3. Regularisierte Lineare Modelle

Early approach "unit weighting": Cohen, J. (1990). Things I Have Learned (So Far).

- "Yet as a practical matter, most of the time, [in regression] we are better off using unit weights: +1 for positively related predictors, -1 for negatively related predictors, and 0, that is, throw away poorly related predictors.
- The catch is that the betas come with guarantees to be better than the unit weights only for the sample on which they were determined.
- But the investigator is not interested in making predictions for that sample – he or she knows the criterion values for those cases. The idea is to combine the predictors for maximal prediction for future samples."

Regularisierung (lernen wir heute)

in-sample performance

out-of-sample performance

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304, S. 1309.

Agenda für heute

- Wiederholung: Lineare und Logistische Regression, Bias-Varianz Tradeoff
- Exkurs: Best Subset Selection
- Regularisierte Lineare Modelle
- Tuning des Regularisierungsparameters mit *k*-fold CV
- Interpretationsbeispiel für Klassifikation

Wiederholung: Lineare und Logistische Regression

- Multiple Lineare Regression (Regressionsmodell: Y_i stetig)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$$

- Multiple Logistische Regression (Klassifikationsmodell: $Y_i = 0$ oder $Y_i = 1$)

$$P(Y_i = 1) = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}$$

$$\hat{Y}_i = \begin{cases} 1, & falls \;\; \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip} \geq 0 \\ 0, & sonst \end{cases}$$

- *Hinweis:* In der prädiktiven Modellierung werden beide Modellklassen manchmal allgemein als (generalisierte) **Lineare Modelle** für Regression bzw. Klassifikation bezeichnet, da die Prädiktoren X_j und die Modellparameter β_j linear miteinander verknüpft sind.

Verhältnis von Stichprobengröße und Anzahl der Prädiktoren

- Lineare Modelle funktionieren am besten, wenn die Stichprobengröße deutlich größer ist als die Anzahl der Prädiktoren ($N >> p$).

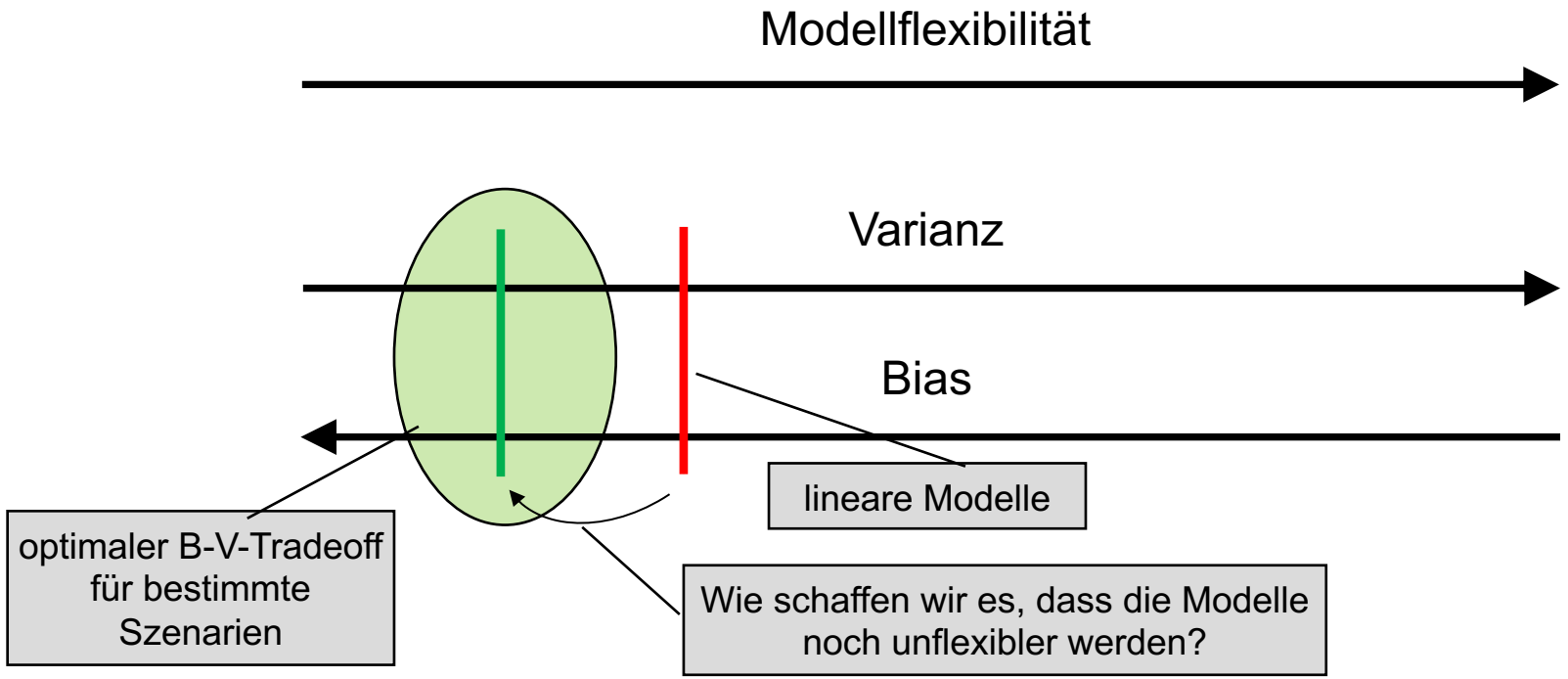
- Mit steigender Anzahl von Prädiktoren (im Verhältnis zur Stichprobengröße) wird die Vorhersagegüte der linearen Modelle typischerweise immer schlechter.

- Ist die Anzahl der Prädiktoren sogar größer als die Stichprobengröße, dann sind die Modelle überhaupt nicht identifiziert und die Modellparameter können nicht mit der Standardmethode geschätzt werden.

- **Regularisierte lineare** Modelle stellen eine Erweiterung dar, die auch bei vielen Prädiktorvariablen ($N \approx p$ oder sogar $N < p$) gute Vorhersagen liefern können.

Bias-Varianz Tradeoff I

- Lineare Modelle sind im Vergleich zu anderen prädiktiven Modellklassen eher **unflexibel**: Der Bias ist hoch, aber die Varianz niedrig.
- Trotzdem gibt es Vorhersageszenarien, in denen normale lineare Modelle immer noch *zu* flexibel sind: Ein lineares Modell mit noch weniger Varianz kann eventuell zu einer höheren Vorhersageleistung führen.



Bias-Varianz Tradeoff I

- Wie schaffen wir es, dass die Modelle noch unflexibler werden?
- Zwei Ideen um die Flexibilität von linearen Modellen und damit die Varianz der Modellklasse weiter zu reduzieren:
  - *Reduktion der Prädiktoranzahl*: Je weniger Prädiktoren bei der Berechnung der Vorhersagen im Modell berücksichtigt werden, desto weniger flexibel kann sich das Modell an die Trainingsdaten anpassen.
  - *Die Regressionsgewichte näher an die 0 ziehen*: Je näher die $\hat{\beta}_j$ an 0 sind, desto weniger Einfluss haben die Prädiktoren auf die Vorhersagen. Das Modell kann sich nicht so flexibel an die Trainingsdaten anpassen, wenn die Vorhersagen nicht so extrem ausfallen können.
- Beide Strategien erhöhen den Bias. Häufig fällt der neue Bias-Varianz Tradeoff aber trotzdem zugunsten einer besseren Vorhersageleistung aus.

Exkurs: Best Subset Selection I

- Eine naheliegende Strategie, um bessere prädiktive Modelle beim Vorliegen von vielen Prädiktoren zu erstellen:
  - Auswahl einer reduzierten Anzahl von „guten“ Prädiktoren
  - Verwendung des prädiktiven Modells nur mit den ausgewählten Prädiktorvariablen

- **Best Subset Selection** ist ein automatisierter Algorithmus für die gerade beschriebene Logik:
  - Schätzung der erwarteten Vorhersageleistung für Modelle mit allen möglichen Kombinationen von Prädiktorvariablen
  - Wahl des Modells mit der höchsten erwarteten Vorhersageleistung (bzw. dem niedrigsten erwarteten Vorhersagefehler).

Exkurs: Best Subset Selection II

- Zu beachten:
  - Die erwartete Vorhersageleistung sollte nicht mithilfe der In-sample Performance geschätzt werden: Die In-sample Performance wird *immer* für das Modell mit allen Prädiktoren am höchsten sein (Erinnerung an R^2 im Bachelor).
  - Es ist also notwendig, eine Out-of-Sample Schätzung zu verwenden. Das funktioniert am besten mithilfe von Resampling-Methoden wie *k*-fold CV. Eine einfachere Alternative stellt die Verwendung von Informationskriterien (AIC oder BIC) dar (→ das schauen wir uns im Sommersemester etwas genauer an).

Exkurs: Best Subset Selection III

- Probleme:
  - Best Subset Selection funktioniert nicht bei $N < p$, da die Modelle mit einer höheren Anzahl von Prädiktoren als die Stichprobengröße gar nicht geschätzt werden können.
  - Bei vielen Prädiktoren gibt es sehr viele mögliche Kombinationen an Prädiktoren (2^p) und daher dauert es sehr lange, die Modelle für alle Kombinationen zu berechnen. Typischerweise ist eine Berechnung ab ca. 40 Prädiktoren auch auf modernen Rechnern nicht mehr möglich.

- **Regularisierte lineare Modelle** sind moderne Alternativen zur Best Subset Selection, die auch bei sehr vielen Prädiktoren eine gute Vorhersageleistung liefern und gleichzeitig sehr effizient berechnet werden können.

Exkurs: Unterschied Lineare und Logistische Regression

- Wir erklären das Prinzip von Parameterschätzung und Regularisierung im Folgenden nur anhand der Multiplen Linearen Regression. Bei der Multiplen Logistischen Regression funktioniert dies äquivalent, nur wird anstatt der quadrierten Abweichungen $\left(Y_i - \hat{Y}_i\right)^2$ eine andere sogenannte Verlustfunktion („Binomial Deviance“) minimiert.

Lineare Regression – „Methode der kleinsten Quadrate“

- Bei der multiplen linearen Regression werden die Modellparameter mit dem Schätzalgorithmus der **Methode der kleinsten Quadrate** geschätzt.
- Dabei werden die optimalen Werte $\hat{\beta}_j$ für die Modellparameter so gewählt, dass der folgende Ausdruck den **niedrigsten** möglichen Wert ergibt:

$$\sum_{i=1}^{N}\left(Y_i - \underbrace{\left(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j X_{ij}\right)}_{\hat{Y}_i}\right)^2$$

Regularisierte lineare Regression

- Bei der regularisierten multiplen linearen Regression, werden die optimalen Werte für die Modellparameter so gewählt, dass der folgende erweiterte Ausdruck den **niedrigsten** möglichen Wert ergibt:

$$\underbrace{\sum_{i=1}^{N}\left(Y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j X_{ij}\right)\right)^2}_{\text{Modellfit}} + \underbrace{\lambda \cdot \sum_{j=1}^{p}\left|\hat{\beta}_j\right|}_{\text{Regularisierung}}$$

- λ heißt Regularisierungsparameter; λ ist immer größer oder gleich 0
- $\sum_{j=1}^{p}\left|\hat{\beta}_j\right|$ ist der sogenannte Regularisierungsterm.

Interpretation der Regularisierungsformel

- $\sum_{j=1}^{p}\left|\hat{\beta}_j\right|$ heißt Regularisierungsterm (oder auch Penalisierungsterm), weil er für hohe positive und negative Werte von $\hat{\beta}_j$ „bestraft“. Der Term „belohnt“ also möglichst kleine Werte von $\hat{\beta}_j$.
- Würde der Regularisierungsterm alleine entscheiden, wie die optimalen Werte $\hat{\beta}_j$ gewählt werden, ergäbe sich immer der Wert 0 für alle $\hat{\beta}_j$. Stattdessen findet aber ein Trade-off zwischen der Regularisierung und dem Modellfit statt.
- Der Regularisierungsparameter λ entscheidet, wie stark der Regularisierungsterm berücksichtigt wird.
  - Je größer λ, desto stärker wird für extreme Werte von $\hat{\beta}_j$ bestraft.
  - Für $\lambda = 0$ fällt der Regularisierungsterm weg und es ergibt sich die Methode der kleinsten Quadrate aus der normalen multiplen linearen Regression.

LASSO und Variablenselektion

- Lineare oder Logistische Regressionsmodelle, die mit der LASSO-Regularisierung geschätzt werden, bezeichnet man häufig allgemein als LASSO-Regression oder einfach nur LASSO.
  - LASSO = „least absolute shrinkage and selection operator“
- Die LASSO-Regularisierung hat **zwei Effekte** auf die Regressionsgewichte:
  - (1) manche Regressionsgewichte werden exakt auf den Wert 0 geschätzt
    - Prädiktoren mit einem Schätzwert von $\hat{\beta}_j = 0$ werden bei der Berechnung von Vorhersagen gar nicht mit berücksichtigt.
    - Damit führt LASSO automatisch zu einer Auswahl von Prädiktoren: Diejenigen mit einem Regressiongewicht von exakt Null „fallen raus“.
  - (2) Diejenigen Gewichte, die „drin bleiben“ (also nicht exakt auf 0 geschätzt werden), werden näher an die Null herangezogen (da kleinere Gewichte vom Regularisierungsterm belohnt werden.

LASSO und Variablenselektion

- Damit ist das Resultat beim LASSO ähnlich wie bei der Best Subset Selection: Man erhält ein vereinfachtes lineares Modell bei dem nur eine reduzierte Anzahl an Prädiktoren verwendet wird. Es lässt sich auch mathematisch zeigen, dass die LASSO-Regularisierung ein sehr ähnliches Optimierungsproblem löst wie die Best Subset Selection.

Bias-Varianz Tradeoff II

- Regularisierte lineare Modelle sind weniger flexibel als normale unregularisierte lineare Modelle (bei gleichem Prädiktorset):
  - LASSO Regularisierung setzt manche Modellparameter auf 0 und entfernt damit der Einfluss der entsprechenden Prädiktoren auf die Vorhersagen.
  - Bei Modellparametern die nicht auf 0 gesetzt werden, sind die Parameterschätzungen näher an 0 als ohne Regularisierung.

- Damit verschiebt sich der Bias-Varianz Tradeoff durch die Regularisierung:
  - Höherer Bias
  - Niedrigere Varianz
- In Situationen mit vielen Prädiktoren im Vergleich zur Stichprobengröße, ist der Trade-Off der regularisierten Modelle oft vorteilhaft für die Vorhersageleistung. (Achtung: In anderen Vorhersageszenarien ist wiederum ein anderer Bias-Varianz-Tradeoff besser).

Exkurs: Standardisierung der Prädiktorvariablen

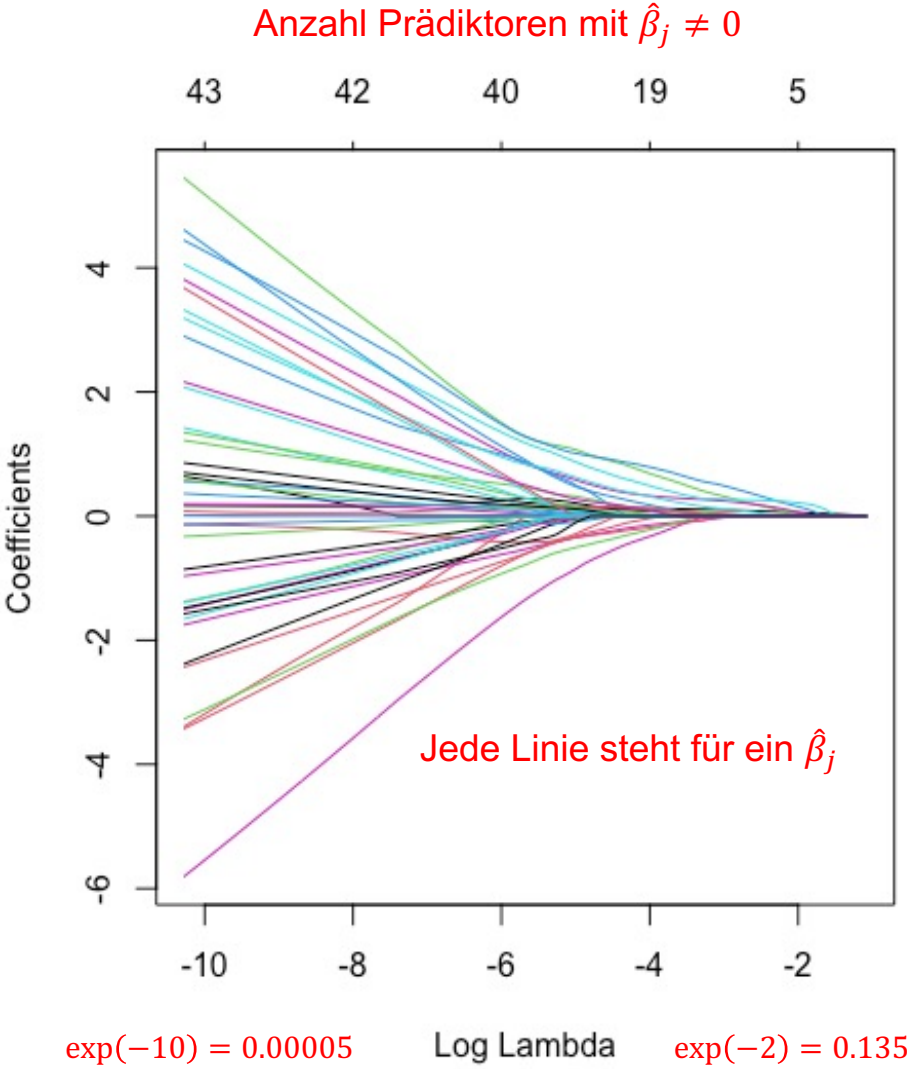
- Die Größe der Modellparameter hängt mit der Einheit der Prädiktoren sowie des Kriteriums zusammen und ist somit in gewissem Maße willkürlich (z.B. wenn $\beta_{Größe\_in\_cm} = 0.1$, dann $\beta_{Größe\_in\_m} = 10$).
- Da der Regularisierungsterm für extreme Werte bestraft und dabei alle Prädiktoren gleich behandelt, *müssen* die Einheiten der Prädiktoren sinnvoll vergleichbar sein. Ansonsten würde ein Parameter stärker regularisiert werden, wenn für den dazugehörigen Prädiktor eine Einheit gewählt wurde die zu größeren β Werten führt.
- Weil Einheiten zwischen den Prädiktoren oft sehr unterschiedlich sind und eine Vereinheitlichung inhaltlich oft nicht möglich ist, werden in den meisten Programmen *vor der Schätzung* der Modellparameter standardmäßig alle Prädiktorvariablen *z*-standardisiert. Die endgültigen Parameterschätzungen werden dann aber wieder in der unstandardisierten Einheit angegeben.

Klassifikationsbeispiel: Depression

- Stichprobe: 183 Personen
- Kriteriumsvariable: ICD10 Diagnose für Depression vorhanden ($Y_i = 1$) oder nicht ($Y_i = 0$)
- Prädiktoren: Alter, Geschlecht, 21 BDI Items, 30 FIE Items
- D.h.: 53 Prädiktoren für 183 Fälle; *SPV* (subject-to-variable ratio) = 3.5
  - Als grobe Faustregel wird von verschiedenen Autoren ein minimales SPV von 10 (Harrell, 2001) bis zu 25 (Jenkins & Quintana-Ascencio, 2020) angegeben.

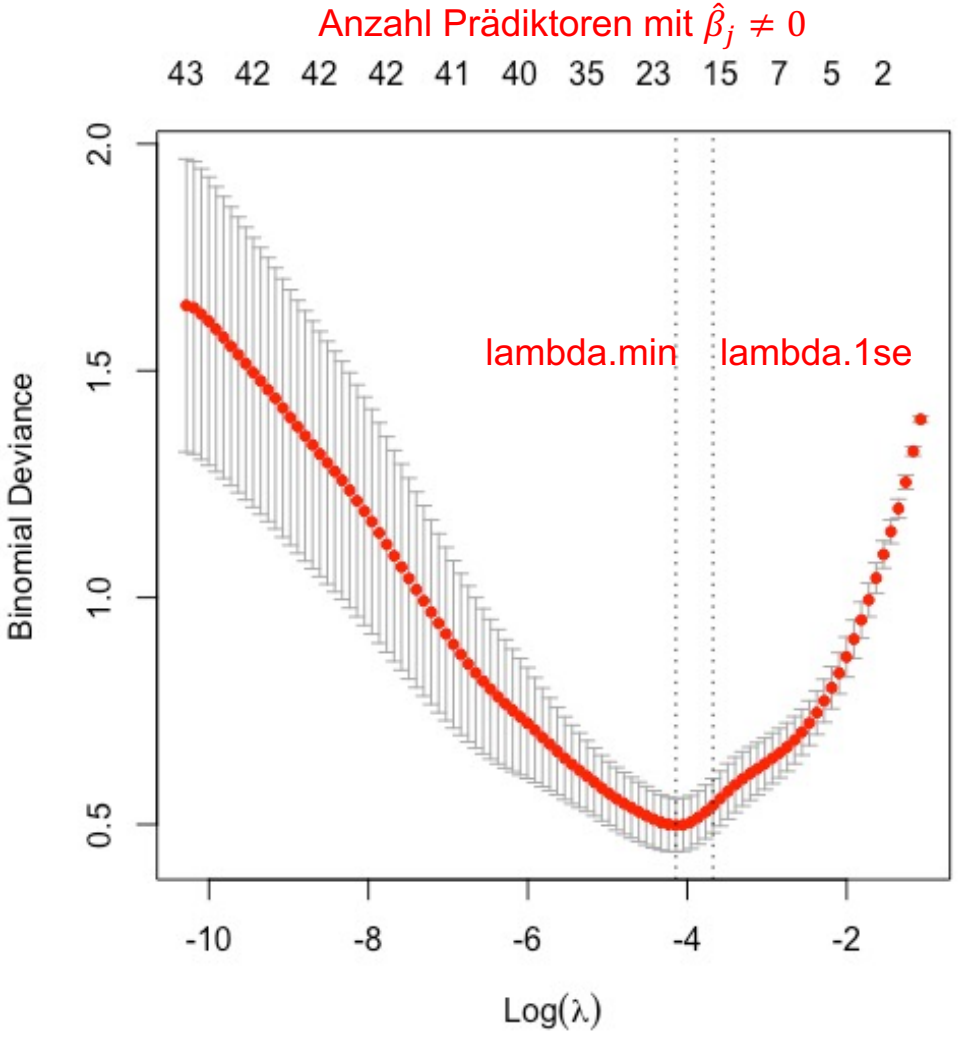
Harrell, F. E. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer New York. https://doi.org/10.1007/978-1-4757-3462-1
Jenkins, D. G., & Quintana-Ascencio, P. F. (2020). A solution to minimum sample size for regressions. PLOS ONE, 15(2), e0229345. https://doi.org/10.1371/journal.pone.0229345

Einfluss von λ auf die Parameterschätzungen $\hat{\beta}_j$



- Für jeden möglichen Wert von λ ergeben sich andere Schätzwerte $\hat{\beta}_j$
- Problem:
  Welches λ soll gewählt werden?
- Lösung: Hyperparameter Tuning
  - Bei welchem λ Wert hat das Modell die beste erwartete Vorhersageleistung?
  - Datengesteuerte Wahl des **Hyperparameters** λ mit Kreuzvalidierung

Hyperparametertuning von λ mithilfe von 10-fold CV



- Aufteilung des Datensatzes in Trainings- und Testsets (z.B. 10-fold CV)
- Parameterschätzungen $\hat{\beta}_j$ mit jedem Trainingssets berechnen für eine Reihe von λ Werten
- Vorhersagen berechnen mit den entsprechenden Testsets
- Wahl des Wertes für λ, bei dem der erwartete Vorhersagefehler (MSE bei Regression, Binomial Deviance bei Klassifikation) am niedrigsten ist
- Konservative Alternative: Minimum + 1 SE

Limitationen der Lasso Regularisierung

- Der Schätzalgorithmus des LASSO ist nicht deterministisch: Mit einem anderen Seed können sich die Parameterschätzungen quantitativ (Höhe der Schätzwerte) aber auch qualitativ ändern (d.h. welche Variablen ausgewählt werden).
- Bei manchen Datensätzen (vor allem bei kleinen Stichprobengrößen) kann die Variablenselektion so instabil sein, dass eine inhaltliche Interpretation der Variablenselektion nicht gerechtfertigt erscheint.
- Die Variablenselektion beim LASSO ist häufig dann instabil, wenn die Prädiktoren hoch miteinander korrelieren.
- (Eine Erweiterung der LASSO Regularisierung, die auch zu einer Variablenselektion führt und besonders bei korrelierten Prädiktoren oft eine bessere Vorhersageleistung aufweist ist die sogenannte **Elastic Net Regularisierung.)**

Interpretation der Modellparameter

- Der Output eines geschätzten LASSO Modells sieht vom Prinzip genauso aus, wie bei der entsprechenden unregularisierten linearen oder logistischen Regression:
  Ein Schätzwert für jedes β_j, wobei manche $\hat{\beta}_j$ den Wert 0 annehmen (zusätzlich sollte der gewählte Wert für λ angegeben werden).
- Modellparameterschätzungen in regularisierten linearen Modellen können genauso interpretiert werden wie in der normalen multiplen linearen und multiplen logistischen Regression.
- Gleichzeitig reduziert die LASSO-Regularisierung durch die Variablenselektion die Anzahl der Prädiktoren, die überhaupt einen Einfluss auf die Vorhersagen haben. Damit sind LASSO Modelle vielleicht sogar noch einfacher zu interpretieren als lineare Modelle ohne Regularisierung.

Interpretationsbeispiel: Depression


```
(Intercept) -4.678
bdi_c        0.693
bdi_b        0.563
bdi_m        0.385
bdi_f        0.295
bdi_u        0.234
bdi_n       -0.184
fie_2       -0.129
alter        0.117
fie_6       -0.117
fie_1       -0.104
bdi_t        0.097
fie_11       0.086
bdi_a        0.077
bdi_s        0.077
bdi_p        0.072
fie_9        0.019
bdi_d        0.000
bdi_e        0.000
. . .        0.000
```

- Optimales λ = 0.025
- Bei 37 Prädiktoren wurde $\hat{\beta}_j$ auf exakt 0 geschätzt; bei 16 auf ≠ 0
- Interpretation von $\hat{\beta}_{bdi\_c}$:
  - Erhöht sich die Itemantwort auf dem BDI Item c um einen Punkt und die Werte auf den anderen Prädiktoren bleiben gleich, dann …
  - erhöhen sich die vorhergesagten **Log-Odds** für das Vorliegen einer Depressionsdiagnose um 0.693
  - erhöhen sich die vorhergesagten **Odds** für das Vorliegen einer Depressionsdiagnose um den Faktor $\exp(0.693) = 2.000$

Ausblick

- Achtung: Wir haben in unserem Depressionsbeispiel noch nicht betrachtet, wie hoch die erwartete Vorhersageleistung (geschätzt mit Resampling Methoden wie 10-fold CV) des auf der letzten Folie dargestellten LASSO Modells (auch im Vergleich zu anderen Modellklassen) ausfällt. Solche Beispiele verschieben wir auf die spätere Sitzung zu **Benchmark Analysen**.
- Wird wie hier der Betrag von $\hat{\beta}_j$ verwendet, handelt es sich um die sogenannte **LASSO-Regularisierung**. Die häufigste Alternative ist die **Ridge-Regularisierung** mit dem Regularisierungsterm $\sum_{j=1}^{p}\left(\hat{\beta}_j\right)^2$.
  - Die Ridge-Regularisierung führt nicht zu einer Variablenselektion. Hier sind Parameterschätzungen eventuell sehr nahe 0, aber nie exakt 0.
  - Die Variablenselektion beim LASSO ist häufig dann instabil, wenn die Prädiktoren hoch miteinander korrelieren. In diesen Fällen liefert die Ridge-Regularisierung häufig eine bessere Vorhersageleistung (allerdings leider keine Variablenselektion).