1 Linear models

$$Y = X\beta + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma_e^2) \qquad \text{or} \qquad Y \sim \mathcal{N}(X\beta, \sigma_e^2)$$

where the model matrix X contains the information on predictors and β includes all coefficients (intercept, slope(s) etc.).

1. **Number of coefficients and degrees of freedom**

   Main effect of a numerical predictor: 1 (slope)
   Main effect of a factor with a categories: $a-1$ (adjustments to intercept)
   Interaction effect of two numerical predictors: 1
   Interaction effect of a numerical predictor and a factor with a categories: $a-1$ (adjustments to slope)
   Interaction effect of two factors, a and b categories each: $(a-1)*(b-1)$ (adjustment to group means)

2. **Assumptions**

   - homogeneous variance of residuals,
   - normal distribution of residuals,
   - correct average relationship $\mathbb{E}(Y) = X\beta$. This may mean a straight linear relationship with a numerical predictor, and/or absence of interaction when interaction effects are not included, etc.

3. **Transformations** of numerical variables: used to meet the assumptions of the linear model or to decrease the influence of points that have high leverage.

   Interpretation when the response is log-transformed:

   - numerical predictor with slope β_1: an increase of 1 unit in the predictor corresponds to an increase of about $(\beta_1 * 100)$ % in the average response (when β_1 is small).
   - factor, with difference β_1 between two categories: switching from one category to the other corresponds to an increase of $(\beta_1 * 100)$ % in the average response (if β_1 has a small value).

   Interpretation when a numerical predictor is log-transformed:

   - non-transformed response: a 1% increase in the predictor corresponds to an increase of the log-predictor by 0.01, and to an increase of $\beta_1 * 0.01$ in the average response.
   - log-transformed response: a 1% increase in the predictor corresponds to an increase by about β_1 % in the average response.

4. **Sums of Squares**

   SSTrt = RSS(model without Trt) − RSS(model with Trt)

   - Type I SS: formula above with a model that progressively includes all predictors, added one by one.
     Type III SS: formula above with a model that uses all predictors, with/without Trt.
   - In traditional designs with complete balance and all categorical predictors, type I SS = type III SS. Otherwise, type III SS and associated ANOVA f-tests are preferred.

5. **ANOVA f-test**

   - To compare nested models, based on the reduction in RSS from the simple model to the complex model.
   - Ex: the complex model has the same list of predictors as the simple model, plus an extra predictor: 'Trt'. H_0: the simple model is correct, i.e. **all** coefficients for Trt are $= 0$.

     Compare the f value with an F-distribution, degrees of freedom for Trt (numerator) and residual df (denominator).
   - Exact test, provided model assumptions are met.

6. **Likelihood Ratio Test (LRT)**

   - To compare nested models, based on the increase in likelihood from the simple model to the complex model.
   - Ex: the complex model has the same list of predictors as the simple model, plus an extra predictor: 'Trt'. H_0: the simple model is correct, i.e. **all** coefficients for Trt are $= 0$.

     Compare the LR statistic $X^2 = 2 * \log L(\text{complex}) - 2 * \log L(\text{simple})$ to a χ^2_k-distribution, with degrees of freedom $k =$ number of coefficients for Trt.
   - Approximate test. Good approximation when the sample size is large.

7. **Testing individual coefficients** for post-hoc comparisons

   Wald t-tests: based on SE for each coefficient. H_0: the coefficient $= 0$.

   - If a predictor uses a single coefficient, the Wald t-test for that coefficient and the f-test for the predictor provide the same p-value.
   - If a predictor is associated with several coefficients, beware of the meaning of these coefficients (differences to which reference level? or differences to an overall mean?)
   - Also beware of multiple testing. F-test preferred for an overall test of the predictor.

8. **Model Selection**

   - Beware of multiple testing: avoid starting with over-complicated models
   - Search methods: stepwise backward, forward, or in both direction
   - Search criteria: AIC (includes more predictors), BIC (includes fewer), f-test or LRT (both provide p-values)

2 Logistic Regression

Binary response (yes/no), binomial distribution assumed:

$$Y_i \sim \mathcal{B}(p_i, n_i) \text{ with } p_i = \frac{e_i^\eta}{1+e_i^\eta} \text{ and } \eta_i = \log(\frac{p_i}{1-p_i}) = X\beta$$

1. **Interpretation**

   - The linear predictor η = log of the odds of 'success'. $\exp(\eta_i)$ = odds of 'success'.
   - Coefficient β_1 of a numerical predictor: at the 50:50 point, an increase of 1 unit in the predictor results in an increase of about $\beta_1/4$ in the proportion of success (approximation okay provided $\beta_1/4$ is not too large).
   - Coefficient β_1 for the difference between 2 categories of a treatment factor: when switching from one category to the other, the odds of 'success' is muliplied by $\exp(\beta_1)$, which corresponds to an increase of about $(\beta_1 * 100)\%$ if β_1 is small.

2. **Likelihood ratio test**

   There is no F-test in logistic regression.

   Only the LR test is available. H_0: the simple model is correct, i.e. **all** extra coefficients in the complex model are = 0. Compare X^2 = Deviance(simple) − Deviance(complex) to a χ^2_k distribution with df k = number of extra parameters in the complex model.

   It is approximate.

3. **Wald test for post-hoc tests of individual coefficients**

   H_0: the coefficient = 0.

   The Wald test is a 'z'-test: the ratio of a coefficient to its SE is compared to a normal distribution.

   It is *very* approximate.

   For a predictor that corresponds to a single coefficient, the Wald test for this coefficient does *not* always agree with the LR test for the predictor. The LR test is to be preferred.

   Check the meaning of each coefficient. Changing the reference level of a categorical predictor can drastically alter the p-values from the Wald tests, because the new coefficients take new meanings.

4. **Model selection**

   Available criteria: LR test, AIC, BIC. Only the LR test provides p-values.

5. **Goodness-of-fit test**

   H_0: the current model is good enough to explain the data.

   - The alternative hypothesis includes any reason that may cause the model to not be good enough, such as: an important predictor is missing, lack of independence within groups causing over-dispersion, presence of one outlier that is not representative of the population, etc.
   - Compare the residual deviance to a χ^2_{dfRes} distribution on df = residual df. Approximate.

- Beware: the p-value obtained that way is a very poor approximation when all $n_i = 1$.
- Alternative: compare the residual deviance to that obtained by parametric bootstrapping (simulate experiments under H_0).

6. **Residuals**

   Response, deviance, or Pearson's residuals.
   Deviance residuals are preferred for residual plots.
   When all $n_i = 1$, there is an expected pattern in residual plots.
   Pearson residuals are preferred for estimating the dispersion.

7. **Dispersion**

   σ^2 under the hypothesis that the variance of observations has the form $\text{var}(Y_i) = \sigma^2 n_i p_i(1-p_i)$.

   The Binomial distribution implies $\sigma^2 = 1$.

   Estimate $\hat{\sigma}^2$ from Pearson residuals. Deviance/dfRes is the estimate from deviance residuals.

   Over-dispersion can be invoked when $\hat{\sigma}^2$ is significantly > 1 and when everything else fails.

8. **Quasi-binomial analysis**

   Based on logistic regression with the extra σ^2 parameter (but does not correspond to a model with a well defined distribution).

   An ad-hoc f-test replaces the LR test. Ad-hoc Wald t-tests replace the Wald z-tests.

3 Random and mixed effect models

Normal distribution of residuals, and some categorical predictors are given random effects:

$$Y = X_{\text{fixed}}\beta + X_{\text{random}}\alpha + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma_e^2) \text{ and } \alpha\text{'s are normally distributed}$$

1. **Parameters**

   Coefficients for fixed effects (intercept and possibly other coefficients), the variance of residuals σ_e^2 and one or more variance component(s).

2. **ML, REML and MS**

   - ML: maximum likelihood method to estimate parameters. Tends to underestimate variance components.
   - REML: restricted maximum likelihood. Less biased or unbiased variance components. Can only be compared across models that share the same fixed effects.
   - MS: mean squares approach. Closed formula for estimating all parameters, but only available in balanced cases. Can result in negative estimated variance components. Same as REML in many cases.

3. **LR test for random effects**

   To compare nested models, when the complex model has extra random effects.
   H_0: the simple model is correct.

   Based on the LR statistic $X^2 = 2\log L(\text{complex}) - 2\log L(\text{simple})$. The REML criterion is preferred, but the ML criterion can be used as well. Use the same criterion for both models.

   - χ^2-based p-value: compare X^2 to a χ^2_k distribution with df = number of extra variance parameters in the complex model.
   - the p-value above tends to be conservative (too large). Dividing it by 2 provides a more accurate p-value in some cases when a single variance parameter is tested.
   - Alternatively, get the exact p-value: compare X^2 to the true null distribution as obtained by parametric bootstrapping (simulate new data under H_0).

4. **LR test for fixed effects**

   To compare nested models, when the complex model has extra fixed effects.
   H_0: the simple model is correct.

   Based on the LR statistic $X^2 = 2\log L(\text{complex}) - 2\log L(\text{simple})$. The ML criterion must be used.

   - χ^2-based p-value: compare X^2 to a χ^2_k distribution with df = number of extra coefficients in the complex model.
   - the p-value above tends to be anti-conservative (too small). Sometimes *very* much so. Other times it is accurate when the data set is large compared to the number of coefficients.
   - Alternatively, get the exact p-value: compare X^2 to the true null distribution as obtained by parametric bootstrapping (simulate new data under H_0).

5. **F test for fixed effects**

   To compare nested models, when the complex model has extra fixed effects.
   H_0: the simple model is correct.

   - F = ratio of Mean Square (MS) values, and based on closed formulas for MS with balanced designs. Based on linear algebra in general.
   - Approximate test: ignores the fact that variance components are not known perfectly.
   - Compare the f-value with an F-distribution. Numerator df = number of extra coefficients in the complex model.
   - Denominator df: can be calculated exactly in balanced designs only. Beware of random effects that are nested in other effects. They may cause the denominator df to differ from the 'residual' df.
   - Can be more powerful than the LR test.

4 Topics for all models

1. **Two-way and higher order interactions**

   - Two-way interaction: when the effect of one predictor depends on the level of the other predictor. Lack of parallelism in profile plot of group means.
   - Three-way interaction: when the two-way interaction effect of two predictors depends on the level of a third predictor.

   Two predictors can be independent and also have an interaction effect on the response.

2. **Correlation among predictors**

   Frequent scenarios:

   - Both predictors are "significant" when included alone in the model, both are non-significant when included together in the model.

     → Their correlation prevents us from teasing out which predictor is really associated with the response. Either predictor can be chosen for inclusion in the model; the other predictor should be dropped. The conclusion should state the uncertainty about which predictor is best.
   - One predictor is significant whether the other predictor is included in the model or not. The significance of the other predictor depends on the inclusion of the first predictor in the model.

     → The first predictor should be included in the model. It can be a confounding variable for the second predictor (and would be lurking if ignored).

3. **Causality versus prediction**

   In designed experiments, *randomization* aims to eliminate any known or unknown confounding variables, by forming treatment groups that are equivalent about everything expect for the treatment(s). Causality can be inferred from designed experiments.

   Otherwise, only prediction can be made, and more assumptions are needed to extend prediction to causality statements.

5 Particular experimental designs

1. **Subsampling**

   Happens when measurements are repeated on the the same experimental unit. Subsampling can be added on top on any design.

   Ignoring the dependence that comes from subsampling is called pseudo-replication. Usually results in false significance.

   Subsampling is accounted for by adding a random effect for the experimental unit from which subsamples are taken.

   Experimental units are nested within treatments (and maybe within blocks as well) because each unit received only one treatment (or belongs to only one block). Therefore...

Beware of **not** using the Subsampling MSerror for F-tests, but instead use the MSError and dfError at the experimental unit level.

2. **CRD: Completely Randomized Design**

   In this design, all levels of the treatment or all combinations of treatment levels are assigned to experimental units completely at random.

   Randomization is most complete, and the associated model is most simple: the only predictors are the treatment(s).

   For a treatment with a categories and n units (plots) per treatment and possibly s subsamples per unit (plot):

Source	df	F-value
Treatment A	$a-1$	MStrt/MSErr
Error	$a(n-1)$	
Total	$an-1$	

or

Source	df	F-value
Treatment A	$a-1$	MStrt/MSPlotErr
Plot Error	$a(n-1)$	
Subsamp. Err.	$an(s-1)$	
Total	$ans-1$	

3. **RCBD: Randomized Complete Block Design**

   Blocks are defined based on biological grounds, so that variability within blocks is expected to be lower than variability between measurements in different blocks.

   $$Y = \text{Treatment} + \text{Block} + \text{error}$$

   No interaction in the usual case when there is no replication within blocks.

   For a treatment with a categories, b blocks, 1 unit (plot) per combination treatment/block and possibly s subsamples per unit (plot):

Source	df	F-value
Trt	$a-1$	MSTrt/MSErr
Block	$b-1$	MSBlock/MSError
Error	$(a-1)(b-1)$	
Total	$ab-1$	

or

Source	df	F-value
Treatment	$a-1$	MSTrt/MSPlotErr
Block	$b-1$	MSBlock/MSPlotErr
Plot Error	$(a-1)(b-1)$	
Subsamp. Err.	$ab(s-1)$	
Total	$abs-1$	

4. **Latin Square**

   Design with double-blocking: $Y = \text{Treatment} + \text{Row} + \text{Column} + \text{error}$

Source	df	F-value
Treatment	$a-1$	MSTrt/MSPlotErr
Row	$a-1$	MSRow/MSPlotErr
Column	$a-1$	MSCol/MSPlotErr
Error	$(a-1)(a-2)$	
Total	a^2-1	

   Multiple Latin squares could be used, in which case a square effect is added. Usually row and column effects are considered nested within squares.

5. **Factorial designs**

For two or more factors. Full factorial: when all combinations of factor levels are tested. When several replicates are included for each combination, the model includes the interaction between the two factors.

$$Y = \text{factor } A + \text{factor } B + A{:}B + \text{error}$$

If A has a levels, B has b levels and n units per combination of treatment levels:

Source	df	F-value
A	$a-1$	MSA/MSErr
B	$b-1$	MSB/MSErr
A:B	$(a-1)(b-1)$	MSA:B/MSErr
Error	$ab(n-1)$	
Total	$abn-1$	

This design could either use complete randomization (CRD, table above) or blocking (RCBD, modify the model and table with a blocking factor).

6. **Split-plot designs**

Happens when one factor (C) is randomized at a finer level than another factor (A). "Whole-plots" are assigned to levels of A and subplots within whole-plots are assigned to levels of B.

$$Y = A + C + A{:}C + \text{ whole-plot } + \text{ error}$$

Whole-plots must be given random effects, because they are nested in A (each whole-plot receives only 1 level of A). Therefore...

Beware of using MS Whole-plot and df Whole-plot when testing the effect of A with the f-test.

If A has a levels, C has c levels, b whole-plots per level of A, and 1 subplot per level of C within each whole-plot: per combination of treatment levels:

Source	df	F-value
A	$a-1$	MSA/MSWPErr
Whole-plot (WP) Err.	$a(b-1)$	
C	$c-1$	MSC/MSSPErr
A:C	$(a-1)(c-1)$	MSA:C/MSSPErr
Subplot (SP) Error	$a(b-1)(c-1)$	
Total	$abc-1$	

Can be combined with blocking whole-plots into blocks, in which case the model needs to account for blocks as well. With b blocks and 1 whole-plot per block per level of A:

Source	df	F-value
A	$a-1$	MSA/MSWPErr
Block	$b-1$	MSBlk/MSWPErr
Whole-plot (WP) Err.	$(a-1)(b-1)$	
C	$c-1$	MSC/MSSPErr
A:C	$(a-1)(c-1)$	MSA:C/MSSPErr
Subplot (SP) Error	$a(b-1)(c-1)$	
Total	$abc-1$	