1 Simple Linear Regression

Properties of the Least Square Estimates

- $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$
- Variances of the estimates

$$Var(\hat{\beta}_0) = \sigma_E^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right)$$

$$Var(\hat{\beta}_1) = \frac{\sigma_E^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

- Precise estimates are obtained with
  - a large number of observations n
  - a good scatter in the predictor x_i
  - an informative/useful predictor, making σ_E^2 small

Simple Regression

- Residual squared error:

$$\hat{\sigma}_E^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n r_i^2$$

- Coefficient of Determination

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

- Confidence interval for the slope

$$\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \hat{\sigma}_{\hat{\beta}_1} \iff \hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \sqrt{\frac{\hat{\sigma}_E^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

- Confidence interval for $E[y \mid x]$

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

- Prediction interval for y

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

Log-Transformation

- For a simple prediction of the y-value on the original scale, we can exponentiate to invert the log-transformation.
- This is an estimate for the median of the conditional distribution $y \mid x$, but not the conditional mean $E[y \mid x]$. If we require unbiased fitted values on the original scale, applying a correction factor is required
  - The link between Gaussian and lognormal distribution

$$\hat{y} = \exp(\hat{y}' + \frac{\hat{\sigma}_E^2}{2})$$

    - The smearing estimator proposed by Duan

$$\hat{y} = \exp(\hat{y}') \cdot \frac{1}{n} \sum_{i=1}^{n} \exp(r_i')$$

- Confidence intervals for the median and prediction intervals by simple transformation:

$$[lo, up] \rightarrow [\exp(lo), \exp(up)]$$

- If one needs a confidence interval for the mean, the theoretical or Duan's smearing correction has to be used

2 Multiple Linear Regression

When to transform?

- If on a relative scale, meaning that an increase from $10 \rightarrow 11$ is not identical to $100 \rightarrow 101$
- Left-closed (with 0 as the smallest possible value), and right-open variables are often relative and require transformation
- If the scatter (i.e. the magnitude of the uncertainty) increases with increasing value, as is often the case for relative scales
- If marginal distribution of the variable (as observed in a histogram) is clearly right-skewed (i.e. extreme values exist)

Estimating the Error Variance

$$\hat{\sigma}_E^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} r_i^2$$

Assumptions on the Error Term

- $E[E_i] = 0$: the hyperplane is correct, no systematic error
- $Var(E_i) = \sigma_E^2$: constant scatter for the error term
- $Cov(E_i, E_j) = 0$: uncorrelated errors
- $E_i \sim \mathcal{N}(0, \sigma_E^2)$: the errors are normally distributed

Coefficient of Determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Adjusted Coefficient of Determination

$$adjR^2 = 1 - \frac{n-1}{n-(p+1)} \cdot (1 - R^2)$$

Confidence Interval for Coefficient β_j

$$\hat{\beta}_j \pm qt_{0.975;n-(p+1)} \cdot \hat{\sigma}_{\hat{\beta}_j}$$

Individual Parameter Tests

- The p-values of the individual hypothesis tests are based on the assumption that the other predictors remain in the model and do not change. Therefore, you must not drop more than one single non-significant predictor at a time.
- It can happen that all individual tests do not reject the null hypothesis, although some predictors have a significant effect on the response. Reason: correlated predictors!
- The multiple testing problem: when doing many tests, the total type I error increases and we may observe spuriously significant predictors. (Overall type I error: $1-(1-p)^m$)

Comparing Hierarchical Models

- The big model must contain all the predictors from small model, otherwise they are not hierarchical and the test does not apply
- F test

$$F = \frac{n-(p+1)}{p-q} \cdot \frac{RSS_{Small} - RSS_{Big}}{RSS_{Big}} \sim F_{p-q,n-(p+1)}$$

- Global F-test

$$F = \frac{R^2}{1-R^2} \times \frac{\mathrm{df}_2}{\mathrm{df}_1}$$

Prediction

- Only interpolation. Prediction within the range of observed y-values works well. Extrapolation yields non-reliable results

Categorical Predictors

- Categorical predictors are often called factor variables. In a linear regression, each level of such a variable is encoded by a dummy variable, so that $(\ell - 1)$ degrees of freedom are spent

Inference with Factor Variable

- In a regression model where factor variables that have more than 2 levels and/or interaction terms are present, the `summary()` function does not provide useful information for variable selection. We have to work with `drop1()` instead!
- `drop1()` performs correct model comparisons and respects the model hierarchy.
- Do not perform individual hypothesis tests on factors that have more than 2 levels, they are meaningless!

Residuals vs. Error

- The residual random variables R_i share some properties of the errors E_i, but not all - there are important differences.
- Even in cases where the E_i are uncorrelated and have constant variance, the residuals R_i feature some estimation-related correlation and non-constant variance.
- The estimation-induced correlation and heteroskedasticity in the residuals R_i is usually very small. Thus, residual analysis using the raw residuals r_i is both useful and sensible.
- One can try to improve the raw residual r_i with dividing it by an estimate of its standard deviation (i.e. standardized)

$$\tilde{r}_i = \frac{r_i}{\hat{\sigma}_E \cdot \sqrt{1-H_{ii}}}$$

- H_{ii} is the diagonal element of hat matrix
- $\hat{\sigma}_E$ is the residual standard error
- Studentized Residuals: $\hat{\sigma}_E$ was estimated by ignoring the i-th data point

Unusual Observations

- There can be observations which do not fit well with a particular model. These are called outliers. The property of being an outlier strongly depends on the model used.
- There can be data points which have strong impact on fitting of the model. These are called influential observations.
- A leverage point is an observation that lies at a different spot in predictor space. This is potentially dangerous because it can have strong influence on the fit.

Leverage

- Leverage points are different from the bulk of data
- The average value for leverage is given by $\dfrac{p+1}{n}$
- We say a data point has high leverage if $H_{ii} > \dfrac{2(p+1)}{n}$

Cook's Distance

$$D_i = \frac{\sum(\hat{y}_k^{[-i]} - \hat{y}_k)^2}{(p+1)\sigma_E^2} = \frac{H_{ii}}{1-H_{ii}} \cdot \frac{\tilde{r}_i^2}{p+1}$$

- Cook's Distance can be computed directly without fitting regression n times
- It measures the influence of a data point and depends on leverage and standardized residual
- Data points with $D_i > 0.5$ are called influential. If $D_i > 1$, the data point is potentially damaging to the regression line

Partial Residual Plots

- The plot of response y vs. predictor x_k can be deceiving
- The reason is that other predictors also influence the response and blur our impression
- We require a plot which only shows the "isolated" influence of predictor x_k on the response y
- We remove the estimated effect of all the other predictors from the response and plot this versus the predictor x_k
$$y - \sum_{k \neq j} x_j \hat{\beta}_j = \hat{y} + r - \sum_{k \neq j} x_j \hat{\beta}_j = x_k \hat{\beta}_k + r$$
- Partial residual plots show the marginal relation between a predictor x_k and the response y after the effect of other variables is accounted for

Checking for Correlated Errors

- If the errors/residuals are correlated, the OLS procedure still results in unbiased estimates of both the regression coefficients and fitted values

- But the OLS estimator is no longer efficient (i.e. there are alternative regression estimators that yield more precise results)
- The standard errors for the coefficients are biased and will inevitably lead to flawed inference results (i.e. tests and confidence intervals). The standard errors can be either too small (majority of cases) or too large
- Durbin-Watson Test
$$DW = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2} \sim \chi^2$$
  – Problem: the DW-test only detects simple correlation. When more complex dependency exists, it has very low power

Residuals vs. Time/Index

When is the plot OK?

- There is no systematic structure present
- There are no longer sequences of pos./neg. residuals
- There is no back-and-forth between pos./neg. residuals
- The p-value in the Durbin-Watson test is > 0.05

Multicollinearity

- If the columns of X are linearly dependent, then $X^T X$ does not have full rank and its inverse does not exist
- Multicollinearity means that there is not perfect dependence among the columns of X but still the columns show strong correlation
- In these cases, there is a (technically) unique solution but it is often highly variable and poorly suited for practice
- The estimated coefficients feature large or even very large standard errors. Hence, they are imprecisely estimated with huge confidence intervals
- Typical case: the global F-test turns out to be significant, but none of the individual predictors is significant
- Extrapolation may yield extremely poor results

Variance Inflation Factor

$$Var(\hat{\beta}_k) = \sigma_E^2 \cdot \frac{1}{1 - R_k^2} \cdot \frac{1}{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}$$

- σ_E^2: error variance
- $\frac{1}{1 - R_k^2}$: variance inflator factor is obtained by determining R-Squared in a regression, where x_k is the response variable and all other predictors main their role
- $VIF \geq 5$ corresponds to a $R_k^2 \geq 0.8$ and has to be seen as a critical multicollinearity
- $VIF \geq 10$ means that $R_k^2 \geq 0.9$ and hence that dangerous multicollinearity is present

Dealing with Multicollinearity

- Amputation: drop all collinear predictors except one
- Generate new variables

Ridge Regression

- Ridge regression requires centered and standardized predictors as the solution is sensitive to location and scale. Furthermore, it is important not to penalize the intercept

Backward Elimination with p-Values

- The removed variables can still be related to the response. If we run a simple linear regression, they can even be significant. In the multiple linear model however, there are other, better, more informative predictors.
- In a step-by-step backward elimination, the best model is often missed. Evaluating more models can be very beneficial for finding the best one.

AIC/BIC Criteria

- $AIC = -2\max(\log likelihood) + 2q = const + n\log(RSS/n) + 2q$
- $BIC = -2\max(\log likelihood) + \log n \cdot q = const + n\log(RSS/n) + \log n \cdot q$
- AIC/BIC allow for comparison of models that are not hierarchical, but they need to be fitted on exactly the same data points and the response variable needs to be identical
- BIC penalizes bigger models harder, with factor $\log n$ instead of factor 2

Variable Selection with AIC/BIC

- Backward Selection
  - The selection stops when AIC/BIC cannot be improved anymore. Predictors do not need to be significant.
- Forward Selection
  - Forward selection is used with big datasets, where there are too many predictors for the number of observations.
- Stepwise Model Search
  - Not much more time consuming but more exhaustive.
- Exhaustive Search
  - It may happen that the AIC-based model search ends at a local AIC-minimum, rather than the global optimum over all models that exist.
  - For finding the model that globally minimizes AIC, a complete search over all 2^p models is required. Depending on n and computing speed, this is feasible for $p \approx 15 - 20$.
  - R function `regsubsets()` of `library(leaps` does the job, but it cannot deal with factor variables correctly.
- Summary
  - Factor variables either remain with all their dummies, or are entirely removed. There is no in-between solution.
  - If interaction terms are present, then the main effects for these variables must be in the model as well.
  - In case of polynomial terms, all lower order terms must be used as well.

LASSO

- In contrast to the OLS and Ridge estimators, there is no explicit solution to Lasso.

- Using the Coordinate Descent procedure in R allows for finding the solution up to problem size around $np \approx 10^6$.
- In contrast to the OLS and Ridge estimators, Lasso is not a linear estimator. There is no hat matrix H such that $\hat{y} = Hy$.
- As a consequence, the concept of degrees of freedom does not exist for Lasso and there is no trivial procedure for choosing the optimal penalty parameterλ.
- Due to the built-in shrinkage property, Lasso is much less susceptible to multicollinearity. However, too many collinear predictors can still hamper model interpretation in practice.
- Inference on the fitted model is at best difficult, or even close to impossible. One can compare standardized coefficients.
- The standard Lasso only works with numerical predictors. Extension to factor variables exist, see Group Lasso.

Cross Validation

- The same response variable is predicted.
- We can perform cross validation on datasets with different number of observations, or even on different datasets.
- The models which are considered in a comparison need not to be hierarchical, and can be arbitrarily different.
- It is possible to infer the effect of response variable transformations, Lasso, Ridge, robust procedures
- Cross Validation should be used if AIC/BIC and Adjusted R-Squared do not work
  - The response variable is transformed: for investigating whether we obtain better predictions from a model with transformed response or not, cross validation is a must.
  - The sample is not identical: if we need to check whether excluding data points from the fit yields better predictions for the entire sample, we require cross validation.
  - For model comparison, neither tests nor AIC-comparison can serve

Significance vs. Relevance

- The larger a sample, the smaller the p-values for the very same predictor effect. Thus do not confuse small p-values with an important predictor effect!
- With large datasets, we can have statistically significant results which are practically useless.
- Most predictors have influence, thus $\beta_j = 0$ hardly ever holds. The point null hypothesis is thus ususally wrong in practice. We just need enough data so that we are able to reject it.
- Absence of Evidence $\neq$ Evidence of Absence
- Measuring the relevance of predictors
  - maximum effect of a predictor variable on the response

$$|\beta_j \cdot (\max_i x_{ij} - \min_i x_{ij})|$$

  - this can be compared to the total span in the response, or it can be plotted vs. the (logarithmic) p-value
- Another way of quantifying the impact of a particular predictor is by standardizing all predictors to mean zero and unit variance. This makes the coefficients β_j directly comparable.

3 Extending the Linear Regression

3.1 Generalized Additive Modelling (GAM)

Polynomial Basis Functions

- Allows for a flexible and data-adaptive fit
- Demonstrates some erratic behavior at boundaries

Cubic Regression Splines

- Space knots evenly
- Set knots at quantiles of unique x values (e.g. 10%,20%, $\cdots$)
- Start with a large enough number and do backward selection, eliminating knots that seem unnecessary

Penalized Regression Splines

- Generalized Ridge Regression: $\hat{\gamma} = (H^T H + \lambda P)^{-1} H^T y$
- Fitted values: $\hat{y} = H(H^T H + \lambda P)^{-1} H^T y = S_\lambda y$
- Empirical Degrees of Freedom: $df = trace(S_\lambda)$
- Generalized Cross Validation Score: $GCV(\hat{f}_\lambda) = n \cdot (trace(I - S_\lambda))^2 \cdot \sum_{i=1}^n (y_i - \hat{f}_\lambda(x))^2$

Smoothing Splines

`smooth.spline()`

- Uses all knots (if `all.knots=TRUE`)
- Spends more degrees of freedom for the fit
- Only works with one single predictor variable

`mgcv::gam()`

- By default uses thin plate regression splines
- Method can be adjusted by `s(xx, "cr")`, not recommended
- By default does not use all knots (i.e. uses a basis that has a limited dimension for saving computing time for obtaining good performance in multiple predictor settings)

Fitting Multiple Additive Models

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + E$$

- Note that there is an identification problem. We either require that $f_1(x_1) + f_2(x_2) = 0$ or need to omit the intercept β_0, otherwise there is no unique solution

Comparing GAMs

- We can directly compare the GCV scores of different models
- AIC (because of the penalized likehood and the smoothing parameter estimation)
- Statistical testing with hierarchical model comparison
  - Comparing GAM against a parametric model via a hierarchical model comparison is not feasible

Variable Selection: Overview

For `gam()`, there is no `step()` function, we thus cannot perform variable selection in the usual manner. The following alternatives exist:

- Manual Backward Elimination, either by analyzing p-values or using the AIC criterion
- Shrinkage Smoother: using basis functions (`bs="ts"` or `bs="cs"`), where by construction also zero degrees of freedom can be awarded (i.e. terms can be eliminated from the model)
- Nullspace Penalization: using argument `select=TRUE`, which means that also parametric, non-smooth terms will be penalized. This works with all basis functions and performs automatic variable selection

Variable Selection: Shrinkage Smoothers

A fundamental drawback of both Shrinkage Smoothers and Nullspace Penalization is that no selection over the parametric terms is possible

- It is possible to automatically select the flexible terms. But this is of limited use as long as unnecessary parametric terms are part of the model – they interfere with the selection
- Performing manual backward elimination on p-values for the parametric terms and later automatic selection for the flexible terms is not really a coherent procedure
- Hence, the overall best procedure for variable selection in GAM is backward elimination based on p-values or AIC

3.2 Generalized Linear Modelling (GLM)

Logistic Regression

- Log-likelihood: $l(\beta) = \sum_{i=1}^{n} (y_i \log(p_i) + (1-y_i)\log(1-p_i))$ where $p = \dfrac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots)}{1+\exp(\beta_0 + \beta_1 x_{i1} + \cdots)}$
- Under mild conditions, the solution exists, but it cannot be written in closed form. Usually, the IRLS algorithm is employed
- Residual Deviance: $D(y, \hat{p}) = -2\sum_{i=1}^{n} (y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)) = -2 \cdot l(\hat{\beta})$
- Null Deviance: the deviance of the simplest possible model that is built from the intercept term only
- Coefficient of Determination: $R^2 = \dfrac{1-\exp(\frac{D_{res}-D_{null}}{n})}{1-\exp(-\frac{D_{null}}{n})}$
- Comparing hierarchical models: $2(ll^{Big} - ll^{Small}) = D(y, \hat{p}_{Small}) - D(y, \hat{p}_{Big}) \sim \chi^2_{(p-q)}$
- If $D(y, \hat{p}_{Null}) - D(y, \hat{p}_{Big}) \gg (p-q)$, then reject H_0
- Pearson Residuals: $R_i = \dfrac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$
  - R_i^2 is the contribution of the i-th observation to the Pearson statistic for model comparison

  - Pearson residuals are scale-free
- Deviance Residuals: $D_i = sign(y_i - \hat{p}_i) \cdot \sqrt{d_i}$ where $d_i = (y_i \cdot \log(\hat{p}_i) + (1 - y_i) \cdot \log(1 - \hat{p}_i))$
- AIC: $AIC = D(y_i, \hat{p}) + 2p$

Binomial Regression

- Log-likelihood: $l(\beta) = \sum_{i=1}^{k} \left[\log \binom{n_i}{y_i} + y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right]$
- For the binomial distribution, the mean determines the dispersion
- Goodness of fit: $D(y, \hat{p}) = 2 \sum_{i=1}^{k} \left[y_i \log \frac{y_i}{\hat{y}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \hat{y}_i}\right]$
- Asymptotics: If Y_i is truly binomial and the n_i are large, the residual deviance is approximately χ^2 distributed. The degrees of freedom is $k - (\# \textit{ of predictors}) - 1$
- $Deviance \gg df$: model is not worth much (check $df \pm 2\sqrt{df}$ or do the exact computation, only apply this test if $n \geq 5$)
- Overdispersion
$$\hat{\phi} = \frac{X^2}{n - q} = \frac{1}{n - q} \cdot \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}$$
  - regression coefficients remain unchanged
  - standard errors will be different: inference

Poisson Regression

- Poisson regression is useful when the response variable is a count. However
  - for bounded counts, the binomial model can be useful
  - for large numbers, the normal approximation can serve
- Poisson regression is a must if
  - the counts are small and/or population size is unknown
  - the population size is big and hard to come by, and the probability of an event, and thus the expected counts are small
- Link function: $\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- Log-likelihood: $l(\beta) = \sum_{i=1}^{n} (y_i \cdot \log(\lambda_i) - \lambda_i - \log(y_i!))$
- There is no closed form solution and we have to resort to the IRLS approach for approximation
- Residual Deviance: $D = 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{\hat{\lambda}_i} - (y_i - \hat{\lambda}_i)\right] \sim \chi^2_{n-(p+1)}$
- Quick check: *residual deviance* $\gg$ *df*
- Pearson residual: $P_i = \dfrac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \quad \text{approx.} \sim \mathcal{N}(0, 1)$
- Overdispersion: estimates are unbiased but standard errors are wrong
  - Dispersion parameter: $\hat{\phi} = \dfrac{\sum_i (y_i - \hat{\lambda}_i)^2 / \hat{\lambda}_i}{n - (p + 1)}$
- Hierarchical model comparison in GLMs with an additional dispersion parameter need to be carried out using F tests, but not via the χ^2-distribution or z tests