Analysis of variance (I)

- In all general linear models (regression, ANOVA, ANCOVA) we strive to ”understand”, explain or account for **variation in our response variable** – by fitting a model to data, which we can use to actually predict the value of Y.
- ANOVA achieves this, by **partitioning total variance** (in Y) into different components or parts, typically presented in an **ANOVA-table**. This is easiest to see in a simple regression type model:

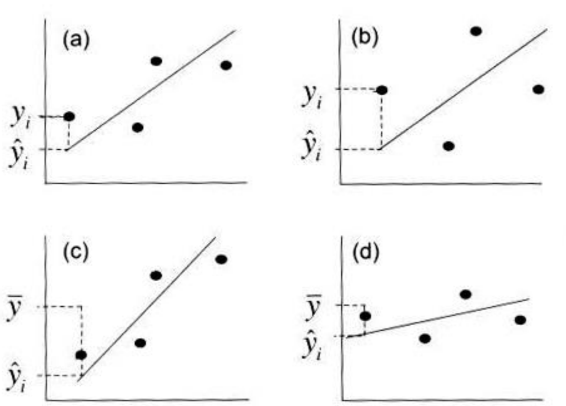
Sums of squares

df here is the number of parameters fitted minus one (intercept and slope minus one = 1)

Table 5.3 Analysis of variance (ANOVA) table for simple linear regression of Y on X

Source of variation	SS	df	MS	Expected mean square
Regression	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{1}$	$\sigma_\varepsilon^2+\beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$
Residual	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-2$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	σ_ε^2
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

SS regression / SS total = r^2



← SS residual

← SS regression

SS residual + SS regression = SS total

- SS is a simple sum, that grows with the number of observations...a poor measure of sample variance!
- By dividing SS with the appropriate df's, we get **mean squares**, which are meaningful estimators of variance.

Mean squares

Variance components estimated by term

Table 5.3 Analysis of variance (ANOVA) table for simple linear regression of *Y* on *X*

Source of variation	SS	df	MS	Expected mean square
Regression	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$\dfrac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{1}$	$\sigma_\varepsilon^2+\beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$
Residual	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-2$	$\dfrac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	σ_ε^2
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

- If our model is "important", then this should affect the relative amount of variance in Y explained by our model as opposed to being error variance.
- This is captured by the variance ratio $MS_{regression}$ / $MS_{residual}$ which can be interpreted as "error variance and variance due to our model slope / error variance".
- This ratio of MSs is called the *F*-ratio, and is ≈ 1 if all true variance is error variance (i.e., here β = 0). The higher the *F* value, the more relative impact of the slope.
- The *F*-ratio follows a well defined probability distribution (the *F*-distribution) and is thus used for hypothesis testing

- The *F*-distribution built upon two df's: the **numerator** (effect/regression df) and the **denominator** (error/residual df) → e.g. "$F_{1,74} = 3.567$" means that the *F*-ratio is 3.567 and that the numerator degrees of freedom is 1 and the denominator df is 74.

Partial F-tests (and multiple partial F-tests) in regression

- Say we have fitted a model to data and know want to know whether the addition of **k** extra variables significantly increases the model fit to data over and above the **p** variables already in our model (H_0: all new β's equal zero).
- $Y = c + \beta_1 x_1 + ... + \beta_p x_p \quad + \quad \beta_{p+1} x_{p+1} + ... + \beta_{p+k} x_{p+k}$
- We compare the ANOVA tables for the **reduced model** (1 through p) and the **full model** (1 through p+k)**.**
- The logic of the F-ratio is (extra MS due to the addition of the new variables) / (MS residual from full model), and this is how it's done:

$$F = \frac{\text{SS residual [reduced]} - \text{SS residual [full]}) / k}{\text{MS residual [full]}}$$

df = k and n-p-k-1, where n is the number of observations (sample size)

- Very useful – called partial F-test when only *one* variable is added (k=1) and multiple partial F-test when *more than one* are added.

One-way ANOVA

- Categorical predictor variables are called **factors**, that has several **levels**
- In many cases, ANOVA used (1) to partition variance in Y into parts attributable to different categorical factors and/or (2) to test null hypothesis relating to means of Y for different factor levels.
- Effects of such factors are modelled slightly differently then regression type models. Single fixed factor linear effects ANOVA model:

  $$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

  where y_{ij} is a particular observation j from a given treatment level i, μ is the overall mean of the response variable (Y) and α_i is the effect of the treatment level i (actually, $\alpha_i = \mu_i - \mu$).
- Cf. similarity with regression type models; difference is that **means of treatment levels are modelled**. Generally OLS fitting.
- Make the same five important **assumptions** about data that regression does! For example, **one common error term (ε_{ij})** for all treatment levels....assumes the same variance for all levels (homoscedasticity) and normality of errors (check residual distribution).
- We can generate **predicted value and residuals** in much the same manner (predicted value is treatment level mean μ_i, and the residual is deviation from this μ_i for a given observation j).

- The ANOVA table is essentially the same as before:

ANOVA table

Source	SS	df	MS
Factor	$n\sum(\bar{y}_i - \bar{y})^2$	a-1	$\frac{n\sum(\bar{y}_i - \bar{y})^2}{a-1}$
Residual	$\sum\sum(y_{ij} - \bar{y}_i)^2$	a(n-1)	$\frac{\sum\sum(y_{ij} - \bar{y}_i)^2}{a(n-1)}$
Total	$\sum\sum(y_{ij} - \bar{y})^2$	an-1	

- Where **a** is the number of factor levels and **n** is the sample size per factor level (number of replicates per cell).

- If a **fixed** factor ANOVA and the homogeneity of variance assumption holds:

  MS Factor estimates $\sigma^2 + n\Sigma(\alpha_i)2/a\text{-}1$

  MS Residual estimates σ^2

  where σ^2 the sample variance (which is assumed to be the same for all ij treatment levels: $\sigma^2 = \sigma^2{}_i = \ldots = \sigma^2{}_j$)

- If a **random** factor ANOVA and the homogeneity of variance assumption holds:

  MS Factor estimates $\sigma^2 + n\sigma^2{}_a$

  MS Residual estimates σ^2

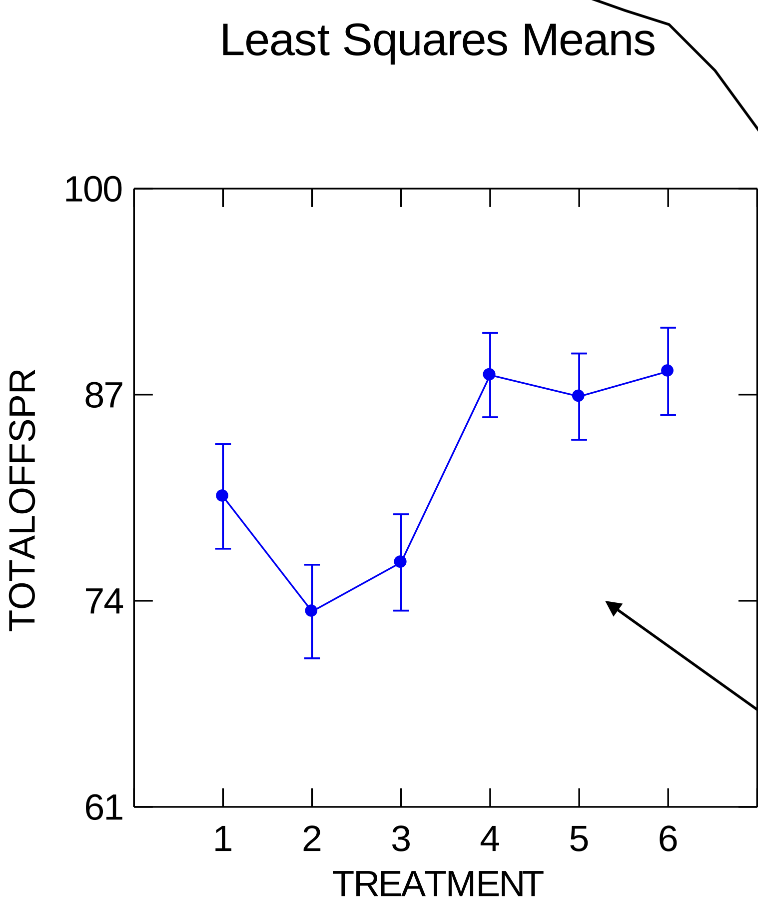
  where σ^2 the sample variance and $\sigma^2{}_a$ is the variance across factor level means.

An example:

```
Dep Var: TOTALOFFSPR N=353 Multiple R: 0.2631512  Squared multiple R: 0.0692485

-------------------------------------------------------------------------------
Analysis of Variance

Source                 Sum-of-Squares  df  Mean-Square      F-ratio      P

TREATMENT              1.26006E+04      5  2.52012E+03      5.1634073    0.0001388

Error/Residual         1.69361E+05    347  488.0729615

("Total" ususally not given....SS and df are additive)
-------------------------------------------------------------------------------
```

F (5,347) = 2520 / 488 = 5.16

Here, df is the number of means/parameters estimated (factor levels) minus one



Note that this ANOVA is still a **linear model** - because of the structure of the model, not the relationship between our factor and the response variable...

- The default H_0: $\mu_1 = \mu_2 = ... = \mu_i = \mu$ i.e.; all the factor level means are the same! (this is what the *F*-test in the ANOVA table refers to....)

- Note that this is equivalent to comparing (with a partial F-test):

  $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ with $y_{ij} = \mu + \varepsilon_{ij}$

- For one-way ANOVA's, random effects and fixed effects models yield the same null hypothesis test results (not true for more complex ANOVAs) **BUT** the interpretations differ! (cf. earlier...).
- How large is the effect of my factor – i.e. what is the effect size?

  For *random effects factors*; proportion of total variance explained by the factor

  For *fixed effects factors*; other measures such as Cohen's f or Omega squared.

Post-hoc tests: comparisons between means

- Given that an effect of our factor is significant (**and only then!**), we may be interested in comparing treatment level means against one another.
- **Controversial topic,** two related problems...

  1) Multiple inferences (increased type I error rate)

  2) Tests not independent, and thus difficult to interpret.... (e.g. given tests of 1 vs 2 and 2 vs 3, the test between 1 vs 3 not independent of the previous).

  → Compensating for multiple tests (problem 1) that are not independent (problem 2) is not easy...
- Two ways: **unplanned post-hoc comparisons** and **planned contrasts**

- In unplanned post-hoc comparisons, all treatment level means are tested against all others – many tests if several factor levels! Several methods available, most common are: Tukey's HSD, Fisher's LSD, Ryan's, Sheffe's and Dunnett's. Most often recomended is **Tukey's HSD** test. For example above:

```
--------------------------------------------------------------------------------
Tukey HSD Multiple Comparisons.

Matrix of pairwise comparison probabilities:
                         1            2            3            4            5
             1       1.0000000
             2       0.5714021    1.0000000
             3       0.9390055    0.9781975    1.0000000
             4       0.4603863    0.0024165    0.0398915    1.0000000
             5       0.6801506    0.0095322    0.1052297    0.9992566    1.0000000
             6       0.4424116    0.0024675    0.0388253    0.9999999    0.9985179

                         6
             6       1.0000000
--------------------------------------------------------------------------------
```

- For planned contrasts or comparisons, you are specifically interested in comparing one/several treatment levels (i.e., groups) with one/several others. Any comparison can be made, by **constructing contrasts** that *sums to zero* across all levels/groups. For example, for the 6 levels/groups in the above example:

  Do level 2 differ from all the others (2 vs 1,3-6)?

  The contrast matrix:

1	2	3	4	5	6 (levels/groups)
-1	5	-1	-1	-1	-1 (constrast)

```
 -----------------------------------------------------------------------------
 Test of Hypothesis
   Source          SS        df       MS             F              P
 Hypothesis  5.46603E+03     1  5.46603E+03     11.1992161      0.0009079
 Error       1.69361E+05   347  488.0729615
-----------------------------------------------------------------------------
```

Do levels 1-3 differ from 4-6?
The contrast matrix:

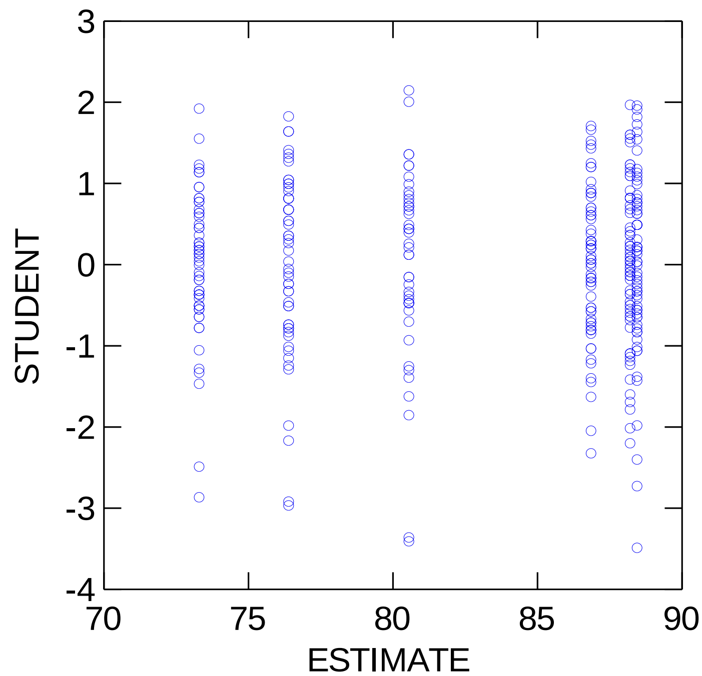
1	2	3	4	5	6 (levels/groups)
-1	-1	-1	1	1	1 (constrast)

```
-------------------------------------------------------------------------------
Test of Hypothesis
  Source         SS        df        MS             F              P
Hypothesis  1.06180E+04     1  1.06180E+04     21.7548827      0.0000044
Error       1.69361E+05   347  488.0729615
-------------------------------------------------------------------------------
```

- In general, try to avoid post-hoc comparisons all together...interpret overall effect of treatment and graphs!
- If you have good reasons to do comparisons, do try to **restrict the number** by performing planned post-hoc contrasts.

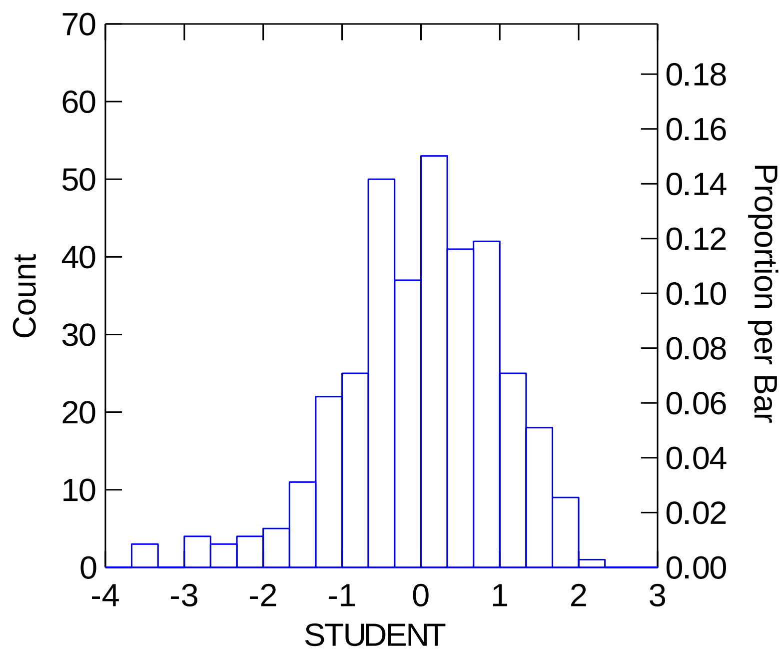
Model diagnostics

- As detailed before, we make a number of assumptions when running ANOVA. Violations can greatly affect our estimates, and should thus be checked.
- The first thing you should do is to do diagnostic plots!
  * Plot residuals against predicted values (i.e. level means) – should show equal spread across levels/groups and no ”trend” (look out for wedge shaped or funnel shaped appearances!)



* Plot residuals against factor level - should show equal spread across levels/groups.

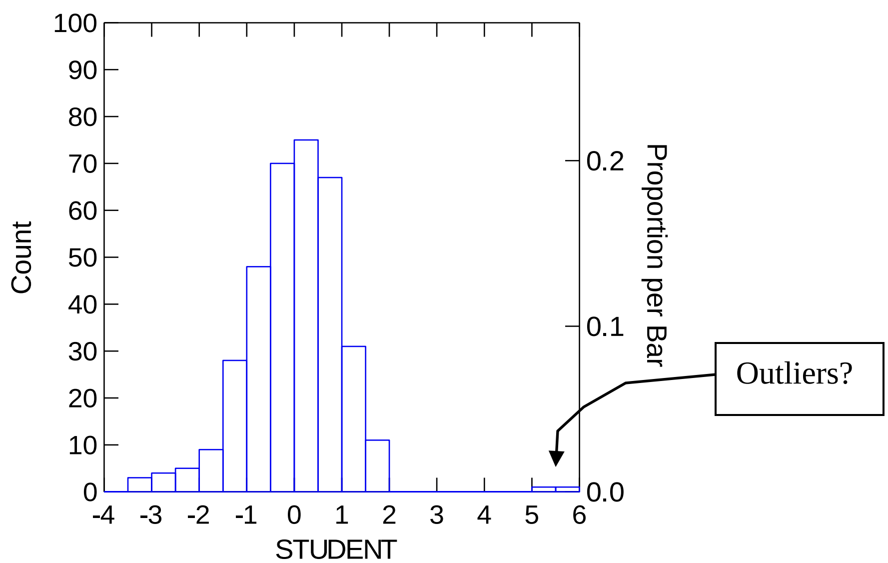
* Frequency plot of residuals – symmetric and normal?



- Then test for a few things!

1. **Homogeneity of variance**: assess residuals graphically and test for equality of variance by a test for homogeneity of variance (Levene's test for homogeneity of variance preferred by many). *Interpret with caution*: sometimes rejects H_0 of equal variances even variance are very similar (if sample size is large) – ANOVA robust at least to *minor* violations of this assumption...
2. **Normality of error**: assess residuals graphically and test whether residuals are normally distributed (Shapiro-Wilk's or Kolmogorov-Smirnov test). Again, *interpret with caution*: tend to rejects H_0 of normality for large samples – again, is ANOVA robust at least to *minor* violations of this assumption...

- **Outliers** are observations that are deviant in the response variable dimension – have ***large residuals***. As a rule of thumb, observations with a |Studentized (i.e., standardized) residual| larger than 3-4 are potential outliers.

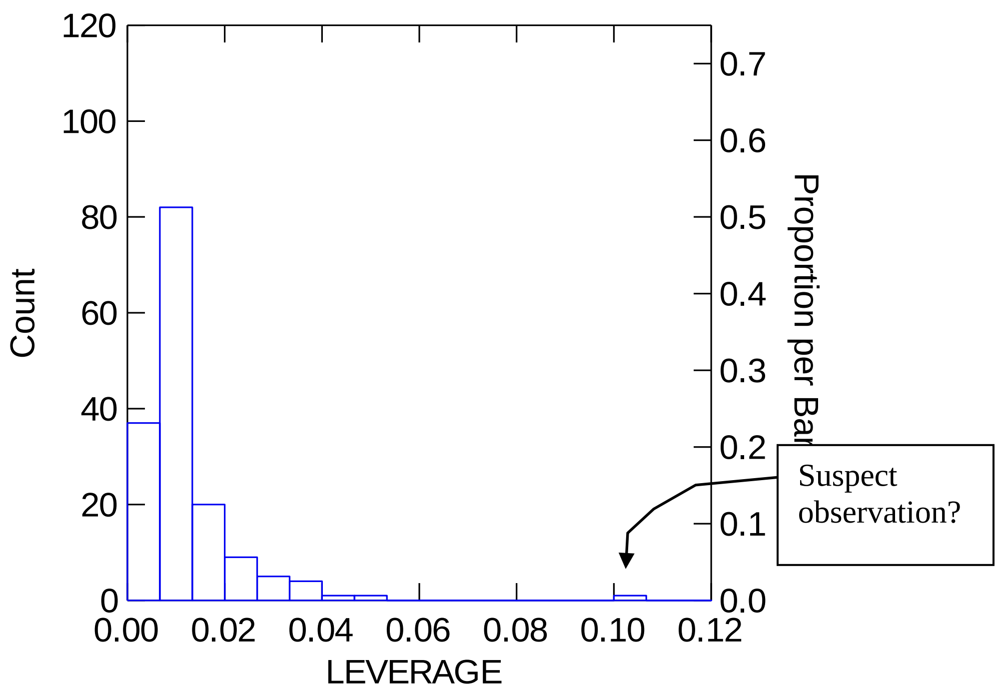


- If outliers are present

  1) Check these observations carefully! Typo when entering data? Something unusual or wrong with this observation? Correct or exclude.

  2) If ”correct”, run model with and without outliers. If the results are the same, your conclusions are not affected...

- **Leverage** is a measure of ”outlier” in the X-dimension. Not interpretable for categorical factors but important for continuous independent variables. Observations with unusually large leverage *affect the model parameter estimates* greatly, and should be checked carefully. Do frequency plot of leverage values – should have no obvious ”stragglers”:



- Unequal sample sizes is a **real problem** if data is heavily unbalanced. Model may yield biassed results and assumptions are hard to check. **Avoid unbalanced data**! If you do have heavily unbalanced data, either use a means model ANOVA or test your model with a resampling (e.g. permutation) test using type III sums-of-squares.

Transformations of data

- Should only be done when needed...***not*** routinely!!!!!
- If our data do not fulfill the assumptions made when fitting a model, we can often improve the situation by **transforming** our response/dependent variable prior to fitting the model.
- Note that some violations of assumptions **cannot** be transformed ”away”: that Y is a continuious variable, that data is random and that observations are independent.
- Can often improve normality of residuals, homogeneity of variances and can reduce the impact of outliers.
- For residual distributions skewed to the right, use either of the following:

  $Y' = log\ (a + Y)$; where a is a constant ≥ 1 (log transformation)

  $Y' = Y^{0.5}$ (square root transformation)

  $Y' = Y^{0.333}$ (cube root transformation)

  $Y' = Y^{0.25}$ (fourth root transformation) or (but keep track of direction – order becomes reversed):

$Y' = 1/Y$ (reciprocal tranformation)

$Y' = 1/Y^{0.5}$ (reciprocal of square root tranformation)

- For residual distributions skewed to the left, use:

  $Y' = Y^{a}$; where a is a constant ≥ 1 (power transformation)

  Alternatively (but keep track of direction – order becomes reversed):

  First, reflect the distribution by $Y_r = a - Y$; where a is a constant larger than the largest value of Y. Second, apply one of the transformations for distributions skewed to the right to Y_r...
- The **Box-Cox transformation** is an iterative procedure for finding an appropriate transformation – sometimes useful.
- When response data is a proportion, people often use:

  $Y' = \arcsin\sqrt{Y}$ (arcsine transformation)

  this stretches both tails and compresses the middle of a distribution bounded between 0 and 1. However, be aware of the fact that proportions are not really continuous variables...and may have other problems...(generalized linear models are preferred).