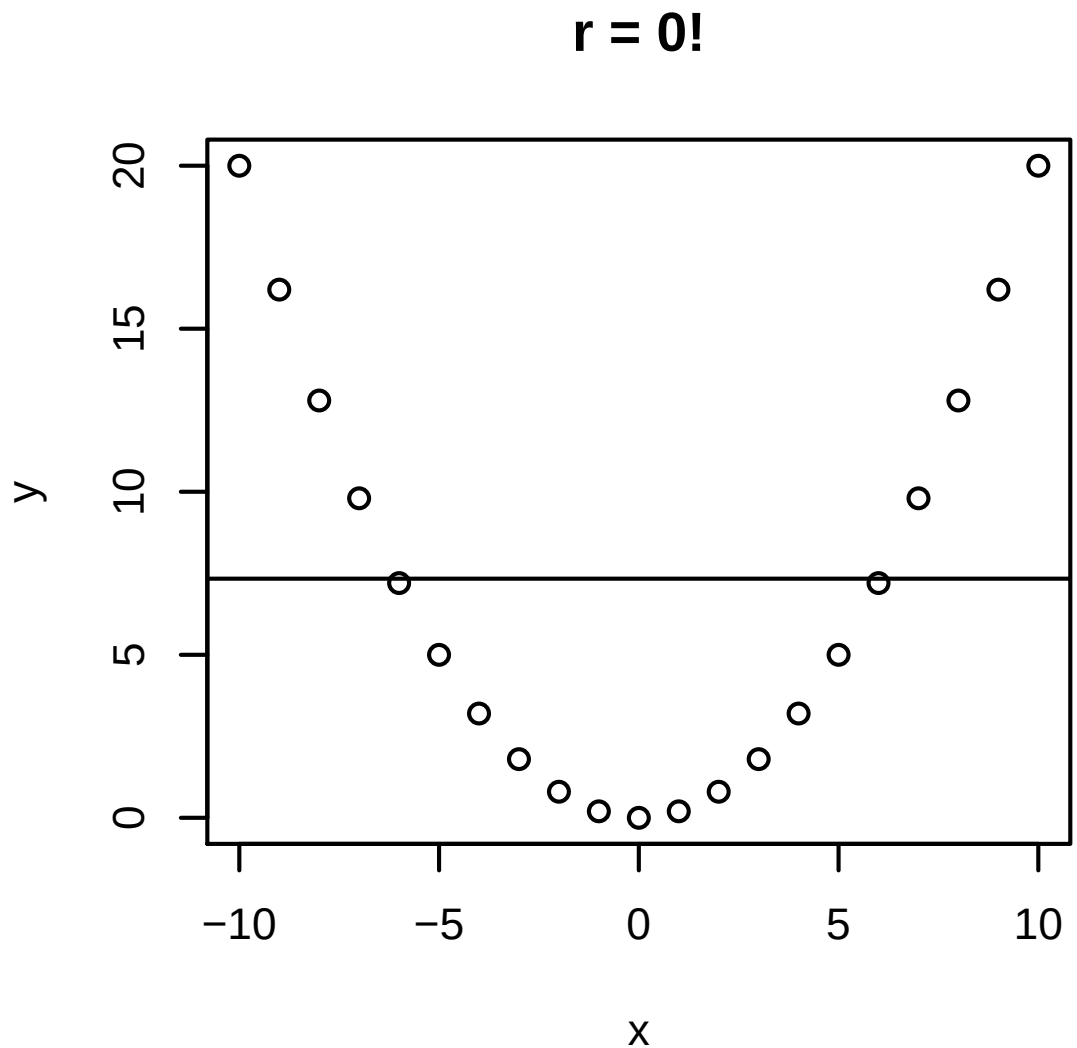
If we let ρ denote the population value of the correlation then $H_0 : \rho = 0$ is equivalent to $H_0 : \beta = 0$.

The usual assumptions apply to testing $H_0 : \rho = 0$ as for linear regression: independent observations, constant variance, normality or large sample size.

Note: the Spearman correlation coefficient is a correlation based on the ranks of the data. It is sometimes (but not often) used in place of the Pearson correlation coefficient when assumptions are in question.

Misinterpretation of 0 correlation

One danger with interpretation of a correlation is that a 0 correlation may be interpreted as implying no association between the two variables. However, this interpretation implicitly assumes linearity. The correlation coefficient only assesses *linear* associations and can completely miss a non-linear association.



Regression analysis of observational studies

The tooth growth experiment used randomization of the experimental units (guinea pigs). Animals were randomly assigned to receive a dose of vitamin C and one of the supplement types. As another example, in the NPK experiment the plots of land were randomly assigned to one of four combinations of N, P, and K within each block.

An *observational* study is one which does not involve random assignment to experimental conditions. Instead, observations are made on naturally occurring processes (it is also called a "natural history" study). In an observational study we need to be more careful about interpretation of results. In general, we cannot make inferences about causal effects, but rather only about associations between variables. This is sometimes expressed by the saying that "correlation is not causation", i.e., just because two variables are correlated does not mean one has a causal effect on the other.

Correlation is not causation

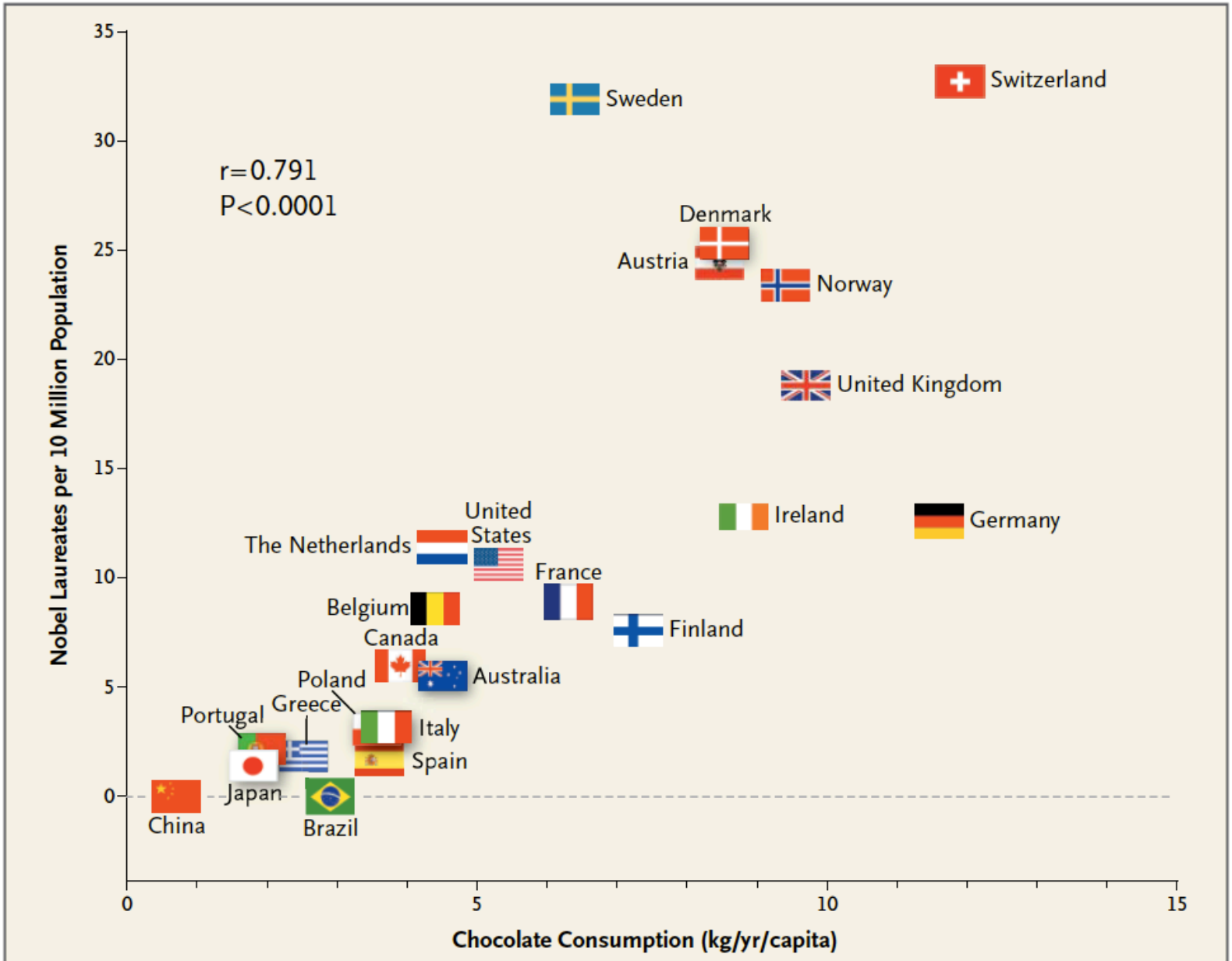


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

The dangers of extrapolation

Another potential pit-fall with regression is extrapolation. The linearity assumption can get us into big trouble if we use it to extrapolate beyond the range of the data.

For example, it would be dangerous to use the results from the tooth-growth data to extrapolate to much higher doses of vitamin C than were used in the experiment.

Such extrapolations are only valid when the linear relationship is true for the entire range of the predictor variable.

Here's an example of what can happen if we take our model too seriously...

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

The Athens Olympic Games could be viewed as another giant experiment in human athletic achievement. Are women narrowing the gap with men, or falling further behind? Some argue that the gains made by women in running events between the 1930s and the 1980s are decreasing as the women's achievements plateau[1]. Others contend that there is no evidence that athletes, male or female, are reaching the limits of their potential[1,2].

In a limited test, we plot the winning times of the men's and women's Olympic finals over the past 100 years (ref. 3; for data set, see supplementary information) against the competition date (Fig. 1). A range of curve-fitting procedures were tested (for methods, see supplementary information), but there was no evidence that the addition of extra parameters improved the model fit significantly from the simple linear relationships shown here. The remarkably strong linear trends that were first highlighted over ten years ago[2] persist for the Olympic 100-metre sprints. There is no indication that a plateau has been

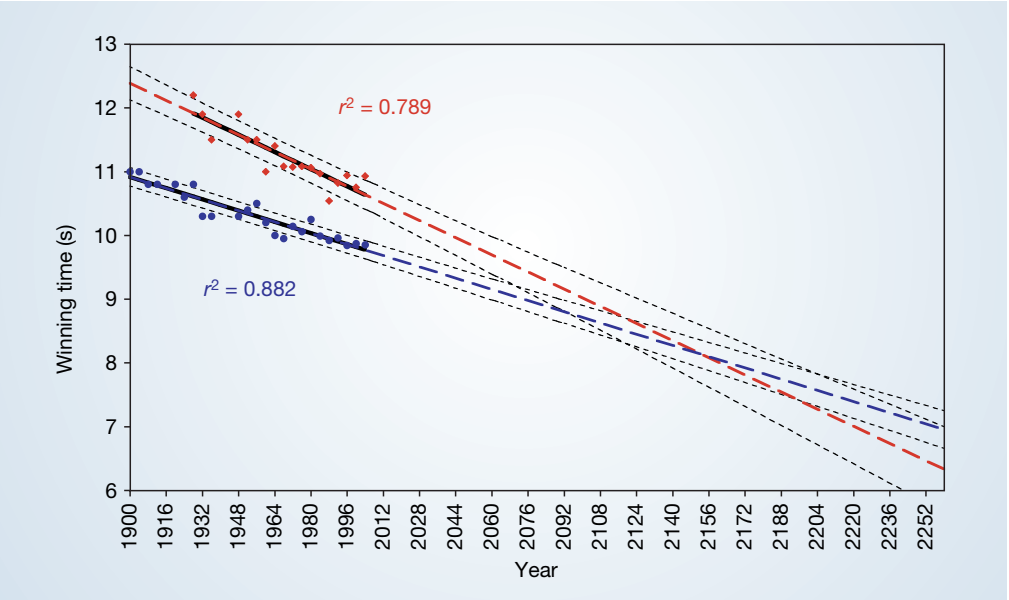


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

say that drug use explains why women's times were improving faster than men's, particularly as that improvement slowed after the introduction of drug testing[1]. However, no evidence for this is found here. By contrast, those who maintain that there could be a continuing decrease in gender gap point out that only a minority of the world's female population has been given the opportunity to compete (O. Anderson,

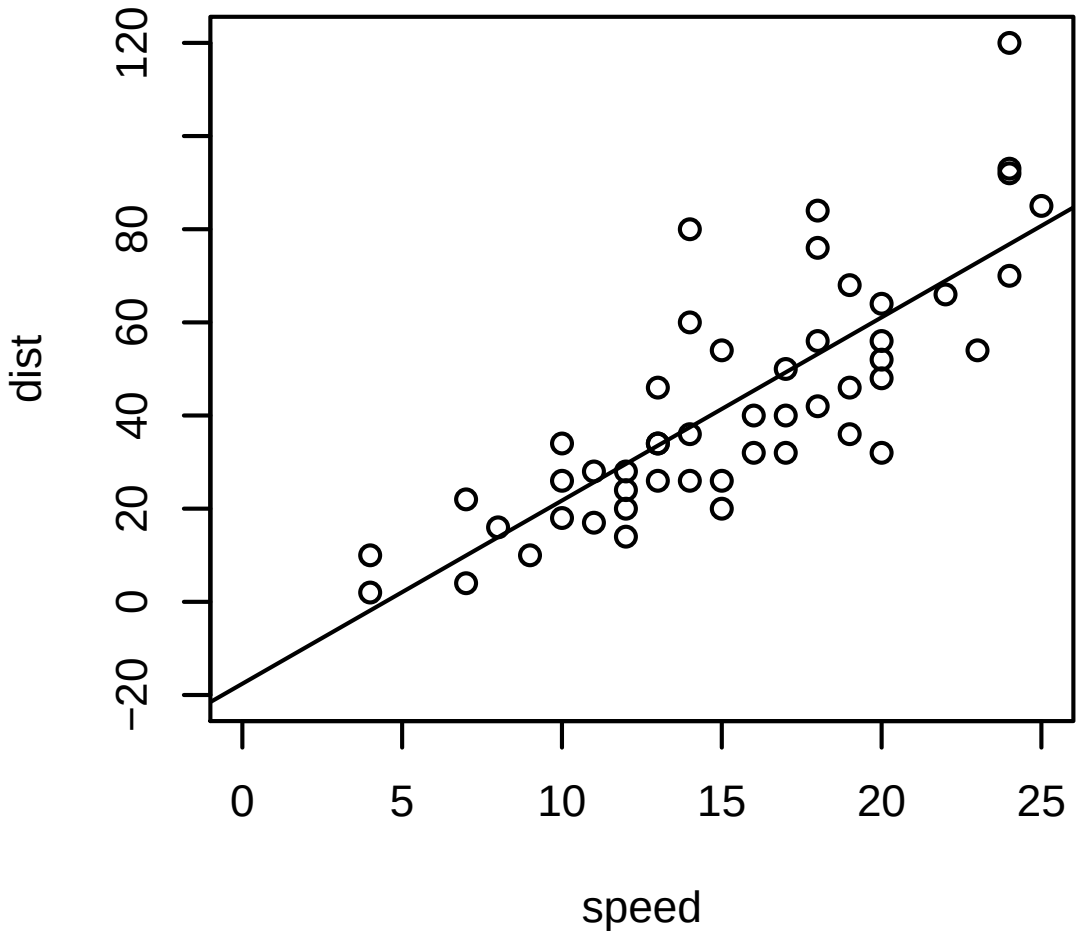
Lung cancer

Intragenic ERBB2 kinase mutations in tumours

The protein-kinase family is the most frequently mutated gene family found in human cancer and faulty kinase enzymes are being investigated as promising targets for the design of antitumour thera-

Extrapolating to 0

Recall that the intercept α is the mean response corresponding to $X = 0$. If 0 is not in the range of the values of X in the data then this involves extrapolation. As a result it can be dangerous to interpret the intercept too literally. For the cars data, the model tells us that the stopping distance for a car travelling 0 mph is -18 ft!



All Models are Wrong!

There is a basic tenet of statistics that models in general should not be interpreted strictly, i.e., that we should never believe our models are correct.

> *All models are wrong - some are useful. . . . George E.P. Box*

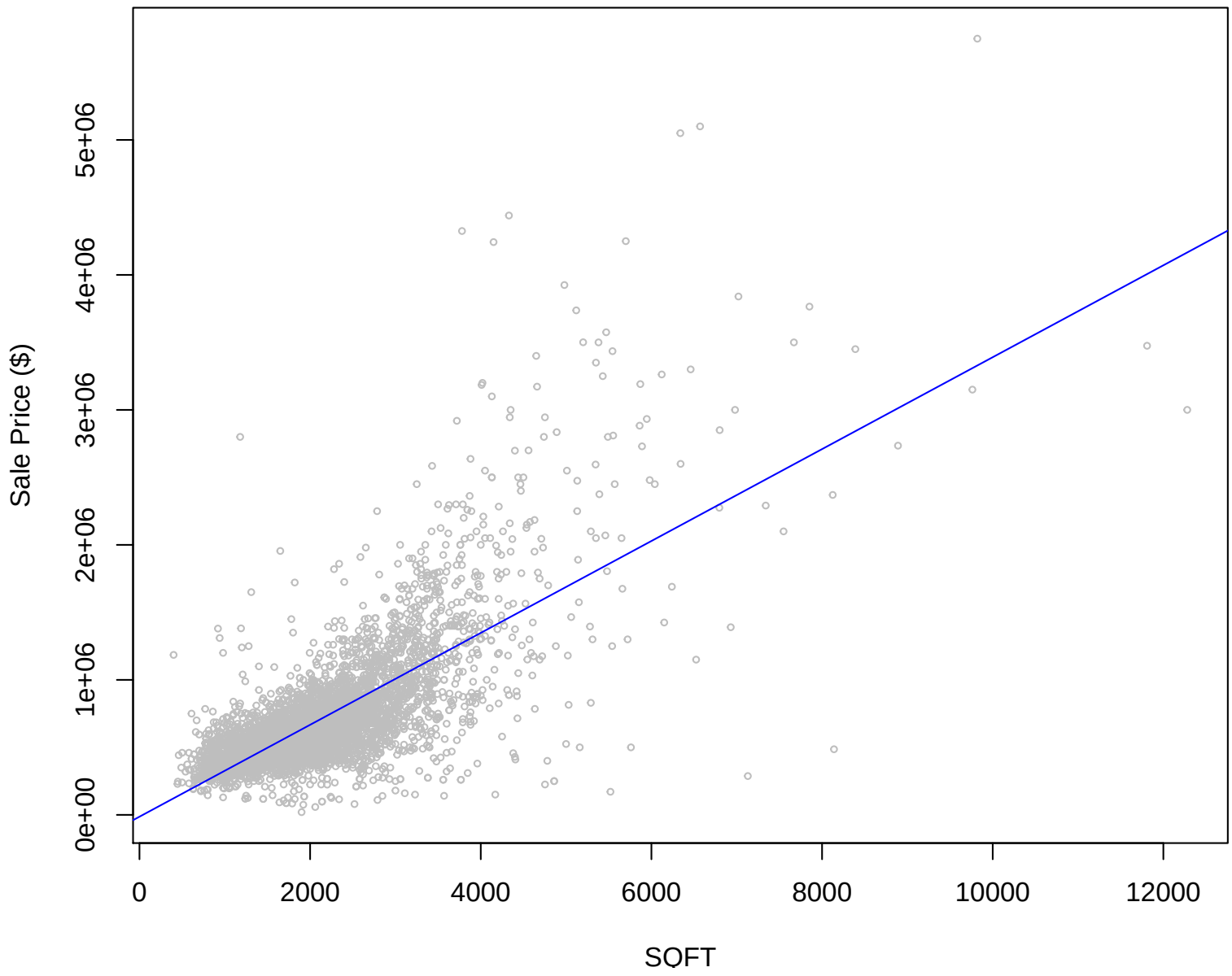
The statistician George Box meant it literally when he said "all models are wrong". What matters is *how wrong* they are.

> *Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. . . . . Since all models are wrong the scientist must be alert to what is importantly wrong.* [1]

[1] Science and statistics, by GEP Box, J. Amer. Stat. Assoc., 76(356):791-799, 1976.

Example of regression analysis of observational data

Data on a sample of house sales in Seattle in 2015-16 ("Sales.csv")



```
               Estimate   Std. Error   t value  Pr(>|t|)
(Intercept) -13574.815 11452.860091 -1.185277 0.2359694
SQFT           340.383     4.793597 71.007842 0.0000000
```

Interpretation:

$$E[\text{Price}] = -13,575 + 340.4X,$$

where X is the square-footage of the house. The interpretation of the slope is *the average difference in price per unit difference in square-feet* or *the average difference in price per square foot*.

Note: the intercept of $-13,575$ is interpreted as the estimated sale price of a house with 0 square-feet. This is another example of how extrapolation to 0 can be difficult to interpret.

Regression through the origin

In this example, it makes sense to force the intercept to be 0. This is called regression through the origin because the fitted regression line goes through the origin (0,0).

```
     Estimate Std. Error  t value Pr(>|t|)
SQFT 335.1952   1.955036 171.4522        0
```

In this model the average difference in price is 335.20 per square foot, which is close to the slope in the previous model but not identical. The slopes in the two models have different interpretations: the no-intercept model says that average price is proportional to square-feet, i.e., $E[\text{Price}] = 335.2X$. Intuitively, this makes sense and allows a literal interpretation as: *the price of a house is* 335.20 *per square foot*. The fitted line goes through the origin, i.e., the point $(0, 0)$. Note that the slope is very close to the previous one because the intercept was very small (relative to the scale of the response).

2. Regression and ANOVA with 2 or More Factors

- 2-way ANOVA
- Interactions
- Regression analysis with interactions
- The overall F-test

The Tooth-Growth Experiment: Factorial Design

We call this a *factorial design*. The factors "supp" and "dose" are called **crossed** because we `cross` each level of supp with each level of dose.[2]

The experimental units (guinea pigs) were randomly assigned to the treatment groups at random, with 10 units per group. This is called a `completely randomized` design (in contrast with a randomized block design).

Blocking might be used in this context if there were large genetic effects that needed to be controlled. In that case, we might take litters of 6 guinea pigs each and assign the 6 treatment combinations to the guinea pigs *within litters*.

[2] Later we will contrast this with "nested" factors.

Statistical Models for ANOVA

The model for ANOVA with 1 factor (called 1-way ANOVA):

$$X_{ij} = \mu_i + \epsilon_{ij}$$

where μ_i is the population mean for the ith group and the random variable ϵ_{ij} represents the "error" for the given observation, i.e., how far the observation is away from its predicted value μ_i.

An equivalent form of the 1-way ANOVA model:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

This model is equivalent to $\mu_i + \epsilon_{ij}$, where μ_i is written as $\mu + \alpha_i$. We call μ the overall mean and α_i the effect of treatment level i, i.e., $\alpha_i = \mu_i - \mu$.

Overparametrization of ANOVA models

The second form of the 1-way ANOVA model has 1 more parameter than can be estimated from the data.

Example: 1-way ANOVA with 3 groups

Model parameters:

1. μ=overall mean
2. α_1 = difference between Group 1 mean and overall mean
3. α_2 = difference between Group 2 mean and overall mean
4. α_3 = difference between Group 3 mean and overall mean

Thus, we use 4 parameters to describe 3 group means. The model is overparametrized. To address this we impose a constraint on the model parameters such as $\sum_i \alpha_i = 0$ (or sometimes $\alpha_1 = 0$.)

The 2-way ANOVA Model

If there are 2 experimental factors, we use 2-way ANOVA for the analysis. The statistical model for 2-way ANOVA:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

where i is the index for the level of factor A, j is the index for the level of factor B, X_{ijk} is the outcome for the kth unit with $A = i$ and $B = j$.

- μ is the overall mean
- α_i is the effect of level i of factor A
- β_j is the effect of level j of factor B

As with 1-way ANOVA we impose constraints on the parameters because the model is overparametrized. We will revisit parametrization of models when we study linear regression and its relationship to ANOVA.

The ANOVA decomposition for 2-way ANOVA

Just like for 1-way ANOVA, there is a decomposition of the variability in the observations into different sources. In this case, the total variability is decomposed into variability due to factor A, variability due to factor B and error variability. The 'anova' function is used to calculate these.

```
anova(lm(len ~ factor(dose)+supp, data=ToothGrowth))
```

```
Analysis of Variance Table

Response: len
             Df  Sum Sq Mean Sq F value    Pr(>F)
factor(dose)  2 2426.43 1213.22  82.811 < 2.2e-16 ***
supp          1  205.35  205.35  14.017 0.0004293 ***
Residuals    56  820.43   14.65
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the variable 'dose' must be designated as a factor so that it will not be treated as numeric. The degrees of freedom are divided up as follows:

- 2 df for dose because there are 3 dose groups
- 1 df for supp because there are 2 'supp' groups

The df for Residuals (or "Error") are calculated by subtraction using the fact that the df for the 3 sources must add up to the total df, which is $n - 1$, i.e., 2+1+56=60-1.

Hypothesis Testing

For each factor we wish to test the null hypothesis of no differences between group means. For dose, we test $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. The F-statistic is calculated just as in 1-way ANOVA, i.e., the mean-square for the dose factor divided by the error mean-square: $F = 1213.22/14.65 = 82.811$. This is compared with the $F_{2,56}$ distribution, yielding $p < 0.0001$.

Similarly, for `supp` we have $F = 205.35/14.65$ which is compared with the $F_{1,56}$ distribution again yielding a highly statistically significant result ($p = 0.0004$).

Conclusion: there is strong evidence that both `dose` and `supp` have effects on the outcome variable.

2-way ANOVA versus separate 1-way ANOVAs for each factor

What would happen if we ran separate analyses for each factor using 1-way ANOVA?

```
summary(aov(len ~ factor(dose), data=ToothGrowth))
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(dose)  2   2426    1213   67.42 9.53e-16 ***
Residuals    57   1026      18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(len ~ supp, data=ToothGrowth))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
supp         1    205  205.35   3.668 0.0604 .
Residuals   58   3247   55.98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2-way ANOVA versus 1-way ANOVA

1. The MS's for each factor are the same in the 1-way ANOVAs as in the 2-way ANOVA. This is a consequence of the balanced factorial design: equal numbers of observations for all 6 treatment combinations. In general, the MS for a factore can change when other factors are added to the model.
2. The error MS is higher for the 1-way ANOVA because the variability due to the missing factor becomes part of the `error` variability. Therefore, the F-statistics are much smaller (and even lose statistical significance for the `supp` factor). This illustrates the value of 2-way ANOVA.

Interactions

One of the advantages of a factorial design, is that there is another hypothesis we can test in addition to testing the effects of the 2 factors: we can test for *interaction* between the two factors. In other words, does the effect of one factor depend on the level of the other factor?

We can explore the interactions descriptively using the group sample means:

```
     0.5     1     2
OJ 13.23 22.70 26.06
VC  7.98 16.77 26.14
```

Now look at the differences between OJ and VC by dose:

```
  0.5     1     2
 5.25  5.93 -0.08
```

There are large differences between OJ and VC are large for doses 0.5 and 1 but not for dose 2. This suggest an interactive effect.

2-way ANOVA with interaction

```
anova(lm(len ~ factor(dose)*supp, data=ToothGrowth))
```

```
Analysis of Variance Table

Response: len
                  Df  Sum Sq Mean Sq F value    Pr(>F)
factor(dose)       2 2426.43 1213.22  92.000 < 2.2e-16 ***
supp               1  205.35  205.35  15.572 0.0002312 ***
factor(dose):supp  2  108.32   54.16   4.107 0.0218603 *
Residuals         54  712.11   13.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is evidence for an interaction between dose and supp (p=0.02).

Note that the MS's for dose and supp are unchanged from previous analyses – this is again a result of the balanced factorial design and will not always be true.

The 2-way ANOVA Model with Interaction

To represent interactions we add a new term to the model:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

In this model, the term γ_{ij} describes the interaction between the two factors. For the vitamin C experiment there are 6 of these parameters ($i = 1, 2, 3; j = 1, 2$). As before constraints are imposed on the parameters because the model is overparametrized.

The terms α_i and β_j in the interaction model are called the `main effects` of the dose and supp factors.

Interpretation of the main effects

We can think of the main effects for a factor as representing the average effects of the factor, averaged over the levels of the other factor.

For example, the results show that the main effect of supp is significant. This represents a comparison of the overall difference between OJ and VC averaged over the levels of dose. To examine these effects we look at the sample means for OJ and VC ignoring the dose level.

```
   OJ    VC
20.66 16.96
```

There are large differences between the groups overall.

Is this still a valid inference when we have evidence of an interaction? Yes, but we have to be careful to remember that the main effect represents the average effect of a factor. If the levels of dose used in the experiment are representative of the range of doses typically used in practice, then the main effect of supp is meaningful.

Using the no-interaction model

In some situations we ignore the possibility of interactions, and use the no-interaction model to make inferences on the main effects.

When is this valid? There are several considerations:

1. Are the levels of the factors typical of those used in practice?
2. Is there reason to believe that interactions, if present, would not be substantial (e.g., from previous experiments)?
3. Is it a balanced factorial design? If not, we have to consider the possibility of confounding - we will come back to this in the context of linear regression.

The NPK experiment

A 3-way factorial experiment on the growth of peas.

The 3 factors are:

1. N: indicator (0/1) for the application of nitrogen.
2. P: indicator (0/1) for the application of phosphate.
3. K: indicator (0/1) for the application of potassium.

The outcome is

yield: Yield of peas, in pounds/plot.

Factorial design

There were 3 observations for each of the 8 combinations of factor levels.

```
, , K = 0

   P
N   0 1
  0 3 3
  1 3 3

, , K = 1

   P
N   0 1
  0 3 3
  1 3 3
```

3-way ANOVA for the NPK experiment

Ignoring the blocking in this experiment for now:

```
summary(aov(yield ~ N*P*K,data=npk))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
N            1  189.3  189.28   6.161 0.0245 *
P            1    8.4    8.40   0.273 0.6082
K            1   95.2   95.20   3.099 0.0975 .
N:P          1   21.3   21.28   0.693 0.4175
N:K          1   33.1   33.14   1.078 0.3145
P:K          1    0.5    0.48   0.016 0.9019
N:P:K        1   37.0   37.00   1.204 0.2887
Residuals   16  491.6   30.72
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3-way ANOVA without interactions

```
summary(aov(yield ~ N+P+K,data=npk))
```

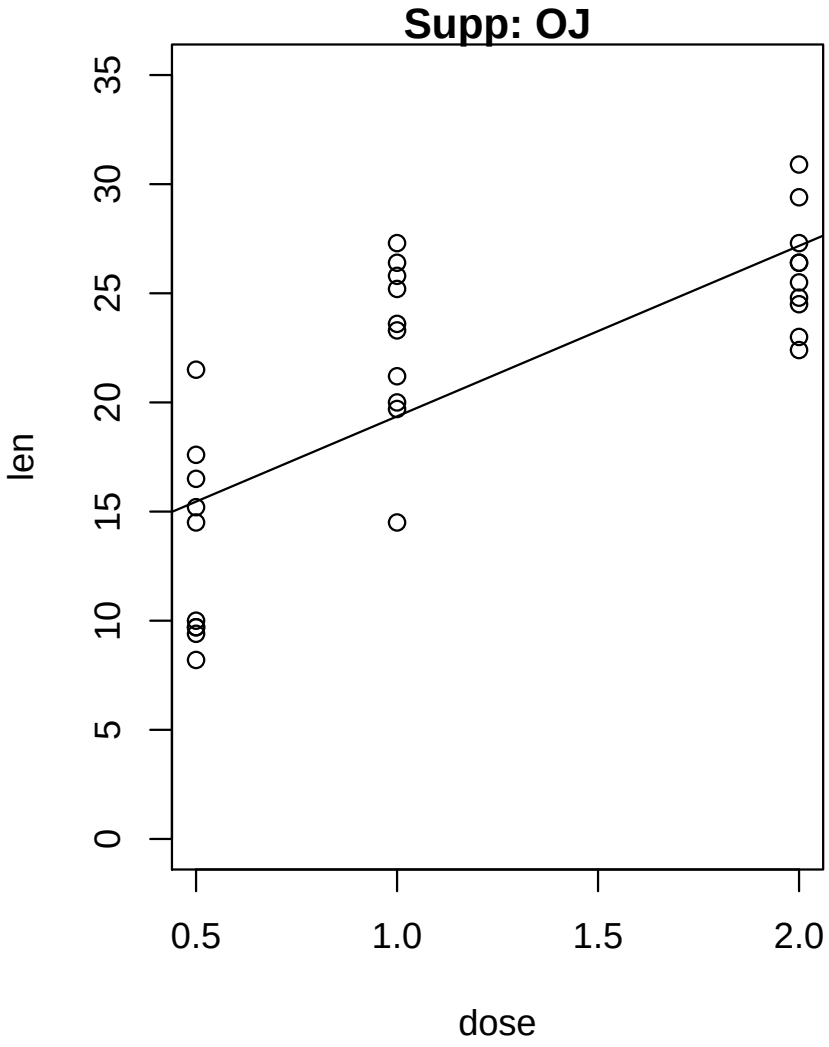
```
            Df Sum Sq Mean Sq F value Pr(>F)
N            1  189.3  189.28   6.488 0.0192 *
P            1    8.4    8.40   0.288 0.5974
K            1   95.2   95.20   3.263 0.0859 .
Residuals   20  583.5   29.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

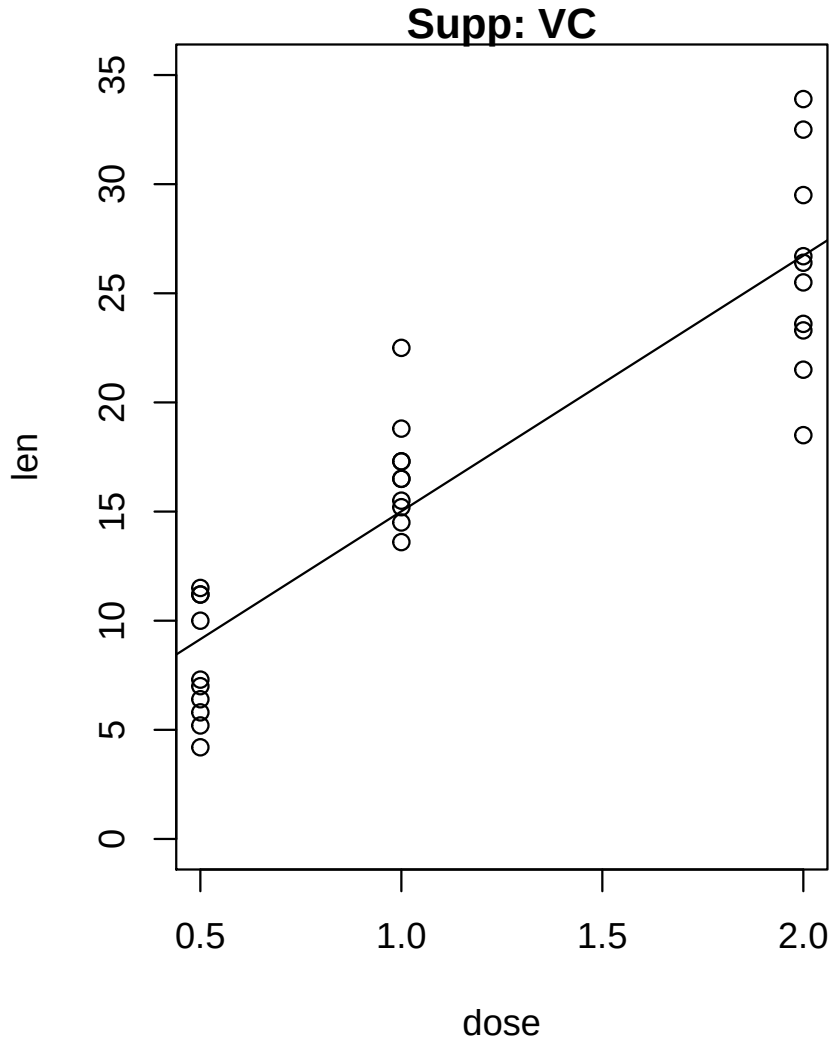
Regression Analysis of Experiments with 2 or More Factors

We can also fit linear regression models that account for both the dose and supplement type factors. As before when we compared ANOVA and regression with 1 factor, the difference between ANOVA and regression is that for ANOVA, dose is treated as a categorical variable, whereas with regression it is treated as a quantitative variable and linear relationships are modeled.

We will review separate linear regression analyses for the two supplement groups and then see how to combine them.

Separate analyses of the two groups





Fitting the regression models for each group separately

This is an experiment with 2 factors: supp (OJ or VC), and dose ($X = 0.5, 1, 2$). Previously we fit dose-response models to the two supp groups separately.

$$\text{Group VC}: Y = \alpha_{VC} + \beta_{VC}\ X + \epsilon_{VC}$$

```
summary(lm(len ~ dose, data=d, subset=(supp=="OJ")))$coef
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 11.550000   1.721951 6.707508 2.788784e-07
dose         7.811429   1.301673 6.001070 1.824801e-06
```

$$\text{Group OJ}: Y = \alpha_{OJ} + \beta_{OJ}\ X + \epsilon_{OJ}$$

```
summary(lm(len ~ dose, data=d, subset=(supp=="VC")))$coef
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.29500   1.427060  2.308943 2.854201e-02
## dose        11.71571   1.078756 10.860392 1.509369e-11
```

Fitting two regression lines with a single model

There are two ways in which this is done in practice. Because we want to combine the two models into one model, we rewrite the two models using a comment error term ϵ (note that this is making an extra assumption):

$$\begin{aligned} Y &= \alpha_{OJ} + \beta_{OJ}\, X + \epsilon, \text{ if group OJ,} \\ &= \alpha_{VC} + \beta_{VC}\, X + \epsilon, \text{ if group VC.} \end{aligned}$$

To write the models with one equation define two indicator variables: $I_{VC} = I\{Group = "VC"\} = 1$, for group VC and 0 for group OJ, and similarly for $I_{OJ} = I\{Group = "OJ"\}$. Then the two regression models can be written as

$$\begin{aligned} Y &= I_{VC}(\alpha_{VC} + \beta_{VC}\, X) + I_{OJ}(\alpha_{OJ} + \beta_{OJ}\, X) + \epsilon \\ &= \alpha_{VC} I_{VC} + \beta_{VC} I_{VC} I_{VC} X + \alpha_{OJ} I_{OJ} + \beta_{OJ} I_{OJ} X \\ &= \alpha_{VC} I_{VC} + \beta_{VC} Z_{VC} + \alpha_{OJ} I_{OJ} + \beta_{OJ} Z_{OJ}. \end{aligned}$$

$$Y = \alpha_{VC} I_{VC} + \beta_{VC} Z_{VC} + \alpha_{OJ} I_{OJ} + \beta_{OJ} Z_{OJ}$$

This model is an example of a *multiple regression model*, which means that it has multiple predictor variables.

This model has four predictor variables: I_{VC}, $Z_{VC} = I_V C \times X$, I_{OJ}, and $Z_{OJ} = I_{OJ} \times X$.

Note that the model has no intercept! In the simplest type of regression model of the form $\alpha + \beta X$, the intercept term (α) is a parameter without any predictor variable attached to it. All 4 terms in the above model include a coefficient and a predictor variable. Note that we can think of the intercept as being the coefficient for the predictor variable which is equal to 1 for all observations, i.e., $\alpha = \alpha \times 1$.

Fitting the model

To suppress the intercept we use the notation "-1" in the 'lm' function.

```
d$I_OJ=as.numeric(d$supp=="OJ")
d$I_VC=as.numeric(d$supp=="VC")
d$Z_OJ=d$I_OJ*d$dose
d$Z_VC=d$I_VC*d$dose
summary(lm(len ~ -1 + I_OJ + Z_OJ + I_VC + Z_VC, data=d))$coef
```

```
      Estimate Std. Error  t value     Pr(>|t|)
I_OJ 11.550000   1.581394 7.303681 1.089558e-09
Z_OJ  7.811429   1.195422 6.534454 2.027753e-08
I_VC  3.295000   1.581394 2.083604 4.177218e-02
Z_VC 11.715714   1.195422 9.800486 9.442117e-14
```

The coefficient estimates are exactly the same as those obtained from fitting the models separately. However, the estimated SEs are slightly different because the combined model assumes the same error variance for both groups, whereas the separate models are not constrained in that way.

Linear regression with interaction

The second method of fitting the two regression lines with one model is by using an interaction model. This is just a different way of combining the two regression lines into one model:

$$Y = \beta_0 + \beta_1 X + \beta_2 I_{VC} + \beta_3 I_{VC} X + \epsilon$$

This model equation is described as a different *parametrization* of the model. In this form of the model, the first part $\beta_0 + \beta_1 X$ represents the OJ group, and the second part $\beta_2 I_{VC} + \beta_3 I_{VC} X$ represents the **difference** between the VC group and the OJ group. The correspondence of the parameters is as follows:

$\beta_0 = \alpha_{OJ}$,

$\beta_1 = \beta_{OJ}$,

$\beta_2 = \alpha_{VC} - \alpha_{OJ}$

$\beta_3 = \beta_{VC} - \beta_{OJ}$

Fitting the interaction model

The syntax for an interaction is A:B for the interaction between 2 variables A and B.

```
summary(lm(len ~ dose + I_VC + dose:I_VC, data=d))$coef
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 11.550000   1.581394  7.303681 1.089558e-09
dose         7.811429   1.195422  6.534454 2.027753e-08
I_VC        -8.255000   2.236429 -3.691152 5.073393e-04
dose:I_VC    3.904286   1.690582  2.309433 2.463136e-02
```

It can also be done with the "*" notation: "dose*I_VC"

Comparing results with the results of the separate model fits:

$\hat{\beta}_0 = 11.55 = \hat{\alpha}_{OJ}$

$\hat{\beta}_1 = 7.811 = \hat{\beta}_{OJ}$

$\hat{\beta}_2 = -8.255 = \hat{\alpha}_{VC} - \hat{\alpha}_{OJ} = 3.295 - 11.55$

$\hat{\beta}_3 = 3.904 = \hat{\beta}_{VC} - \hat{\beta}_{OJ} = 11.71571 - 7.811429$

Interpretation of the parameters in the interaction model

β_0: mean of Y *for group OJ*, dose 0

β_1: difference in mean of Y per unit difference in dose *for group OJ*

β_2: difference between mean Y for group VC and group OJ *for dose 0*

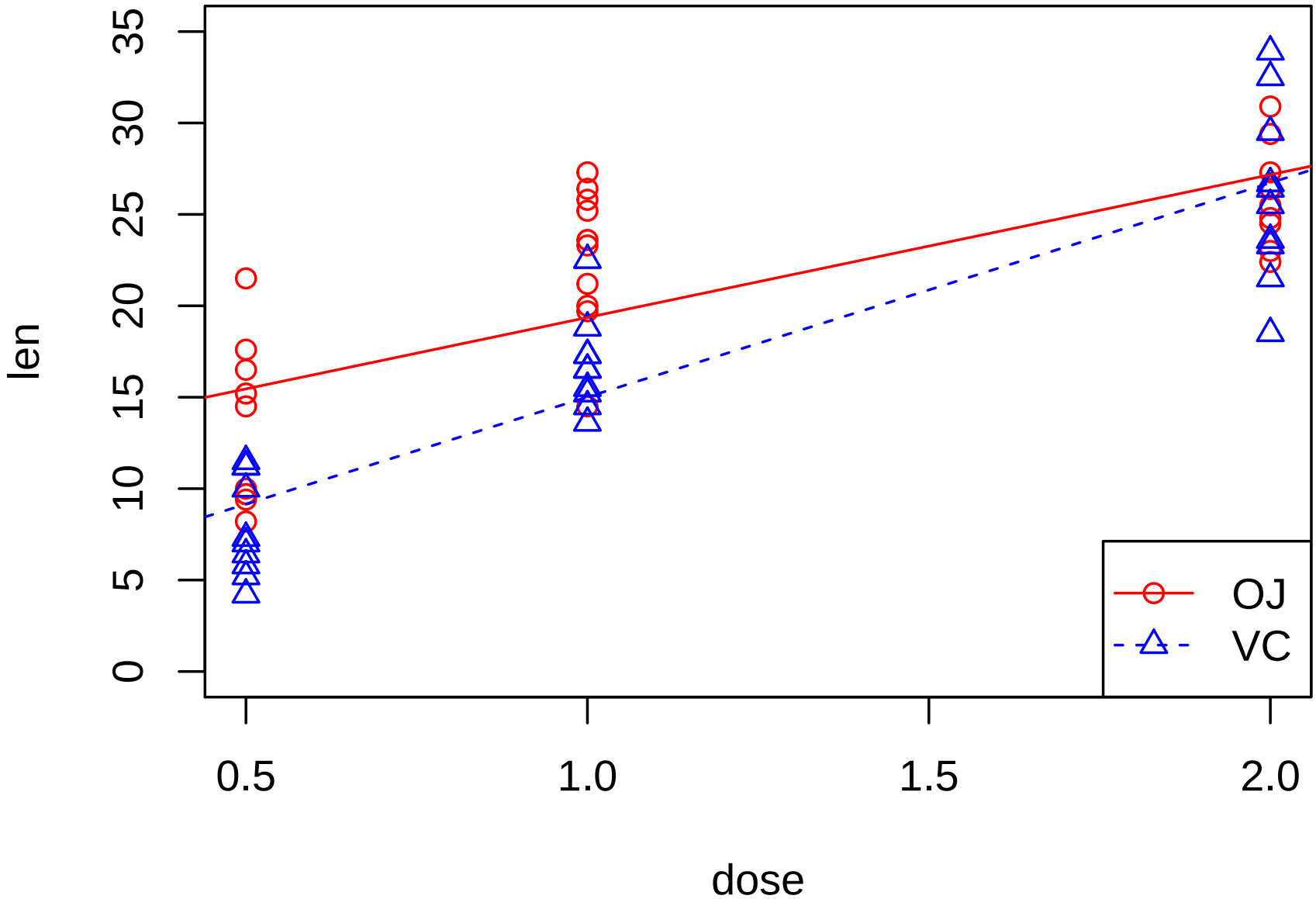
β_3: difference between regression coefficient for dose for group VC and regression coefficient for dose for group OJ

Terminology: the OJ group here plays the role of the "reference" group. The intercept (β_0) and the coefficient of dose (β_1) pertain to the reference group. The other terms involve contrasts between the other group and the reference group.

Note that the interpretation of the intercept is consistent with the general definition of an intercept as the mean response for all variables in the model set equal to 0.

Graphical interpretation of the interaction model

Interaction means non-parallel lines. The slope of the line for the VC group is greater than the slope of the line for the OJ group by the amount $\hat{\beta}_3 = 3.9$.



Advantages (and a disadvantage) of the combined model

Advantage of the combined model

Using the combined regression model allows us to answer two types of questions:

1. What is the effect of dose on tooth growth for each specific type of vitamin C?
2. Is the effect of dose on tooth growth **different** for the two types of vitamin C? We can estimate the difference between effects of dose in the two groups as well as perform confidence intervals and hypothesis tests for this difference.

Advantage of 2 separate models:

A (minor) disadvantage of the combined model is that we were forced to assume a constant error variance for the two groups, whereas the separate models can have different error variances. We will see how to avoid this problem with the combined model using robust SEs.

Comparison of ANOVA and regression for the interaction model

Comparing results from ANOVA with linear regression for this data, we see that the results are similar. Both models provide some evidence for an interaction.

If the true relationship between dose and response is linear, then linear regression can yield more powerful tests by taking advantage of the linearity.

The full regression output

```
summary(lm(len ~ supp*dose, data=ToothGrowth))
```

```
Call:
lm(formula = len ~ supp * dose, data = ToothGrowth)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2264 -2.8462  0.0504  2.2893  7.9386

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.550      1.581   7.304 1.09e-09 ***
suppVC        -8.255      2.236  -3.691 0.000507 ***
dose           7.811      1.195   6.534 2.03e-08 ***
suppVC:dose    3.904      1.691   2.309 0.024631 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-squared:  0.7296,    Adjusted R-squared:  0.7151
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

The overall F-test

The F-test for this model is called the "overall" F test because it is assessing the significance of the entire model (not one specific term in the model). The null hypothesis for this F-test is that all coefficients in the model (except for the intercept) are equal to 0. In this case it is unsurprisingly highly statisically significant because we already knew that some of the terms in the model are highly significant.