

## 6.2: Analysis of regression

### Single line

Analysis of correlation allows to determine if variables are dependent and calculate the strength and sign of the dependence. However, if the goal is to understand the other features of dependence (like direction), and, even more important, predict (extrapolate) results (Figure 6.2.1) we need another kind of analysis, the *analysis of regression*.

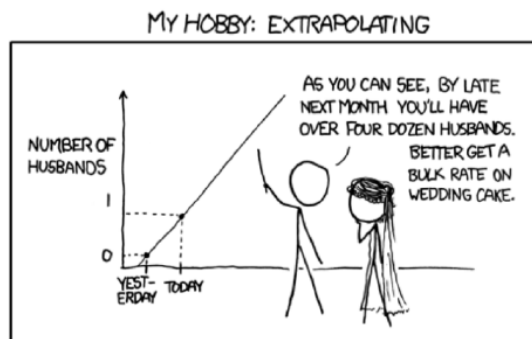


Figure 6.2.1 Extrapolation (taken from XKCD, <http://xkcd.com/605/>).

It gives much more information on the relationship, but requires us to assign variables *beforehand* to one of two categories: *influence* (predictor) or *response*. This approach is rooted in the nature of the data: for example, we may use air temperature to predict ice cream sales, but hardly the other way around.

The most well-known example is a simple *linear regression*:

$$\text{response} = \text{intercept} + \text{slope} \times \text{influence}$$

or, in R formula language, even simpler:

$$\text{response} \sim \text{influence}$$

That model estimates the average value of *response* if the value of *influence* is known (note that both effect and influence are *measurement* variables). The differences between observed and predicted values are model *errors* (or, better, *residuals*). The goal is to *minimize residuals* (Figure 6.2.3); since residuals could be both positive and negative, it is typically done via squared values, this method is called *least squares*.

Ideally, residuals should have the normal distribution with zero mean and constant variance which is not dependent on effect and influence. In that case, residuals are homogeneous. In other cases, residuals could show heterogeneity. And if there is the *dependence* between residuals and influence, then most likely the overall model should be non-linear and therefore requires the other kind of analysis.

Linear regression model is based on the several assumptions:

- **Linearity of the relationship.** It means that for a unit change in influence, there should always be a corresponding change in effect. Units of change in response variable should retain the same size and sign throughout the range of influence.
- **Normality of residuals.** Please note that normality of data is not an assumption! However, if you want to get rid of most other assumptions, you might want to use other regression methods like LOESS.
- **Homoscedasticity of residuals.** Variability within residuals should *remain constant* across the whole range of influence, or else we could not predict the effect reliably.

The null hypothesis states that *nothing* in the variability of response is explained by the model. Numerically, *R-squared* coefficient is the the degree to which the variability of response is explained by the model, therefore null hypothesis is that R-squared *equals zero*, this approach uses F-statistics (Fisher's statistics), like in ANOVA. There are also checks of additional null hypotheses that both *intercept* and *slope* are zeros. If all *three* p-values are smaller than the level of significance (0.05), the whole model is statistically significant.

Here is an example. The embedded `women` data contains observations on the height and weight of 15 women. We will try to understand the dependence between weight and height, graphically at first (Figure 6.2.2):

(Here we used function `Cladd()` which adds *confidence bands* to the plot<sup>[1]</sup>.)

Let us visualize residuals better (Figure 6.2.3):

To look on the results of model analysis, we can employ `summary()`:

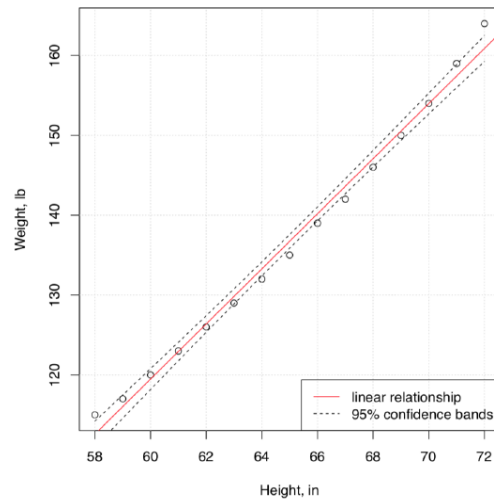


Figure 6.2.2 The relation between height and weight.

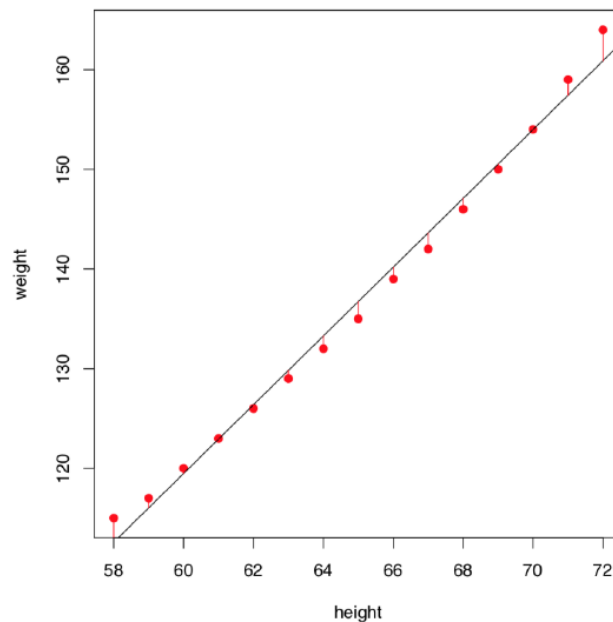


Figure 6.2.3 Residuals of the women weight vs. height linear model.

This long output is better to read from bottom to the top. We can say that:

- The significance of relation (reflected with R-squared) is high from the statistical point of view: F-statistics is 1433 with overall **p-value: 1.091e-14**.
- The R-squared (use **Adjusted R-squared** because this is better suited for the model) is really big,  $R^2 = 0.9903$ . This means that almost all variation in response variable (weight) is explained by predictor (height). R-squared is related with the coefficient of correlation and might be used as the measure of *effect size*. Since it is squared, high values start from 0.25:
- Both coefficients are statistically different from zero, this might be seen via “stars” (like \*\*\*), and also via actual p-values **Pr(>|t|): 1.71e-09** for intercept, and **1.09e-14** for **height**, which represents the slope. To calculate slope in degrees, one might run:
- Overall, our model is:  

$$\text{Weight (estimated)} = -87.51667 + 3.45 * \text{Height},$$
 so if the height grows by 4 inches, the weight will grow on approximately 14 pounds.
- The maximal positive residual is 3.1167lb, maximal negative is -1.7333lb.

- Half of residuals are quite close to the median (within approximately  $\pm 1$  interval).

On the first glance, the model summary looks fine. However, before making any conclusions, we must also *check assumptions* of the model. The command `plot(women.lm)` returns four consecutive plots:

- First plot, *residuals vs. fitted values*, is most important. Ideally, it should show *no structure* (uniform variation and no trend); this satisfies both linearity and homoscedasticity assumptions.
- Unfortunately, `women.lm` model has an obvious trend which indicates non-linearity. Residuals are positive when fitted values are small, negative for fitted values in the mid-range, and positive again for large fitted values. Clearly, the first assumption of the linear regression analysis is violated.
- 
- To understand residuals vs. fitted plots better, please run the following code yourself and look on the resulted plots:
- On the the next plot, standardized residuals do not follow the normal line perfectly (see the explanation of the QQ plot in the previous chapter), but they are “good enough”. To review different variants of these plots, run the following code yourself:
- Test for the normality should also work:
- The third, *Scale-Location* plot, is similar to the residuals vs. fitted, but instead of “raw” residuals it uses the square roots of their standardized values. It is also used to reveal trends in the magnitudes of residuals. In a good model, these values should be more or less randomly distributed.
- Finally, the last plot demonstrates which values exert most influence over the final shape of the model. Here the two values with most leverage are the first and the last measurements, those, in fact, that stay furthest away from linearity.

(If you need to know more about summary and plotting of linear models, check help pages with commands `?summary.lm` and `?plot.lm`. By the way, as ANOVA has many similarities to the linear model analysis, in R you can run same diagnostic plots for any ANOVA model.)

Now it is clear that our first linear model *does not work well* for our data which is likely *non-linear*. While there are many non-linear regression methods, let us modify it first in a more simple way to introduce non-linearity. One of simple ways is to add the *cubed term*, because weight relates with volume, and volume is a cube of linear sizes:

(Function `I()` was used to tell R that `height^3` is arithmetical operation and not the part of model formula.)

The quick look on the residuals vs. fitted plot (Figure 6.2.4) shows that this second model fits much better! Confidence bands and predicted line are also look more appropriate :

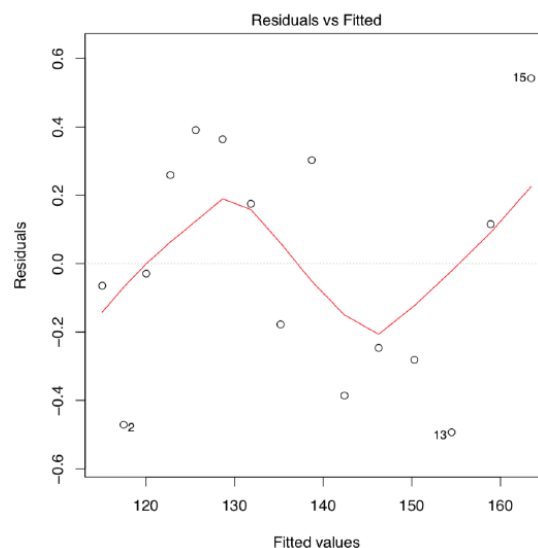


Figure 6.2.4 Residuals vs. fitted plot for the women height/weight model with the cubed term.

You may want also to see the *confidence intervals* for linear model parameters. For that purpose, use `confint(women.lm)`.

Another example is from egg data studied graphically in the second chapter (Figure 2.9.1). Does the length of the egg linearly relate with with of the egg?

We can analyze the assumptions first:

The most important, residuals vs. fitted is not perfect but could be considered as “good enough” (please check it yourself): there is no obvious trend, and residuals seem to be more or less equally spread (homoscedasticity is fulfilled). Distribution of residuals is close to normal. Now we can interpret the model summary:

Significance of the slope means that the line is definitely *slanted* (this is actually what is called “relation” in common language). However, intercept is not significantly different from zero:

(Confidence interval for intercept includes zero.)

To check the magnitude of effect size, one can use:

This is a really large effect.

Third example is based on a simple idea to check if the success in multiple choice test depends on time spent with it. Data presents in `exams.txt` file which contains results of two multiple choice tests in a large class:

First variable is the number of test, two others are order of finishing the work, and resulted number of points (out of 50). We assume here that the order reflects the time spent on test. Select one of two exams:

... and plot it first (please check this plot yourself):

Well, no visible relation occurs. Now we approach it inferentially:

As usual, this output is read from bottom to the top. First, statistical significance of the relation is absent, and relation (adjusted R-squared) itself is almost zero. Even if intercept is significant, slope is not and therefore could easily be zero. There is no relation between time spent and result of the test.

To double check if the linear model approach was at all applicable in this case, run diagnostic plots yourself:

And as the final touch, try the regression line and confidence bands:

Almost horizontal—no relation. It is also interesting to check if the other exam went the same way. Please find out yourself.

) . Please find which morphological measurement characters are most correlated, and check the linear model of their relationships.

) plant. Please find which pair of morphological characters is most correlated and analyze the linear model which includes these characters. Also, check if length of leaf is different between the three biggest populations of sundew.

As the linear models and ANOVA have many in common, there is no problem in the analysis of multiple groups with the default linear regression methods. Consider our ANOVA data:

This example shows few additional “tricks”. First, this is how to analyze several response variables at once. This is applicable also to `aov()`—try it yourself.

Next, it shows how to re-level factor putting one of proximal levels first. That helps to compare coefficients. In our case, it shows that blonds do not differ from browns by weight. Note that “intercepts” here have no clear relation with plotting linear relationships.

It is also easy to calculate the effect size because *R-squared is the effect size*.

Last but not least, please check assumptions of the linear model with `plot(lm(...))`. At the moment in R, this works only for singular response.

Is there the linear relation between the weight and height in our ANOVA `hwc` data?

## Many lines

Sometimes, there is a need to analyze not just linear relationships between variables, but to answer second order question: *compare several regression lines*.

In formula language, this is described as

`response ~ influence * factor`

where factor is a categorical variable responsible for the distinction between regression lines, and star (\*) indicates that we are simultaneously checking (1) response from influence (predictor), (2) response from factor and (3) response from *interaction* between influence and factor.

This kind of analysis is frequently called ANCOVA, “ANalysis of COVAriation”. The ANCOVA will check if there is any difference between intercept and slope of the first regression line and intercepts and slopes of all other regression lines where each line corresponds with one factor level.

Let us start from the example borrowed from M.J. Crawley’s “R Book”. 40 plants were treated in two groups: grazed (in first two weeks of the cultivation) and not grazed. Rootstock diameter was also measured. At the end of season, dry fruit production was measured from both groups. First, we analyze the data graphically:

As it is seen on the plot (Figure 6.2.5), regression lines for grazed and non-grazed plants are likely different. Now to the ANCOVA model:

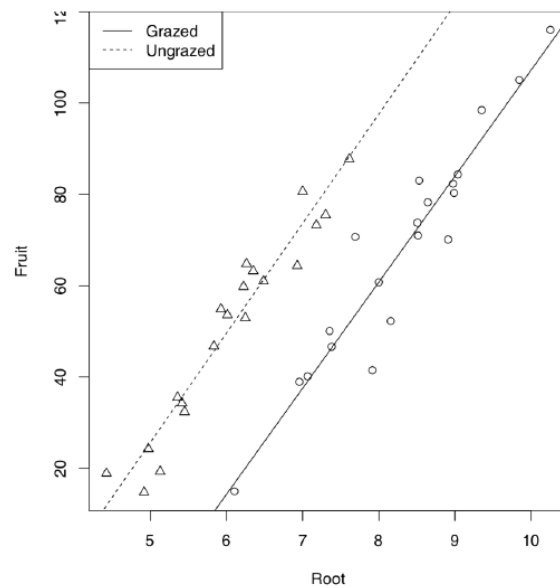


Figure 6.2.5 Grazed vs. non-grazed plants: linear models.

Model output is similar to the linear model but one more term is present. This term indicated *interaction* which labeled with colon. Since **Grazing** factor has two level arranged alphabetically, first level (**Grazed**) used as default and therefore (**Intercept**) belongs to grazed plants group. The intercept of non-grazed group is labeled as **GrazingUngrazed**. In fact, this is not even an intercept but difference between intercept of non-grazed group and intercept of grazed group. Analogously, slope for grazed is labeled as **Root**, and difference between slopes of non-grazed and grazed labeled as **Root:GrazingUngrazed**. This difference is interaction, or how grazing affects the shape of relation between rootstock size and fruit weight. To convert this output into regression formulas, some calculation will be needed:

$$\text{Fruit} = -125.174 + 23.24 * \text{Root} \text{ (grazed)} \quad \text{Fruit} = (-125.174 + 30.806) + (23.24 + 0.756) * \text{Root} \text{ (non-grazed)}$$

Note that difference between slopes is not significant. Therefore, interaction could be ignored. Let us check if this is true:

First, we updated our first model by removing the interaction term. This is the *additive* model. Then `summary()` told us that all coefficients are now significant (check its output yourself). This is definitely better. Finally, we employed AIC (Akaike's Information Criterion). AIC came from the theory of information and typically reflects the entropy, in other words, adequacy of the model. The smaller is AIC, the better is a model. Then the second model is the unmistakable winner.

By the way, we could specify the same additive model using plus sign instead of star in the model formula.

What will the AIC tell about our previous example, women data models?

Again, the second model (with the cubed term) is better.

It is well known that in the analysis of voting results, dependence between attendance and the number of people voted for the particular candidate, plays a great role. It is possible, for example, to elucidate if elections were falsified. Here we will use the [elections.txt](#) data file containing voting results for three different Russian parties in more than 100 districts:

To simplify typing, we will `attach()` the data frame (if you do the same, do not forget to `detach()` it at the end) and calculate proportions of voters and the overall attendance:

Now we will look on the dependence between attendance and voting graphically (Figure 6.2.6):

So the third party had a voting process which was suspiciously different from voting processes for two other parties. It was clear even from the graphical analysis but we might want to test it inferentially, using ANCOVA:

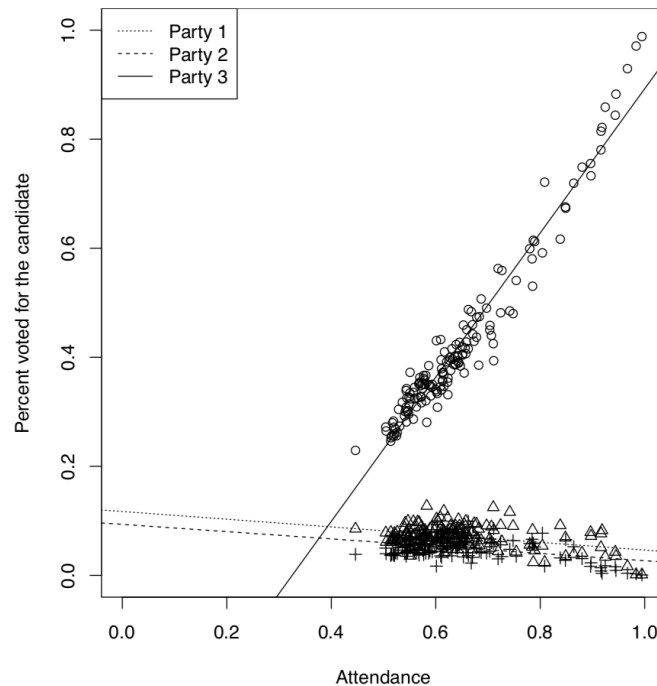


Figure 6.2.6 Voting results vs. attendance for every party.

(Here we created and checked the new data frame. In `elections2`, all variables are now `stack()`'ed in two columns, and the third column contains the party code.)

Here `(Intercept)` belongs specifically to the model for first party. Its p-value indicates if it differs significantly from zero. Second coefficient, `atten`, belongs to the continuous predictor, attendance. It is not an intercept but slope of a regression. It is also compared to zero.

Next four rows represent differences from the first party, two for intercepts and two for slopes (this is the traditional way to structure output in R). Last two items represent interactions. We were most interested if there is an interaction between attendance and voting for the third party, this interaction is common in case of falsifications and our results support this idea.

*Figure 6.2.7 Heterostyly in primroses: flowers from the different plants of one population.*

## More then one way, again

Armed with the knowledge about AIC, multiplicative and additive models, we can return now to the ANOVA two-way layouts, briefly mentioned before. Consider the following example:

(To start, we converted `dose` into factor. Otherwise, our model will be ANCOVA instead of ANOVA.)

Assumptions met, now the core analysis:

Now we see what was already visible on the interaction plot (Figure 5.3.4: model with interactions is better, and significant are *all three terms*: dose, supplement, and interaction between them.

Effect size is really high:

*Post hoc* tests are typically more dangerous in two-way analysis, simply because there are much more comparisons. However, it is possible to run `TukeyHSD()`:

The rest of comparisons is here omitted, but `TukeyHSD()` has plotting method allowing to plot the single or last element (Figure 6.3.1):

## References

1. Function `Cladd()` is applicable only to simple linear models. If you want confidence bands in more complex cases, check the `Cladd()` code to see what it does exactly.

This page titled 6.2: Analysis of regression is shared under a [Public Domain](#) license and was authored, remixed, and/or curated by Alexey Shipunov via source content that was edited to the style and standards of the LibreTexts platform.