# Syntax: Motivation

Syntaxgebrauch ermöglicht Analysen zu

- replizieren und reproduzieren
- kommunizieren und diskutieren
- dokumentieren zwecks Transparenz und Verständlichkeit
- archivieren auf lange Sicht und unmittelbar,
  z. B. um die zahlreichen, oft fehleranfälligen und zeitaufwändigen Datenkorrekturen und –transformationen zu automatisieren



# Syntax

- Beispieldateien: C:/Programme/SPSS/Tutorials/sample_files
- `Einfügen`–Button in den meisten Dialogfeldern
- Prozeduraufbau: Name Argumente /Subkommando .
- Kommentarbeginn: `*` oder `COMMENT`
  Kommentarende: `.`
  Kommentarblock innerhalb eines Befehls: `/* ..... */`
- `TITLE`, `SUBTITLE` Befehle zur Strukturierung des Outputs, etwa Seitenumbruch mit `TITLE ' '`.
- Ergebnisdatei soll abgearbeitete Syntax beinhalten,
  `Bearbeiten > Viewer`: eingeblendeter Anzeigestatus des Log–Fensters und Befehle im Log–Fenster anzeigen

# Syntax

- `F1` ”listet” alle zum über Cursorposition ausgewählten Kommando vorhandenen Unterbefehle auf, so wie sie die Befehlssyntaxreferenz (PDF) enthält.
- `Ausführen` einschränkbar auf Auswahl oder aktuellen Befehl allein bzw. bis zum Ende des vorhandenen Codes
- `CRTL + R` oder Submit–Symbol zum Ausführen des aktuellen Befehls
- `EXECUTE.` sofortige Ausführung von Transformationsbefehlen
- `DATASET NAME <name>.` und `DATASET ACTIVATE <DatenSet>.` zur Benennung und kontrollierten Aktivierung eines Datensatzes



# Einfache lineare Regression

- Zurückführen/Rückgang auf das mittlere Niveau
- sachlogische Argumentation, keine Kausalitätsanalyse
- hier: Analyse der linearen statistischen Abhängigkeit zweier metrischer Variablen
  prinzipiell auch: Analyse zwischen metrischer und kategorieller Variable (ANOVA)
- Modell:
  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

# Einfache lineare Regression: Modell

Matrixnotation: $\boldsymbol{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

| | |
|---|---|
| $\boldsymbol{y}$ | Regressand, abhängige/Ziel-/Antwortvariable, Response |
| $\mathrm{X} = (\mathbf{1}\ \boldsymbol{x})$ | Regressionsmatrix mit Regressoren, erklärenden/unabhängigen/Einfluss-/Ko-Variablen, Prädiktoren |
| $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ | Regressionskonstante/-koeffizient; unbekannte interessierende Größen (Effektgrößen/Effekte) |
| $\boldsymbol{\varepsilon}$ | Residuum, additive zufallsbedingte Rest-/Störvariable; nicht direkt beobachtbar |



# Einfache lineare Regression: Annahmen

- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

  $\rightarrow$ $\boldsymbol{y}$ ist Zufallsvariable
  - unabhängige Zufallsgrößen $\varepsilon_i$ bzw. $y_i$
  - Homoskedastizität: $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 \equiv \sigma^2$
  - Heteroskedastizität: $\longrightarrow$ gewichtete lineare Regression oder Datentransformation

- weitere unbekannte Parameter?

- Schätzung über Kleinste–Quadrate–Methode basierend auf Stichprobe $(y_i, x_i), i = 1, \ldots, n$

# Regression: terminologische Konvention

- in der Statistik:
  - Einflussvariable nimmt eine über die Fälle variable Ausprägung an
  - Parameter sind unbekannt, müssen geschätzt werden
- in der Technik:
  - Einflussgrößen stellen Stellschrauben dar, z. B. Parameter einer MR–Aufnahmesequenz
  - Regressionskoeffizienten werden oft als Variablen bezeichnet, da hier die Werte mit dem Experiment noch nicht feststehen



# Einfache lineare Regression

- `Analysieren > Regression > Linear...`
- Beispiel anhand von `salat.sav`:
  Inwieweit lässt sich das Gewicht durch Körpergröße erklären?
- Auswahlvariable, z. B. Geschlecht
- mit/ohne Fallbeschriftungen
- Statistiken, Diagramme, Save, Optionen
- $R^2$ für Modell mit/ohne Intercept auf dieselbe Weise interpretierbar?

## Einfache lineare Regression: Output

- Modellzusammenfassung mit $R^2$ — Was sagt $R^2$ aus? — und korrigiertem $R^2$ (Principle of Parsimony) — Was bedeutet korrigiert?

- ANOVA-Tabelle:
  Ist das Modell fähig, Variation im Response zu erkennen?
  Zusammenhang mit dem Bestimmtheitsmaß?

- Koeffiziententabelle mit
  - B und Standardfehler
  - Beta: unabhängig von Messeinheiten $\rightarrow$ absolute Bedeutsamkeit eines Regressors, Vergleich multipler Regressoren
  - T und Signifikanz



## Einfache lineare Regression versus Korrelation

- $\rho^2 = R^2$

- Korrelation ist symmetrisch, Regression asymmetrisch

- **Übung:**
  1. Betrachten Sie die Streudiagramme ”Gewicht vs. Größe” und ”Größe vs. Gewicht”.
  2. Vergleichen Sie die zugehörigen Regressionsgeraden.
  3. Führen Sie beide Regressionsmodelle mit standardisierten Variablen aus. Hinweis zur Standardisierung:
     `Analysieren > Deskriptive Statistiken > Deskriptive Statistiken...`

## Regression: Modelldiagnose

- graphisch anhand von Residualplots, meist $\hat{\varepsilon}_i$ versus $x_i$
  bei systematischer Streuung kann Datentransformation helfen
- Streudiagramm dient auch zur Überprüfung auf gleiche Varianzen
- numerisch anhand der Anpassungsgütemaße
- Einfluss einzelner Datenpunkte auf die Koeffizientenschätzer und die Gesamtanpassung mittels Cook’s Distance und Hebelwerten (Leverage Points)



## Regression: Konfidenzintervalle und Prognose

- `Statistik: Konfidenzintervalle` Intervallschätzer der Parameter
- `SAVE: Konfidenzintervalle individuell und Mittelwert bezogen`
- Prognose der Zielvariablen für einen neuen Wert $x_0$ des Regressors $X$.
  - neue Fälle zum Datensatz hinzufügen, z. B. Größe von Christian Heumann.
  - Kreiere Filtervariable über `not(missing(Gewicht))`.
  - Modellanpassung mit `Auswahlvariable Gewicht=1` und Speichern der vorhergesagten Werte. .
  - Heumann müsste xx kg wiegen. Verlässliche Prognose?

# Regression

- `Analysieren > Allgemeines lineares Modell > Univariat`
- `Analysieren > Verallgemeinerte lineare Modelle > Verallgemeinerte lineare Modelle...`
  `Verteilung: Normal, Verknüpfungsfunktion: Identität`
- einfache Spezifikation von Interaktionstermen im Aufruf zum verallgemeinerten linearen Modell
  Voraberzeugung von Interaktionstermen über `Transformieren > Variable berechnen` für `Analysieren > Regression > Linear...`
- (Multi–)Kollinearität der Regressoren:

$\longrightarrow$ große Standardfehler der Regressionskoeffizienten
$\longrightarrow$ geringere Power der Prädiktoren
$\longrightarrow$ schlecht konditionierte bis zu singulärer Kovarianzmatrix
$\longrightarrow$ Interpretierbarkeit der Regressionskoeffizienten?



# Übung

1. Lesen Sie den Datensatz `kurse.asc` ein. Siehe auch http://www.statistik.lmu.de/service/datenarchiv/aktien/aktien.html. Importieren Sie zunächst die Datumsvariable im Stringformat.

2. Formatieren Sie die Datumsangabe gemäß '`MMM JJ`' Schreibweise mit Hilfe des Menüpunktes `Variable berechnen`.
   Verwenden Sie dazu die Befehle `DATE.DMY`, `NUMBER` und `SUBSTR(2)`.
   Die finale Formatzuweisung ist nur über das Syntaxfenster möglich.
   Konsultieren Sie hierzu die Hilfe zu `FORMATS`.

3. Berechnen Sie die Rendite sowohl für die Aktie der Münchner Rückversicherung als auch für den Gesamtmarkt (DAX) gemäß der Formel:

$$\text{rendite}(t) = \log\left(\frac{\text{kurs}(t)}{\text{kurs}(t-1)}\right)$$

# Übung

4. Aggregieren Sie jeweils beide Kurse und berechneten Renditenverläufe gemäß dem monatlichen Median und das formatierte Datum gemäß dem Ersten Wert in einen neuen Datensatz. Legen Sie gegebenenfalls eine numerische Breakvariable an.

5. Öffnen Sie den Datensatz `renditem.asc` und formatieren Sie die Datumsvariable erneut.

6. Untersuchen Sie den Einfluss der Marktrendite auf die Rendite der MRU–Aktie für den Zeitraum Juni’91 bis Dez’93 mittels einfacher linearer Regression.

7. Trifft die NV–Annahme zu?

8. Kann von Homoskedastizität ausgegangen werden?



# Übung

9. Überprüfen Sie die Zielvariable auf Autokorrelation, d. h. zwischen $(Y_i, Y_{i+1}), i = 1, \ldots, n$.

10. Wie lautet die Regressionsgleichung?

11. Wie lautet die Schätzung für $\sigma^2$?

12. Überprüfen Sie die Hypothesen $H_{00} : \beta_0 = 0$ und $H_{01} : \beta_1 = 1$ anhand der 95%–Konfidenzintervalle und beurteilen Sie das Risiko der MRU Aktie gegenüber dem Marktrisiko anhand der Regressionskoeffizienten.

13. Prognostizieren Sie die Rendite der MRU-Aktie für die Marktrenditen 0.042, 0.046 und 0.005. Wie verhalten sich die zugehörigen Prognoseintervalle?