Theory Defence

1 Linear regression

In the thesis we use linear regression models in order to model the CLR transformed counts for the different cell types. This is a summary of what was seen in the linear modeling course by Els Goetghebeur.

1.1 Simple linear regression

1.1.1 Notation

- Capitalized letters (e.g. Y) are random variables. They have a mean and variance.
- Small letters (e.g. x, y) are fixed constants or observed values.
- Indices are donoted by i or j.

1.1.2 Meaning and interpretation

Regression only shows association between the dependent and independent variables. It is not because two things tend to go in the same direction, that a change in one is *causing* the change in the other. It is possible that they are both causally responding to a third factor that is changing. Causal relationships need additional assumptions and/or data to be interpreted.

1.1.3 Regression model formally

A regression model describes a statistical relationship between a random variable Y and a fixed x or random X. It describes how the distribution of the random variable Y varies with given values of x, and the association between Y and x.

The simple regression model captures this by a parametric form for the distribution Y for every level of x (e.g. the normal distribution with μ_x and σ_x^2), and a functional form for the means of these distributions μ_x, which may depend on unknown parameters (e.g. $\mu_x = \beta_0 + \beta_1 x$).

1.1.4 The simple linear regression model

The Simple Linear Regression model takes the following form (Formula 1).

$$Y_i = \beta_0 + \beta_1 x_1 + \epsilon_i \; with \; \epsilon_i \text{ iid } N(0, \sigma^2) \tag{1}$$

The model is characterized by the following elements:

- $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$
- $Var(Y_i|x_i) = \sigma^2$
- $E(\epsilon_i) = 0$
- $Var(\epsilon_i)$ is constant
- The observations are independent (no correlation between any Y_i, Y_j)
- The data is normally distributed

1.1.5 Interpretation of the coefficients

β_0 is the intercept. It shows the intersection with the y-axis and often does not have an interpretation. It does have an interpretation when x is centered or standardized, as it then signifies the Y for the average x. β_1 is the slope of the regression line. Per unit increase in x, the Y is expected to increase with an amount equal to β_1.

1.1.6 Estimation of the regression function

The regression coefficients are determined using *ordinary least squares*. This methods finds the optimal values of β_0 and β_1 that minimize the q-function (Formula 2). This q-function is the sum of the squared residuals.

$$q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{2}$$

One can try to solve this minimization problem numerically, however it is much feasible in an analytical way. For this, the normal equations are used (Formula 3), which result in point estimates for β_0 and β_1. These normal equations are acquired by taking the derivative of the q-value towards β_0 and β_1, respectively.

$$\begin{gathered}
\begin{cases} \sum y_i = n\beta_0 + \beta_1 \sum x_i \\ \sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \end{cases} \\
\begin{cases} \beta_0 = \bar{y} - \beta_1 \bar{x} \\ \beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)} \end{cases}
\end{gathered} \tag{3}$$

1.1.7 Estimation of the variance

We know that the expected value of y_i equals μ and the variance equals σ^2. In an ideal world, this is an unbiased estimator, however, we do not know the value of μ in real life. For this reason we can calculate the variance by taking the squared difference over the real and predicted value, and then dividing by the number of samples. This estimator is biased though, as we do need to account for the unknown μ. By dividing by the number of samples minus 2 (as we use two times y in the formula), we get an unbiased estimator for the variance, which we call the MSE (Formula 4).

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} E_i^2}{n-2} \tag{4}$$

1.1.8 Parameter estimation using MLE

Maximum Likelihood Estimation can also be used in order to fit the parameters of the linear regression model. This works by maximizing the probability density function of Y given a model (Formula 5), in this under the assumption that Y is normally distributed. From the density distribution, we see that we have to estimate three parameters: β_0, β_1 and σ^2. Although, to do this, we need to calculate the loglikelihood, as maximizing a sum is far easier than maximizing a product. By taking the partial derivatives towards the parameters of interest, we can derive an estimator for them. As can be seen, for β_0 and β_1, we get the same normal equations (Formula 3) as seen previously.

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[y-(\beta_0+\beta_1 x)]^2}{2\sigma^2}\right),$$
$$\mathcal{L}(\beta_0,\beta_1,\sigma^2) = \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[\frac{Y_i-\beta_0-\beta_1 x_i}{\sigma}\right]^2\right)\right\}, \tag{5}$$

$$\text{where} \quad \begin{cases} \dfrac{d\log(\mathcal{L})}{d\beta_0} = \dfrac{1}{\sigma^2}\sum(Y_i-\beta_0-\beta_1 x_i) = 0, \\ \dfrac{d\log(\mathcal{L})}{d\beta_1} = \dfrac{1}{\sigma^2}\sum x_i(Y_i-\beta_0-\beta_1 x_i) = 0, \\ \dfrac{d\log(\mathcal{L})}{d\sigma^2} = -\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\sum(Y_i-\beta_0-\beta_1 x_i)^2 = 0 \end{cases} \tag{6}$$
$$\text{with results:} \quad \hat{\beta}_0 = B_0, \quad \hat{\beta}_1 = B_1, \quad \hat{\sigma}^2 = \frac{\sum(Y_i-\hat{Y}_i)^2}{n}.$$

1.2 Inference on the coefficients

1.2.1 Confidence intervals

The confidence interval is the range in which the true value of β_1 is in, with 95% certainty.

$$CI = \left[\beta_1 \pm t(\frac{1-\alpha}{2}; n-2)S(\beta_1)\right] \tag{7}$$

1.2.2 Two-sided t-test

In order to test whether the estimate for β_1 is different from zero. For this we set up the following hypotheses:

$$H_0 : \beta_1 = 0 H_1 \qquad : \beta_1 \neq 0 \tag{8}$$

In order to assess this, we calculate the t-statistic using the following formula: $t = \frac{\beta_1 - 0}{S(\beta_1)}$. A p-value can be found by computing the area under the probability density function with $n-2$ degrees of freedom until $\alpha/2$ and subtracting it from 1.

1.2.3 Prediction interval

The prediction interval is the interval in which 95% of the predictions will end up being. This is calculated similarly as the confidence interval.

$$PI = \left[\hat{Y_h} \pm t(\frac{1-\alpha}{2}; n-2) S(\hat{Y_h})\right] \tag{9}$$

1.3 Multiple Linear Regression

1.3.1 The multiple linear regression model

When we add more variables to the model, we get a multiple linear regression model. This can be written as seen in Formula 10.

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon_i \quad = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ik} + \epsilon_i \tag{10}$$

The β coefficients can be interpreted as follows (e.g. β_1): β_1 is the change in mean response when x_1 is increased by a unit and all other variables remain constant.

Often there can exist interaction between variables, which can be added to the model by incorporating an interaction term $\beta_{12} x_1 x_2$. This can be interpreted as the additional effect on y, when x_1 is held constant and x_2 increases with one unit (and vice versa). For categorical variables, this is the effect of a certain combination of categories on y.

1.3.2 Estimation of the regression function

In order to fit the regression line, we again minimize the q-function (Formula 11)

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i,p-1})^2 \tag{11}$$

To do this, we solve the least squares normal equations, that take the form of $X'XB = X'Y$, with estimators $B = (X'X)^{-1}(X'Y)$

$$\begin{cases} \dfrac{dQ}{d\beta_0} = \sum_{i=1}^{n}(y_i + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}) = 0 \\ \dfrac{dQ}{d\beta_1} = \sum_{i=1}^{n}(y_i + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}) x_{i1} = 0 \\ \cdots \\ \dfrac{dQ}{d\beta_1} = \sum_{i=1}^{n}(y_i + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}) x_{i,p-1} = 0 \end{cases} \tag{12}$$

1.4 Weighted least squares

In weighted least squares, the idea is that weights can be used to deal with heteroscedastic data. Weights are user defined (e.g. the inverse of the variance of the variables). The weighted least squares estimator is defined as in Formula 13. Deriving which weights to use is often the hardest part.

$$\hat{\beta}_{WLS} = \arg\min_{\beta} \sum_{i=1}^{n} \epsilon_i^{*2} \quad = (X^T W X)^{-1} X^T W Y \tag{13}$$

2 Compositional data

The major issue we face in the thesis, is that the data is compositional in nature. The observed cell counts from the sequencing experiment only capture relative information as the number of cells observed for a particular sample does not reflect the total number of cells in the sample, but is constrained to an arbitrary sum. Due to this, we need to use techniques from the field of compositional data analysis to perform corret inference on the unobserved absolute cell counts using the observed data that only contains relative information.

Compositional data exists in a simplex space with 1 fewer dimensions than components. Analysis of relative data as if they were absolute data often leads to erroneous results due the following reasons. First, most statistical models assuming independence between features. However, due to the inverse correlation between the features vialotes this assumptions, as the features are mutually dependent. Next, distances between samples are misleading and sensitive to arbitrary inclusion or exclusion of components. Lastly, components can appear correlated, even when they are statistically independ.

Quinn et al. propose to call each sample in the data a composition and each feature (in our case cell type) a component. They further propose three methods of dealing with the analysis of compositional data. First, they discuss the 'normalization-dependent' approach. In this approach, normalization is used in order to reclaim absolute abundances. However, most methods have assumptions that are not valid outside of thightly controlled experiments. The second approach is 'transformation-dependent'. The idea is to transform the data with regard to a reference to make statistical inference relative to that reference. Lastly, the 'transformation-independ' approach performs calculations directly on the components or component ratios.

In the thesis, we use a transfomration-dependent method: the centered-log-ratio (CLR) transformation (Formula 14). The reference to which the data is transformed, is the geometric mean of the sample vector.

$$\begin{aligned} clr(x_j) &= \left[ln\frac{x_{1,j}}{g(x_j)}, \dots, ln\frac{x_{D,j}}{g(x_j)} \right] \\ g(x_j) &= \left(\prod_{j=0}^{D} x_j \right)^{1/D} \end{aligned} \tag{14}$$

Other modeling approaches that we do not get into with the thesis are the following. People have tried count models for each population, often modeling them using NB models. The total number of observed cells is added as an offset, however this model does not account for compositionlity. Next, compositional count models have been used, most commonly the Dirichlet-multinomial model. This accounts for compositionality as the multinomial parameters must sum to

1. Dirichlet allows Multinomial parameters to change by sample, and it is a Bayesian model.

3 VoomCLR

3.1 Limma-Voom

Limma-Voom is a method applied in order to perform DE analysis on RNA-seq data. It estimates the mean-variance relationship of the log-counts, generates a precision weight for each observation and enters these into the limma empirical Bayes analysis pipeline.

The read-counts are first log-cpm transformed (Formula 15) as a means of normalizing the data. However, the probability distribution for counts are naturally heteroscedastic, with larger variances for larger counts. *Law et al.* conclude that the log-cpm values generally show a smoothly decreasing mean-variance trend with count size, and that the log-cpm transformation roughly de-trends the variance as a function of count size for genes with higher counts.

$$log_2 CPM = log2\left(\frac{count * 10^6}{\sum_{j=1}^{N} count_j}\right) \tag{15}$$

To analyse the RNA-seq data, the log-cpm values are inputted into the limma software. However, due to the mean-variance trend in lower counts, there should be a correction applied. First, gene-wise linear models are fitted using the normalized log-cpm values, taking into account experimental design, treatment conditions, replicates and so on (Formula 16). The residual standard deviations for each gene is plotted, and modeled on an observation-level in a non-parametric way using a lowess fit as a function of the average log-count (Figure 1). Using the fitted log-CPM value for each observation, a predicted count is generated and used to interpolate a predicted standard deviation for that observation. In the end, the inverse squared predicted standard deviation for each observation, becomes the weight for that observation in that gene model that are now fitted using weighted least squares.

$$logCPM(gene_j) = \beta_0 + \beta_1 x_1 + \cdots + \epsilon \tag{16}$$

This procedure accounts for the systematic relationship between the mean ex-

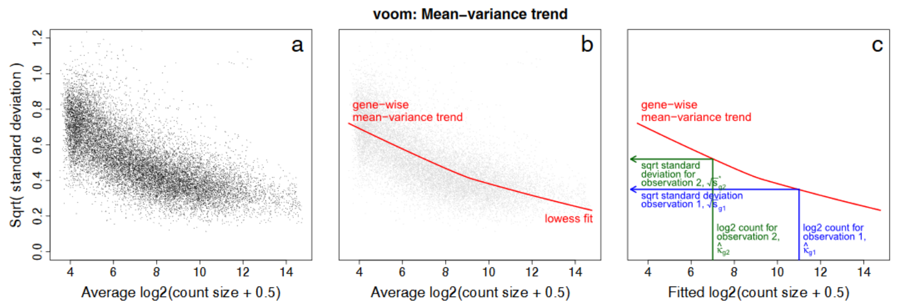

Figure 2 voom mean-variance modeling. (a) Gene-wise square-root residual standard deviations are plotted against average log-count. **(b)** A functional relation between gene-wise means and variances is given by a robust LOWESS fit to the points. **(c)** The mean-variance trend enables each observation to map to a square-root standard deviation value using its fitted value for log-count. LOWESS, locally weighted regression.

Figure 1:

pression level and the residual variance. However, this does only provide precision-weights that depend on the mean-variance trend and do not stabilize gene-wise variance estimates. The counts are still heteroscedastic. Some genes may still have highly noisy or extreme variances due to the limited sample sizes, high biological variability or other random effects. To solve this, the emperical Bayes method comes in. It targets the residual variance from each gene-wise models by borrowing information across all genes to estimate the pooled prior variance, and shrinks individual gene-wise variances towards this pooled prior. T- and F-tests can than be performed using this shrunk variance, for which we call them *moderated* tests.

The main take away is that voom corrects for the systematic mean-variance trend across genes and samples, making model fitting more accurate, while eBayes stabalizes the individual gene-wise variance estimates by borrowing strength across genes, which is critical for robust hypothesis testing.

3.2 VoomCLR

Next, we show how VoomCLR uses the Limma-Voom framework in order to account for compositional bias and the bias on the effect sizes. The first thing VoomCLR does, is CLR transforming the cell counts (Formula 14). This is performed to deal with the composition bias. However, compositional transformations do not stabilize the variance (it is a function of the mean). For this reason, the transformed counts are input into the Limma framework to first account for the mean-variance structure using heteroscedasticity weights. Next, the authors of the LinDA paper saw that the CLR transformation in combination with linear models result in biased effect sizes in respect to the effect sizes one would obtain based on the absolute abundances. In order to correct for this, the mode of the effect size across cell types is used.

Next, we will explain why this is used as a correction. If we consider a log-linear model on the absolute abundance X_{ip}, we get 17. By assuming that the CLR transformed relative abundance Y_{ip} can be written as the CLR transformed absolute abundance plus an estimation error e_{ip}, we can say that the model looks as follows (Formula 19).

$$\log X_{ip} = C_i^T \beta_p + \epsilon_{ip} \tag{17}$$

$$\begin{aligned} \log \left\{ \frac{Y_{ip}}{\left(\sum_p Y_{ip}\right)^{i/p}} \right\} &= \log X_{ip} + e_{ip} - \left\{ \frac{1}{p} \sum_p \log(X_{ip} + e_{ip}) \right\} && (18) \\ &= C_i^T (\beta_p - \bar{\beta}) + (e_{ip} - \bar{\epsilon}_i) && (19) \end{aligned}$$

From this follows that when modeling CLR-transformed data, you are provided with estimates for $\beta_p - \bar{\beta}$, while we want estimates for β_p. Next, an additional assumption is taken: the mode of β_p across p is zero. Given this assumption, $\bar{\beta}$ can be estimated by shifting the histogram of our estimates of $\beta_p - \bar{\beta}$ such that it has a mode of zero. The shift is our estimate for $\bar{\beta}$, i.e., $\hat{\bar{\beta}}$.

The now bias corrected estimates are used in the models. The residual variance of the linear model is shrunken using Empirical Bayes and using the shrunk variance, moderated t- and F-statistics are calculated on the parameter. The t-test tests to the null hypothesis that the coefficient is equal to zero. The F-test tests to the null hypothesis that the model has at least one non-zero coefficient. The F-statistic is calculated as $\frac{MSR}{MSE}$, with MSR the variation explained by the model and MSE the residual unexplained variation.

4 Causal Inference

4.0.1 CI vs regular regression analysis

One major distinction that we need to make is the difference between causal inference and regression analysis. Regression analysis statistically models the relationship between a dependent variable and a set of independent variables. However, this does not imply that these relationships are causal. Furthermore, regression analysis does not take into account the direction of the relationship. Demonstration of causality is a logical and experimental, rather than statistical, problem. An apparently strong relationship between variables could stem from many causes, including the influence of other currently unmeasured variables. The causal inference methods try to balance the data in such a way that the confounders are equally distributed between the groups. This can be done by matching, IPW, using counterfactuals...

4.0.2 Identifiability assumptions

One set of assumptions that is taken to identify causal effects in observed data, are the identifiability assumptions: consistency, positivity, and exchangeability. These are important, as this enables the identification of causal effects in data that is not (perfectly) randomized.

- Exchangeability: The treatment groups need to be similar to each other in every aspect (except for the treatment). In observational studies this is often relaxed to conditional exchangeability, as the groups have different characteristics by design. This does mean that there should be no unmeasured confounders.
- Consistency: The treatment of interest should be clearly defined and applied uniformly across individuals
- Positivity: There should be a non-zero probability of being assigned to each treatment group for every individual given its characteristics.

4.0.3 IPW

Inverse Probability weighting is a method used to minimize confounding when estimating treatment effects from observational data. By calculating a propensity score by using logistic regression on the treatment using the confounders, a weight can be derived. This propensity score is the chance of an observation to belong to the treated class given its covariates. In the thesis we get an ATE by calculating the weighted mean both populations (Formula 20).

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{20}$$

4.0.4 G-computation

G-computation is a causal inference technique that applies outcome modeling instead of treatment modeling. In this technique, a model is fitted to the outcome variables. In our case, this was a simple multivariate linear regression. Next, counterfactuals are constructed. This is the exact same dataframe twice: once with all individuals artificially put as 'diseased' and once with all individuals artificially set as 'healthy'. The model is used to predict the outcomes of the counterfactual individuals, which are then used to get an estimate of the ATE.

4.0.5 TMLE

Lastly, the TMLE technique is applied, which combines treatment modeling and outcome modeling in a very clever way. The analysis contains the following steps:

- Fit an outcome model.
- Predict counterfactuals.
- Fit a treatment model.
- Calculate the IPW for each observation, which they call the clever covariate. H_A will be the corresponding IPW weight if the individual was treated or untreated.
- The fluctuation parameter is calculated based on the outcome model estimate (for the right treatment) and the clever covariate. This is calculated using a logistic regression model using the logit transformed outcome as an offset. The information in the fluctuation parameter tells us how much to change or fluctuate the initial outcome estimates.
- The initial outcome estimates are updated using the fluctuation parameter.
- Calculation of the ATE using these updated outcomes for both populations.

4.1 Terms

- **Lowess fit** = Stands for *LOcally WEighted Scatterplot Smoothing.* It fits simple models localized to subsets of the data in order to build up a function that describes the deterministic part of the variation in the data, point by point.

- **Degrees of Freedom** = the maximum number of logically independent values, which may vary in a data sample.

- **Standard deviation** = The interval of one standard deviation of the mean (on each side) in a normal distribution contains 68% of the data. 2σ contains 95%, and 3σ contains 99.7%.