Serie 2

1. Hier wollen wir ein paar der in der Vorlesung erwähnten univariaten Kenngrössen berechnen.

Ein erster Datensatz besteht aus den Bevölkerungszahlen der Schweizer Kantone im Jahr 1987. `d.Kt` ist der Name des "Kantonsdaten"-`data.frame`. Mit `summary(..)` erhalten Sie wieder einen ersten Überblick. Mit `attach(d.Kt)` haben Sie die einzelnen Variabeln als S-Objekte zur Verfügung. Berechnen Sie Mittel, Median, und gestutzte Mittel für die `Einw.87` Variable (S-plus: `mean(...., trim = ..)`,`median`. Zeichnen Sie auch ein "Kisten-Diagramm" (Box plot) mit `boxplot(..)`.

Wie und warum unterscheiden sich Median und Mittel ?
Was zeigt der Box plot ?

2. Nun machen wir Regression. Die Hauptfunktion dazu ist `lm(.)`. Überfliegen Sie die Dokumentation, die Sie mit `?lm` erhalten.

Wir arbeiten wieder mit den Kantonsdaten. Die einfache Regression

`lm(Scheid. ~ Einw.87)`

(*Resultat "abspeichern", d.h.* `lm.SE <- lm(...)`, *auch* `summary(lm.SE)` *benützen!*)

gibt einen hohen R^2 Wert[1].

Ist sie deswegen schon „gut“[2]? Betrachten Sie dazu eine graphische Darstellung (mit `plot(..)`). Die Gerade können Sie mit `abline(`*lm–Resultat*`)` einzeichnen.

3. Führen Sie auch eine einfache Regression mit den logarithmierten Daten (`log10(.)`) durch. Mittels `10^predict(`*lm–Resultat*`)` erhalten Sie die geschätzte Anzahl Scheidungen zurück (wieso ?). Mittels

```
> ii _ order(Einw.87)
> lines(Einw.87[ii], 10 ^ predict( lm-Resultat)[ii], lty=2)
```

können Sie diese auch in eine ursprüngliche Graphik einzeichnen.

4. Betrachten Sie die Residuen `residuals(.)` der beiden Regressionen.
Welche sind „zu gross“ ?

Mittels `qqnorm(`*Residuen*`)` können Sie in einem „Normal Plot“ die (angenommene) Normalverteilung überprüfen (Punkte sollten ungefähr auf einer Geraden liegen).
Ist die Annahme erfüllt? Wo allenfalls nicht?

Sehen Sie andere Strukturen in den Residuen?

Abgabe: Mo, 18.Nov. 94.
Präsenz: .... im HG E 27
Hilfe auch via e-mail: To: `maechler@stat.math.ethz.ch`
oder sonst: Tel.: (63) 2-3408

[1] $R^2 := 1 - \left(\sum_i (y_i - \hat{y}_i)^2\right) / \left(\sum_i (y_i - \bar{y})^2\right)$ ist ein Gütemass, $= r_{xy}^2$, r_{xy} = Korrelation$(x_., y_.)$

[2] $\hat{\sigma}$, die geschätzte Standardabweichung der zufällige Fehler, wird mit `Residual standard error` bezeichnet