AdvStDaAn, Summaries of all weeks

Michael Lappert

29 April, 2022

Contents

Week 1: Review of Multiple Linear Regression I

- We introduced the linear regression model $Y = \boldsymbol{X}\beta + E$ with $E\ N(0, \sigma\boldsymbol{I})$
- We introduced Tukey's first-aid transformations as a good start when building a data-driven model
- We fitted the regression model to data (i.e., we estimated β)
  - by the least squares algorithm
  - is motivated by the maximum likelihood principle
- We performed hypthesis testing on marginal parameters (i.e., β_k) and compared nested models
- We ensured that the regression model fits the data adequately by perfomring a residual and sensitivity analysis

Week 2: Review of Multiple Linear Regression II

- AIC can be used for the **variable selection**
- Also **categorical variables** can be used as predictors
- Even published data analyses do not necessarily yield models that adequately describe the data
- **Multicollinearity** may prevent a sound interpretation
  - Often the **Variance Inflation Factor (VIF)** is udes to detect multicollinearity
- **Cross-validation** can be used to evaluate the prediction performance of a model (*In regression analysis, for example **PRESS** a leave-one-out method, can be used. It can be calculated using the results of a single least squares fit for all n observations*)
- The **prediction interval** is used to quantify the accuracy of a single prediction
- There are some **implicit assumptions** under which only predictions can work:

  - The learning the dataset and the future cases must be from the same population
  - I nmost cases, the prediction is only reliable within the range of the leraning data
  - The relationship between the response and predictors must remain invariant (e.g., no time dynamics)
- A regression model does not imply a **cause-effect relationship** between the variables

Week 3: Some Advanced Topics in Linear Regression Modelling

- We have learnt a number of techniques for building a statistical model:
- **Residual Analyis**: We check the requirements independency, constant variance, liinearity (i.e., constant expected value), and normal distribution as best as we can. We also try to identify influential observations. - Robust methods are suitable to do this analysis very efficiently.
- **Transformations**: Depending on the situation, it makes sense to transform the response variable as well as the explanatory variables, to form new wxplanatory variables from several or to extend the models with factor variables.
- **Variable Selection**: Nowadays it is easy to do it automatically. The fact that this is not always appropriate has been explained.
- But it is not obvious in which order the tools need to be applied in practice.
- A good genereci solution, but not the ultimate, always-optimal strategy may be the following:

  1. Clarify the tast (purpose? goal? prediction vs. exploration); are there already model approaches?
  2. Data screening and processing
  3. Set up model using Tukey's first-aid transformations or GAM's
  4. Model fitting: preferably with robust methods
  5. Residual and sensitivity analysis; does this dataset help solving the task? *(Eventually backk to 3., 2. or 1.)*
  6. Variable selection: treat collinearities if necessary *(eventually back to 3.)
  7. Checking model adequacy

  - Residual and sensitivity analysis with selected model(s)
  - Check plausibility; match model(s) with subject matter expertise
  - Out-of-sample validation, especially if the model is to be used for prediction *(eventually back to 1., 2., 3., . . .)*

  8. Reporting: It is key to be honest and openly report all data manipulations and decisions that were made

Week 4: Binary Response

- The Logistic Regression Model was motivated and introduced:
  - It is an adequate model when the response is binary or the result of aggregates binary data (-> 'binomial data')
  - The model approach can be motivated by latent variables and their distributions
  - Logistic regression is build from three elements: The assumption about the distribution of the response (i.e., binomial distribution), the linear predictor and the link function
- It is fitted to the data by Maximum Likelihood estimation resulting in an system of nonlinear equations
- The coefficients are interpreted by log-odds
- The fitted coefficients are asymptotically gaussian distributed -> we can interpret the summary output in the same way as that of a least squares fit, except that the statements about test results and confidence interval are only approximate

Week 5: A Unifying Model Family

- We learned that members of the two parameters exponential family can be written in a common form
- A compilation of the key elements of the two parameter exponential distribution family for some of the most popular distributions was introduced (check slide 17 of week 5)
- The link functin which is identical to $b(\mu)$ is called the canonical link
- The two main elements that define the GLM are
  - *Distributional element:* The distributin of the response Y_i, given the explanatory variables x_i, belongs to the **two parameter exponential family** with expectation $E(Y_i|x_i) = \mu_i$ and **dispersion** ϕ.
    The responses $Y_i, i = 1, ..., n$ are independently distributed
  - *Structural element*: The eypectation μ_i is related to the **linear predictor** $\eta_i = x_i^T\beta$ of the explanatory variables by applying a (posibly nonlinear) function g() on μ_i:
    $g(\mu_i) = \eta_i = x_i^T\beta$
    The functin g() is called the **link function**
- The estimator β is derived based on the **maximum likelihood priciple**
  - it is defined by a nonlinear system of equations
  - it can be solved by the IRLS algorithm
- $\hat{\beta}$ is **asymptotically** (i.e. approximately) **Gaussian distributed**

Week 6: Model Simplification and Inference

- **Residual deviance** is the building block to statistically compare nested GLMs
- The GLMS generalistaion of the F test is the **deviance test**
- **Null deviance** is the residual deviance of 'Y ~ 1'
- If $\phi = 1$, the residual deviance is $\sim \chi^2_{n-p}$ except the resonse is binary
- Variable selection is done with **Aikaike Information Criterion (AIC)**
- Use **deviance-based (profiling)** insted of Wald-based confidence intervals (CI), i.e., use *confint(...)* in R
- **CI for expected response:** Transform CI determined on the linear predictor back into the response space
- Use CI based on Taylor approximation only in special cases such as MAC

Week 7: Diagnostics and Model improvement

- Applying the standard residual and diagnostic plot as well as the Bootstrap simulation we can check the model adequacy and determine which model assumptions, if any, are violated
- Again GAMs are used o find suitable transformations of the explanatory variables

Week 9: What is Bayesian Data Analysis?

- New terminology: prior, posterior, conditional, credible interval
- Review of known distributions, beta distribution is newly introduced
- Recipe of Bayesian Data Analysis:
  - To perform a Bayesian Data Analyis it requires
    - observed **data**, you want to draw conclusions from
    - a **generative model** or data generating process
    - to quantify the uncerainty of parameters in your model, i.e. include **prior information** before looking at your data

- With a clear understanding of the random process leading to your data, you have to follow three steps (gelman et al. 2014)
  - set up a full probability model and include all types of uncertainty
  - condition on observed data
  - evaluate the fit of the model and posterios's implications