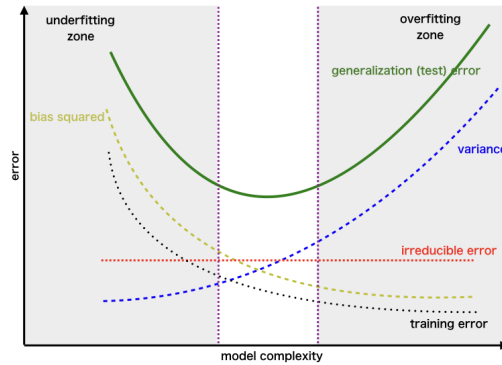


# 1 Grundlagen

- Grundlagen
  - $Y = f(x) + \epsilon$
  - $Y = \text{Zielgröße}$ ,  $f()$  = unbekanntes/wahres Modell,  $X = \text{Prädiktoren}$ ,  $\epsilon$  Nicht reduzierbarer Fehler
  - $\hat{Y} = \hat{f}(X) + \epsilon$
  - $\hat{Y}$  = Schätzung der Zielgröße,  $\hat{f}$  = Schätzung des Modells
  - Ziel: Möglichst genaue Schätzung finden
- Ziel:
  - Prediction (Vorhersage von Werten)
  - Inference (Ursachenanalyse, wie wirken sich Änderungen aus)
- Bias-Variance Tradeoff
  - Bias: Fähigkeit des Modells die eigentliche Beziehung der Daten abzubilden
  - Variance: Fähigkeit des Modells auf anderen Subsets gleich gute Modelle zu erzeugen
  - TrainingsError: Wird immer kleiner, da Modell sich immer besser anpasst
  - TestError: Wird erst kleiner, steigt dann aber wieder (Overfitting)
  - Nichtreduzierbarer Error: Bleibt immer gleich (Messfehler etc.)



## 2 Regression

- Modellgüte:
  - Schätzung der Parameter  $\beta_0$  und  $\beta_1$  über kleinste Quadrate
  - \*  $\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
  - \*  $\beta_0 = \bar{y} - \beta_1 \bar{x}$
- Qualitative Prädiktoren:
  - **Prädiktoren mit 2 Ausprägungen:**
  - DummyVariable aka 0(No) oder 1 (Yes)
  - **Achte auf Normalausprägung von R**
  - $\hat{y} = \beta_0 + \beta_1 * x_i$
  - Koeffizient  $\beta_1$  kürzt sich je nach Ausprägung raus
  - **Prädiktoren mit k Ausprägungen:**

- Erstelle  $k - 1$  Dummyvariablen
- Andere ist Normalzustand
- Interaktionseffekte:
  - Synergieeffekte zwischen zwei oder mehreren Variablen
  - $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$
  - Auswirkung erkennen durch Umformung:
    - $\hat{y} = \beta_0 + \beta_2 x_2 + (\beta_1 + \beta_3 x_2) * x_1$
    - Erhöht man  $x_1$  um eine Einheit erhöht sich  $\hat{y}$  um  $\beta_1 + \beta_3 x_2$  Einheiten
    - $x_1$  moderiert  $x_2$  und Vice versa
    - Signifikanz über p-value feststellen
    - Interaktion zwischen Quali und Quanti Variablen:
    - Kürzt sich komplett raus (wenn 0) oder ist \* 1 (wenn 1)

## 3 Klassifikation

## 4 Resampling

## 5 Modellauswahl

## 6 R - Hilfe

- `set.seed(X)` Setzt Seed für random Number Generator
- `c(1,2,3,4)` Vektor mit Zahlen 1-4

- `df[2,3]` Greift auf Element der 2.Reihe und 3.Spalte des DFs zu
- `df[, -3]` Entfernt 3. Spalte
- `head()` Zeigt erste X Zeilen von DF an
- `summary()` gibt Zusammenfassung von Modellen (DF, Modelle etc.)
- **Modelle:**
  - `lm(A ~ B + poly(C, 2) + B * C)` Lineare Regression für A mit Interaktivität von B und C mit Exponent 2
  - `predict(Modell, DataFrame, interval = , type = )`
    - \* DF: `data.frame(x1 = c(2), x2 = c(3))`
    - \* `interval` Konfidenzintervall(confidence), Prognoseintervall(prediction)
    - \* **type OFFEN!**
  - `coef()` Zeigt Koeffizienten des Modells
  - `confint()` Zeigt Konfidenzintervalle für Koeff.
- **Plots:**
  - `pairs()` Zeigt Pärchenplott aller quantitativer Variablen
  - `plot()` Zeigt X/Y Plot zweier Variablen
  - `abline(Modell, col = "red")` Zeigt Regressionslinie
  - `qplot()` aus ggplot2 für quickplot