# [GK] DaSci 7.3 Statistische Modellierung – 4h

Done: View | Done: Make a submission | To do: Receive a grade

**Opened:** Monday, 18 September 2023, 12:00 AM
**Due:** Wednesday, 20 December 2023, 12:00 AM

## Data Science ``Statistische Modellierung'' - Taskdescription

### Einführung

Die Theorieeinheiten sollen als Einführung in die Statistische Modellierung dienen. Zum Umgang mit R und RStudio oder Python ist der Link aus den Theorie Einheiten empfohlen.

### Ziele

Das Ziel ist es das Grundwerkzeug zur Statistischen Modellierung zu verstehen, einsetzen und interpretieren zu können.

### Voraussetzungen

- Kenntnis von Markdown, LaTeX und Grundkenntnisse in R und Python
- funktionstüchtige Installation von R, RStudio, Python auf eurem Rechner, virtueller Maschine etc.
- Kenntnis von Scatterplot Matrix, linearer Regression, Quality Plots; Residuen, Regressionskoeffizienten; Regressionsmodell, Modellgleichung;
- Kenntnis von logistischer Regression

### Aufgabe 1:

1. Lade den Datensatz 'state.x77' in R. Beschreibe die Daten anhand der internen Hilfe.
2. Ermittle ein lineares Regressionsmodell, dass die Mordrate ('Murder') durch die unabhängigen Variablen Population, Income, Illiteracy,und Life Exp(ectancy) erklärt. Schreibe die Modellgleichung an und interpretiere die Werte der Koeffizienten im Kontext.
3. Führe alle fünf für dieses Regressionsmodell geltenden Modellvoraussetzungen an und überprüfe diese Voraussetzungen nachweislich anhand der Zusammenfassung (summary), Quality Plots der Regression und der pairwise Scatterplot Matrix. Erkläre, ob diese Modell überhaupt gültig ist. Falls es gültig ist, gib die Qualität der Erklärung durch das Modell an.
4. Führe eine Modellselektion der relevanten erklärenden Variablen durch.
5. Vergleiche die Paramaterschätzungen und Anpassungen des linearen Regressionsmodells aus Punkt 1) mit Modellen, die LASSO Regression, Ridge Regression und elastic net Regression (bei unterschiedlichen Parametereinstellungen für lambda) anpassen. (EK)

### Aufgabe 2:

1. Installiere das Package 'MASS' mithilfe der Funktion install.packages. Lade den Datensatz 'Pima.tr' in R. Beschreibe die Daten anhand der internen Hilfe.
2. Ermittle ein logistisches Regressionsmodell, dass das Auftreten von Diabetes ('type') durch die übrigen unabhängigen Variablen Alter (age), Anzahl der Schwangerschaften (npreg), BMI, Glukosespiegel (glu), Blutdruck (bp), familiäre Häufung von Diabetesfällen (ped) und Hautfaltendickemessung am Oberarm (skin) erklärt. Schreibe die Modellgleichung an und interpretiere die Werte der Koeffizienten im Kontext.
3. Führe eine Modellselektion der relevanten erklärenden Variablen durch, die LASSO Regression, Ridge Regression und elastic net Regression (bei unterschiedlichen Parametereinstellungen für lambda) anpassen. (EK)
4. Ermittle die prädiktive Qualität des Modells mithilfe einer Receiver Operating Characteristic (ROC) Kurve. Führe auch die False Positive, False Negative, True Positive und True Negative Raten in einer Tabelle an. (EK)

### Abgabe

Das Protokoll ist als PDF-Dokument abzugeben vorzulegen, welches die graphischen Darstellungen und Interpretationen in ganzen deutschen Sätzen enthält.

### Bewertung

Gruppengrösse: 1 Person

### Anforderungen überwiegend erfüllt

- aktuelle Markdown- oder LaTeX-Protokollvorlage aus Github bzw. Moodle verwendet
- grundlegende Beschreibung und Verwendung der im Unterricht angeführten Informationen und Begriffe: Scatterplot Matrix, lineare Regression, Quality Plots; Residuen, Regressionskoeffizienten; Regressionsmodell, Modellgleichung; logistische Regression
- Codebeispiele referenziert

### Anforderungen zur Gänze erfüllt

- zusätzliche zu den grundlegenden Aufgabenstellungen vertiefende Aufgabenstellungen zu den einzelnen Kapitel durchgeführt (EK)
- Verbale Beschreibung und Erklärung aller angeführter Begriffe: LASSO Regression, Ridge Regression und elastic net Regression; Receiver Operating Characteristic (ROC) Kurve; Confusionmatrix, False Positive, False Negative, True Positive und True Negative und deren Anwendung in konkreten Beispielen in vollständigen deutschen Sätzen
- ausführliche Codebeispiele und Visualisierungen dokumentiert

### Quellen

Edit submission | Remove submission

## Submission status

| | |
|---|---|
| **Submission status** | Submitted for grading |
| **Grading status** | Not graded |
| **Time remaining** | Assignment was submitted 34 days 12 hours late |
| **Last modified** | Tuesday, 23 January 2024, 12:53 PM |
| **File submissions** | Statistische-Modellierung.pdf 23 January 2024, 12:53 PM |
| **Submission comments** | Comments (0) |