**MVZV** $F_X(x_1,\dots,x_p) = P(X_1 \le x_1, \dots, X_p \le x_p) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_p} f_X(t_1,\dots,t_p)\, dt_p \dots dt_1$ Verteilungsfunktion, Gemeinsame Dichte von $X \leftarrow X$ p-dim ZV

- Randverteilung: $F_{X_i}(x) = P(X_i \le x) \leftarrow$ Bsp: $(X_1, X_2)$, $F_{X_1}(x) = P(X_1 \le x) = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f(t_1,t_2)\, dt_1 dt_2$
- Unabhängigkeit: $F_X = \prod_1^p F_{X_i}(x_i) \ \forall x_i$
- Bedingt: $f_X(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)}$ stetig, $P(X=x|Y=y) = \frac{P(X=x \cap Y=y)}{P(Y=y)}$ diskret
  - $\overset{u}{\hookrightarrow} f_X(x)$ $\qquad$ $\overset{u}{\hookrightarrow} P(X=x)$
- E: $E(X|Y=y) = \int x f_X(x|y)\, dx$ stetig, $E(X|Y=y) = \sum x_i P(X=x_i|Y=y)$ diskret
  - $E(X|Y) \leftarrow E(X|Y)(\omega) = E(X|Y=Y(\omega)) \Rightarrow E(E(X|Y)) = E(X)$, $E(Bx+b) = BE(x)+b$, $B$ $(m \times p)$, $b \in \mathbb{R}^m$, $a \in \mathbb{R}^p$

$E(X+Y|Z=z) = E(Y|Z=z) + E(X|Z=z)$
$E(X|X=x) = x$
$E(X|Y=y) \overset{u}{=} E(X)$

**Varianz** $Cov(aX+b, cY+d) = ac\, Cov(X,Y)$, $Cov(X+Y, Z) = Cov(X,Z) + \dots$ $\quad |\Sigma| = \det(\Sigma)$

$E(X) = (E[X_i])$

$Cov(X) = \Sigma(X) = (\sigma_{ij})$, $\sigma_{ij} = Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$ sym

$Cor(X) = \rho(X) = (\rho_{ij})$, $\rho_{ij} = Corr(X_i, X_j) = Cov(X_i, X_j)/(\sigma_{X_i}\sigma_{X_j})$ sym

$Cov(BX+b) = B\Sigma B^T$, $Var(a^T X) = a^T \Sigma a$

- Standardisiert: $X^* = \left(\frac{x_i - E(x_i)}{\sigma_{x_i}}\right) = D^{-1}(X - E(x))$, $D = eye(\sigma_i)$
- Unabhängig $X \sim N_p(\mu, \Sigma)$: $X_i$ u $\Leftrightarrow \Sigma$ diag $\Leftrightarrow X_i$ unkorreliert
- Schätzer: $\bar{X} = \frac{1}{n}\sum X_i$, $S = \frac{1}{n-1}\sum (X_i - \bar{X})(X_i - \bar{X})^T = (S_{ij})$, $R = (R_{ij}) = (S_{ij}/\sqrt{S_{ii}S_{jj}})$

Normalverteilung
$\hookrightarrow f_X(x) = (\sigma 2\pi)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
$\hookrightarrow f_X(x) = (2\pi)^{-p/2} |\Sigma|^{-0.5} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$
$\hookrightarrow X_i$ u stdnormv $\Rightarrow AX+b$ multinormv mit $\mu = b$, $\Sigma = AA^T$

Diskret Unabhängig

| | | $a_1$ | $X$ | $a_n$ | gdw |
|---|---|---|---|---|---|
| $Y$ | $b_1$ | $p_{11}$ | | $p_{1n}$ | $p_{11} = p_1 \cdot$ |
| | $b_n$ | $p_{n1}$ | | $p_{nn}$ | |
| | | $p_1$ | | $p_n$ | |

- Verlustfunktion: $L(y, \hat{f}(x))$, Trainingsfehler: $\overline{err}(L) = \frac{1}{n}\sum L(y, \hat{f}(x_i))$ Bsp: $L = (y - \hat{f}(x))^2$
  - Testfehler: $E(L) = \int [L(y, \hat{f}(x))] f_{X,Y}(x,y)\, dx\, dy$

**Multi LinReg**

- Modell: $Y = f(X) + \varepsilon = Xb + \varepsilon \Leftrightarrow \begin{pmatrix} Y_1 \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{1p} \\ 1 & X_{n1} & X_{np} \end{pmatrix} \begin{pmatrix} b_0 \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_n \end{pmatrix}$ $\left\{ \begin{array}{l} Cov(\varepsilon_i, \varepsilon_j) = 0 \\ E(\varepsilon_j) = 0 \\ Var(\varepsilon_j) = \sigma^2 \end{array} \right.$

Erklärte Varianz: $V(Y)/\sum V(X_i) \leftarrow$ Var (Variable) / Var Gesamt

$\hookrightarrow \hat{b} = (X^T X)^{-1} X^T y$ $\quad \hookrightarrow E(Y|X=x) = b_0 + \sum_1^p x_k b_k \leftarrow \hat{b}_0 + \sum_1^p x_k \hat{b}_k = (1, x^T)\hat{b}$ ist Schätzer

$\hookrightarrow \hat{y} = X\hat{b}$ geschätzte Werte $\quad \hookrightarrow \hat{y} = Hy$ mit $H = X(X^T X)^{-1} X^T$ sym hat Matrix

$\hookrightarrow \hat{\varepsilon} = y - \hat{y}$ Residuenvektor $\Rightarrow SSE = \hat{\varepsilon}^T \hat{\varepsilon} = \sum (y_i x_i^T \hat{b})^2$ Residual sum of squares $\leftarrow$ wird durch LS min

$\hookrightarrow \hat{\sigma}^2 = \frac{1}{n-(p+1)} \hat{\varepsilon}^T \hat{\varepsilon} = \frac{1}{n-(p+1)} \sum (\hat{\varepsilon} - \bar{\varepsilon})$, $\bar{\varepsilon} = 0$, Residual variance $\Rightarrow RSE = \sqrt{\hat{\sigma}^2}$ Residual standard error

$\hookrightarrow E(\hat{b}) = b$, $Cov(\hat{b}) = \sigma^2 (X^T X^{-1})$, $E(\hat{\sigma}^2) = \sigma^2$, $E(\hat{\varepsilon}) = 0 \leftarrow \hat{\sigma}^2$ Approximation: $\widehat{Cov}(\hat{b}) = \hat{\sigma}^2 (X^T X)^{-1}$

- $VIF = 1/(1 - R_j^2)$ Steigerungsfaktor von $b_j$ im Vergleich zu $X_j$ unkorreliert $\begin{cases} <4 \text{ unkritisch} \\ >10 \text{ kritisch} \end{cases}$

- Statistische Tests. Annahme: $\varepsilon \sim N_n(0, \sigma^2 I_n)$ verteilt
  - $\hookrightarrow \hat{b} \sim N_{p+1}(b, \sigma^2 (X^T X)^{-1}) \longrightarrow \frac{\hat{b}_j - b_j}{\hat{\sigma}\sqrt{v_j}} \sim N(0,1)$ mit $v_j = spur[(X^T X)^{-1}]_j$
  - $\hookrightarrow CI: [\hat{b}_j \pm z_{1-\alpha/2}\, \hat{\sigma} \sqrt{v_j}]$ $\quad \hookrightarrow H_0: \{b_j = 0\} \rightarrow \hat{b}_j/(\hat{\sigma}\sqrt{v_j}) \sim t_{n-(p+1)}$

$R^2$ Anteil der Variabilität der $Y_i$ die durch das Model erklärt wird

- Anpassungsgüte
  - $\hookrightarrow SST = \sum (y_i - \bar{y})^2$ total sum of squares
  - $\hookrightarrow SSE = \sum (y_i - \hat{y}_i)^2$ residual sum of squares
  - $\hookrightarrow SSM = \sum (\hat{y}_i - \bar{y})$ model sum of squares

$R^2 = \frac{SSM}{SST} = \frac{SST - SSE}{SST}$ $\quad SST = SSE + SSM$

$R^2 = Cor(Y, \hat{Y})^2$

$\overset{p=1}{\longrightarrow} Cor(Y, X)^2$

- Modelparameter [p Kovariaten $\Rightarrow$ p+1 Modellparameter]
  - $\hookrightarrow AIC = n \log(\frac{1}{n} SSE) + 2(p+1)$
  - $\hookrightarrow BIC = n \log(\frac{1}{n} SSE) + \log(n)(p+1)$
  - $\hookrightarrow C_p = SSE/\hat{\sigma}^2 - (n - 2(p+1))$, $\hat{\sigma}^2$ Varianzschätzer aus Modell mit allen Kovariaten

$\hookrightarrow Adj\ R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$ $\quad X_1, \dots, X_p$; $X_1, \dots, X_q$ $\quad q < p$

$\hookrightarrow F = \frac{(SSE^q - SSE^p)/(p-q)}{SSE^p/(n-p-1)} \sim F_{p-q, n-p-1}$ unter $H_0: \{b_{q+1} = \dots = b_p = 0\}$

- Variablen Selektion $\leftarrow$ subset: alle, berechnen; Forward: p=0 und dann nach best hinzufügen — backward: p=voll nach best entfernen

**Klassifikation** $X: (\Omega, A, P) \rightarrow \mathbb{R}^p$, $\delta: X \rightarrow \{1, \dots, K\}$, $\delta(x) = \begin{cases} 0, & b^T x \le 0 \\ 1, & b^T x > 0 \end{cases}$

- Konfusion Matrix $\quad y = h(x) = \sum_0^p b_i x^i \leftarrow b_i$: Diskriminanzkoeff

| | | $\hat{Y}$: t | f |
|---|---|---|---|
| $Y$ | t | TP | FN |
| | f | FP | TN |

TP = True positive $\quad$ FP = False positiv, Fehler 1. Art
TN = True negative $\quad$ FN = False negativ, Fehler 2. Art

Accuracy = (TP + TN)/Total
Missclassification Rate = (FP + FN)/Total

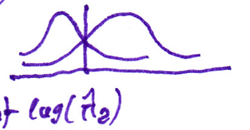
**Lin Diskriminanz** $[X|Y = k \sim N_p(\mu_k, \Sigma)$ $\forall k = 1, \dots, K]$

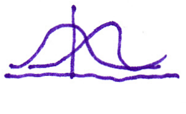
$\hookrightarrow$ Annahme: Gemeinsame Kovarianz, für Dichten der multivariaten Normalverteilung

$\hookrightarrow \hat{\pi}_k = n_k/n$

$\hookrightarrow K = 2$ $\begin{cases} \delta(x) = 1 \Leftrightarrow h(x) > 0 \\ h(x) = x^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) \end{cases}$ $\quad \mu_2$ analog

Bsp $\pi_0 = 0.5, \pi_1 = 0.5$ $\qquad$ $\pi_0 = 0.3, \pi_1 = 0.7$





$\hookrightarrow$ Quadratisch: $h(x) = \log(\pi_1) - \log(\pi_2) = -\frac{1}{2}\log|\hat{\Sigma}_1| - \frac{1}{2}(x - \hat{\mu}_1)^T \hat{\Sigma}_1^{-1}(x - \hat{\mu}_1) - \dots + \log(\hat{\pi}_1) - \log(\hat{\pi}_0)$

**Multitest** $p = P_{H_0}(T(X_1,\dots,X_n) > T(x_1,\dots,x_n))$
$p \leq \alpha \Leftrightarrow T(x_1,\dots,x_n) > c_\alpha \Leftrightarrow H_0$ verwerfen

| | # Ablehnung | # Nicht | |
|---|---|---|---|
| # wahre $H_0$ | $V$ | $m_0 - V$ | $m_0$ |
| # falsche $H_0$ | $R-V$ | $m_1 - R + V$ | $m_1$ |
| | $R$ | $m - R$ | $m$ |

- $FWER = P(V>0)$
- $FDR = E(V/\max(R,1))$ ← p-Werte unabhängig

Bonferri: $p_{v_i} \leq \frac{\alpha}{m}$ | Sidak: $p_{v_i} \leq 1-(1-\alpha)^{\frac{1}{m}}$ | Holm: $p_{v_i} \leq \frac{\alpha}{m}$, p-Werte sortiert, $m = m-1$

**Linkage** $d(C_i, C_j)$ Clusterabstand
avg: $\frac{1}{|C_i||C_j|}\sum d$
Zentroid

**Cluster** $d(i,j) = d(j,i)$; $d(i,j) \geq 0$; $d(i,i) = 0$ | Agglomerativ: n cluster → nearest merge
k-means: k cluster → d min zuweisen → $C_k$ mittelpunkte neue $C_k$ → repeat bis no change
Single: $\min_{x \in i, y \in j} d(x,y)$
Comp: $\max_{x \in i, y \in j} d(x,y)$

**PCA** $\Sigma = Cov(X) = A \Lambda A^T$, A normierte EV Ladungsmatrix, $\Lambda = diag(\lambda_i)$ $\lambda_1 \geq \dots \geq \lambda_p \geq 0$
$Y = A^T X \Rightarrow y = A^T x$, $y = A_i^T x$ i-te Hauptkomponente zu X
$Cov(Y_i, Y_j) = 0$, $Var(Y_i) = \lambda_i$, $\sum Var(Y_i) = spur(\Lambda) = s$ Gesamtvarianz $\Rightarrow \frac{\lambda_i}{s}$ erklärte Varianz

Kaiser-Kriterium
$\lambda_i \geq \frac{1}{p}\sum \lambda_i$

$Cov(Y_i, Y_j)$
$= Cov((A^T X)_i, (A^T X)_j)$
$= Cov(A^T X)_{ij}$
$= (A^T Cov(X) A)_{ij}$

**Multi LinReg** Überprüfung der Modellvoraussetzungen

| | Modellannahme | Konsequenzen aus Verletzung | Überprüfung Residuenplot |
|---|---|---|---|
| A | Ein Zusammenhang zwischen X, Y | Verzerrung der Schätzer | Residuenplots ($\hat{y}$ vs $\hat{\varepsilon}$ / $\hat{\varepsilon}^*$, $x_j$ vs $\hat{\varepsilon}, \hat{\varepsilon}^*$) |
| B | $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ | Verzerrung der Schätzer | |
| C | Unkorreliertheit der Fehlerterme | Verzerrung der Schätzer | zeitliche Strukturen |
| D | Normalverteilung der Fehlerterme | stat. Tests und CI halten nicht das Niveau | QQ-Plots der Residuen |
| E | $Rang(X) = p+1 \Rightarrow x_j$ lin unab | Parameterschätzer nicht eindeutig | Korrelationsplots |
| F | Multikolinearität | instabile Schätzer, hohe SE | Korrelationsplots, VIF |

- A+B+C ⇒ best linear unbiased estimator ⇒ unverzerrte Schätzer
- Residuenplots sollten keine Struktur aufweisen
- Verletzungen in A, B, D können durch Transformation behoben werden
- Verletzung in F können Kovariaten ~~mgs~~ entfernt werden (PCA)

**Shapiro-Wilk** Test auf Normalverteilung
$p > \alpha$ → Normalverteilung

**Log Likelihood** $T = 2(LL_{M_2} - LL_{M_1}) \underset{H_0}{\sim} \chi^2_{1-\alpha}$ ← Variable aufnehmen, wenn $T > \chi^2_{1-\alpha}$

**R-Output**
- Multiple $R^2$: Anpassungsgüte. Wie nahe die Punkte der Geraden sind. Beschreibt die erklärte Varianz und steigt mit der Anzahl an Kovariaten.
- Adjusted $R^2$: $R^2$ von allen statistisch signifikanten Kovariaten
- F-Statistic: model MSE/residual MSE. Test der Nullhypothese, dass alle Parameter 0 sind.

| ANOVA | DF | sum squares | mean sum squares | F-Test |
|---|---|---|---|---|
| Regression (Modell) | $p$ | SSM | $SSM/p$ | $F = \frac{SSM/p}{SSE/[n-(p+1)]}$ |
| Residuen (Error) | $n-(p+1)$ | SSE | $SSE/[n-(p+1)]$ | |
| Gesamt (Total) | $n-1$ | SST | $SST/(n-1)$ | |

**Variablen Selektion** [Best: min AIC, max $R^2 \Leftrightarrow$ min SSE]
Subset selection: Alle Modelle berechnen ⇒ choose best
Forward selection: Starte mit $p=0$ ⇒ choose best new Kovariate
Backward selection: Start mit $p$ = voll ⇒ choose worst Kovariate ↳ Best model

**Diskriminanz**

- Bayes-Classifier: $\pi_k(x) = P(Y=k|X=x) = f_k(x)\pi_k / [\sum_{l=1}^{K} f_l(x)\pi_l] \Rightarrow \hat{y} = \arg\max_K \hat{\pi}_k(x)$
  ↳ $\hat{\pi}_k(x)$ Schätzer durch $f_l(x) = \hat{f}_X(\cdot | X=l)$
- Logistische Regression [Y binär {0,1}, X p-dim, $x, b \in \mathbb{R}^p$]
  ↳ $P(Y=1|X=x) = \exp(b^T x)/[1+\exp(b^T x)]$, $P(Y=0|X=x) = 1/[1+\exp(b^T x)]$
  ↳ $L(b) = \prod_{i=1}^{n} P_b(Y_i = k_i | X = x_i)$ ↳ $odds(x) = \exp(b^T x)$ ← $\exp(b_i)$ relative Veränderung der odds bei Veränderung von Merkmal i

$f(x) = \frac{\exp(x)}{1+\exp(x)}$ logistische Funktion

Bsp: $\exp(b_i) = 1.5$ ⇒ odds für Y=1 vs Y=0 steigt um 50%

- Diskriminierung
  ↳ Güte anhand der Trennschärfe
  ↳ Fehlerrate: $P(Y \neq \hat{Y})$ ↳ Sensitivität: $P(\hat{Y}=1|Y=1)$ ↳ Spezifität: $P(\hat{Y}=0|Y=0)$
  ↳ positiver prädikativer Wert: $P(Y=1|\hat{Y}=1)$ ↳ negativer prädikativer Wert: $P(Y=0|\hat{Y}=0)$

AUC (Area under curve)