Stochastik für die Naturwissenschaften

Dr. C.J. Luchsinger

10. ANOVA/Regression

Literatur Kapitel 10

* Statistik in Cartoons: Kapitel 11
* Stahel: Kapitel 12.1, 12.2 und Kapitel 13 (ohne 13.4, 13.8 und 13.9)

10.1 Einfache Varianzanalyse (ANOVA (Analysis of Variance))

Der einfachste Fall ist die einfache Varianzanalyse oder Einweg-Varianzanalyse (Single-Factor Experiment), komplizierter Zweiweg-Varianzanalysen und mehr.

Wir entwickeln die Theorie zur ANOVA anhand eines Beispiels: In einem Agro-Konzern wird das Wachstum von Pflanzen unter vier klar kontrollierten Bedingungen untersucht. Wir nennen diese vier verschiedenen Bedingungen I-IV. Diese können zum Beispiel unterschiedliche Mengen von Düngemittel sein. Wir erhalten dann zum Beispiel folgende Tabelle des Wachstums der Pflanzen über einen gegebenen Zeitraum:

I	II	III	IV
33.3	35.5	29.6	38.5
47.8	35.4	33.4	42.4
44.4	47.6	32.8	45.5
42.9	38.8	38.8	38.9
40.9		42.8	38.9
35.5			44.5

Tabelle 4.3: Wachstum von Pflanzen unter 4 verschiedenen Bedingungen (1 Faktor); Lese-Beispiel: die vierte Pflanze unter Bedingungen I wuchs im beobachteten Zeitraum 42.9 cm hoch.

Die drei wichtigsten Skizzen in ganz MAT 183:

Wir werden jetzt ein mathematisches Modell mit k Gruppen aufstellen (oben $k = 4$):

$$Y_{ij} = \mu_j + \epsilon_{ij} \quad (1 \leq j \leq k; 1 \leq i \leq n_j). \tag{10.1}$$

Dabei bezeichne n_j die Anzahl Messungen in Gruppe j (oben (6,4,5,6)). Die totale Anzahl Beobachtungen ist

$$n = \sum_{j=1}^{k} n_j \ .$$

Die Zahl 42.9 in der Beschreibung ist dann die Realisation von Y_{41}. Wir fassen die Messgrössen also als Realisationen von Zufallsgrössen auf. Die ϵ_{ij} sind iid $\mathcal{N}(0, \sigma^2)$-verteilt. **Achtung: Gleiche Varianz in allen Gruppen!** Damit gilt

$$Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$$

und die n_j Realisationen in Gruppe j sind unabhängig voneinander. Wir interessieren uns jetzt für die Frage, ob die k Mittelwerte gleich sind oder nicht. Wir wollen in obigem Beispiel ein hohes Wachstum erzielen und den Dünger auswählen, der das höchste Wachstum hervorbringt. Die Nullhypothese lautet

$$\mu_1 = \mu_2 = \ldots = \mu_k. \qquad (\mathcal{H}_0 - \text{Hypothese})$$

Sicher werden wir aus den Daten die unbekannten Mittelwerte schätzen. Wir werden also den Mittelwert μ_1 in Gruppe 1 durch

$$\hat{\mu}_1 := \bar{Y}_{\cdot 1} := \frac{\sum_{i=1}^{n_1} Y_{i1}}{n_1}$$

schätzen und alle weiteren nach der allgemeinen Formel

$$\hat{\mu}_j := \bar{Y}_{\cdot j} := \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j}$$

wo $1 \leq j \leq k$. Damit haben wir jetzt k geschätzte Mittelwerte - sie werden mit Wahrscheinlichkeit 1 alle verschieden sein. Sollen wir jetzt einfach den grössten nehmen und die entsprechende Gruppe zum Sieger erklären? Nein! Aus unserer bisherigen Erfahrung in Kapitel 9 wissen wir, dass auch unter der Nullhypothese immer einer am grössten sein

wird *und* dass dieser eine signifikant grösser sein muss, damit wir die Nullhypothese verwerfen. Aber was heisst signifikant? Je nachdem, wie gross σ^2 ist, ist auch mit grösseren Abweichungen selbst bei Gültigkeit der Nullhypothese zu rechnen. Zudem kennen wir σ^2 gar nicht. Es scheint hoffnungslos zu sein - aber: Wir können ja auch σ^2 schätzen. Dies geschieht in 2 Schritten und wird uns (eher unerwartet) durch geschickte Umformung gleich die Lösung des Problems liefern.

1. Unter der Nullhypothese haben wir eine iid-Stichprobe und wir können alle n Datenpunkte gleichberechtigt zur Berechnung eines **G**rand **M**ean GM einsetzen: Mit $GM := \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} Y_{ij}}{n}$ können wir in einem

2. Schritt einen Schätzer für die Varianz (genauer das $(n-1)$-fache davon) angeben mit:

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-GM)^2 &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}\big((\bar{Y}_{.j}-GM)+(Y_{ij}-\bar{Y}_{.j})\big)^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{.j}-GM)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2 \\
&\quad + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{.j}-GM)(Y_{ij}-\bar{Y}_{.j}) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{.j}-GM)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2 \\
&\quad + 2\sum_{j=1}^{k}(\bar{Y}_{.j}-GM)\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j}) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{.j}-GM)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2 \\
&\quad + 2\sum_{j=1}^{k}(\bar{Y}_{.j}-GM)*0 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{.j}-GM)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2.
\end{aligned}$$

Wir haben also zusammengefasst:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-GM)^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{\cdot j}-GM)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\cdot j})^2. \qquad \text{(Fundi ANOVA)}$$

Diese Gleichung wird auch als Fundamentalgleichung der Varianzanalyse bezeichnet (deshalb "Fundi"). Was sagt sie aus?

Die gesamte Summe der quadratischen Abweichungen der einzelnen Beobachtungen vom Grand Mean (linke Seite) lässt sich aufspalten ("+") in Summe der quadrierten Abweichungen der Behandlungsmittelwerte vom Grand Mean ("zwischen den Behandlungen", erster Summand) und der Summe der quadrierten Abweichungen der einzelnen Beobachtungen vom jeweiligen Behandlungsmittelwert ("innerhalb der Behandlung", zweiter Summand).

Zwischenfrage ans Publikum: Angenommen, die Annahme gleicher Mittelwerte ist verletzt. Welcher der beiden Summanden auf der rechten Seite von (Fundi ANOVA) wird tendenziell grösser im Verhältnis zum anderen? Gehen Sie bei solchen Überlegungen in die Extreme!

Damit bietet sich als Test-Statistik folgende Grösse an:

$$V:=\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{\cdot j}-GM)^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\cdot j})^2/(n-k)}=\frac{\sum_{j=1}^{k}n_j(\bar{Y}_{\cdot j}-GM)^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\cdot j})^2/(n-k)}.$$

Diese Test-Statistik hat unter $\mathcal{H}_0$ eine $F_{k-1,n-k}$-Verteilung.

Wir werden den Ablehnungsbereich der Nullhypothese einseitig dort wählen, wo diese Statistik gross ist.

Wenn wir nur $k=2$ Gruppen haben, so kommen wir in eine bereits bekannte Situation. In welche und warum?

```
> growth<-c(33.3,47.8,44.4,42.9,40.9,...,42.4,45.5,38.9,38.9,44.5)
> dung<-rep(LETTERS[1:4],c(6,4,5,6))
> dung<-factor(dung)
> agro<-data.frame(dung,growth)
> analyse<-aov(growth ~ dung, agro)
> analyse
Call:
aov(formula = growth~dung, data = agro)
Terms:
                         dung       Residuals
Sum of Squares         113.7990     408.5105
Deg.  of Freedom         3          17
Residual standard error:  4.902043
Estimated effects may be unbalanced
> summary(analyse)
            Df     Sum Sq     Mean Sq     F value     Pr(>F)
dung         3     113.80     37.93       1.5786     0.2313
Residuals   17     408.51     24.03
> summary(agro)
dung     growth
A:6     Min.  :29.60
B:4     1st Qu.:35.50
C:5     Median :38.90
D:6     Mean :39.44
        3rd Qu.:42.90
        Max.  :47.80
```

10.2 Regression

10.2.1 Einfache Regression (simple regression)

Die einfache Regression ist das einfachste nichttriviale Beispiel eines komplexen Modells, in dem Kapitel 8 und 9 vorkommen. Schätzen und Testen sind Hauptthema für alle komplexen Modelle.

In diesem Kapitel werden wir eine etablierte Methode kennenlernen, mit der wir eine (lineare) Beziehung zwischen 2 Variablen untersuchen können (z.B. Körpergrösse von Vater (x) und Sohn (y)). Im (theoretischen) Modell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10.2}$$

werden wir die x-Variable als fest betrachten und versuchen, die Y-Variable möglichst genau durch x vorherzusagen oder durch x zu erklären (stören wird uns dabei der Störterm ϵ). Es wird dann mit Daten (x_i, y_i), $1 \leq i \leq n$, darum gehen, β_0 (engl. intercept) und β_1 (engl. slope) zu schätzen und zu testen, ob nicht z.B. $\beta_1 = 0$ gilt.

Gründe für Regressionsanalyse sind jeweils:

* Y vorhersagen (Interpolation (interessierendes $x \in [x_{(1)}, x_{(n)}]$) und Extrapolation (interessierendes $x \notin [x_{(1)}, x_{(n)}]$, heikel))

* Zusammenhang erklären anhand bisheriger Daten

10.2.1.1 Motivierendes Beispiel - Gefahren

Wir schauen den Tschernobyl-Datensatz mitsamt den dazugehörigen Plots an:

ort	regen	dist	kBq/m^2
Pripjet	0	3.16	5300.00
Chistogalovka	0	5.62	2000.00
Lelev	0	7.94	2100.00
Tschernobyl	0	14.13	2000.00
Rudki	0	15.85	800.00
Orevichi	0	28.18	2000.00
Kiew	0	89.13	21.00
Tschernikow	0	125.89	55.00
Tscherkassy	0	281.84	12.00
Minsk	0	316.23	20.00
Donezk	0	707.95	6.00
Wien	0	1000.00	3.00
Oesterreich	1	1000.00	53.00
Stockholm	0	1122.02	1.50
Gaevle	1	1258.93	31.00
SuedBayern	1	1258.93	81.00
Konstanz	1	1584.89	31.00
Irland	1	1584.89	16.00
Stuttgart	0	1590.54	1.50
Chilton	0	1995.26	1.50
Schottland	1	1995.26	17.00
Japan	0	12589.25	0.15
Japan2	1	12589.25	0.85

* **Rohdatenplot** Distanz versus Radioaktivität > nicht informativ

* Häufig muss man in der Statistik die Daten vorgängig transformieren (zum Beispiel Kehrwert nehmen, Logarithmieren, etc; mehr dazu im Stahel). Welche Datentransformation könnte Abhilfe verschaffen? Wie findet man hier diese? Entweder hat man praktische Erfahrung aus der Statistik beziehungsweise theoretische Kenntnisse aus dem Fachgebiet (hier Meteorologie, Ausbreitung von Gasen), oder wir machen folgende allgemeine Überlegung: global gesehen haben wir hier eine Punktquelle, deren Wirkung sich (anfänglich) radial (kugelförmig) ausbreitet. Wir erinnern an die beiden Transformationen, welche wir in MAT 182 kennengelernt haben:

Wir wenden hier also den Logarithmus an, sowohl für die x-, wie auch für die y-Achse. Schauen wir **Pairs-Plot** mit den transformierten Daten an.

Jetzt versuchen wir die Radioaktivität mit der Distanz allein zu erklären (oder vorherzusagen). Wir wollen uns vorerst auf die Schätzungen der Parameter in $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ konzentrieren. Mit einem noch zu besprechenden Verfahren (Kleinste Quadrate - siehe 10.2.1.3), geben wir in R ein (Befehl ist R-spezifisch)

NurDistanz<- lm(log(bq) $\sim$ log(dist))

Dabei seien die Daten bereits als "dist", "regen" und "bq" in R eingelesen worden. "lm" steht für **L**inear **M**odel, log(bq) nennt man die "Response Variable" (abhängige Variable) und log(dist) eine erklärende Variable oder einen "Predictor". Wenn wir also Daten haben und in Modell (10.2) die Parameter schätzen, erhalten wir folgende Regressionsgerade:

$$\log(bq) = 9.803 - 1.091 * \log(dist)$$

Wegen des negativen Koeffizienten (-1.091) haben wir also einen negativen (linearen) Zusammenhang: je grösser die Distanz zum Unglücksreaktor, desto kleiner die Radioaktivität.

Wenn wir jetzt noch versuchen, die Radioaktivität mit dem Regen allein zu erklären (oder vorherzusagen), gibt es eine Überraschung; wir erhalten hier nämlich erstmals:

$$\log(bq) = 3.7784 - 0.8247 * \text{regen}$$

Entgegen unseren Erwartungen haben wir auch hier einen negativen (linearen) Zusammenhang. Wieso ist das hier so (beachten Sie auch die spätere Relativierung dieses Resultats)?

Auf dem **Computer-Ausdruck hinten** finden Sie viele Zahlen, welche wir jetzt zum ersten Mal anschauen.

10.2.1.2 Welche Modellannahmen werden wir machen?

Wie überall in der Statistik werden wir ein intensives Wechselspiel zwischen Daten (x_i, y_i) und (vermuteten) Zufallsgrössen haben. Wenn wir uns auf der Ebene der Zufallsgrössen befinden, machen wir folgende Modellannahmen $(1 \leq i \leq n)$:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10.2}$$

$\epsilon_1, \ldots, \epsilon_n$ sind iid $\mathcal{N}(0, \sigma^2)$-verteilte Störterme. Auch σ^2 ist ein zu schätzender Parameter. In vielen Anwendungen subsumiert man in der residualen Grösse ϵ_i kleine Effekte wie Messfehler, Rundungsfehler, zufällige Schwankungen, kleinste Einflüsse, welche man nicht in das Modell einbauen will, um es einfach zu halten.

10.2.1.3 Schätzen von β_0, β_1 und σ^2: OLS

Wir werden jetzt eine Methode kennenlernen (OLS), mit der man die unbekannten Parameter in (10.2) schätzen kann. Wir werden die Schätzungen der unbekannten Parameter mit $\hat{\beta}_0$ und $\hat{\beta}_1$ bezeichnen. Wir definieren damit die geschätzten Punkte

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, 1 \leq i \leq n,$$

die geschätzte Gerade ist dann $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Die Aufgabe ist lediglich: wir suchen eine Gerade durch die Punktwolke $(x_i, y_i)_{i=1}^n$ derart, dass die Summe der quadrierten Fehler (Sum of Squared Errors)

$$SSE := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 := \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

minimal ist (OLS=Ordinary Least Squares; kleinste Quadrat-Schätzung). Das ist eine sinnvolle Aufgabe, denn wir wollen den Fehler (die Abweichung der geschätzten Werte $\hat{y}_i$ von den beobachteten Werten y_i) klein halten.

Es gibt genau eine solche Gerade; wir werden Sie gleich bestimmen.

Wir definieren wichtige Summen:

$$SSR := \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

"R" steht dabei für Regression, dann noch die 3 Summen

$$SS_{xx} := \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \qquad \text{(entspricht etwa } nV(X))$$

$$SS_{yy} := \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \qquad \text{(entspricht etwa } nV(Y))$$

und

$$SS_{xy} := \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}. \qquad \text{("Kreuzsumme")}$$

Schreiten wir jetzt zur OLS-Schätzung: Wir haben von SSE die Summe

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{10.3}$$

bzgl. β_0 und β_1 zu minimieren (die x_i, y_i sind fest!). Von der Analysis her wissen wir, dass man bei eindimensionalen Optimierungsproblemen die erste Ableitung gleich 0 setzt (und die zweite Ableitung noch überprüft). Dies ist auch bei mehrdimensionalen Problemen so (bei der zweiten Ableitung ist es ein bisschen schwieriger - der Dozent weiss aber, dass es so gut geht).

Die Ableitung von (10.3) nach β_0 bzw. β_1 gleich 0 gesetzt ergibt:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0; \qquad \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0.$$

Dies ist äquivalent zu

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

und

$$SS_{xy} + n\bar{x}\bar{y} - \beta_0 n\bar{x} - \beta_1 (SS_{xx} + n\bar{x}^2) = 0.$$

Nach einfachen Umformungen erhalten wir

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{10.4}$$

und

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{10.5}$$

Für die Schätzung von σ^2 nehmen wir die *beobachteten Residuen* $e_i := y_i - \hat{y}_i, 1 \leq i \leq n$, (die wahren sind ja unbeobachtet) und definieren:

$$\hat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^{n} e_i^2. \tag{10.6}$$

Warum "$(n-2)$" im Nenner? Mit dieser Wahl wird der Schätzer für σ^2 erwartungstreu. Sie beweisen als freiwillige HA, dass auch $\hat{\beta}_0, \hat{\beta}_1$ erwartungstreu sind. Wir besuchen nochmals unseren Computerausdruck und identifizieren dort die neu gefundenen Grössen (10.4), (10.5) und (10.6).

Nachtrag zu e_i und ϵ_i:

Eine kleine Betrachtung, welche uns für den Modelfit mit R^2 ("R-Squared") eine leistungsfähige Kenngrösse liefern wird: Wir haben vielen Summen definiert: $SS_{xx}, SS_{yy}, SS_{xy}$; dazu SSE und SSR. Überraschenderweise gilt folgende Beziehung (den Beweis machen Sie bitte als kleine HA, wobei Sie Schätzungen (10.5) für β_0 und (10.4) für β_1 benutzen):

$$SS_{yy} = SSR + SSE,$$

also

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{10.7}$$

Wie wird dies interpretiert? SS_{yy} ist ein Mass für die Variabilität in den uns interessierenden Daten y (die Radioaktivität!). Wir wollen die Radioaktivität möglichst präzis vorhersagen oder erklären. Mittel dazu sind die Anzahl (1 oder 2) und Auswahl (Distanz und Regen) der erklärenden Variablen. Dann wollen wir die SSE, die Sum of Squared Errors, klein machen. Das haben wir mit OLS gemacht. Je nach Anzahl und Auswahl der erklärenden Variablen wird SSE mehr oder weniger klein. Wenn wir ein kleines SSE erhalten, ist es gut. Da SS_{yy} gegeben ist (es sind die Messwerte) und in der Form $SS_{yy} = SSR + SSE$ aufgespalten wird, ist entsprechend ein grosses SSR gut. Statistiker sagen dann: "Die Variation in den y (SS_{yy}) lässt sich aufspalten in einen Anteil, der durch die Regression erklärt wird (SSR) und eine residuale Summe (SSE). Dies lädt ein, mit

$$R^2 := \frac{SSR}{SSR + SSE} \in [0, 1]$$

eine Kenngrösse anzugeben, welche sagt, wie gut der Fit des Modells an die Daten ist. Grosse Werte bedeuten einen guten Fit (Modell nur mit Distanz), kleine Werte einen schlechten Fit (Modell nur mit Regen). Sobald wir mehr erklärende Variablen einbauen, wird der Fit automatisch besser (siehe später und auch im vollen Modell). Deshalb hat man mit "R^2-adjusted" noch eine weitere Kenngrösse definiert, welche viele erklärende Variablen bestraft. Auch dies suchen wir jetzt in unserem Computerausdruck. Informativ sind auch die beiden Bilder a) mit den Boxplots der Residuen und b) der Cartoon.

10.2.1.4 Testen ob $\beta_1 = b$

Wenn wir uns im Spezialfall der einfachen Regression nur für die Steigung der Geraden interessieren, gibt es eine einfache Herleitung eines statistischen Tests ob

$$\mathcal{H}_0 : \beta_1 = b$$

gilt oder nicht, wo b eine beliebige Zahl (oft 0; default in R). Wieso treten solche Fragen überhaupt auf? Dazu zwei Skizzen, $\mathcal{H}_0$ und $\mathcal{H}_1$:

Für die jetzige Untersuchung setzen wir voraus, dass (10.2) gilt. Dann können wir folgendermassen argumentieren:

1. Die Schätzformel für β_1 lautet (vgl. (10.4)):

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

2. Weil $\sum_{i=1}^n (x_i - \bar{x}) = 0$ gilt auch $\sum_{i=1}^n \bar{y}(x_i - \bar{x}) = 0$ und damit

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

3. Schätzer sind (vor der Realisation) Zufallsgrössen. Wir setzen für die Datenpunkte y jetzt die Zufallsgrössen Y ein:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) Y_i. \tag{10.8}$$

Da $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), 1 \leq i \leq n$, unabhängig verteilt, können wir schliessen (kleine Rechnungen für Personen, welche die Mathe lieben), dass gilt:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Jetzt machen wir die Z-Transformation im Fall $\mathcal{H}_0 : \beta_1 = b$:

Das kleine Problem ist, dass wir σ^2 (wieder mal) nicht kennen. Aber genau wie in früheren Kapiteln können wir ja eine Schätzung von σ^2 (hier Formel (10.7)) zu Hilfe nehmen. Unter $\mathcal{H}_0$ haben wir $\beta_1 = b$ und damit hat unsere Teststatistik

$$T_{n-2} := \frac{\hat{\beta}_1 - b}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} = \frac{\hat{\beta}_1 - b}{\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

die t-Verteilung mit $n-2$ Freiheitsgraden. Wir werden die $\mathcal{H}_0$-Hypothese verwerfen, wenn diese Teststatistik Werte annimmt, welche weiter als die kritischen Werte von 0 entfernt sind. Wir schauen nochmals unseren Computer-Ausdruck an.

10.2.1.5 Probleme & Diagnostic Checking

Wir geben hier nur einen kurzen Überblick über mögliche Probleme, Gefahren und die Methoden, welche man unter "Diagnostic Checking" zusammenfasst. Zusätzliche Informationen dazu in Stahel.

* Ausreisser (engl. Outlier)

* ungleiche Varianzen

* verbleibende Muster (z.B. quadratisch) in den e_i's

* ϵ_i's nicht unabhängig

* ϵ_i's nicht normalverteilt

10.2.1.6 Warum ist die lineare Regression mit OLS so wichtig, bekannt und erfolgreich?

* wird auch von Nicht-MathematikerInnen/Nicht-StatistikerInnen verstanden

* theoretisch einfach zu berechnen

* einfach auch zur multiplen Regression erweiterbar

* früher war in Statistik-Paketen oft nur diese Regression programmiert (heute kaum mehr als Argument relevant)

warum speziell linear

* Mensch kann nur lineare Zusammenhänge gut erfassen

* viele nichtlineare Abhängigkeiten können durch Transformation zu linearen Problemen gemacht werden (ist aber auch umstritten: "Man foltert die Daten bis sie gestehen"). Vor allem: viele Phänomene mit exponentiellem (oder geometrischem, falls diskret) Wachstum: Wirtschaft (gesamte Volkswirtschaft und einzelne Firmen), Pflanzen (Zellteilung), Ausbreitung Bekanntheitsgrad von Websites, Ausbreitung von Epidemien

warum speziell OLS

* früher EDV-Probleme bei alternativen Vorschlägen (heute kaum relevant)

* OLS ist auch der BLUE (Best Linear Unbiased Estimator), auch BLUE im multivariaten Fall

10.2.2 Ausblick multiple Regression

Wir werden jetzt versuchen, die Radioaktivität mit Distanz *und* Regen zu erklären (oder vorherzusagen). Damit verlassen wir Modell (10.2) und gehen zu

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \ . \tag{10.9}$$

Wir werden die **OLS-Schätzung** in dieser Vlsg nur für die einfache Regression anschauen. Aber auch in der multiplen Regression kann man eine OLS-Schätzung machen; sie ist einfach komplizierter. Mit der OLS-Schätzung erhalten wir hier

$$\log(\text{bq}) = 10.522 - 1.360 * \log(\text{dist}) + 2.723 * \text{regen}$$

Residual standard error: 0.6166197

Dies ist *auf die Schnelle* wohl die beste Datenanalyse: der Regen hat (wie in der ersten Stunde vermutet und theoretisch erwartet) einen "Wash out" mit erhöhter Radioaktivität zur Folge.

Die geschätzte Varianz des Fehlerterms ist hier kleiner als bei der einfachen Regression, da noch ein Predictor dazugekommen ist (siehe auch nachfolgend bei den Bemerkungen zur multiplen Regression).

Wir gehen noch kurz darauf ein, wie dann mit zweifacher Regression oder sogar mit noch mehr erklärenden Variablen vorgegangen wird. Dazu ein Schema zu "Top Down" und "Bottom Up":

Lose Bemerkungen zur multiplen Regression:

* wenn man in der Forschung oder Industrie Probleme untersucht, weiss man manchmal überhaupt nicht, welche Variablen (erklärenden Faktoren) relevant sind (zB Regen, Distanz als Erklärung der Radioaktivität).

* Fantasie, Kreativität, Ideen sind dann zentral wichtig; wir können dann Hypothesen bilden (Forschungsfreiheit!) und sollten uns selbstkritisch keine Anmassung von Wissen erlauben.

* danach Hypothesen an Daten überprüfen (Popper: nur allfällige Falsifizierung möglich)

* Welche und wieviele Daten: n=Stichprobengrösse und k=Anzahl erklärende Variablen (Regen und Distanz sind $k = 2$, Intercept wird nicht gezählt)

* ganz allgemein sollte gelten: $n >> k$

* ein grosses n ist erwünscht (abgesehen von Kosten/Zeitaufwand); ein grosses k ist prinzipiell auch erwünscht, aber die Analyse ist mit Fallstricken und Problemen behaftet.

* Was geschieht, wenn wir eine weitere erklärende Variable hinzufügen möchten; zum Beispiel von $k = 1$ zu $k = 2$? Anstelle von

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2, \tag{10.4}$$

welches wir bzgl. β_0 und β_1 minimieren tritt neu

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2,$$

welches wir bzgl. β_0, β_1 und β_2 minimieren. Die neue Summe ist ganz sicher kleiner als diejenige in (10.4): wir sind in der gleich guten Situation, wenn wir $\beta_2 = 0$ wählen, können aber neu auch noch β_2 variieren und kommen damit ganz sicher in eine bessere Situation. Der Fit wird besser, die Residuen insgesamt kleiner. Das hat auch einen Nachteil: selbst wenn wir völlig unsinnige Grössen als erklärende Faktoren einführen, wird der Fit besser.

Bei $k = n$ haben wir im Normalfall sogar einen perfekten Fit. Um den Einbau von unsinnigen erklärenden Faktoren zu bestrafen, gibt es neben dem "Multiple R-Squared" auch das "Adjusted R-squared". Das "Adjusted" bestraft zu viele erklärende Variablen und ist deshalb eine aussagekräftigere Grösse.

10.2.3 Welches ist jetzt das richtige Modell?

Nach diesen verschiedenen Modellen fragen wir uns vielleicht: "Welches ist jetzt *das richtige Modell*?" Ausser in Simulationen, wo man genau weiss, woher die Daten stammen, ist diese Frage nicht so leicht zu beantworten. Je mehr Daten wir aber haben, desto klarer wird die Modellwahl generell sein.

```
Radioaktivitaet mit Distanz allein erklaeren (einfache Regression):
> NurDistanz<-lm(log(bq) ~ log(dist))
> aov(NurDistanz)
Call:
aov(formula = NurDistanz)
Terms:
                      log(dist) Residuals
Sum of Squares       152.87486   34.39141
Deg.  of Freedom       1         21
Residual standard error:  1.279721
Estimated effects may be unbalanced
> summary(NurDistanz)
Call:
lm(formula = log(bq) ~ log(dist))
Residuals:
Min       1Q     Median  3Q     Max
-1.8592 -1.1348 -0.1033 1.1650 2.3797
Coefficients:
            Estimate    Std.  Error     t value     Pr(>|t|)
(Intercept) 9.8027         0.7022         13.960     4.26e-12 ***
log(dist)   -1.0911         0.1129       -9.662     3.53e-09 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  1.28 on 21 degrees of freedom
Multiple R-Squared:  0.8164, Adjusted R-squared:  0.8076
F-statistic:  93.35 on 1 and 21 DF, p-value:  3.527e-09
```

```
Radioaktivitaet mit Regen allein erklaeren (einfache Regression):
> NurRegen<-lm(log(bq) ~ regen)
> aov(NurRegen)
Call:
aov(formula = NurRegen)
Terms:
                      regen   Residuals
Sum of Squares       3.31193   183.95434
Deg.  of Freedom         1       21
Residual standard error:  2.959684
Estimated effects may be unbalanced
> summary(NurRegen)
Call:
lm(formula = log(bq) ~ regen)
Residuals:
Min       1Q     Median  3Q     Max
-5.6755 -2.3332 -0.1205 2.1735 4.7970
Coefficients:
            Estimate     Std.  Error     t value     Pr(>|t|)
(Intercept)   3.7784       0.7399       5.107     4.66e-05 ***
regen         -0.8247       1.3412     -0.615     0.545
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  2.96 on 21 degrees of freedom
Multiple R-Squared:  0.01769, Adjusted R-squared:  -0.02909
F-statistic:  0.3781 on 1 and 21 DF, p-value:  0.5452
```

```
Volles Modell (zweifache Regression):
> FullModel<-lm(log(bq) ~ log(dist)+ regen)
> aov(FullModel)
Call:
aov(formula = FullModel)
Terms:
                      log(dist)     regen     Residuals
Sum of Squares         152.87486     26.78701     7.60440
Deg.  of Freedom 1 1 20
Residual standard error:  0.6166197
Estimated effects may be unbalanced
> summary(FullModel)
Call:
lm(formula = log(bq) ~ log(dist) + regen)
Residuals:
Min       1Q     Median  3Q     Max
-1.37021 -0.37287 -0.05441 0.21542 1.61984
Coefficients:
              Estimate     Std.  Error     t value     Pr(>|t|)
(Intercept)   10.52246     0.34904         30.147     < 2e-16 ***
log(dist)     -1.36027     0.06316         -21.536    2.62e-15 ***
regen         2.72254       0.32436         8.394    5.51e-08 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  0.6166 on 20 degrees of freedom
Multiple R-Squared:  0.9594, Adjusted R-squared:  0.9553
F-statistic:  236.3 on 2 and 20 DF, p-value:  1.219e-14
```

10.2.4 Korrelation (*ein* Mass für die *lineare* Gleichläufigkeit; kein Kausalzusammenhang)

Mit der einfachen Regression haben wir also eine Möglichkeit, den linearen Zusammenhang zwischen 2 Grössen (x, y) zu untersuchen. Dabei wird vorausgesetzt, dass das Modell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10.2}$$

gilt. Es gibt jedoch auch eine Masszahl, welche nicht zwingend (10.2) voraussetzt und für alle 2 dimensionalen Zusammenhänge berechnet werden kann: der Korrelationskoeffizient. Wir werden in dieser Vorlesung lediglich den empirischen Korrelationskoeffizienten anschauen. Betrachten wir dazu folgende Darstellung:

Der empirische Korrelationskoeffizient zwischen den Daten $(x_i)_{i=1}^n$ und $(y_i)_{i=1}^n$ ist also definiert als

$$r_{xy} := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}.$$

Mathematiker/innen haben gezeigt, dass diese Masszahl zwischen -1 und $+1$ liegt. Zudem ist $|r_{xy}|$ genau dann 1, wenn die Punkte alle auf einer Geraden liegen. Ist $r_{xy} = -1$, so hat diese Gerade negative Steigung - ist $r_{xy} = 1$, so hat diese Gerade positive Steigung. Wir machen dazu jetzt noch ein paar Bilder, um ein bisschen ein Gefühl für diese Masszahl zu erhalten:

Wichtig:

1. Lesen Sie jetzt in "Statistik in Cartoons" Kapitel 11 und/oder in Stahel: Kapitel 12.1, 12.2 und Kapitel 13 (ohne 13.4, 13.8 und 13.9).
2. Gehen Sie in die Übungsstunde. Drucken Sie das Übungsblatt dazu *vorher* aus, lesen Sie *vorher* die Aufgaben durch und machen sich erste Gedanken dazu (zum Beispiel, wie man sie lösen könnte).
3. Dann lösen Sie das Übungsblatt: zuerst immer selber probieren, falls nicht geht: Tipp von Mitstudi benutzen, falls immer noch nicht geht: Lösung von Mitstudi anschauen, 1 Stunde warten, versuchen, aus dem Kopf heraus wieder zu lösen, falls immer noch nicht geht: Lösung von Mitstudi abschreiben (und verstehen - also sollte man insbesondere keine Fehler abschreiben!).
4. Lösen Sie die entsprechenden Prüfungsaufgaben im Archiv.