# Regression I — Serie 4

**1.** Betrachten Sie das Data-Frame `cancer`. In der Variable `Pop` ist die Anzahl der erwachsenen weissen Frauen im Jahre 1960 für 301 Bezirke in North Carolina, South Carolina und Georgia abgespeichert. Die Variable `BCM` enthält die Anzahl der Frauen, die zwischen 1950 und 1960 an Brustkrebs gestorben sind.

**a)** Erstellen Sie ein Streudiagramm von `BCM` gegen `Pop` und führen Sie eine lineare Regression durch den Nullpunkt durch. Welche Voraussetzungen für die Durchführung der Kleinste-Quadrate-Methode sind nicht erfüllt?
**R-Anleitung:** Die lineare Regression durch den Nullpunkt erhält man mittels
`regr(BCM ~ Pop -1, data = cancer)` ("`-1`" = kein Achsenabschnitt).
Zeichnen Sie die Residuen gegen die angepassten Werte (Tukey-Anscombe-Plot) und kommentieren Sie. Was soll man tun? Erstellen Sie auch einen QQ-Plot.
**R-Anleitung:** Entweder "von Hand" oder durch `plot(`*`regr-Objekt`*`)` (Setzen Sie als plot-Argument `panel=panel.smooth`.)

**b)** In **a)** sehen wir eine klare Zunahme der Varianz; d.h. die Varianz ist nicht mehr konstant. Eine oft sinnvolle (varianzstabilisierende) Transformation ist z.B. die Wurzeltransformation oder die log-Transformation. Führen Sie eine lineare Regression durch den Nullpunkt von $\sqrt{\texttt{BCM}}$ gegen $\sqrt{\texttt{Pop}}$ durch und erstellen Sie einen Tukey-Anscombe-Plot und einen QQ-Plot für diese Regression. Kommentar?
**R-Anleitung:** Benützen Sie den Befehl `regr(sqrt(BCM) ~ sqrt(Pop)-1, data = cancer)`.

**c)** Führen Sie eine lineare Regression von `log(BCM)` gegen `log(Pop)` ohne den 2. Datenpunkt durch (dieser ist 0 und verursacht beim Logarithmieren Schwierigkeiten) und erstellen Sie einen Tukey-Anscombe-Plot und einen QQ-Plot für diese Regression. Kommentar?
**R-Anleitung:** Benützen Sie den Befehl `regr(log(BCM) ~ log(Pop), data = cancer[-2,])`.

**d)** Welche andere Transformationen würden Sie probieren?
(Freiwillig: ausprobieren mit *R*!)

(Quelle: Rice (1995))

**R-Hinweis:**
Um Graphiken und Outputs wie im Regressions-Skript zu erzeugen, brauchen Sie die von Werner Stahel geschriebenen und im Skript beschriebenen R-Funktionen `regr`, `g.res2x` und `plot.regr` und `print.regr`. Diese Funktionen finden Sie auf dem Internet unter:
`ftp://stat.ethz.ch/NDK/Source-NDK-9/R/regr.R`
Mit `source("ftp://stat.ethz.ch/NDK/Source-NDK-9/R/regr.R")` können Sie diese Funktionen laden. Natürlich können Sie sich die Funktionen auch auf Ihren Computer/Laptop runterladen, damit Sie später nicht immer Netzzugang brauchen, wenn Sie sie für Regressionsanalysen verwenden wollen.

**2. Residuen-Analyse:**
Die Baukosten von Atomkraftwerken hängen von vielen Grössen ab, von denen man einige einfach erfassen kann. Der Datensatz `nuk` enthält unter anderem die folgenden Variablen:

| | |
|---|---|
| `K` | Baukosten |
| `G` | Grösse |
| `D` | Datum der Baubewilligung |
| `KG` | partielle Kostengarantie des Generalunternehmers |

Die Daten wurden zusammengestellt, um abzuklären, ob die "Kostengarantie" ($\texttt{KG} = 1$) zu niedrigeren Baukosten führe. Dabei sind indirekte Effekte zu berücksichtigen: Wenn beispielsweise die Kraftwerke mit $\texttt{KG} = 1$ generell kleiner wären als die anderen, würde dies zu niedrigeren Kosten führen, ohne dass die Kostengarantie ein Grund dafür wäre. Diese Überlegung führt dazu, in einem Regressionsmodell, das neben `KG` alle sinnvollen und erfassten anderen Einflussgrössen enthält, den Koeffizienten von `KG` gegen 0 zu testen; dadurch sind indirekte Einflüsse soweit als möglich ausgeschlossen.

Wir wollen zunächst nur das Modell $\texttt{K} = \beta_0 + \beta_1\texttt{G} + \beta_2\texttt{D} + \beta_3\texttt{KG} + E$ untersuchen und prüfen, ob

**a)** die Fehler normalverteilt sind,

**b)** die Varianzen der Fehler gleich sind.

**R-Hinweise:**
Wenn Sie allfällige Ausreisser im Tukey-Anscombe-Plot identifizieren wollen, gehen Sie wie folgt vor:

```
> plot(fitted(name des fit), resid(name des fit))     # Tukey-Anscombe-Plot
> identify(fitted(name des fit), resid(name des fit))
```

Punkte, welche nach dem Aufruf von `identify` mit der linken Maustaste angeklickt werden, werden mit ihrem Index auf dem Plot markiert. Um `identify` zu beenden, klickt man auf die rechte Maustaste, währenddem sich der Cursor im Grafikfenster befindet.

**c)** Betrachten Sie auch die Scatterplots der Residuen gegen die erklärenden Variablen (Rg1-Skript 4.6.a).

**d)** Welche Änderungen am Modell schlagen Sie vor? Hat `KG` nach der Änderung des Modells einen Einfluss auf die Kosten?

Die obigen Fragen werden durch Betrachten der graphischen Darstellung geklärt. Formale Tests sind hier nicht sehr sinnvoll.

**e)** Führen Sie die gleichen Aufgaben mit den nun zur Verfügung stehenden Funktionen (z.B. `regr` und `plot.regr`) von Werner Stahel aus. Überlegen Sie sich, was die einzelnen Plots und Tabellen aussagen.

**R-Hinweis:**

```
r.lnuk <- regr(log10(K) ~ log10(G) + D + KG, data = d.nuk)
plot(r.nuk)
```