Modern statistics in medical and genetic research

Supervisors: Dr. Firneisz Gábor, Dr. Németh László

Somogyi Dalma

Background in genetics

To understand what genetic research investigates, I had to review the foundations of molecular genetics. In my understanding, all biological information necessary for the functioning and reproduction of living organisms is encoded. Deoxyribonucleic acid (DNA) has a particularly outstanding role in coding and storing the information. DNA molecule, or deoxyribonucleic acid, is the hereditary material in organisms, structured as a double helix (twisted ladder) made of repeating units called nucleotides, which contain a sugar, a phosphate, and one of four chemical bases: Adenine (A), Thymine (T), Guanine (G), or Cytosine (C). The specific sequence of these nucleobases determines the biological information stored in the DNA. A gene is a genomic unit of the DNA that frequently encodes proteins by which every three consecutive nucleotides (a codon) specify a particular amino acid. Typically the genes are composed of coding (exon) and non-coding (intron) DNA regions, and introns are not transcribed and translated into proteins. Humans have approx. 20000 protein-coding genes. The genetic information – in humans due to that we are diploid organisms – is packed into 23 chromosome pairs. The human genome consists around 3.2 billion nucleotide base pairs.

Genetic variation across individuals arises because these gene sequences are not identical for everyone. The most common gene variations are called single-nucleotide polymorphisms (SNPs). There are many bi-allelic gene variants and it means that there is a major allele and (another) minor allele and this is the source of the genetic variation. The specific allele setup of the individual is termed genotype and in case of a bi-allelic polymorphism it results in 3 possible genotypes: homozygous for the major allele, heterozygous, homozygous for the minor allele. Certain genotypes over specific SNPs (gene variant) are associated with increased risk of disease development. There are different genetic models to assess the genetic effects out of which in my work the dominant genetic model (binary) and the additive model will be used.

In the early 2000s, the human genome was mapped, producing large-scale datasets containing millions of SNPs. Although this provided unprecedented information about genetic variation, much less was known about the biological roles of most genes. This gap motivated the development of genome-wide association studies (GWAS), which investigate whether specific genetic variants are associated with observable traits, such as diseases. The total number of recorded human gene variants to date is over 700 million, however most of them are rare variants. In addition many large GWASs were already completed for epidemiologically important diseases, such as type 2 diabetes mellitus, hypertension, etc. In the view of these exceptionally large studies that were already done in order to end-up with novel or even ground-breaking results the approach and/or the outcomes should have a potential for novelty and it is also essential that the most advanced statistical methods are to be applied.

Setup

In my work the study group at the Semmelweis University I joined have access to a unique international database that contains genotype and phenotype information for nearly 5000 mother-neonate pairs. The number of SNPs that were directly genotyped prior for this project are 600 000 – 1 200 000. Furthermore - with the use of advanced data science multi-step algorithms and imputed genotypes based on international genetic data resources the No of SNPs is reaching approx. 20 million. Subsequently we are planning to perform GWASs for novel outcomes, such as the cord blood C-peptide level (C-peptide means "connecting peptide" and it occurs in a strictly proportional amount to the circulation insulin level). We are also currently working on novel conceptual approaches that are facilitated by the fact that we can work with dyads. In such GWAS studies it is critically important to adjust for potential confounders and covariates and given that we do have access to a large number of clinical data both from the mother and the newborn it also requires a thorough statistical approach.

The model

In genetic studies, the main goal is to see if a given SNP has significant effect in the development of a disease. In order to that, we build a model which describes this effect.
Generally, the independent variables are the SNPs and other clinical parameters, and the dependent variable is whether the disease develops or not.
The particular SNP being studied is called the explanatory variable which is binary (1 if it is the variation we test the connection to the disease and 0 otherwise), and the other clinical parameters are called covariates, which can be binary classifications (such as whether the data is for a first pregnancy) or continuous variables (such as C-peptide level). The dependent variable is target- or response variable which is also a binary classification (1 if the disease is diagnosed and 0 if not).
There may be additional variables, called confounders that are correlated to both the covariates and target variable, thus their effect should be taken into account upon inference.
In binary classifications, logit regression is typically used.

Definition: Model
Let the domain of the independent variables be $\mathcal{X} = \mathcal{X}^{(1)} \times \mathcal{X}^{(2)} \times ... \times \mathcal{X}^(k)$ (where $\mathcal{X}^{(i)}$ the domain of the ith variable) and of the target variable $\mathcal{Y}$. Let the random variable of independent variables be $X = (X^{(1)}, ..., X^{(k)})$ where $X^{(i)}$ is the random variable of the ith variable, $X^{(1)}$ beig the exploratory variable, and of the target variable Y. We assume there is an underlying true distribution of $X \times Y$.
A sample is a set of input-output data $((x_1, y_1), ..., (x_n, y_n)) \in (X_1 \times Y_1) \times ... \times (X_n \times Y_n)$, where n is the sample size, $X_i \times Y_i$ is the domain of the ith input-output pair.
A model class $\mathcal{M}$ is a set of $\mathcal{X} \to \mathcal{Y}$ functions, its elements are the models, the possibilities of how the target variable depend on the independent variables.
We intend to choose the best model $f \in \mathcal{M}$ in the sense that expectedly its result differ the least from the real outcome, i.e. minimizes the risk $R(f) = \mathbb{E}[\ell(f(x), y)]$ for some $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ loss function. (It is typically the squared loss, but for logistic regression it is $\ell(f, x, y) = -y \log(f(x)) - (1-y) \log(1 - f(x)))$.

Definition: Logistic Regression
The logistic regression aims to estimate connection between the target variable and the independent

variable as the following:
We assume $\mathbb{P}[Y_i = 1|X_i] = p_i$,
and $\text{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta X_i$, where $\beta = (\beta^{(1)}, ..., \beta^{(k)} \in \mathcal{X}^{(1)} \times ... \times \mathcal{X}^(k)$ (here X_i is taken as a column vector).
The inverse of the logit function is $\text{expit}(t) = \frac{1}{1+e^{-t}}$.

When building the model, we don't know the true β (marked β^*), so we have to use an estimation. Typically, it is the maximum-likelihood estimate.

Definition: Maximum likelihood estimate
$L(\beta) = \mathbb{P}[Y_1 = y_1, ..., Y_n = y_n|\beta, X_1 = x_1, ..., X_n = x_n]$ is the likelihood function of the sample. The maximum likelihood estimate of $\hat{\beta}$ is $\arg\max_\beta L(\beta)$.
Claim: Under certain regularity conditions (which for this model are met), $\hat{\beta} \to \beta^*$ almost surely and $\sqrt{n}(\hat{\beta} - \beta^*) \to N(0, \mathcal{I}(\beta^*)^{-1})$ in distribution (where $\mathcal{I}$ is the Fisher-information) as n tends to infinity. Hence for large enough sample sizes, it is a close estimate of the true parameter.

Dimension reduction

In many cases k can be very large, so due to the physical bounds of computers (executing the computations), dimension reduction is often applied, i.e. we have to choose what variables to include in the model and what to exclude without significant loss of information.
One way to do that is variable selection, and another is principal component analysis.

Variable selection

The aim of the variable selection procedures based on (typically regression) models is to determine the right covariates in order to receive a better estimation for the explanatory variable's effect.

Definition: forward and backward selection
In forward selection, we build up our model step by step from zero covariates, always choosing the one that best improves a criterion measuring model accuracy.
In backward selection, we start from the full set of covariates and step by step remove the one, whose exclusion leads to the best improvement of the criterion.
The iterations stop after reaching a given threshold of criteria or number of iterations.

In addition to p-values (see in sec. Corrections), AIC is often used as the criterion in regression models:
Definition: Akaike information criterion (AIC)
For $k+1$ estimated parameters (in our case, $\beta_0, ..., \beta_k$), the Akaike information criterion is $2(k+1) - 2\ln(\hat{L})$, where $\hat{L} = \max L(\beta) = \max \prod_{i=1}^{n} \text{expit}(\beta_0 + \beta X_i)^{y_i}(1 - \text{expit}(\beta_0 + \beta X_i))^{1-y_i}$, the maximum of the log-likelihood function.
The interpretation of AIC is that the number of parameters, i.e. the complexity of the model increases it, and the maximum of the likelihood, i.e. the accuracy of the model decreases it. Thus we aim to obtain low AIC values, building as accurate and as little complex model as possible.

Principal component analysis

Definition: Principal component analysis (PCA)
Principal component analysis is another useful tool to conceptualize large samples.
We define a new coordinate system so that the greatest variance of the sample lies on the first coordinate, which is called the principal component (PC):
$u_1 = \arg\max(\mathbb{D}[u^T X] : |x| = 1)$, $u_k = \arg\max(\mathbb{D}[u^T X] : |x| = 1, u \perp \{u_1, ..., u_{k-1}\})$.
Generally, the first few components responsible for a large margin of variance in the data. This allows us to use this procedure as a dimension reduction tool.
Thus, looking at the data in this coordinate system, the PC (hopefully) distinguishes the data according to some trait. If we look at the second, third,.. components, those will distinguish further.
Thus by this, we can identify confounders.

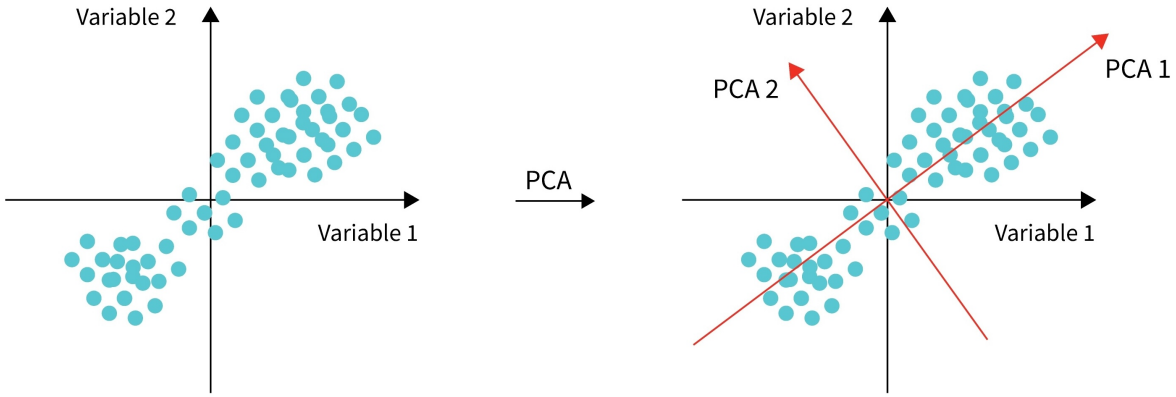

Fig.1

Corrections

For a given SNP as exploratory variable, β_1 is considered the measure of its effect on the target variable. So for the jth SNP, the null-hypothesis is $H_j : \beta_1 = 0$ (SNP has no effect) and the alternative hypothesis is $H'_j : \beta_1 \neq 0$ (SNP has effect).

Definition: p-value
The p-value of the jth SNP is $\mathbb{P}[|T| \geq |t_{\text{obs}}||H_0]$ for some appropriate statistic T with t_{obs} observed value.
Typically, Wald-test is used, where the statistic is $W_n(\hat{\beta}) = \frac{\hat{\beta_1}^2}{[\mathcal{I}(\hat{\beta})^{-1}]_{1,1}}$ (n indicates the sample size).
Claim: Under the null-hypothesis, W_n has χ^2_1 distribution (as n tends to infinity).
This way, for $1-\alpha$ level if significance, the null-hypothesis is rejected if $W_n(\hat{\beta}) > \chi^2_{1,1-\alpha}$ where $\chi^2_{1,1-\alpha}$ is the quantile of χ^2_1 distribution.
The p-vale is $\mathbb{P}[\chi^2_1 \geq W_n(\hat{\beta})|H_0]$.

Since again, we don't know the true β, in practice, p-value is used as the measure of evidence to-

wards the alternative hypothesis: if $P_j < \alpha$, then H_0 is rejected, i.e. the jth SNP has an effect in the development of the disease. $1-\alpha$ is a pre-determined level of significance. $\alpha = 5 \cdot 10^{-8}$ is a standard choice.

The P_j's are typically presented on a so called Manhattan-plot. On its x-axis are the SNPs (grouped by which chromosomes they are on) and on its y-axis is the $-\log_{10}$ of the p-values. So the SNPs with effect on the disease are the ones with $-\log_{10}$ of P-values above $-\log_{10}(\alpha)$.

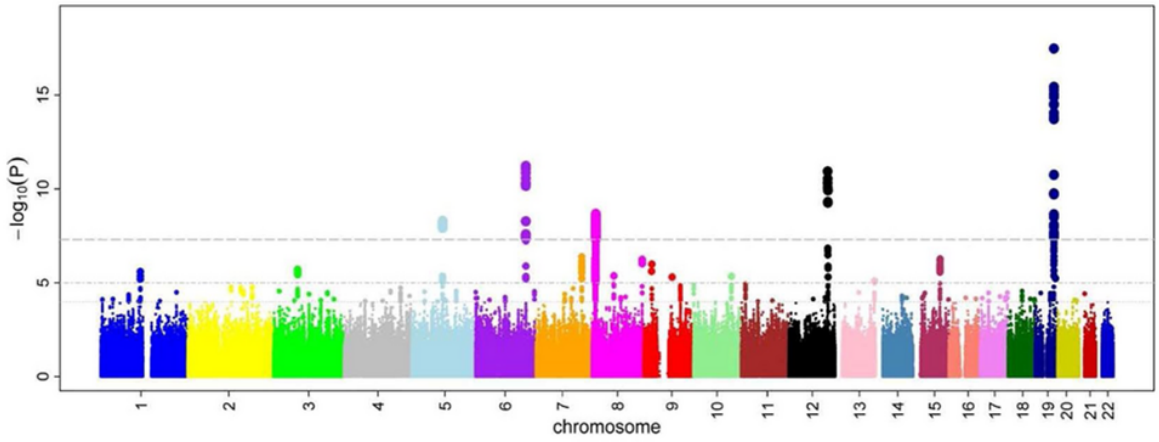

Fig.2

During GWAS studies, we test for millions of SNPs as explanatory variables, so multiple hypoteses are tested. Let m be the number of tests and m_0 be the number of true null-hypotheses. The probablities of type I errors (false discoveries, i.e. we reject the null-hypothesis, but it is true) can add up to concerning heights. For the jth test, $\mathbb{P}[\text{type I error}] = \mathbb{P}[W_n(\hat{\beta}) > \chi^2_{1,1-\alpha}] = \alpha$.

This add-up of type I errors can be measured by family wise error rate or false discovery error rate. To compensate, different types of corrections can be introduced.

Definition: Family wise error rate (FWER)

The FWER is the probability of making at least one type I error among all m tests.

Definition: Bonferroni correction (p-correction)

If we would like to achieve $1-\alpha$ level of confidence, Bonferroni correction adjusts it to $1-\alpha/m$, i.e. rejects H_j at $P_j \leq \alpha/m$.

Claim: This way, FWER$= \mathbb{P}[\bigcup_{H_j \text{ is true}} P_j \leq \frac{\alpha}{m}] \leq \sum_{H_j \text{ is true}} \mathbb{P}[P_j \leq \frac{\alpha}{m}] \leq m_0 \frac{\alpha}{m} \leq \alpha$.

Definition: False discovery error rate (FDR)

Let Q be the ratio between false discoveries and discoveries. FDR is the expected value of Q, $\mathbb{E}(Q)$.

Definition: Benjamini-correction

For FDR level α, we reject all $H_1, ..., H_k$ and accept $H_{k+1}, ..., H_m$ where $k = \max(l : p_l \leq l\frac{\alpha}{m})$.

Claim: $\mathbb{E}(Q) \leq m_0 \frac{\alpha}{m} \leq \alpha$.

Future goals

In the focus of our group's new study will be the analysis of the interactions between multiple SNPs as effects on Gestational Diabetes Mellitus.

Literature

Fig.1: https://www.scaler.com/topics/nlp/what-is-pca/ [2025.12.14.]
Fig.2: https://en.wikipedia.org/wiki/Manhattan_plot [2025.12.14.]
1, Balázs Csanád Csáji : Statistical Learning Theory [notes for statistical learning theory, 2025]
2, Balázs Csanád Csáji : Matematikai statisztika [notes for Mathematical Statistics, 2025]
3, https://www.statlect.com/fundamentals-of-statistics/Wald-test
4, https://en.wikipedia.org/wiki/Family-wise_error _rate
5, https://en.wikipedia.org/wiki/False_discovery_rate
4, Gábor Firneisz et al.: Association Study with 77 SNPs Confirms the Robust Role for the rs10830963/G of MTNR1B Variant and Identifies Two Novel Associations in Gestational Diabetes Mellitus Development
5, Alan Kuang et al.: Multi-ancestry genome-wide association analyses: a comparison of meta- and mega-analyses in the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study
6, Jinyoung Byun et al.:Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure
7, Nobuyuki Horita, Takeshi Kaneko: Genetic model selection for case-control study and meta-analysis
8, Dr. Firneisz Gábor, Nemes A. Botond, Dr. Németh László, Dr. Nádasdi Ákos, Prof. Dr. Benyó Zoltán: Association of Mode of Delivery and Birth Order with Neonatal Cord Blood Hyperinsulinemia [preprint]