Übung 06 - Datenanalyse und Statistik WS 2008/2009

```
Laden Sie die Dateien "lohndata.txt" und "emdata.txt" von der Website
herunter, von wo Sie auch die Übungen beziehen und lesen Sie die
Datensätze in R ein!
```

Aufgabe 1: Die meisten Menschen besitzen gewisse Vorstellungen darüber, wie hoch ihr Gehalt einmal sein sollte; Muss man, um diese zu erfüllen, Ingenieur werden? Könnte man nicht ebenso gut auf Bali unter Palmen sitzend Schildkröten aufpäppeln? Anwalt sein in Havanna?
In *lohndata* finden sich Angaben von Angehörigen entsprechender Berufsgruppen über ihren durchschnittlichen Monatslohn (in US-Dollar) und dem Bruttoinlandprodukt (in Millarden US-Dollar) des jeweiligen Landes...

(1) Sie vermuten zunächst einen einfachen, linearen Zusammenhang zwischen der Höhe des Einkommens und dem Bruttoinlandsprodukt. Welches Verfahren könnten Sie unter der Annahme, dass alle nötigen Voraussetzungen erfüllt sind, einsetzen, um diesen Zusammenhang näher zu beschreiben?

---

(2) Stellen Sie eine allgemeine Gleichung für obigen Zusammenhang auf!

---

(3) Nachfolgender R-Code führt das in Aufgabe 1 zu benennende Verfahren durch. Wie groß ist der Anteil der Varianz, den dieses Modell erklärt? Der relative Anteil welcher Varianz ist dies? Worüber gibt dieser Anteil Auskunft (Tip: Warum betrachte ich ihn überhaupt)? Existiert ein signifikanter Zusammenhang zwischen dem Einkommen und dem BIP?

```
model <- formula(lohn ~ bip)
rsq   <- 1- var(resid(lm( model, data=lohndata))) / var(lohndata$lohn)
lm( model, data=lohndata)
anova(lm( model, data=lohndata))
cat("R^2=",rsq)
```

(4) **Stellen Sie unter Berücksichtigung ihrer Ergebnisse aus Aufgabe 4 ein lineares Modell auf, welches zusätzlich die Möglichkeit betrachtet, dass manche Berufsgruppen (unabhängig vom BIP) einfach mehr Geld bekommen als andere.**

Tip: Cool bleiben und Nerven bewahren!

---

(5) **Unten wird das in Aufgabe 4 verlangte Modell angewendet. Geben Sie eine begründete Aussage, ob dieses Modell besser angepasst ist, als das erste Modell! Sind beide Einflüsse signifikant?**

```
model <- formula(lohn ~ beruf + bip)
rsq   <- 1- var(resid(lm( model, data=lohndata))) / var(lohndata$lohn)
lm( model, data=lohndata)
anova(lm( model, data=lohndata))
cat("R^2=",rsq)
```

(6) **Betrachten Sie die ausgegebenen Parameterwerte. Wo sind die Anwälte abgeblieben? Stellen Sie die ermittelten Modellgleichungen einmal für die Anwälte und einmal für die Ingenieure auf!**

---

(7) **Visualisieren Sie einmal skizzenhaft, wie die Daten entsprechend dem Modell aus Aufgabe 3 verteilt sind. Vergleichen Sie dies mit einem Streudiagramm (Lohn in Abhängigkeit vom BIP) - worin besteht der qualitative Unterschied?**

---

(8) Stellen Sie zunächst die allgemeine Gleichung zu nachfolgendem linearen Modell auf! Stellen Sie dann die angepassten Gleichungen für die Berufsgruppen der Anwälte und der Biologen auf!

```
model <- formula(lohn ~ beruf + bip + beruf*bip)
rsq   <- 1- var(resid(lm( model, data=lohndata))) / var(lohndata$lohn)
lm( model, data=lohndata)
anova(lm( model, data=lohndata))
cat("R^2=",rsq)
```

(9) Ist das letzte Modell besser angepasst als die beiden vorherigen? Besteht zwischen den Berufsgruppen ein signifikanter Unterschied im Zusammenhang von Einkommen und BIP?

Tip:

Aufgabe 2: In dem Datensatz *emdata* sind Auskünfte von jungen ($\leqq$ 65 Jahre) und junggebliebenen ($>$ 65 Jahre) Menschen zusammengetragen worden, die Auskunft über ihre Hörgewohnheiten bezüglich bestimmter Musiker gaben. Protokolliert ist, wie viele Stunden im Monat jeweils dem Rapper Eminem, dem Volksmusiker Hansi Hinterseer sowie Herrn Florian Silbereisen (ebenfalls Volksmusiker) gelauscht wurde.

(1) Versuchen Sie zunächst einmal unter der Annahme, das alle nötigen Voraussetzungen erfüllt sind, die Hördauer von Eminem mit der Hördauer von Hansi Hinterseer zu prognostizieren! Wie lautet das allgemeine lineare Modell hierzu? Was sagt das Bestimmtheitsmaß?

```
model <- formula(eminem ~ hansi)
rsq   <- 1- var(resid(lm( model, data=emdata))) / var(emdata$eminem)
lm( model, data=emdata)
anova(lm( model, data=emdata))
cat("R^2=",rsq
```

(2) Ziehen Sie die Hördauer von Florian Silbereisen als zusätzliche, erklärende Variable hinzu, ohne jedoch eine Interaktion zwischen Hansi und Florian anzunehmen. Wie lautet hierzu das lineare Modell?

---

(3) Es folgt die Berechnung des Modells aus Aufgabe 2. Welche der folgenden Aussagen sind richtig (wurden statistisch signifikant nachgewiesen)?

- **O** Die Hördauer von Eminem ist unabhängig von der von Herrn Hinterseer
- **O** Es besteht ein Zusammenhang zwischen der Hördauer von Eminem und der von Florian Silbereisen
- **O** Die Hördauer von Eminem ist unabhängig von der von Herrn Silbereisen
- **O** Es konnte kein Zusammenhang zwischen der Hördauer von Eminem und der von Florian Silbereisen nachgewiesen werden

```
model <- formula(eminem ~ hansi + silber)
rsq   <- 1- var(resid(lm( model, data=emdata))) / var(emdata$eminem)
lm( model, data=emdata)
anova(lm( model, data=emdata))
cat("R^2=",rsq)
```

(4) Betrachten Sie das Streudiagramm: Hördauer Eminem in Abhängigkeit von Hördauer Hansi Hinterseer. Ist es wahrscheinlich, dass Sie eine Interaktion zwischen zwei Variablen benötigen, um die Hördauer von Eminem vorherzusagen? Worauf gründen Sie ihre Erkenntnis?

---

(5) Formulieren Sie ein möglichst gut angepasstes lineares Modell!

---

(6) Geben Sie basierend auf nachfolgendem Modell die angepassten Gleichungen für beide Altersgruppen an!

```
model <- formula(eminem ~ age + hansi + age*hansi)
rsq   <- 1- var(resid(lm( model, data=emdata))) / var(emdata$eminem)
lm( model, data=emdata)
anova(lm( model, data=emdata))
cat("R^2=",rsq)
```

(7) Geben Sie unter Bezug auf Aufgabe 6 eine begründete Vorhersage der durchschnittlichen Hördauer von Eminem für einen 31 jährigen Menschen an, welcher etwa 80 Stunden im Monat Herrn Hansi Hinterseer lauscht.

Tip: Achten Sie auch auf realistische Werte!
