1 Simple Regression

- The LOESS-Smoother is better, more flexible and more robust than the Gaussian Kernel Smoother.
- Log-transformation on both response and predictor: 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!

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

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

- 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 Regression

- Necessary but not sufficient condition for the full rank of X: $p < n$
- Gauss-Markov Theorem: The OLS regression coefficients are unbiased, and they have minimal variance among all estimators that are linear and unbiased.
- If the errors are i.i.d. and follow a Normal distribution, the estimated regression coefficients and the fitted values will also be normally distributed.
- Adjusted R-squared is always (but often irrelevantly) smaller than multiple R-squared.
- 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 effort on the response. **Reason**: correlated predictors!
  - Multiple testing problem: when doing many tests, the total type I error (i.e. false rejection) increases and we may observe spuriously significant predictors. (Overall type I error: $1 - (1 - p)^m$)
- Prediction: only interpolation (i.e. prediction within the range of observed y-values works well); extrapolation yields non-reliable results.
- 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.
- Inference with Categorical Predictors
  - Do not perform individual hypothesis tests on factors that have more than 2 levels, they are meaningless! We use `anova()`.
- Residuals vs. Errors
  - 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

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

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

- Tukey-Anscombe-Plot: Residuals vs. Fitted
  - If $E[E_i] \neq 0$, the response/predictor relation might be nonlinear or some important predictors/interaction may be missing
  - If non-constant variance of E_i, the smoother deviates from 0
  - When is the plot OK?
    * the residual scatter around the x-axis without any structure (i.e. constant variance)
    * the smoother line is horizontal without systematic deviation (i.e. zero expectation)
    * there are no outliers
  - Systematic error (i.e. the smoother deviates from the x-axis and hence $E[E_i] \neq 0$)
    * Log-transformation on the response and/or predictors
    * Omitted variables (novel variables, higher polynomials, interaction terms)
  - Non-constant variance
    * Transformations!

- Normal Plot
  - Identifying non-iid or non-Gaussian errors
  - When is the plot OK?
    * No systematic deviation from line which leads to the 1st and 3rd quantile
    * A few data points that are slightly "off the line" near the boundaries are often encountered and usually tolerable
    * Long-tailed but symmetrical residuals are not optimal either, but often tolerable. Alternative: robust regression!

- Scale-Location-Plot: $\sqrt{|\tilde{r}_i|}$ vs. $\hat{y}_i$
  - Identifying heteroscedasticity
  - If $Var(E_i) \neq \sigma_E^2$, use transformations or weighted regression!
  - When is the plot OK?
    * The smoother line runs horizontally along the x-axis, without any systematic deviations.

- Leverage-Plot: $\tilde{r}_i$ vs. H_{ii}
  - Leverage: H_{ii}
    * Leverage points are different from the bulk of data
    * The average value for leverage is given by $\frac{p+1}{n}$
    * We say a data point has high leverage if $H_{ii} > \frac{2(p+1)}{n}$
  - Standardized residuals: $\tilde{r}_i$
  - 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 the regression n times.
    - It measures the influence of a data point and depends on leverage and standardized residuals.
    - Data points with $D_i > 0.5$ are called influential. If $D_i > 1$, the data point is potentially damaging to the regression problem.
  - Identifying outliers, leverage points, influential observations and uncritical data points at one and the same time
  - When is the plot OK?
    - No extreme outliers in y-direction
    - High leverage (i.e. $H_{ii} > \frac{2(p+1)}{n}$) is always potentially dangerous, especially if it appears in conjunction with large residuals
  - How to deal with influential observations (i.e. $D_i > 0.5$)?
    - Check the data for misprints, typos
    - Influential observations often appear if the input is not suitable (i.e. predictors are extremely skewed, log-transformations were forgotten)
    - Simply omitting these data points is always a delicate matter and should at least be reported openly
    - Influential data points tell much about the benefits and limits of a model and create opportunities and ideas for how to improve the model

- Partial Residual Plots
  - The plot of response y vs. predictor x_k can be deceiving because other predictors also influence the response and thus blur our impression
  - Partial residual plots show the marginal relation between a predictor x_k and the response y, after/when the effect of the other variables is accounted for
  - When is the plot OK?
    - If the red line with the actual fit from the linear model, and the green line of the smoother do not show systematic differences
  - What to do if not OK?
    - Using transformations; including additional predictors; adding interaction terms
    - Using GAM

- Correlated Errors
  - If the errors/residuals are correlated
    - OLS procedure still leads to unbiased estimates of both regression coefficients and fitted value
    - OLS estimator is no longer efficient
    - The standard errors for the coefficients are biased and wiill inevitably lead to flawed inference results (i.e. hypothesis tests and confidence intervals)
    - The standard errors can be either too small (majority of cases), or too large
  - Residuals vs. Time/Index
    - When is the plot OK?
      - No systematic structure
      - No long sequences of positive/negative residuals
      - No back-and-forth between positive/negative residuals
      - The p-value in Durbin-Watson test is > 0.05
    - What to do if not OK?
      - Adding omitted variables
      - Using generalized least squares method (GLS)
      - Estimated coefficients and fitted values are unbiased but confience intervals and tests are biased!
  - Autocorrelation of Residuals
  - Durbin-Watson-Test

$$DW = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2}$$

    * Under the null hypothesis, no correlation. The test-statistic has a χ^2-distribution. The p-value can be computed.
    * The DW-test is somewhat problematic, because it can only detect simple correlation structure. When more complex dependency exists, it has very low power.

- Multicollinearity
  - The columns X show strong correlation (but not perfect dependence).
    * In these cases, there is a (technically) unique solution, but it is often highly variable and poor for practical use
    * The estimated coefficients feature large standard errors (wider 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
  - Analyzing all pairwise correlation coefficients among predictors will not identify all situations where there is multicollinearity because multicollinearity is not always a pairwise phenomenon
  - Dealing with multicollinearity
    * Amputation: among all colinear predictors, all except one will be discarded
    * Generating new variables
    * For factor variables, use generalized VIFs. If the squared $GVIF^{1/(2df)}$ exceed the usual threshold rules (i.e. > 5, > 10), then we have overly redundant factor variables among our predictors

- 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!
  - `lm.ridge()` does neither provide fitted values nor residuals. These have to be sorted out by oneself via the design matrix
  - Residual plots have to be generated by hand too

- Variable selection is not a method! The search for the best predictor set is an iterative process. It also involves estimation, inference and model diagnostics

- 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 and more informative predictors
  - In a step-by-step backward elimination, the best model is often missed.

- 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

- Backward Elimination with AIC/BIC
  - The selection stops when AIC/BIC cannot be improved anymore. Predictors do not need to be significant

- Exhaustive Search
  - 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 variable correctly

- Guidelines for Variable Selection
  - 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 models as well
  - In case of polynomial terms, all lower order terms must be used as well

  - Stick to these rules when using manual selection procedures
  - `regsubsets()` does not obey to these rules, while `step()`, `drop1()`, `add1()` do

- LASSO
  - Coefficients are artificially shrunken towards zero and hence biased. The benefit is that they are less variable (smaller standard errors)
  - No explicit closed-form solution to LASSO.
  - In contrast to the OLS and Ridge estimators, the solution is found via numerical optimization. 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 result, the concept of degrees of freedom does not exist for LASSO and there is no trivial procedure for choosing the optimal penalty parameter λ
  - Inference on the fitted model is difficult, or even impossible (biased). One can compare standardized coefficients
  - The standard LASSO only works with numerical predictors. Extension to factor variables exist (see Group LASSO)

- Cross Validation
  - The only key point is that the same response variable is predicted.
  - We can perform cross validation on datasets with different number of observations, or even on different datasets
  - Cross validation allows for comparison of models that are not hierarchical, and that can be arbitrarily different
  - It is possible to infer the effect of response variable transformations, LASSO, Ridge, robust procedures
  - AIC/BIC and Adjusted R-squared do not work if
    * We want to investigate whether we obtain better prediction from a model with transformed response or not
    * We want to check whether excluding data points from the fit yields better predictions from the entire sample
    * We want to compare performance of alternative methods such as LASSO, Ridge or Robust Regression

- Significance vs. Relevance
  - The larger a sample, the smaller the p-values for the very same predictor effect. Do not confuse small p-values with an important predictor effect!
  - With large datasets,
    * statistically significant results which are practically useless
    * most predictors have influence, thus $\beta_j = 0$ hardly ever holds
    * the point null hypothesis is thus usually wrong in practice
  - Absence of Evidence $\neq$ Evidence of Absence
  - Measuring the relevance of predictors
    * Maximum effect of a predictor variable on the response:

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

    * Standardizing all predictors to mean zero and unit variance, which makes the coefficients β_j directly comparable
    * More sophisticated approach - LMG criterion

3 Extending the Linear Model

- Penalized Regression Splines
  - Fitting Penalized Regression Splines is a parametric problem that can be solved analytically in closed form via Generalized Ridge Regression
  - Choosing λ such that the Generalized Cross Validation Score (GCV) is minimized
- `smooth.spline()` vs. `gam()`
  - `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)

4 Exam S21

- Confidence Intervals vs. Prediction Intervals
  - The CI characterizes the variability in the fitted value, not the future response value
  - The PI says that if we use another sample from the same population, the response values will be lying within this interval
  - The PI accounts for the scatter of the data points around the regression line, thus PI is always wider than the corresponding CI
- `gam()`: which variable is penalized the most?
  - The variable with smaller `edf` is penalized harder
- In a regression model where factor variables that have > 2 levels and/or interaction terms are present, the `summary` function does not provide useful information for variable selection. Alternatives: `drop1()`, `anova()`
- Because the first model is fully parametrized with all pairwise interactions terms, hence can be separated into models for each species.
- It is possible that the model obtained with `direction='both'` is larger than the one obtained by using `direction='forward'`
- Under Gaussian errors the OLS estimator is the maximum likelihood estimator.
- Split the data into K roughly equal-sized parts in cross validation (True).
- For any $\lambda \in (0, \infty)$, the minimizer of the smoothing splines is always given by cubic splines with knots at each data point.
- Logistic regression model can be viewed as a model where we try to find a relation between the conditional expected value of response y and the predictors.

5 Exam S20

- exp makes the distribution right-skewed and therefore the mean is bigger than the median
- By comparing this formula to the one for the AIC (above), one can notice that for any $n > 7$, $\log(n) > 2$, hence BIC penalizes for model size stronger than AIC
- If the observed response of a data point is increased by 1, then the new fitted value of this data point is equal to the old fitted value times the leverage of the data point
- The `anova` function can be employed to test two Gaussian Kernel smoother fits with different bandwidths.
- When fitting a model using regression splines, setting a knot at each data point would result in overfitting (True)
- Generally, for a twice differentiable piecewise cubic fit with k knots, there are $k + 4$ basis functions.

6 Exam W20

- A global test of the fitted model against the null model in logistic regression - χ^2_{p-q}
- If we incorporate the dispersion parameter in binomial model, the following features of fitted object do not change:
  - the estimated coefficients
  - the fitted values
  - any prediction
  - degrees of freedom
  - residual deviance
  - AIC score of the model
  - log-oods
  - $\hat{y}(= \hat{p}n)$
- The running mean smoother is not robust to outliers
- The function `step()` can not perform variable selection based on p-values.
- We do not need values of the response variable to compute the leverages of the individual data points
- The Cook's Distance tells us how strongly a data point may force the regression line to run through it (False)
- It is mandatory to use a non-robust smoother in the TA plot in logistic regression
- Type I and Type II error

7 Exam S19

- Even in a well specified least squares regression we expect the residuals to be correlated.
- Diagnostic plots

8 Exam W19

- VIFs are strictly more informative than pairwise correlation plots; If predictors are correlated pairwise, their VIF is high, but the reverse implication does not hold.
- Ridge regression is not appropriate as it does not force predictors out from the model and hence does not change the VIFs.
- Leverage of point i is given by H_{ii}, where H is the hat matrix.
- Sum of the leverages of all data points:

$$\sum_i H_{ii} = Tr(H) = Tr(X(X^T X)^{-1} X^T) = Tr((X^T X)^{-1} X^T X) = Tr(I) = p$$

- Interaction term between categorical variable and numerical variable