M. KNIGGE

INTERPLAY BETWEEN GUT MICROBIOME AND IMMUNE SYSTEM

IMPORTANT TERMS & CONCEPTS

α-parameter ×

ℓ_1-penalty ×

ℓ_2-penalty ×

λ ×

R^2 √

Adjusted R^2 √

ANOVA √

Bias ×

Coefficient ×

Collinearity √

Cross-Validation ×

Degrees of freedom ×

Dependent Variable √

Differentiation ×

Elastic Net Regularization ×

Error ×

Error due to Bias √

Error due to Variance √

Estimator ×

F-distribution ×

F-test √

Gauss Markov Theorem ×

Genetic Variant ×

Heteroscedasticity √

High-Dimensional Data ×

Independent Variable √

Interaction Terms √

Intercept ×

Least Absolute Shrinkage and Selection Operator (LASSO) ×

Level-Level Regression Specification ×

Linear Regression √

Log Transformations ×

Logistic Regression ×

Log-Level Regression Specification ×

Log-Log Regression Specification ×

Mean Square Error (MSE) √

Multiple Linear Regression √

One-Sample Mendelian Randomization (MR) √

Ordinary Least Squares (OLS) √

Outlier Detection ×

Population mean √

Population variance ×

Residuals √

Ridge Bias Constant ×

Ridge Regression ×

Ridge Trace ×

Sample mean √

Sample Variance ×

Scale in Ridge Regression ×

Shrinkage Estimators √

Standard Deviation (SD) ×

Standard Error of the Mean (SE) ×

Two-Stage Least Squares (2SLS) Regression ×

Variable Selection ×

Variance ×

Variance inflation factors (VIF) √

α-parameter

ℓ_1-penalty

ℓ_2-penalty

λ

R^2

The coefficient of determination. This is a measure of the goodness of fit for a linear regression model. It is the percentage of the dependent variable variation that is explained by a linear model. R^2 = explained variation / total variation. The R^2 is always between 0 and 100%. Zero percent indicates that the model explains none of the variability of the dependent variable around its mean. 100 percent indicates that the model explains all the variability of the dependent data around its mean.

Low R^2 values are not inherently bad. In some fields it is expected that the R^2 values will be low. For example in predicting human behavior. If the R^2 is low, but the still the predictors are significantly it is still possible to draw conclusions in how changes in the predictor values are associated with changes in the response value. And regardless of the R^2, the significant coefficients still represent the mean change in the response for one unit change in the predictor. The number of independent variables in the model will increase the value of R^2, regardless whether the variables offer an increase in explanatory power.

Adjusted R^2

In response to the phenomena where the number of independent variables will increase the R^2, no matter if these bring an increase in explanatory power, the adjusted R^2 metric can be utilized to penalize a model for having to many variables.

ANOVA

This method is used to compare differences between means of more than two groups. This is done by looking at the variation in the data and where that variation lies. **ANOVA** compares the amount of variation between groups with the amount of variation within groups. When samples are taken from a population, it is expected that each **sample mean** is different because samples are taken instead of the whole population: this is called the sampling error, or the effects of chance. It is always expected to find differences between means from different samples. With ANOVO it is tested if there is true difference in the **population mean** and **sample mean**. Is the difference between groups greater than expected to occur by chance?

ANOVA: $X_{ij} = \mu_i + \epsilon_{ij}$. Here X are the individual data points in the sample. i, and j denote the group and the individual observations. ϵ is the unexplained variation and the parameters of the model (μ) are the population means of each group. Each data point ($i,\ j$) is its group mean plus error. ANOVA is used to calculate a statistic called the F-ratio with which the probability of obtaining the data assuming the null hypothesis. A significant P-value suggest that at least one group mean is different from the others.
-- Null hypothesis: all population means are equal.
-- Alternative hypothesis: at least one population mean is different from the rest.

ANOVA splits the variation from the data sets into two parts: between-group and within-group. This variation is called the sums of squares, which is done in three steps: (I) **variation between groups**, (II) **variation within groups**, and (III) the **F ratio**. To test that all population means are equal or that at least on is different the **F ratio** must be considered. For this the following terms are important: **Grand Mean** (**GM**), **Total Variation** (), **Between Group Variation** (Sum of Squares Between: **SSB**), **Within Group Variation** (Sum of Squares within: **SSW**), **F statistic**, **Mean Square Between groups** (**MSB**), **Mean Square Within groups** (**MSW**).

Grand Mean

The grand mean of a set of samples is the total of all the data values divided by the total sample size. This requires all of the sample data available, which is usually the case, but not always. It turns out that all that is necessary to find perform a one-way analysis of variance are the number of samples, the sample means, the sample variances, and the sample sizes: $\bar{X}_{GM} = \frac{\sum x}{N}$. Here, $\bar{X}_{GM}$ is the **grand mean**, N is the total sample size, and x individual observations. Another way to find the grand mean is to find the weighted average of the sample means. The weight applied is the sample size: $\bar{X}_{GM} = \frac{\sum n\bar{x}}{\sum n}$. Here, $\bar{x}$ is the sample mean, and n is the sample size.

Total Variation

The total variation (not variance) is the sum of the squares of the differences of each mean with the **grand mean**. There is the between group variation and the within group variation. The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance. If the variance caused by the interaction between the samples is much larger when compared to the variance that appears within each group, then it is because the means are not the same. $SST = \sum (x - \bar{X}_{GM})^2$

Between-group variation, or between-group sums of squares (**SSB**) is calculated by comparing the mean of each group with the overall mean of the data: $SSB = \sum n(\bar{x} - \bar{X}_{GM})^2$. Here n is the sample size, and $\bar{x}$ is the sample mean. This is the variation due to interaction between the samples. If the sample means are close to each other this will be small. There are k samples involved with one data value for each sample (sample mean), so there are k-1 degrees of freedom. The variance due to the interaction between the samples is denoted **MSB** for **Mean Square Between groups**. This is the between group variation divided by its degrees of freedom. It is also denoted by s_b^2.

Within-group variation is used in **ANOVA** tests. It refers to variations caused by differences within individual groups (or levels). In other words, not all the values within each group are the same. These are differences not caused by the independent variables. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: $df = N - k$. SSW can be calculated as follows: $SSW = \sum df s_b^2$. The variance due to the differences within individual samples is denoted **MSW** for **Mean Square Within groups**. This is the within group variation divided by its degrees of freedom. It is also denoted by s_w^2. It is the weighted average of the variances (weighted with the degrees of freedom).

F statistic

The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k). If the average difference between **SSB** is similar to **SSW** the **F ratio** is close to one. If the difference becomes larger between **SSB** and **SSW** the **F ratio** will be greater than one. Thus, variables with a higher **F ratio** have higher explanatory power: $F = \frac{s_b^2}{s_w^2}$. This can be tested by obtaining a p-value, this can be tested against **F-distribution** of a random variable with the degrees of freedom associated with the numerator and denominator of the ratio. The p-value describes the probability of getting that **F ratio** or a greater one. Large **F ratios** gives smaller p-values.

Bias

Coefficient

Collinearity

This is the not favored situation where the **correlations** among the **independent variables** are strong. In some cases, **multiple regression** can reflect a significant relation when there is none. For example, the model can fit the data very well, when even none of the X variables have a significant impact on explaining Y. This is possible when two X variables are highly correlated, they both bring the same information, and when this happens, the X variables are **collinear** and the results show **multicollinearity**. This is a problem because it inflates the **standard errors** of the **coefficients**, which makes some variables not significant when they should be.

Cross-Validation

Degrees of Freedom

Dependent Variable

Dependent variables depend on the values of independent variables. The **dependent variables** represent the output or outcome whose variation is being studied.

Differentiation

Elastic Net Regularization

Error

The **error** of an observed value is the deviation of the observed value from the true value of a quantity of interest (for example, a population mean). For example, if the mean height in a population of 21-year-old men is 1.75 meters, and one randomly chosen man is 1.80 meters tall, then the **error** is 0.05 meters; if the randomly chosen man is 1.70 meters tall, then the **error** is −0.05 meters. The expected value, being the mean of the entire population, is typically not observable, and the statistical error cannot be observed either.

Error due to Bias

Error because of bias is taken as the difference between the expected or average prediction of the model and the correct value which the model tries to predict.

Error due to Variance

Error because of variance is taken as the variability of a model prediction for a given data point. The entire model building process is repeated multiple times for example, then the variance is how much predictions for a given data point vary between different realizations of the model.

Estimator

F-test

Gauss Markov Theorem

Heteroscedasticity

Linear regression using **OLS** has the assumption that **residuals** are identical distributed across every X variable. If this condition holds true, the error terms are **homogeneous**, which means that the errors have the same scatter regardless of the value of X. When the scatter of the errors is different, varying depending on the value of one or more of the **independent variables**, the error terms are **heterogeneous**.

A collection of random variables is **heteroscedastic** if there is a sub-population that has different variabilities from others. Here "variability" could be quantified by the variance or any other measure of statistical dispersion. Thus **heteroscedasticity** is the absence of **homoscedasticity**. The existence of **heteroscedasticity** is a major concern in the application of regression analysis, including the analysis of **variance**, as it can invalidate statistical tests of significance that assume that the modeling errors are uncorrelated and uniform—hence that their variances do not vary with the effects being modeled.

High-Dimensional Data

Independent Variable

The **independent variables** represent inputs or causes, potential reasons for variation or, in the experimental setting, the variable controlled by the experimenter. Models and experiments test or determine the effects that the i**ndependent variables** have on the **dependent variables**.

Interaction Terms

Adding **interaction terms** to a regression model can improve the understanding of the relationship among variables in the model and allows for more hypotheses to be tested. The presence of a significant interaction indicates that the effect of one **independent variable** on the **dependent variable** is different at different values of the other **independent variables**. This is tested by adding a term to the model in which the two **independent variables** are multiplied: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \epsilon$

Intercept

Least Absolute Shrinkage and Selection Operator (LASSO)

Level-Level Regression Specification

Linear Regression

Regression analysis is the science of fitting straight lines to patterns of data, it used for prediction and forecasting. In a linear regression model, the variable of interest is (**dependent variable**) predicted from a single or more variables (**independent variable**) using a linear formula. Regression analysis is also used to understand to find out which among the the **independent variables** are related to the **dependent variables**, and to explore the forms of these relations. The simple regression model can be represented as follows: $Y = \beta_0 + \beta_1 X_1 + \epsilon$. The β_0 is the Y **intercept** value, the **coefficient** β_1 is the slope of the line, the X_1 is an **independent variable** and ϵ is the **error term**. The error term is used for correcting a prediction error between the observed and predicted value. This can be interpreted as for each unit increase in X_1, Y will increase by some value.

Log Transformations

Logistic Regression

Log-Level Regression Specification

Log-Log Regression Specification

Mean Square Error (MSE)

Measures the average of the squares of the errors, that is, the difference between the estimator and what is estimated. **MSE** is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.

Multiple Linear Regression

A multiple linear regression is the same as a simple linear regression but, there can be multiple **coefficients** and **independent variables**: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \epsilon$

One-Sample Mendelian Randomization (MR)

This is the basic implementation of **MR** on SNPs, exposure, and outcome from a single data set, with the addition of genetic data from participants. The causal effect of the exposure on the outcome can be estimated by a two-step regression analysis: Two-Stage Least Squares (**2SLS**) regression. In the first step, a linear regression is fitted on the **genetic variants** from the exposure. In the second step,the genetic variants from the outcome are regressed over the predicted values from step one by linear or **logistic regression**. The **coefficient** from the second stage can be interpreted as the change in the outcome per unit increase in the exposure. If this **coefficient** is significant, than the exposure is causal (or protective, depending on the sign of the coefficient) for the disease.

Ordinary Least Squares (OLS)

This is a method for estimating/predicting parameters that are unknown in a **linear regression** model. **OLS** chooses the parameters of a linear function in a set of **independent variables** by minimizing the sum of the squares of the differences between the observed **dependent variable** (values of the variable that will be predicted) in the given data set and those predicted by the linear function. This can be seen as a square drawn between the predicted and observed (**dependent variable**). The sum of Squares are a representation of the error in the **OLS** regression model. In **linear regression** models, prediction errors can be decomposed into two main sub-components: **error due to bias**, and **error due to variance**. Understanding bias and variance is key to understanding the behavior of prediction models. It is important to understand how different sources of error lead to **bias** and **variance** and this can help improve data fitting. There is a trade-off between a models ability to minimize **bias** and **variance**. The best model is at the level of complexity where the increase in **bias** is equivalent to the reduction of **variance**. **Bias** is reduced and **variance** is increased in relation to the complexity of a model. As more parameters are added to the model, the complexity of the model rises and the **variance** becomes the primary concern, and the **bias** falls. The **Gauss Markov Theorem** states that among all linear unbiased estimates, **OLS** has the smallest **variance**. What implies that **OLS** estimates have the smallest **MSE** among linear estimators.

-- But, is it possible that there can be a biased estimator with a smaller **MSE**?

--This is where **Shrinkage Estimators** are used.

Outlier Detection

Population mean

The **population mean** is an average of a group characteristic. The group could be a person, item, or thing, like "all the people living in the United States" or "all dog owners in Georgia". A characteristic is just an item of interest. It is actually rare that the **population mean** can be calculated. That's because asking an entire population about something is usually cost prohibitive or too time consuming. For example, one veterinary practice probably keeps weight records of all the pets that come in the door, enabling you to calculate the average weight of a dog for that practice (i.e. the population mean for that practice). But if you were working for a pet food company who wanted to know the average weight of a dog, you wouldn't be able to track down all the 70 to 80 million dogs in the US and weigh them.

Residuals

The **residual** of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean). A **residual**, is an observable estimate of the not observable statistical error. Consider the example with men's heights and suppose we have a random sample of n people. The **sample mean** could serve as a good estimator of the **population mean**. Then we have: the difference between the height of each man in the sample and the not observable **population mean**, this is the statistical error, whereas the difference between the height of each man in the sample and the observable **sample mean** is a residual.

Residual plot

This is a scatterplot of the **residuals** (difference between the actual and predicted value) against the predicted value. A good model will show a random pattern of the residuals with no shape. Residual plots can be used for diagnostics and assumption testing in linear regression.

Ridge Bias Constant

Ridge Regression

Ridge Trace

Sample Mean

Simply the mean or average when the context is clear, is the sum of a collection of numbers divided by the number of numbers in the collection.

Sample Variance

Scale in Ridge Regression

Shrinkage Estimators

This an estimator that, either explicitly or implicitly, incorporates the effects of **shrinkage**. This means that a raw estimate is improved by combining it with other information. The term relates to the concept that the improved estimate is made closer to the value supplied by the 'other information' than the raw estimate. Estimators can be improved, in terms of **MSE**, by shrinking them towards zero (or any other fixed constant value). Assume that the expected value of the raw estimate is not zero and consider other estimators obtained by multiplying the raw estimate by a certain parameter. A value for this parameter can be specified so as to minimize the **MSE** of the new estimate. For this value of the parameter, the new estimate will have a smaller **MSE** than the raw one. Thus it has been improved. An effect here may be to convert an unbiased raw estimate to an improved biased one.

The standard **estimator** of the **population mean** is the **sample mean**, which is not biased. Building an **estimator** by shrinking the **sample mean** results in a biased **estimator**, with an expected value less than the **population mean**. Shrinkage always reduces the estimator's **variance** and reduce the **MSE**. For example, a variable Y, which has a population of values. When taken a sample from Y, the observations drawn are $Y_1 .... Y_n$. The **population mean** is the "true value" μ_Y (the estimand): the thing to be estimated. The **sample mean** is $\bar{Y}$. The **population variance** is σ^2. There can be more than more than one **estimate** (multiple dimensions), the number of **estimates** are denoted by k. For example three dimensions with **independent variables** X_1, X_2, and X_3.

Standard Deviation (SD)

Standard Error of the Mean (SE)

Two-Stage Least Squares (2SLS) Regression

Variable Selection

Variance

Variance inflation factors (VIF)

This measures how much **variance** of the estimated **coefficients** are increased over the case of not having correlation between X variables. If there are two variables correlated, how to decide which one should be removed? To determine the best one to remove, remove each one separately from the model en perform regression and select the regression equation that explains the most **variance**: the highest R^2.