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 $X \in \mathbb{X}$ 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 (↳ $f_X(x)$), $P(X=x|Y=y) = \frac{P(X=x \cap Y=y)}{P(Y=y)}$ diskret (↳ $P(X=x)$)
- E: $E(X|Y=y) = \int x f_X(x|y)\,dx$ stetig, $E(X|Y=y) = \sum_i x_i P(X=x_i|Y=y)$ 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 $\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y)$, $\mathrm{Cov}(X+Y, Z) = \mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z)$

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

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

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

$\mathrm{Cov}(BX+b) = B\Sigma B^T$, $\mathrm{Var}(a^TX) = a^T\Sigma a$

- Standardisiert: $X^* = \left(\frac{x_i - E(x_i)}{\sigma_{x_i}}\right) = D^{-1}(X - E(X))$, $D = \mathrm{eye}(\sigma_i)$
- Unabhängig $X \sim N_p(\mu, \Sigma)$: X_i u $\Leftrightarrow$ Σ diag $\Leftrightarrow$ X_i unkorreliert

$|\Sigma| = \det(\Sigma)$

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

- $\bar{X} = \frac{1}{n}\sum X_i$, $S = \frac{1}{n-1}\sum(X_i-\bar{X})(X_i-\bar{X})^T = (s_{ij})$, $R = (R_{ij}) = \left(s_{ij}/\sqrt{s_{ii}s_{jj}}\right)$

Diskret Unabhängig

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

□ 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 Lin Reg

□ Modell: $Y = f(X) + \varepsilon = Xb + \varepsilon \Leftrightarrow \begin{pmatrix} y_1 \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{1p} \\ 1 & x_{n1} & x_{np} \end{pmatrix}\begin{pmatrix} b_0 \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_n \end{pmatrix}$ $\begin{cases} \mathrm{Cov}(\varepsilon_i,\varepsilon_j) = 0 \\ E(\varepsilon_i) = 0 \\ \mathrm{Var}(\varepsilon_i) = \sigma^2 \end{cases}$

Erklärte Varianz: $V(\hat{Y}) = \sum V(\hat{Y}_i)$ ← Varianz (Variable) / zur Gesamt

↳ $\hat{b} = (X^TX)^{-1}X^Ty$ ↳ $E(Y|X=x) = b_0 + \sum_1^p x_k b_k \leftarrow \hat{b}_0 + \sum_1^p x_k\hat{b}_k = (1, x^T)\hat{b}$ ist Schätzer
↳ $\hat{y} = x\hat{b}$ geschätzte Werte ↳ $\hat{y} = Hy$ mit $H = X(X^TX)^{-1}X^T$ sym hat Matrix
↳ $\hat{\varepsilon} = y - \hat{y}$ Residuenvektor $\Rightarrow SSE = \hat{\varepsilon}^T\hat{\varepsilon} = \sum(y_i - x_i^T\hat{b})^2$ Residual sum of squares ← wird durch LS min
↳ $\hat{\sigma}^2 = \frac{1}{n-(p+1)}\hat{\varepsilon}^T\hat{\varepsilon} = \frac{1}{n-(p+1)}\sum(\hat{\varepsilon}-\bar{\varepsilon})^{\bar{\varepsilon}=0}$ Residual Variance $\Rightarrow RSE = \sqrt{\hat{\sigma}^2}$ Residual standard Error
↳ $E(\hat{b}) = b$, $\mathrm{cov}(\hat{b}) = \sigma^2(X^TX^{-1})$, $E(\hat{\sigma}^2) = \sigma^2$, $E(\hat{\varepsilon}) = 0$ ← $\hat{\sigma}^2$ Approximation: $\hat{\mathrm{Cov}}(\hat{b}) = \hat{\sigma}^2(X^TX)^{-1}$

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

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

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

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

$R^2 = \frac{SSM}{SST} = \frac{SST - SSE}{SST}$ $SST = SSE + SSM$
$R^2 = \mathrm{Cor}(Y, \hat{Y})^2$
$p=1 \rightarrow \mathrm{Cor}(Y,X)^2$

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

↳ Adj $R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$ $x_1,\dots,x_p$ / $x_1,\dots,x_q$ $q<p$
↳ $F = \frac{(SSE^q - SSE^p)/(p-q)}{SSE^p/(n-p-1)} \sim F_{p-q,n-p-1}$ unter $H_0: \{b_{q+1} = \dots = b_p = 0\}$

□ Variablen Selektion ← subset: alle Kombinationen / Forward: p=0 und dann nach best hinzufügen — backward: p=voll nach best entfernen

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

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

y \ ŷ	F	T
T	TP	FN
F	FP	TN

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

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

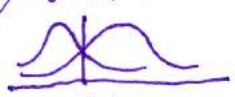
Lin Diskriminanz $[X|Y=k \sim N_p(\mu_k, \Sigma)\ \forall k = 1,\dots,K]$

↳ Annahme: Gemeinsame Kovarianz, fk Dichten der multivariaten Normalverteilung
↳ $\hat{\pi}_k = n_k/n$

↳ $K=2$: $\delta(x) = 1 \Leftrightarrow h(x) > 0$
$h(x) = x^T\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) - \frac{1}{2}(\hat{\mu}_0 + \hat{\mu}_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right)$ μ_2 analog

↳ Quadratisch: $h(x) = \log(\pi_1) - \log(\pi_2) = -\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \log(\pi_2) + \log(\pi_1)$

Bsp $\pi_0 = 0.5$, $\pi_1 = 0.5$



$\pi_0 = 0.3$, $\pi_1 = 0.7$

Multitest $p = P_{H_0}(T(X_{11},\dots,X_{1n}) > T(x_{11},\dots,x_{1n}))$
$p \le \alpha \Leftrightarrow T(x_{11},\dots,x_{1n}) > c_\alpha \Leftrightarrow H_0$ 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: $p_{V_i} \le \frac{\alpha}{m}$ | Sidak: $p_{V_i} \le 1-(1-\alpha)^{\frac{1}{m}}$ | Holm: $p_{V_i} \le \frac{\alpha}{m}$, p-Werte sortiert, $m = m-1$

Linkage $d(C_i, C_j)$ (Clusterabstand)
single: $\min_{x\in i, y\in j} d(x,y)$
comp: $\max_{x\in i, y\in j} d(x,y)$
avg: $\frac{1}{|i|\cdot|j|}\sum d$
Zentroid

cluster $d(i,j) = d(j,i)$; $d(i,j) \ge 0$; $d(i,i) = 0$ | agglomerativ: n cluster → nearest merge
k-means: k cluster → d min zuweisen → C_k Mittelpunkte neue C_k → repeat bis no change

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

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

Multi LinReg Überprüfung der Modellvoraussetzungen

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

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

- A+B+C $\Rightarrow$ best linear unbiased estimator $\Rightarrow$ unverzerrte Schätzer
- Residuenplots sollten keine Strukturen aufweisen
- Verletzungen in A, B, D können durch Transformation behoben werden
- Verletzung in F können Kovariaten entfernt werden (PCA)

Shapiro-Wilk Test auf Normalverteilung
$p > \alpha \rightarrow$ Normalverteilung
Log-Likelihood ← Variable aufnehmen, wenn $T > \chi^2_{1-\alpha}$
$T = 2(LL_{un} - LL_{Ho}) \overset{H_0}{\sim} \chi^2_{1-\alpha}$

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: $\pi_k(x) = P(Y=k \mid X=x) = f_k(x)\pi_k / [\sum_{l=1}^K f_l(x)\pi_l] \Rightarrow \hat{y} = \underset{k}{\mathrm{argmax}}\ \hat{\pi}_k(x)$
  - $\hat{\pi}_k(x)$ Schätzer durch $f_k(x) = \hat{f}_X(\cdot \mid X=k)$
- Logistische Regression [Y binär $\{0,1\}$, X p-dim, $x, b \in \mathbb{R}^p$] ← $f(x) = \frac{\exp(x)}{1+\exp(x)}$ logistische Funktion
  - $P(Y=1 \mid X=x) = \exp(b^T x)/[1+\exp(b^T x)]$, $P(Y=0 \mid X=x) = 1/[1+\exp(b^T x)]$ ← Bsp: $\exp(b_i) = 1.5 \Rightarrow$ odds für Y=1 vs Y=0 steigt um 50%
  - $L(b) = \prod_{i=1}^n P_b(Y_i = k_i \mid X = x_i)$
  - $\mathrm{odds}(x) = \exp(b^T x)$ ← $\exp(b_i)$ relative Veränderung der odds bei Veränderung von Merkmal i
- Diskriminierung
  - Güte anhand der Trennschärfe
  - Fehlerrate: $P(Y \ne \hat{Y})$
  - Sensitivität: $P(\hat{Y}=1 \mid Y=1)$
  - Spezifität: $P(\hat{Y}=0 \mid Y=0)$
  - positiver prädikativer Wert: $P(Y=1 \mid \hat{Y}=1)$
  - negativer prädikativer Wert: $P(Y=0 \mid \hat{Y}=0)$

AUC (area under curve)