Statistical Modelling
Emanuel Nussli, page 1 of 2

1 Linear Models

Partial residuals
$\epsilon_{xj,i} = y_i - x_i^T\hat\beta + \hat\beta_j x_{ji}$ $Y = \beta_j X_j + \sum_{i\neq j}(\beta_i - \hat\beta_i)X_i + \epsilon \approx \beta_j X_j + \epsilon$. All effects of the other covariates on y are regressed out.

Normal equations
$LS(\beta)$ typically strict convex in β and $LS(\beta) = (Y-X\beta)^T(Y-X\beta) = 2(-X)^T(Y-X\beta) = 0 \to \beta = (X^TX)^{-1}X^Ty$ given (X^TX) is invertible, ie. columns of X are lin. independent and $n>p$. X^TX are the scalar products of the columns of X. If X has full rank, e.g $rank(X)=p$, the LS solution is unique.

Orthogonalization
$\tilde Y = Y - \overline{Y}$ and $\tilde X = X - \overline{X}$ s.t the are mean-zero and then a regression w.o intercept which gives us the same slope as with ordinary calculation.

Regression through the mean
$\frac{y-\overline{y}}{\hat\sigma_y} = \hat\rho_{x,y}\frac{x-\overline{x}}{\hat\sigma_x}$ which means that y is closer to its mean than x

Projections
$r = \hat\epsilon = y - \hat y = (I-P)y = Qy$ with Q being idempotent and $tr(Q) = n-p$ and $QP = PQ = 0$ as they are antiprojections. Further: $\hat Y = X\hat\beta = X(X^TX)^{-1}Y = PY$ with P also being a projection and $tr(P) = tr(I_{p\times p}) = p$. $r \perp X^{(j)}$ s.t $X^Tr = 0$ and $X^T(Y-X\hat\beta)=0$. We have $P = X(X^TX)^{-1}X^T$ with $P^n = P$ and $P^T = P$ and its a projection from $\mathbb{R}^n$ to $\mathbb{R}^p$.

Link to MLE
$f(y_1,...y_n) = \prod_{i=1}^n 1/\sigma\phi\left(\frac{y_i - \sum_{i=1}^p \beta_i Xij}{\sigma}\right)$ which is the joint density of the standardised ys and that is the liklihood function if treated as function of β,σ^2. The MLE estimators solve the minimization of $-\log f(y_1,...y_n)$. We get the same for β if the errors are i.id Gaussian but a different result for σ^2 ($\frac{1}{n}$), which is biased. Optimization is done seperately, which is called Gaussian decoupling.

Partial correlations
$\beta_j = parcor(Y, X^{(j)} \mid \{X^{(h)}; h\neq j\}\frac{\Omega_{jj}^{-1}}{\Omega_{yy}^{-1}}$ with Ω being the covariance matrix and the last term being a scaling factor. β_j measures effect of $X^{(j)}$ on Y which is not explained by the other Xs by the part of $X^{(j)}$ that is not explained by the other Xs. If the Xs are orthogonal, we can run multiple single regressions. Recipe to get β_j: (i) regress X^j on all the other Xs to get residuals Z^j (ii) regress Y on all the other Xs to get residuals R^j (iii) regress R^j on Z^j s.t $\hat\beta_j = \frac{(Z^j)^TR^j}{(Z^j)^TZ^j} = \frac{(Z^j)^TY}{(Z^j)^TZ^j}$

Properties of LS
- $\mathbb{E}[\hat\beta] = \beta; \mathbb{E}[\hat\epsilon] = 0; \mathbb{E}[\hat Y] = \mathbb{E}[Y] = X\beta$
- $Cov(\hat\beta) = \sigma^2(X^TX)^{-1}; Cov(\hat Y) = \sigma^2P$ and $Cov(\hat\epsilon) = \sigma^2Q = \sigma^2(I-P)$ which means that the residuals are correlated and not constant
- $\hat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^n\hat{\epsilon_i^2}$ which is unbiased opposed to the ML estimator.
- If constant Gaussian errors: $\hat y \sim N_n(X\beta, \sigma^2P)$; $\hat\beta \sim N(\beta, \sigma^2(X^TX)^{-1})$; $\hat\epsilon \sim N(0,\sigma^2Q)$; $\frac{\sum_{i=1}^n r_i^2}{\sigma^2} \sim \chi^2_{n-p}$; $\hat\beta$ and $\hat{\sigma}^2$ are independent.

Tests and CIs
$\frac{1}{\sigma}(X^TX)^{0.5}(\hat\beta - \beta) \xrightarrow{distr.} N_p(0,1)$ as $\hat\beta \xrightarrow{distr.} N_p(\beta, \sigma^2(X^TX)^{-1})$ (standardisation). As σ^2 is not known, we plug in $\hat{\sigma^2}$ and it becomes t_{n-p}-distributed. Proof of t_{n-p}-distribution: $Z\sim N(0,1)$ and $U\sim\chi^2_{n-p}$ and are independent, then: $\frac{Z}{\sqrt{U/n-p}} \sim t_{n-p}$. Rewrite using a orthogonal transformation to prove.

- Global F test: testing all parameters (including intercept): $\frac{(\hat\beta-\beta)^TX^TX(\hat\beta-\beta)}{p\hat{\sigma^2}} = \frac{\|X(\hat\beta-\beta)\|^2}{p\hat{\sigma^2}} \sim F_{p,n-p}$ which is the estimated error of the regression surface. For test stat T under H_0, just leave away the βs above.
- Partial F-Test: $\hat v = B\hat\beta$ with $rank(B) = q \le q$ s.t $\frac{(\hat v - v)^T(B(X^TX)^{-1}B^T)^{-1}(\hat v - v)}{q\hat{\sigma^2}} \sim F_{q,n-p}$ which is very relevant if the regressors are highly correlated. $dim(B) = q\times p$
- Golbal F-test: Intercept not included and is found in R software output; $dim(B) = (p-1)\times p$ as the intercept is not tested.
- Comparison of models: $\frac{(SSE_0 - SSE)/p-q}{SSE/n-p} \sim F_{p-q,n-p}$
- $\sqrt{q_{1-\alpha;F_{1,n-p}}} = q_{1-\alpha/2;t_{n-p}}$
- new observation: $\frac{\hat y_0 - \mathbb{E}[y_0]}{\hat\sigma\sqrt{x_0^T(X^TX)^{-1}x_0}} \sim t_{n-p}$ and plus 1 in the denominator for $\hat{y_0}$.
- expectation of i-th observation: $\frac{\hat{y_i}-\mathbb{E}[y_i]}{\hat\sigma\sqrt{P_{ii}}} \sim t_{n-p}$

ANOVA
$\frac{(B\hat\beta - b)^T(B(X^TX)^{-1}B^T)^{-1}(B\hat\beta - b)}{(p-q)\hat{\sigma^2}} \sim F_{p-q,n-p}$ if we test $p-q$ restrictions, i.e $dim(B) = (p-q)\times p$ and $dim(b) = p\times 1$ and B is of full rank. $SSE/n-p$ is an unbiased estimator for σ^2 such as $(SSE_0 - SSE)/(p-q)$ is aswell. If H_0 is true, we expect this difference to be larger than σ^2. We thus arrive at the foundation of ANOVA: $\|y-\hat y^{(0)}\|^2 = \|y - \hat y\|^2 + \|\hat y - \hat y^{(0)}\|^2$ **Anova table** Regression has $p-1$ dfs, error has $n-p$ df and the total around overall mean has $n-1$ dfs and the mean square being their l_2-norm diveded by the dfs. $\mathbb{E}[MSE_{Reg}] = \sigma^2 + \frac{\|\mathbb{E}[y] - \mathbb{E}[\overline{y}]\|^2}{p-1}$

Coefficient of Determination
$R^2 = \frac{\|\hat y - \overline{y}\|^2}{\|y-\overline{y}\|^2} = \frac{SSReg}{SSTotal} = \frac{SSTotal - SSError}{SSTotal} = 1 - \frac{SSE}{SST} = max\{\hat\rho(Y,\hat Y)^2\} = \hat\rho^2_{XY}$ in a simple regression.

Fisher Transformation
Transformation of the distribution of the correlation into $N(0,1)$ as $\hat\rho \approx N(\cdot,\cdot)$ being no pivot as the variance depends on ρ itself. Variance stabilizing transformation: $z = tanh^{-1}(\hat\rho) = 0.5\log\left(\frac{1+\hat\rho}{1-\hat\rho}\right) \approx N(tanh^{-1}(\rho), \frac{1}{n-3})$ and $CI(z) = 1 \pm z_{1-\alpha/2}\frac{1}{\sqrt{n-3}}$ and $tanh(CI)$ for $CI(\rho)$. Z-transform compresses in the middle and stretches the edges. If ρ is near 0, the variance of $\hat\rho$ is big and vice versa.

Correlations
$r_{Spear} = 1 - \frac{6\sum_{i=1}^n rank(X_i) - rank(Y_i)}{n(n^2-1)}$ if X_i, Y_i are integers. Pearsons correlation coefficient between ranked variables. $r_{Kend} = 2\frac{T_k - T_d}{n(n-1)}$ or $\frac{T_k - T_d}{\binom{n}{2}}$ where T_k =number of pairs with $(X_i - X_j)(Y_i - Y_j) > 0$ and vice versa.
Partial Correlations: correlation of X and Y conditional on Z with $\rho_{XY\cdot z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho^2_{XZ}}\sqrt{1-\rho^2_{YZ}}}$ which corresponds to the regression coefficient up to a scaling factor, i.e $Y = \beta X + \gamma Z + \epsilon$ and $\rho_{XY\cdot Z} = \beta\frac{\sqrt{\sum_{11}^{-1}}}{\sqrt{\sum_{22}^{-1}}}$ with the scaling being strictly positive and $\sum$ the covariance matrix. Use the Fisher transform for inference, i.e $tanh^{-1}(\hat\rho_{XY\cdot Z}) \approx N(tanh^{-1}(\rho_{XY\cdot Z}), \frac{1}{n-3-1})$. To condition for another variable, use iteratively $\rho_{XY\cdot Z_1, Z_2} = \frac{\rho_{XY\cdot Z_1} - \rho_{XZ_2\cdot Z_1}\rho_{YZ_2\cdot Z_1}}{...}$ and for the variance another -1. The partial correlations are symmetric and scaled in contrast to regression. We also have $sign(\hat\rho) = sign(\hat\beta)$.

Analyzing Residuals
QQ-plots: $u = F_n(x) = \frac{1}{n}\#\{X_i \le x\}$ which is a step function that converges to the true distribution for large n. If the X_i are normally distributed, we have that $F_n(x) \to \phi(\frac{x-\mu}{\sigma})$ and hence $z = \phi^{-1}(F_n(x))$ and hence $z \approx \frac{x-\mu}{\sigma}$ which is a linear function for sufficiently large n. Heavy tailed distributions have tails below the line, skewed distributions are curved and outliers can also be found in the QQ-plots. **Tukey-Anscombe plot:** plot residuals against fitted values where we always have $\sum_{i\in I} r_i\hat y_i = 0$, i.e a sample correlation of zero (if there is an intercept). We want to see no structure. If the spread of the residuals increases linearly with the fitted values, we want to use a logarithmic transformation. If the residuals increase with the sqrt of the fitted values, apply a sqrt transformation to Y. If there is a parabolic behaviour, include a quadratic term. **Durbin Watson test** $T = \frac{\sum_{i=1}^{n-1}(r_{i+1}-r_i)^2}{\sum_{i=1}^n r_i^2} \approx 2\left(1 - \frac{\sum_{i=1}^{n-1} r_ir_{i+1}}{\sum_{i=1}^n r_i^2}\right)$ where we have the serial correlation on top.If independent, we expect a value of 2 and otherwise lower. Only focuses on residuals that immediately follow. **Run-test:** cont. runs of residuals with the same sign; independence assumes Bernoulli with $p=0.5$.

Generalized least squares; weighted Regression
$Y = X\beta + \epsilon, \epsilon \sim N(0,\sigma^2\sum)$ and hence $\hat\beta \sim N(\beta, \sigma^2(X^TX)^{-1}X^T\sum X(X^TX)^{-1})$ and $\sum$ being known and positive definite. We propose $\tilde Y = \sum^{-1/2}Y$ s.t $\tilde Y = \tilde X\beta + \tilde\epsilon$ which we solve to get $\hat\beta_{GLS} = \underset{\beta}{\operatorname{argmin}}(Y-X\beta)^T\sum^{-1}(Y-X\beta) = (\tilde X^T\tilde X)^{-1}\tilde X^T\tilde Y = (X^T\sum^{-1}X)^{-1}X^T\sum^{-1}y$. It is the same as performing LS on the orginal data but with a different scalar product, s.t all the properties remain. We have $\hat\beta \sim N_p(\beta, \sigma^2(X^T\sum^{-1}X)^{-1})$. If $\sum \neq I$, GLS method has a smaller variance than the standard least squares. **Heteroskedasticity:** $\sum = diag(v_1,...,v_n)$ with $\sum^{-1} = diag(v_1^{-1},...,v_n^{-1})$ reduces the problem to weighted LS as we can rewrite $\hat\beta_{GLS=WL} = \underset{\beta}{\operatorname{argmin}}\sum_{i=1}^n v_i^{-1}(y_i - X_i^T\beta)^2$ with $w_i = v_i^{-1}$ being the weights and the importance of X_i being small, if $Var(\epsilon_i)$ is big as $Var(\epsilon_i) = \sigma^2v_i$, where we are still left with n unknowns for the covariance matrix. We therefore try a parametric model for the matrix: $\sigma_i^2 = v_\theta = \exp(v + \sum_{j=1}^p\gamma_j\log|X_{i,j}|)$ where $j=2$ if there is an intercept. We can estimate that via nonlinear LS, e.g $\hat\theta = \underset{\theta}{\operatorname{argmin}}\sum_{i=1}^n(r_i^2 - v(x_i))^2$ or by linearizing the parametric model and running LS via $\hat\theta = \underset{\theta}{\operatorname{argmin}}\sum_{i=1}^n(\log(\max\{\delta^2, r_i^2\}) - (v + \sum_{j=1}\gamma_j\log|X_{i,j}|))^2$ with δ being a tuning parameter because of log. We then use the **alternating algorithm** to do WLS with the estimated weights. **Estimating** $\sum$: requires estimating $n(n+1)/2$ parameters which is impossible with n data points.

Toeplitz-Covariance: $\begin{pmatrix} 1 & \rho & \rho^2 & .... \\ \rho & 1 & & \\ \rho^2 & & & \end{pmatrix}$ which shows exponential decay of ρ as e.g time processes.

Alternating Optimization:
(i) compute $\hat\beta$ via OLS and residuals (ii) estimate $\hat\sum$ via e.g MLE (iii) GLS with $\hat\sum$ and compute new $\hat{\beta_{GLS}}$ and residuals (iv) repeat until convergence

Huber-White Sandwhich for Heteroskedasticity $\hat\sum = diag(r_1^2,...,r_n^2)$ One can show that $n(\widehat{Cov}(\hat\beta) - Cov(\hat\beta)) \to 0$ in probability meaning that $\frac{\hat\beta - \beta}{\widehat{Cov}(\hat\beta)}$ converges in probability to $N_p(0,I)$. The covariance estimator is thus $\widehat{Cov}(\hat\beta) = (X^TX)^{-1}X^T\hat DX(X^TX)^{-1}$. Not always used as normal LS yields smaller variances.

Gauss Markov Theorem
BLUE: best linear unbiased estimator: $\hat{\beta_{GLS}}$ is BLUE w.r.t Y and $\tilde Y$ if zero-mean, constant errors, linear data and $rank(X) = p$.
UMVU: assume Gaussian constant errros on top of assumptions above: uniformly minimum variance unbiased: no linearity needed!
(i) if errors non-gaussian: non-linear estimators can be way better (ii) if distribution of errors is known up to unknown parameters, e.g $\epsilon_i \sim \sigma t_v$ with v unknown, MLE is asymptotically better (nonlinear in Y) (iii) if distribution of error is unknown we use robust methods. $\hat\beta_{GLS}$ is also UMVU if the errors are constant Gaussian.

Model Selection
Loss-function for submodel M: $\mathbb{E}[\|Y^{\hat M} - X\beta\|^2/n] = \mathbb{E}[\|X(\hat\beta^M - \beta)\|^2/n] = n^{-1}\sum_{i=1}^n\left(\mathbb{E}[X_i^{MT}\hat\beta^M] - X_i^T\beta\right)^2 + n^{-1}Var(X_i^{MT}\hat\beta^M) = \text{bias}^2 + \sigma^2\frac{q}{n}$ with $dim(M) = q$ by the bias variance decomposition. That describes the Bias-Variance tradeoff (which holds for non-linear models aswell).
Stepwise Regression: greedy algorithms
Forward: start with empty model, choose $X^{(j)}$ with smallest p-value with t-test (equivalent to comparing models with partial F-test). Do until no significant covariate is found.
Backwards: Same as above but backwards. **Remarks**: dangerous as significance of variables has got nothing to do with best model; forward is computationally somewhat more efficient than backward and can be used if $p >> n$; problem if variables are correlated; p-values no longer valid (selection and multiple testing); $2^p - 1$ possible models which is NP with $\mathcal{O}(exp)$, forward and backward may give entirely different results. Backwards doesnt work if $p\ge n$. Forward and backward can be combined using different significance levels.

Mallow's Cp statistic
$\mathbb{E}[\|\hat Y^M - \mu\|^2] = SMSE = \sigma^2|M| + \text{bias}^2$ which is the same as above but w.o n^{-1}, which is not important. We know want to estimate σ^2 and bias^2 to get at our loss function. $\mathbb{E}[SSE_M] = \mathbb{E}[\|Y - Y^{\hat M}\|^2] = \sum_{i=1}^n Var(y_i - \hat{y_i}) + \sum_{i=0}^n(\mathbb{E}[\hat{y_i}^M] - \mu_i]^2 = \sigma^2(n-|M|) + \text{bias}^2$. We can thus write bias$^2 = SSE - \sigma^2(n-|M|)$ and we see that the **bias is unbiased** as $\mathbb{E}[\text{bias}^2] = \text{bias}^2$.
We have en estimator for SMSE: $\widehat{SMSE} = \sigma^2|M| + SSE - \sigma^2(n-|M|)$ and standardized $\frac{SMSE_M}{\sigma^2} = \frac{SSE_M}{\hat{\sigma_p^2}} + 2|M| - n = \hat\Gamma_p(M)$ where $\mathbb{E}[\Gamma_p(\hat M)] \approx \Gamma_p(M) = |M|$ if M is true as we have no bias. Only approx as $\mathbb{E}[\hat{\sigma^2}] \neq \sigma^2$. **Summary** We want a small Γ and one thats roughly $|M|$ and $\Gamma >> |M|$ suggests that we use the wrong model. Minimizing SMSE or Γ_p or SPSE always leads to the same model. C_p is an estimate of Γ_p. Mallows C_p can only lay underneath $|M|$ due to random fluctuations or misspecification. **AIC** $AIC(\alpha) = -2\hat l(k) + \alpha k \to$ MLE territory, which means $\frac{1}{n}$ for the variance and BIC: $\alpha = \log(n)$. Γ_p and the AIC are very similar (Taylor-expansion of AIC) if $SSE(M)/n$ is very near of σ^2 used in computing Mallows C_p or here Γ_p. **Taylor expanded AIC:** $\approx n\log(\sigma^2) + \frac{SSE}{\sigma^2} - n + 2|M|$.

Generalized linear models
Logistic Regression
$Y_i \in \{0,1\}$ independent with $Y \sim Bern(p_i)$ with two choices for link functions, i.e logit: $[0,1]\to\mathbb{R}$ and $p \to \underbrace{\log\left(\frac{p}{1-p}\right)}_{\text{logit}}$ and thus $p(x_i) = \frac{exp(x_i^T\beta)}{1+exp(x_i^T\beta)}$ or probit: $[0,1]\to\mathbb{R}$ and $p\to\phi^{-1}(p)$ and hence $p(x_i) = \phi(X_i^T\beta)$. There is not a big difference with logit more popular and the variable in logit having to be rescaled by $Var(x_i)^{-1} = \sqrt{3/\pi^2}$ to be able to compare. Logit is comp. easier and has canonical link function.
Estimation of β MLE gives
$\hat\beta = \underset{\beta}{\operatorname{argmin}}\sum_{i=1}^n\left(\mathbb{1}_{Y_i=1}X_i^T\beta + log(1+exp(X_i^T\beta))\right)$ which is typically strictly (strictly if $(X^TX)^{-1}$ exists) convex in β and can thus easily be solved. $\hat\beta$ satisfies $\sum_{i=1}^n(y_i - \mathbb{P}[Y_i = 1\mid X_i])X_i = 0$
IRLS
Idea: computing $\hat\beta$ via IRLS. (i) initialize $\hat{p_i} = 0.99$ if $Y_i = 1$ and complement if not. (ii) Taylor expansion of $logit(\hat{y_i})$ around $\hat{p_i}$ with $z_i = logit(\hat{p_i}) + logit'(\hat{p_i})(y_i - \hat{p_i}) \approx X_i^T\hat\beta + \frac{1}{\hat{p_i}(1-\hat{p_i})}(y_i - \hat{p_i})$ which is a simple regression with non-constant error variance (iii) Do WLS with weights $w_i = v_i^{-1} = \hat{p_i}(1-\hat{p_i})$ and compute $\hat\beta$ (iv) compute $\hat{p_i} = logit^{-1}(X_i^T\hat\beta)$ (v) repeat until convergence and see that $\beta_{IRLS} = \beta_{MLE} \approx N(\beta, (X^TWX)^{-1})$ with $W = diag(p_i(1-p_i))$
Tests and CIs
$CI(\beta_j) = \hat{\beta_j} \pm \phi^{-1}(1-\alpha/2)\hat{s.e}(\hat{\beta_j})$ with $\hat{s.e}(\hat{\beta_j}) = \sqrt{(X^T\hat WX)^{-1}_{jj}}$ and testing as usual as standardized RV is asymptotically $N(0,1)$ under the Null. The asymptotics get very bad as $p(x_i)$ gets steep (deterministic). **Analogue of Partial F Test**: $T = 2(l(\hat{\beta_p}) - l(\hat{\beta_q})) \approx \chi^2_{p-q}$ under the Null with $q = dim$(small model). Called the log-likelihood ratio test. **Null deviance:** full model has $p=n$; $-2l(\hat\beta_{intercept})$ with $n-1$ df and **residual deviance:** $-2l(\hat\beta_{full})$ with $n-p$ df. $-2loglikelihood$ is a goodness of fit measure analogue to RRS in linear models. **Pearson-residuals:** $\frac{y_i - \hat{p_i}}{\sqrt{\hat{p_i}(1-\hat{p_i})}}$ which are standardized residuals. **Deviance residuals:** $D : i = \sqrt{-2(y_i\log(\hat{p_i}) + (1-y_i)\log(1-\hat{p_i}))}s_i$, where $s_i = 1\{y_i=1\}$ which is the square root of the i-th term of $-2l(\cdot)$ and $\sum_{i=1}^n D_i^2 = -2l(\hat\beta)$ where D_i come straight from the MLE estimation. **R working residuals:** $\frac{y_i - \hat{p_i}}{\sqrt{\hat{p_i}(1-\hat{p_i})}}$ from the last iteration of IRLS, but they are weird as we will obviously have structure in the residuals.

General case of GLMs
We want to model $\mathbb{E}[Y_i] = \mu_i$. We have a link function g s.t $g(\mu_i) = X_i^T\beta$. Link functions: Poisson: $\log(\lambda_i) = X_i^T\beta$, as we have to transfer to the positive real numbers. Negative Binomial: $\log(\mu_i) = X_i^T\beta$, which is not canonical. **Exponential family:** $p(y_i\mid x_i) = y_i\beta_i + c(\beta_i) + log(h(y_i)), i = 1,...,n$. Its holds for these distributions that $\mathbb{E}[Y_i] = \mu(\beta_i) = -c'(\beta_i)$. If we have that $g = \mu^{-1}$, we have a canonical link function (like in linear model with Gaussian errors, logistic regression). Exponential family is very important as MLE is very much about the score (derivative -log-likelihood), which is convenient for the exponential family. **Poisson vs. Negative Binomial:** $\mathbb{E}(Y_i) = Var(Y_i) = \lambda$ if Poisson distribution and $\mathbb{E}[Y_i] = \frac{rp}{1-p} = \mu$ and $Var(Y_i) = \mu + \mu^2/r$, meaning $Var > E$, which is called **overdispersion**. The factor $\frac{1}{\alpha} = r$ is the dispersion parameter. Negative Binomial models the nr. of successes until r failures depending on r and p. Take-Away: If one has overdispersion in practice, use negative binomial and not Poisson models.

Cox Regression
Response is a survival or failure time T_i (could use GLM with Exp. or Gamma distribution). Cox Regression is semi-parametric (GLM part and nonparametric part). Let $h(t) = \lambda(t) = \lim_{h\downarrow 0}\frac{1}{h}\mathbb{P}(t\le T\le t+h\mid T\ge t) \overset{Bayes}{=} \frac{f(t)}{1-F(t)} = -\frac{d}{dt}\log(1-F(t))$ and hence $F(t) = 1 - exp\left(\int_0^t h(u)du\right)$. Model for $h(t)$: $h(t) = exp(X_i^T\beta)\times h_0(t)$ with the latter term being the base rate failure which is model free in t. Proportional hazard: $\frac{h_i(t)}{h_j(t)} = exp((x_i - x_j)\beta))$ should be constant w.r.t time and is testable.

Partial Likelihood:$\hat\beta = \underset{\beta}{\operatorname{argmin}}\prod_{i=1}^n\frac{exp(X_i^T\beta)}{\sum_{j;t_j\ge t_i}exp(X_j^T\beta)}$ as $h_0(t)$ would enter MLE which is impossible to solve and partial likelihood uses only the order of failures. **Censoring:** we observe $\min\{T_i, C_i\}$, which is called right censoring. No intercept present, as it would be absorbed into $h_0(t)$. A **strictly monotonous and differentiable** transformation of the survival times transform a Cox model into another Cox model with the same parameters and a different $h_0(t)$.

2 Non-Parametric Models

Nonlinear Least Squares
$Y = f(x_i;\beta) + \epsilon_i$ with f being a known non-linear function but the parameters β are unknown. The dimension of the parameter need no longer be of dimension of the explanatory variables. Same assumptions as in LM. Let $\hat\beta = \underset{\beta}{\operatorname{argmin}}\sum_{i=1}^n(y_i - f(x_i;\beta))^2$ and $\hat{\beta_{LS}} = \hat{\beta_{MLE}}$ if the errors are gaussian. Problem: $S(\beta)$, the lossfunction is often non-convex and thus hard to optimize (no closed form; iterative methods with good starting values used). **Estimated error variance:**$\hat\sigma^2 = (n-p)^{-1}S(\hat\beta)$ as we have no intercept. **CIs and tests by asymptotics** as assumptions of normal errors os not enough:$f(x_i;\beta) \approx f(x_i;\beta_0) + a(\beta_0)^T(\beta - \beta_0)$ with $a(\beta_0) = \left(\frac{\partial}{\partial\beta_j}f(x_i;\beta), j = 1,...,p\right)^T$ with with $dim(\cdot) = 1\times p$ and the β thing with $dim(\cdot) = p\times 1$. Under assumptions: $\hat\beta \approx N(\beta_0, \sigma^2\left(A(\beta_0)^TA(\beta_0)\right)^{-1})$ with $A(\beta_0)$. One can compute $a(\beta_0)$ for all $i = 1,...,n$ and stack them into a matrix to get $A(\beta_0)$ with $dim(\cdot) = n\times p$. $CI(\hat{\beta_j}) = \hat{\beta_j} \pm t_{n-p;1-\alpha/2}\sqrt{\hat\sigma^2(A(\hat\beta)^TA(\hat\beta))^{-1}_{jj}}$ but we could use the normal distribution as we are doing asymptototics anyway. **More precise CIs and tests:** $\hat\beta^{-(j)}(\beta_j^*) = \underset{\beta;\beta_j=\beta_j^*}{\operatorname{argmin}}S(\beta)$ which is constrained nl LS at $\beta_j = \beta_j^*$ giving us the **CI:** $CI(\beta_j) = \left\{\beta_j^*; \sqrt{S(\hat\beta^{(-j)})(\hat{\beta_j^*}) - S(\hat\beta} \le t_{n-p;1-\alpha/2}\hat\sigma\right\}$. We arrived at the CI by the test statistic $\tau_j(\beta_j^*) = \frac{sign(\beta_j^* - \hat{\beta_j})}{\hat\sigma}\sqrt{S(\hat\beta^{(-j)}) - S(\hat\beta)}$ while plotting τ_j against β_j^* gives us an idea of the degree of non-linearity. **Contour plots:** plot $S(\beta)$ against β_1 and β_2 which should give us ellipses. The further away from circular, the higher the dependence of the variables with perfect circle if independent. **Profile plot:** $\hat{\beta_1}^{-1}(\beta_2^*)$ versus β_1^* and vice versa should give straight lines if the model is linear as S is quadratic and the derivative therefore linear. The lines intersect at β and are orthogonal if they are independent.Gradient (profile plots) is perpendicular to the contours, i.e profile plots and contours intersect perpendicular. For higher dimensions, only profile plots are done as the

derivation of the profile likelihood is computationally very extensive as it involves $\binom{p}{2}$ optimizations and the profile stuff only p optimization problems with one constraint.

Non-Parametric Regression

$Y = f(x_i) + \epsilon_i$ with f unknown but reasonably smooth (much weaker assumptions). **Kernel:** $\kappa(\cdot)$ is a pdf on $\mathbb{R}$ and often symmetric around 0 and has support $[-1, 1]$ or decreases rapidly. The weights are defined by the kernel, i.e $w_i(x) = \kappa\left(\frac{x-x_i}{h}\right)$ with h being the bandwidth. If covariates X_i are chosen at random, the number of observations with non-zero weights can vary strongly as x varies, which is why one might try KKN.

Nadaraya-Watson kernel estimator:

$\hat{f}_h(x) = \frac{\sum_{i=1}^n \kappa\left(\frac{(x-x_i}{h}\right)Y_i}{\sum_{i=1}^n \kappa\left(\frac{x-x_i}{h}\right)}$ with large h making the curve smoother.

Gasser-Müller estimator:

$\hat{f}_h(x) = \sum_{i=1}^n \int_{s_{i-1}}^{s-i} \frac{1}{h}\kappa\left(\frac{x-x_i}{h}\right)du$ which weights x more if there are not a lot of xs around. Works with the ordered statistics $x_1 \le x_2 \le ... \le x_n$ and $s_0 = -\infty, s_i = (x_i + x_{i+1})/2$ and $s_n = \infty$.

Nearest Neighbors (non-kernel based)

$w_i(x) = \mathbb{1}_{x - x_i \text{ is among the k smallest values}}$. Connection to Kernel-approaches: define $D_i(x) = |x - x_i|$ and h_x is the distance of the k-th nearest neighbor to x (k-th order statistics of $D_i(x)$). We then see that it can be written as $\hat{f}_h(x) = \frac{\sum_{i=1}^n \tilde{\kappa}\left(\frac{x-x_i}{h_x}\right)Y_i}{\sum_{i=1}^n \tilde{\kappa}\left(\frac{x-x_i}{h_x}\right)}$ with $\tilde{\kappa}(u) = \mathbb{1}\{|u| \le 1\}$ or 0.5 to fulfill the assumptions of a density. h_x is thus a variable bandwidth which is small if the xs are dense around x.

Local polynomial estimation

Nadaraya-Watson solves the locally constant local polynomial problem, which can be generalized, i.e $\hat{\beta}(x) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \kappa\left(\frac{x-x_i}{h}\right)\left(Y_i - \sum_{j=0}^p \beta_j (x_j - x)^j\right)^2$ which locally does weighted LS. Can deduce the following easily: $\hat{f}_h(x) = \hat{\beta}_0(x)$ and $\hat{f}'_h(x) = \hat{\beta}_1(x)$ and so on. Does better at boundaries than locally constant Nadaraya-Watson. p is mostly chosen to be odd as that works better (espec. at boundaries).

Smoothing Splines: Repr. Kernel Hilbert Space very popular in ML

$\hat{f}(x) = \underset{f(\cdot)}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i - f(x_i)\right)^2 + \lambda \int_{\mathbb{R}} \left(f''(x)\right)^2 dx$ with $\hat{f}$ being a cubic spline with knots at x_i which can be parametrized. Is a form of penalized LS and linear at fringes (boundaries) and it converges to LS if $\lambda = 0$. It can be restricted to $f(x) = \sum_{j=1}^n J_j(x)\beta_j$ with $N_{k+1} = d_k(x) - d_{k-1}(x)$ with $d_k(x) = \frac{(x-x_k)_+^3 - (x-x_h)_+^3}{x_h - x_k}$. STacking the N_i into a $n \times n$ matrix, we get the optimization problem $\|Y - N\beta\|^2 + \lambda\beta^T\Omega\beta$ with $\Omega_{jk} = \int N_j(x)'' N_k(x)'' dx$ with $dim(\cdot) = n \times n$, which gives $\hat{\beta}_\lambda = (N^T N + \lambda\Omega)^{-1} N^T y$

Bias-Variance Tradeoff

$MSE(x) = \mathbb{E}[(\hat{f}_h(x) - f(x))^2] = (\mathbb{E}[\hat{f}_h(x)] - f(x))^2 + Var(\hat{f}_h(x))$

Asymptotics for Nadaraya-Watson (X_i equispread $\in [0, 1]$, f twice cont. diff., symmetric pdf kernel) gives us:

- $Bias(x) \sim h_n^2 f''(x) C_1(\kappa)$ as long as second derivative exists
- $Var(\hat{f}_h(x)) \sim \sigma_\epsilon^2 \frac{1}{nh_n} C_2(\kappa)$
- combine the two for MSE and take derivative w.r.t h for optimal bandwidth to get $h_{opt} \sim n^{-1/5}\left(\frac{\sigma_\epsilon^2 C_2(\kappa)}{4(f''(x))^2 C_1(\kappa)}\right)$ meaning the bandwidth decreases with rate $(\cdot)^{-1/5}$ as n goes to infinity. By plugging that optimal bandwidth into the MSE, we see: $MSE_{opt}(x) \sim n^{-4/5} C(x)$ with constant depending on f'' compared to linear model with MSE decreasing with rate n^{-1}.
- all cont. kernels are nearly as good as each other
- df of Nadaraya-Watson: $trace(H) = \sum_{r=1}^n \kappa(0/h) / \sum_{s=1}^n \kappa((x_r - x_s)/h)$ which goes for all non-parametric estimators.

Cross-Validation

LOOCV:

$\hat{h}_{CVlo} = \underset{h}{\operatorname{argmin}}\, n^{-1} \sum_{i=1}^n \left(y_i - \hat{f}_h^{(-i)}(x_i)\right)^2$

CV K-fold:

$\hat{h}_{CVf} = \underset{h}{\operatorname{argmin}}\, K^{-1} \sum_{k=1}^K \frac{1}{|I_k|} \sum_{i \in I_k} \left(y_i - \hat{f}_h^{(-Ik)}(x_i)\right)^2$

with K being the number of equal sized $\lfloor n/K \rfloor$ partitions called $I1, I_2, ..., I_k$. CV is an estimator $\sigma^2 + n^{-1}\sum_{i=1}^n \mathbb{E}[(\hat{f}_h(x_i) - f(x_i))^2]$, hence optimizes MSE not only at one point. The error for the Kfold algorithm is typically larger than the one from LOOCV as we leave out more than one data point, making $\hat{f}_h^{(-i)}(x_i) \approx \hat{f}_h(x_i)$ a worse approximation. $Var(CV(h))$ is difficult because the different estimated functions are very much correlated but its a smooth curve as a function of h and $Var(CV_{kfold})$ is typically smaller than $Var(CV_{loocv})$.

Estimating the variance for non-parametric regressions Under the assumptions of Gaussian errors, fixed design and a linear nonparametric estimator $\hat{f}$, we get that $\hat{f}(x)([\hat{f}(x)], Var(\hat{f}(x))$ which is after rescaling and centering w.r.t $f(x)$: $Var(\hat{f}(x))^{-1/2}(\hat{f}(x) - f(x)) \sim N(B(x), 1)$ with $B(x) = Var(\hat{f}(x))^{-1/2}([\hat{f}(x)] - f(x))$. One can show that the bias (if X are equispread and f twice cont. diff., converges in probability to $\mathbb{E}[\hat{f}_{hn}(x)] - f(x) \sim h^2 f''(x) C_1(K)$ and $Var(\hat{f}_{hn}(x)) \sim \sigma^2(nh_n)^{-1} C_2(K)$ and thus the bias term $B(x) \sim h_n^{5/2} n^{1/2} C(f''(x), K)$. If we now choose h_n according to the optimal rate $n^{-1/5}$, we see that the term does not go to zero as $n \to \infty$. We thus choose h smaller, which leads to **undersmoothing**. If we choose $h_n << n^{-1/5}$, we get $Var(\hat{f}(x))^{-1/2}(\hat{f}_{hn}(x) - f(x)) \xrightarrow{d} N(0, 1)$ which also holds for a large class of non-Gaussian errors. By plugging in $\hat{\sigma}^2$, we have a variance to do inference. The estimated variance also converges to the true one as $n \to \infty$ s.t $\widehat{Var(\hat{f}_{hn}(x))}^{-1/2}(\hat{f}_{hn}(x) - f(x)) \xrightarrow{d} N(0, 1)$ which we use to construct **pointwise** CIs.

Curse of Dimensionality

Is dimension increases, the bigger the points are apart, the worse the estimation is. We can show that $MSE_{opt} \sim n^{-\frac{4}{4+p}}$ if f is twice cont. differentiable, which gets bad as p increases, so for $p \ge 4$, its no bueno as the probability of laying iside the unit sphere is $vol(\text{unit sphere}) \times \frac{1}{2^p}$ which gets really small as p gets large. The ratio of diamaters of sphere and cube is $\frac{1}{\sqrt{p}}$. **Dealing with curse of dimensionality:** e.g fit an additive model, s.t $Y_i = \sum_{j=1}^p f_j(x_i^{(j)}) + \epsilon_i$ where the whole function is a sum of p one-dimensional functions. The $MSE(x)$ goes down at a rate of $\sim pn^{-4/5}$ while e.g the linear model at a rate of $\sim pn^{-1}$.

3 High-dimensional models

$p > n$, s.t $\hat{\beta} = (X^T X)^{-1} X^T y$ cant be computed as $rank(X) \neq p$, which means $(X^T X)$ cant be inverted since it has not got full rank. Remember **Smoothing Splines:** $\hat{f}(x) = \underset{f(\cdot)}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i - f(x_i)\right)^2 + \lambda \int_{\mathbb{R}} \left(f''(x)\right)^2 dx$ which is solved by a natural cubic spline of the form $f(x) = \sum_{i=1}^n N_j(x)\beta_j$ which can be solved to get $\hat{\beta} = (N^T N + \lambda\Omega)^{-1} N^T y$ with it being a solution for $\|Y - N\beta\|^2 + \lambda\beta^T\Omega\beta$ with $\Omega = \int N_j''(x) N_k''$ dx and Ω being known.

Ridge Regression

$\hat{\beta}_{RIDGE} = \underset{\beta}{\operatorname{argmin}}\left(\|Y - X\beta\|^2 + \lambda\|\beta\|^2\right)$ which gives $\hat{\beta}_{RIDGE} = (X^T X + \lambda I)^{-1} X^T y$ with $\min\{\text{eigenvalues}(X^T X)\} = 0$ and hence makes sure that $\min\{\text{eigenvalues}(X^T X + \lambda I)\} = \lambda > 0$ if $p > n$ and that the estimator exists therefore. **Singular Value Decomposition:** $X = UDV^T$ with $dim(U) = n \times p$ and $dim(D) = p \times p$ and $dim(V^T) = p \times p$ and $U^T U = I_n$ and $V^T V = I_p$ if $p < n$. If $p \ge n$, we have the same decomposition but $dim(U) = n \times n$, $dim(D) = n \times n$ and $dim(V^T) = n \times p$ and $U^T U = UU^T = I$ and $V^T V = I$ and columns of U spaning $col(X)$ and columns of V spaning $row(X)$. That gives $\hat{\beta}_{RIDGE} = Vdiag(\frac{d_i}{d_i^2 + \lambda})U^T y$ and $\lim_{\lambda \downarrow 0} \hat{\beta}_{RIDGE} = \hat{\beta}_{GLS}$ or $\frac{d_i}{d_i^2 + \lambda} = \frac{1}{d_i}\left(1 - \frac{\lambda}{d_i^2 + \lambda}\right)$ with the latter in the braces being the shrinkage factor η which is small if d_i is small or λ large. D is a diagonal matrix with the singular values. **Bias of Ridge:** $\mathbb{E}[\hat{\beta}_{RIDGE}] = Vdiag(d_i/(d_i^2 + \lambda))U^T \mathbb{E}[Y] = Vdiag(d_i^2/(d_i^2 + \lambda))V^T \beta \overset{\lambda \downarrow 0}{\to} VV^T\beta \neq \beta$ but a projection of β onto the row space of X, if $p > n$. Otherwise, $row(X)$ spans everything because $n > p$ and we have an unbiased estimator. **Shrinkage:** $\|VV^T\beta\|^2 \le \|\beta\|^2$ because of the projection and usually much smaller if $dim(\beta) >> dim(row(X))$. If λ is large, $\mathbb{E}[\hat{\beta}_{RIDGE}]$ is even more shrunken towards zero than $V^T V\beta$. **Covariance matrix of** β_{RIDGE}: $Cov((X^T X + \lambda I)^{-1} X^T Y) = \sigma^2(X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}$ and $\hat{\sigma}^2$ being estimated as normal with $df = trace(H)$, which gives us $\hat{\sigma}^2 = \frac{n}{1 - trace(H)} \sum_{i=1}^n (y_i - x_i^T\hat{\beta}(\lambda))$ with $H = X\hat{\beta} = X(X^T X + \lambda I)^{-1} X^T$. We can obtain no CIs for the Ridge (only for $V^T V\beta$) as its biased and inconsistent if $p >> n$ and thus $X\hat{\beta}_{RIDGE}$ is inconsistent for $X\beta$.

Lasso Regression: Least Absolute Shrinkage and Selection Operator

$p >> n$ regularization w.r.t sparsity; we assume many entries of β to be 0, which can be extended to weak sparsity. $\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}}\left(-X\beta\|_2^2/n + \lambda\|\beta\|_1\right)$ with $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ which has no closed form solution but is a convex optimization problem. (i) sparsity: $\hat{\beta}_j = 0$ for many js which positively depends on λ and (ii) $\hat{S}(\lambda) = \{j; \hat{\beta}_j(\lambda) \neq 0\}$ is the support, e.g $supp(\hat{\beta}(\lambda))$ and (iii) $\hat{\beta}(\downarrow 0) = \underset{\beta}{\operatorname{argmin}}\|\beta\|_1\, s.t Y = X\beta$ which typically means that $|\hat{S}(\lambda \downarrow 0)| = n$. **Estimation of** $\hat{\sigma}^2$**:** normal construction but $df \neq trace(H)$ as LASSO is not a linear estimator and has no Hat-matrix and one chooses $df = |\hat{S}(\lambda)|$. The LASSO leads to a spase solution, i.e $|\hat{S}(\lambda)| \le \min\{n, p\}$ One can not derive a closed form solution for $\mathbb{E}[\hat{\beta}_{LASSO}]$ and $Var(\hat{\beta}_{LASSO})$ **orthonormal design** Let $X^T X/n = I$ with $p = n$, s.t we have square matrices. We then have $\hat{\beta}_{OLS} = \left((X^T X)^{-1} X^T y\right)_j = \left(IX^T Y/n\right)_j = X^{(j)} Y/n = Z_j$ which is the inner product and the empirical correlation if the variables are mean centered. We have an explicit connection of the Lasso to OLS, i.e $\hat{\beta}_{Lasso}(\lambda) = g_{soft, \lambda/2}(Z_j)$ with $g_{soft, \tau} = sign(z)(|z| - \tau)_+$ meaning $\hat{\beta}_{Lasso}$ is zero for all $|z| \le \tau = \lambda/2$. We said that Lasso exhibits a bias even for large $|Z_j|$, which could be evaded with the hard-thresholding as the $\hat{\beta}_{Lasso}$ is on the line of the OLS outside the region of τ. That is not done because its computationally very extensive and is in general the solution of the following optimization problem: $\hat{\beta}_{l0}(\lambda) = \underset{\beta}{\operatorname{argmin}}\left(|Y - X\beta|_2^2/n + \lambda supp(\beta)\right)$. From the picture, we see that LASSO (under soft-thresholding) has a downward bias of $\tau = \lambda/2$ even for large z_j. If $\beta_j > 0$, we have $\hat{\beta}_j = Z_j - \lambda/2$ and $\hat{\beta}_j = Z_j + \lambda/2$ if $\beta_j < 0$ and $|Z_j| \le \lambda/2$ if the minimum is at zero. We also have $sign(Z_j) = sign(\beta_j)$. **Model selection** Using e.g C_p or AIC for variable selection is a NP-problem as there are $\sum_{k=0}^{\lfloor 0.8n \rfloor}\binom{p}{k}$ possible submodels if we restrict the maximal cardinality to be $\lfloor 0.8n \rfloor$. Alternative: $\hat{S}(\lambda) = \{j; \hat{\beta}_j \neq 0\}$ which is with high probability equal to S if (i) Xs interpretability (irrepresentability condition on X) given (ii) sparsity of β: $|S| << \frac{n}{\log(p)} = o\left(\frac{n}{\log(p)}\right)$ as $p > n \to \infty$ and (iii) beta-min condition $\min\{|\beta_j|; j \in S\} >> |S|\sqrt{\frac{\log(p)}{n}}$ with probability 1. **Screening as more practical tool:** When do we have $\hat{S}(\lambda) \supseteq S$ for certain λ? We need (i) compatibiltity condition on X (weaker than irrepresentability) (ii) sparsity and (iii) beta-min-condition. If $\hat{S}(\lambda) \supseteq S$ indeed held in practice, we would be able to reduce dimension without loosing any information! That is not true but its still done in practice to get a dimension of $|\hat{S}| \le \min\{n, p\}$ and than fit an OLS model, whose inference cant be trusted because of the post-selection inference problem, which is why one does sample splitting. **Oracle Inequality:** for $\lambda = \sigma\sqrt{\frac{\log(p)}{n}}C$ with C being suff. large and positive, with prob. 1 as $p > n \to \infty$, we have: $\|X(\hat{\beta} - \beta)\|^2/n + \lambda\|\hat{\beta} - \beta\|_1 \le 4\lambda^2|S^0|/\phi_0^2$ with ϕ_0^2 being a condition on the design of X (smallest l_1-eigenvalue or copatibility constant) which has a positive relation with the design of X. That implies (1) $\|X(\hat{\beta} - \beta)\|^2/n \le const|S^0|\log(p)/n$ which only differs by $\log(p)$ compared to OLS, which is not bad for not knowing the support $|S^0|$ and (2) $\|\hat{\beta} - \beta^0\|_1 \le const|S^0|\sqrt{\frac{\log(p)}{n}}$ and (3) if beta-min-condition holds, we have proper screening, i.e $\hat{S}(\lambda) \supseteq S^0$. The RHS of the last equation converges to 0 if $|S^0| << \sqrt{\frac{n}{\log(p)}}$.

4 Robust Methods

L_1 Regression

$\hat{\beta}_{L_1} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n |y_i - X_i^T\beta|$, s.t the large residuals only enter linearly. Price to be paid is that no closed form solution exists so we have to do the convex optimization problem. Location model, i.e $p = 1$ and $x_i = 1$ yields the median of the data, which is less precise than the mean (50% more data for same precision).

Huber Methods

$\hat{\beta}_H = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \rho_c\left(y_i - X_i^T\beta\right)$ with $\rho_c(u) = 0.5u^2(|u| \le c)$ and $\rho_c(u) = c(|u| - c/2)(|u| \ge c)$, i.e we have a quadratic loss function for all values smaller than c and a linear for all values bigger than c, which is a combination of OLS and L_1-norm regression. We solve for $\hat{\beta}$ by differentiation. We need $\psi_c(u) = \rho_c'(u) = sign(u)\min\{|u|, c\}$ which is linear with slope u if residuals are inside c and otherwise constant. Realizing that c should depend on the error variance (if high variance, we also want larger residuals to be under a quadratic loss function as they are not outliers but legitimate observations), we get the two following equations: (1) $\sum_{i=1}^n \psi_c\left(\frac{y_i - X_i^T\hat{\beta}}{\hat{\sigma}}\right)X_i = \vec{0}$ and (2) $\sum_{i=1}^n \chi\left(\frac{y_i - X_i^T\hat{\beta}}{\hat{\sigma}}\right) = 0$ where $\chi(u)$ is chosen s.t $\int_{\mathbb{R}} \chi(u)\phi(u)du = 0$ with $\phi(u)$ being a pdf of a $N(0, 1)$, making $\hat{\sigma}$ a valid estimator for Gaussian errors. χ is Hubers Proposal 2. Could alternatively choose $1/\beta \times median(\text{absolute residuals})$. Then, it can be derived that (asympt.): $\sqrt{n}(\hat{\beta} - \beta) \to N_p(0, V)$ with $V = \frac{\mathbb{E}[\psi_c(\epsilon_i/\sigma)^2}{\mathbb{P}[|\epsilon_i| \le c\sigma]^2}\sigma^2(XX^T)^{-1}$ with the factor in front of the normal expression $<< 1$ if the errors are non Gaussian (big win) and only a little bigger than 1 (inverse to c) if the errors are Gaussian, e.g almost 1 if $c \ge 4$. One thus can gain a lot using robust regressions and loose little: good stuff! **Influence of observations:** Influence of large values of y are bounded for fix X, i.e $\Delta\hat{\theta} \approx \frac{1}{n\mathbb{P}[|\epsilon_i|\sigma]}(\mathbb{E}(X^T X)^{-1} X\psi_c\left(\frac{y - X^T\theta}{\sigma}\right)\sigma$. **Schweppe's proposal:** limit the influence of the Xs downweights (y_i, X_i) if X_i is far out unless the residual r_i is very small. **High-Breakdown point** How many points to drag to $\pm\infty$, e.g $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ \text{median}(y_i - X_i\theta)^2$ has a 50% breakdown point. In words: among all pairs of parallel hyperplanes sandwiching 50% of the observations, we look for the pair whose distance along the y-axis is minimal. Very inefficient estimator as the rate only is $n^{-1/3}$. **No closed form solution** exists for L_1 and Huber but L_1 can be reduced to an interior point LP. Huber is erformed by IRWLS using weights $w_i \propto \frac{\psi_c(y_i - x_i^T\hat{\beta})/\hat{\sigma}}{y_i - x_i^T\hat{\beta}} \propto \min\{1, \frac{c\hat{\sigma}}{|y_i - x_i^T\hat{\beta}|}\}$ until stabilization.

5 Nifty Take-Aways from Exercises

Linear transformations of Variables

Let $Y \sim X$ $X' = X - 10$: slope unchanged, $\hat{\alpha'} = \hat{\alpha} + 10\hat{\beta}$, fitted values and SSE unchanged, R^2 and ρ unchanged. The same goes for $X' = 10X$. $Y' = 5Y$: $\hat{\beta}' = 5\hat{\beta}$, $\hat{\alpha}' = 5\hat{\alpha}$; everything is $\times 5$ apart from $SSE' = 5^2 SSE$, R^2 and ρ are unchanged. With $Y' = Y + 10$, only the intercept changes. General: R^2 and ρ are not changed by a linear transformation.

Box-Cox Transformation

$X \to X^{(p)} = \frac{X^p - 1}{p}$ with $X^{(0)} = \log(x)$

Interesting Stuff

If we have power or exponential dependence, e.g $Y_i = \alpha x_i^\beta + \epsilon_i \leftrightarrow \log(Y_i) = \log(\alpha) + \beta\log(x_i) + \epsilon$ and then transforming back, we see that the errors are rather multiplicative than additive. **Empirical Correlation of two standardized predictors** $x^{(1)T} x^{(2)}$ **Rank** maximal nr. of linearly independent columns of X; while full rank means that $rank(X) = \min\{\#columns, \#rows\}$. A matrix is invertible if it has full rank **Trace** $trace(A + B) = trace(A) + trace(B)$ and $trace(A^T) = trace(A)$. **Row and Columns of X:** X_i^T denotes rows and $X^{(i)}$ columns. **Bias Variance Decomposition:** $\mathbb{E}[(z - c)^2] = \mathbb{E}[((Z - \mathbb{E}(Z)) + (\mathbb{E}(Z) - c))^2] = Var(Z) + (\mathbb{E}(Z) - c)^2 + 2 \times 0$ **LS and robustness:** the level of tests and CI of LS is robust but their power not. **Cook's Distance:** $D_i = \frac{1}{p}\frac{r_i^2}{\hat{\sigma}^2(1 - P_{ii})}\frac{P_i i}{1 - P_{ii}}$ describes the change in $\hat{\beta}_{LS}$ when computing it without the i-th data point X_i, Y_i. **Median Regression** Function has got a lot of local minima. **Exponential family parameters:** normal ($\beta = \frac{\mu}{\sigma^2}, c(\beta) = -0.5\sigma^2\beta^2$), binomial ($\beta = \log(p/1 - p), c(\beta) = -n\log(1 + e^\beta)$ and poisson ($\beta = \log(\beta), c(\beta) = -e^\beta$) while we have $\mathbb{E}[Y_i] = -c'(\beta_i)$ by integrating and differentiating the density. **Rank and system of equations** $Ax = b$ has a unique solution if $rank(A) = n$ and infinitely many if $rank(A) < n$ with n being the number of columns of A. **Properties of MLE** the MLE for suitable GLM has asymptotically the smallest variance among all asymptotically unbiased estimators. **Cox Regression** is not a special case of a GLM. **Sample median:** $\hat{\beta} = \underset{\beta}{\operatorname{argmin}}|y_i - \beta|$ which is L_1-regression with $X_i = 1$ and $p = 1$ **No bias in LS! Computing Lasso:** iterative soft-thresholding by componentwise updating is a feasible algorithm **Landau-notation:** if $f \in \mathcal{O}(g)$, then f does not grow substantially faster than g. If $f \in o(g)$, then f grows slower than g. **Unbiasedness of OLS:** we only need the model to be linear in the parameters, no perfect multicollinearity ($X^T X$ has full rank) and the zero conditional mean sumption, i.e $\mathbb{E}[\epsilon_i \mid X) = 0$ ($\mathbb{E}[\epsilon_i]$ is enough. **F-test** and χ^2-test: if σ^2 is known in linear models, one could do a χ^2-test, but the F-test is exact (if errors are Gaussian). **Asymptotic normality of LS:** assume i.i.d non-Gaussian errors and conditions that $eig_{MIN}\{X^T X\} \to \infty$ as $n \to \infty$ and $max_j P_{jj} \to 0$ as $n \to \infty$, we have that $(X^T X)^{1/2}(\hat{\beta} - \beta)$ converges weakly to $N_p(0, \sigma^2 I)$ **Confidence band for entire hyperplane** shape of hyperboloid and $(y_0 - \mathbb{E}[y_0])^2 \le \hat{\sigma}^2(x_0^T(X^T X)^{-1}x_0)pF_{p, n-p}(1 - \alpha)$ **Simple linear regression:** $\hat{SE}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sum_{i=1}^n (x_i - \bar{x})^2}$ **Making QQ-plots** for $n \le 100$, plot e.g $\phi^{-1}\left(\frac{i - 1/2}{n}\right)$ on the horizontal axis. For large n: choose equidistant values of x from the sample. **SPSE** $SPSE = \sum_{i=1}^n \mathbb{E}[(Y_{n+i} - \hat{Y}_i^M)^2] = \sum_{i=1}^n \mathbb{E}[(Y_{n+i} - \mu_i)^2] + \sum_{i=1}^n \mathbb{E}[(\hat{y}_i^M - \mu_i)^2] = n\sigma^2 + SMSE$ **Unbiasedness** not always necessary, e.g in Ridge or Bayesian regression. **Michaelis Menton:** reaction speed and concentration s.t $f(x; \beta) = \frac{\beta_1 x}{\beta_2 + x}$ **Covariance matrix Logistic regression** $V(\beta) = I(\beta)^{-1} = \left(\sum_{i=1}^n x_i x_i^T \frac{\exp(x_i^T\beta)}{1 + \exp(x_i^T\beta))^2}\right)^{-1}$ which is the inverse of the Fisher information. **Conditional Expectation as best predictor:** $\mathbb{E}[(Y_i - f(X_i))^2] \le \mathbb{E}[(Y_i - g(X_i))^2]$ with $f(x_i) = \mathbb{E}[Y_i \mid X_i = x]$ if second moments exist. **Bias-Variance Tradeoff local polynomial** if p is odd $bias^2 \sim \text{const}(K, p)h^{p+1}f^{(p+1)}(x)$ and $Var([\hat{f}(x)] \sim \text{const}(K, p)\frac{\sigma_\epsilon^2}{nh}\left(\frac{1}{nh}\sum_{i=1}^n K((x - x_i)/h)\right)^{-1}$ and hence $MSE \sim \mathcal{O}(\frac{1}{nh}) + \mathcal{O}(h^{2(p+1)})$ and $h_{opt} \sim \mathcal{O}(n^{-1/(2p+3)}$ and $MSE_{opt} \sim \mathcal{O}(n^{-(2p+2)/(2p+3)})$ which is difficult as that would suggest to make p as large as possible but that is not right as that makes estiamting the constants also harder. Its the same as for the Nadaraya-Watson with $p = 1$ and also the same for the Gasser-Müller with $p = 1$ but with Var being $1.5\times$ bigger. For Smoothing splines, it is similar to local polynomials with $p = 3$. **Additivity of errors:** $y_i = \exp(X_i^T\beta + \epsilon_i) \to$ log-transform yields errors that are log-normally distributed with $\mathbb{E}[y_i \mid x_i] = \exp(x_i^T\beta + \sigma^2/2)$ **Modelling nonlinear effects:** $y_i = \beta_0 + \beta_1 f(z_i) + \epsilon_i$ can be transformed into a normal linear model with $x_i = f(z_i) - \bar{f}$ and thus centering the effect of $\hat{\beta}_1 x$ around zero. We can interpret by plotting $\hat{\beta}_1 x$ against z_i. **Dummies** if one does not include a reference group, we have perfect multicollinearity and the regression parameters are not identifiable as we can add a value a to the intercept and subtract it from the βs. **Columns-space of X:** $col(X) = \{u; u = \sum_{j=1}^p \gamma_j X^{(j)}; \gamma \in \mathbb{R}^p\} \subseteq \mathbb{R}^n$. **Orthogonal desgin matrix** if X has orthogonal columns we have $\hat{\beta}_{single, j} = \hat{\beta}_j$ **T-distribution and normal distribution:** $t_v \to N(0, 1)$ as $v \to \infty$. **Latent variable Interpretation of Logit/Probit:** $Y_i = \mathbb{1}_{Z_i > 0}$ and $Z_i = X_i^T\beta + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$ being probit and $\epsilon_i \sim$ logistic being logit. **W for Poisson regression** $W = diag(\mu_i)$ **Nonlinear LS for oxygen-stuff:** $f(x; \beta) = \beta_1(1 - \exp(-\beta_2 x))$ **Lasso Geometry** There is a 1 to 1 correspondence of λ and R ($\|\beta\|_1 > leqR$) that depends on the data.

Assumptions

Linear Model: (i) meaningful data (ii) correct regression, i.e $\mathbb{E}[\epsilon_i] = 0$ (iii) $\mathbb{E}[\epsilon_i\epsilon_j] = 0$ (iv) Xs are exact (v) homoskedasticity (vi) errors follow a joint normal distribution