Kapitel 5

Regression

5.1 Korrelation und empirische Korrelation

Die gemeinsame Verteilung von abhängigen Zufallsvariablen X und Y ist i.A. kompliziert, und man begnügt man sich oft mit einer **vereinfachenden** Kennzahl zur Beschreibung der Abhängigkeit. Die Kovarianz und Korrelation zwischen X und Y sind wie folgt definiert:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \quad \text{(Kovarianz)} \\ \mathrm{Corr}(X,Y) &= \rho_{XY} = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y) \quad \text{(Korrelation)}, \end{aligned}$$

wobei $\sigma_X = \sqrt{\mathrm{Var}(X)}$, und analog für σ_Y.

Die Korrelation ρ_{XY} ist eine dimensionslose, normierte Zahl
mit Werten $\rho_{XY} \in [-1,1]$.

Die Korrelation misst Stärke und Richtung der **linearen Abhängigkeit** zwischen X und Y. Es gilt

$$\begin{aligned} \mathrm{Corr}(X,Y) = +1 &\text{ genau dann wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b > 0, \\ \mathrm{Corr}(X,Y) = -1 &\text{ genau dann wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b < 0. \end{aligned}$$

Überdies gilt:

$$X \text{ und } Y \text{ unabhängig} \quad \Longrightarrow \quad \mathrm{Corr}(X,Y) = 0. \tag{5.1}$$

Die Umkehrung gilt i.A. nicht.

5.1.1 Die empirische Korrelation

In Kapitel 4.2.2 und Abbildung 4.3 haben wir ein Beispiel gesehen, wo die Daten $(x_1,y_1),\ldots,(x_n,y_n)$ als Realisierungen von i.i.d. Zufalls-vektoren $(X_1,Y_1),\ldots,(X_n,Y_n)$ aufgefasst werden können.

Die empirischen Korrelation ist dann

$$\widehat{\mathrm{Corr}}(X,Y) = \hat{\rho}_{XY} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}.$$

Analog zur Korrelation gelten die folgenden Eigenschaften:

$$\hat{\rho}_{XY} \in [-1, 1],$$
$$\hat{\rho}_{XY} = +1 \ \Leftrightarrow \ y_i = a + bx_i \text{ für alle } i = 1, \ldots, n, \text{ und für ein } a \in \mathbb{R} \text{ und ein } b > 0,$$
$$\hat{\rho}_{XY} = -1 \ \Leftrightarrow \ y_i = a + bx_i \text{ für alle } i = 1, \ldots, n, \text{ und für ein } a \in \mathbb{R} \text{ und ein } b < 0.$$

5.2 Einfache lineare Regression

Wir betrachten das folgende Beispiel aus der Chemie. Die Dimerisation von 1,3-Butadien läuft nach einem Reaktionsmodell zweiter Ordnung ab und ist deshalb charakterisiert durch die Gleichung $\frac{d}{dt}C(t) = -\kappa C(t)^2$, wobei C hier den Partialdruck des Edukts und t die Zeit bedeutet. Diese Gleichung hat Lösungen der Form

$$\frac{1}{C(t)} = \frac{1}{C(0)} + \kappa t.$$

Messungen zu verschiedenen Zeitpunkten im Ablauf der Reaktion sind in Abbildung 5.1 dargestellt. Die letzte Gleichung zeigt, dass der Kehrwert des Partialdruckes linear von der Zeit abhängen sollte. Wegen zufälligen Messfehlern und kleinen systematischen Abweichungen vom einfachen Modell liegen die Punkte nicht genau auf einer Geraden.

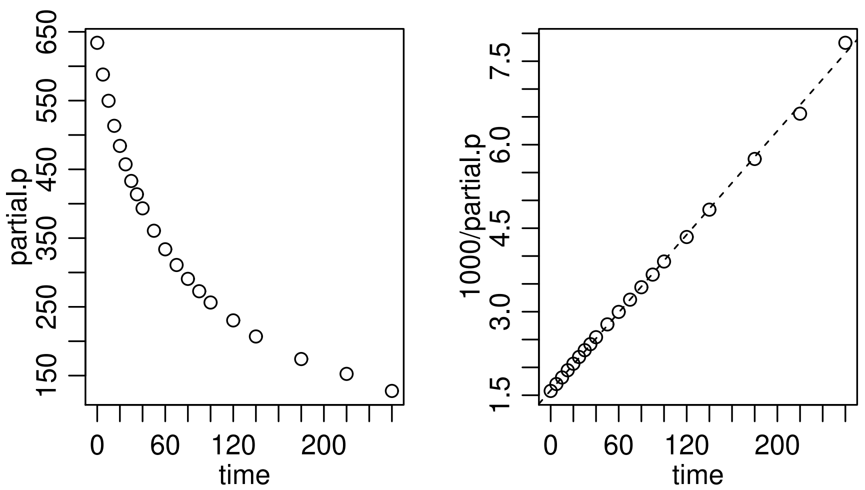

Abbildung 5.1: Partialdruck von Butadien (links) und Kehrwert 1000 $\times$ 1/Partialdruck (rechts), gegen die Zeit aufgetragen

5.2.1 Das Modell der einfachen linearen Regression

Im obigen Beispiel haben wir Daten

$$(x_1, y_1), \ldots, (x_n, y_n),$$

wobei x_i der Zeitpunkt der i-ten Messung und y_i der Kehrwert des Partialdrucks der i-ten Messung bezeichnen. Diese fassen wir auf als Realisierungen des folgenden Modells:

$$Y_i = h(x_i) + E_i \ (i = 1, \ldots, n),$$
$$E_1, \ldots, E_n \text{ i.i.d. }, \ \mathcal{E}(E_i) = 0, \ \mathrm{Var}(\mathcal{E}_i)) = \sigma^2.$$

Die Y-Variable ist die **Zielvariable** (engl: response variable) und die x-Variable ist die **erklärende Variable** oder **Co-Variable** (engl: explanatory variable; predictor variable; covariate). Die Zufallsvariablen E_i werden öfters als Fehler-Variablen oder Rausch-Terme bezeichnet. Sie besagen, dass der Zusammenhang zwischen der erklärenden und der Ziel-Variablen nicht exakt ist. Die erklärenden Variablen x_i $(i = 1, \ldots, n)$ sind deterministisch, hingegen sind die Ziel-Variablen Y_i Zufallsvariablen (wegen den Zufalls-Variablen E_i).

Die Modelle für die Funktion $h(\cdot)$ sind:

$$h(x) = \beta_0 + \beta_1 x \quad : \text{einfache lineare Regression,}$$
$$h(x) = \beta_1 x \quad : \text{einfache lineare Regression durch Nullpunkt.}$$

Meistens betrachten wir das allgemeinere Modell mit einem Achsenabschnitt β_0. Das Modell ist illustriert in Abbildung 5.2, wo für die Fehler-Variable eine $\mathcal{N}(0, 0.1^2)$-Verteilung spezifiziert wurde.

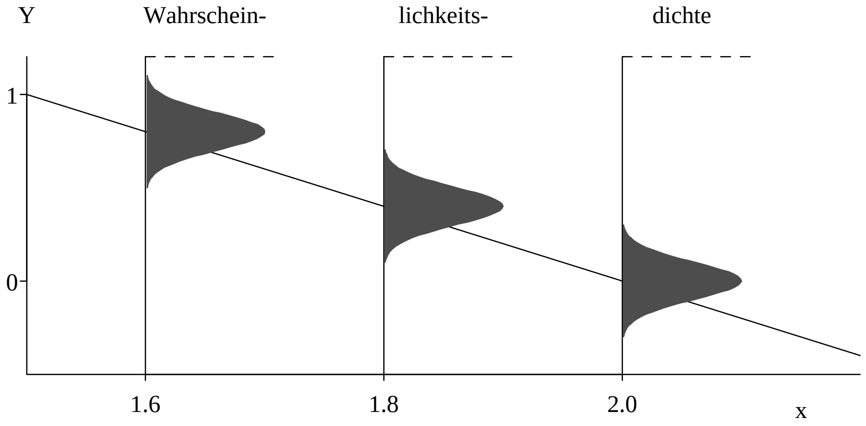

Abbildung 5.2: Veranschaulichung des Regressionsmodells $Y_i = 4 - 2x_i + E_i$ mit $E_i \sim \mathcal{N}(0, 0.1^2)$ für drei Beobachtungen.

5.2.2 Parameterschätzungen

Die unbekannten Modell-Parameter in der einfachen linearen Regression sind β_0, β_1 und auch die Fehlervarianz σ^2. Die Methode der Kleinsten-Quadrate liefert die folgenden Schätzungen:

$$\hat{\beta}_0, \hat{\beta}_1 \text{ sind Minimierer von } \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2.$$

Die Lösung dieses Optimierungsproblem ist eindeutig:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ \hat{\beta}_0 &= \bar{y}_n - \hat{\beta}_1 \bar{x}_n. \end{aligned}$$

Das Prinzip der Kleinsten-Quadrate liefert sogenannt erwartungstreue Schätzungen:

$$\mathcal{E}(\hat{\beta}_0) = \beta_0, \quad \mathcal{E}(\hat{\beta}_1) = \beta_1,$$

das heisst, dass die Schätzungen keinen systematischen Fehler haben (z.B. sind sie nicht systematisch zu gross was bedeuten würde, dass z.B. $\mathcal{E}(\hat{\beta}_1) > \beta_1$).

Für die Schätzung von σ^2 benützen wir das Konzept der Residuen. Falls wir Realisationen der Fehler-Terme E_i beobachten könnten, so könnten wir die empirische Varianzschätzung fur σ^2 verwenden. Hier approximieren wir zuerst die unbeobachteten Fehler-Variablen E_i durch die **Residuen**:

$$R_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \ (i = 1, \ldots, n).$$

Da $E_i = Y_i - (\beta_0 + \beta_1 x_i)$ scheint die Approximation $R_i \approx E_i$ vernünftig. Als Varianzschätzung benutzt man dann:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} R_i^2. \tag{5.2}$$

Dabei ist zu beachten, dass (bei einfacher linearer Regression mit einem Achsenabschnitt β_0) gilt: $\sum_{i=1}^{n} R_i = 0$. Das heisst, die Varianzschätzung in (5.2) ist wie die empirische Varianz bei einer Stichprobe (siehe Kapitel 4.6.1), ausser dass wir den Faktor $1/(n-2)$ anstelle von $1/(n-1)$ nehmen. Dieser Faktor entspricht der folgenden Faustregel: $1/(n-\text{Anzahl Parameter})$, wobei die Anzahl Parameter ohne den zu schätzenden Varianz-Parameter zu zählen ist (in unserem Falle sind dies die Parameter $\beta_0, \ \beta_1$).

Bei einem Datensatz mir realisierten y_i $(i = 1, \ldots, n)$ werden die Schätzungen mit den Werten y_i anstelle von Y_i gerechnet. Zum Beispiel sind dann die realisierten Residuen von der Form $r_i = y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i)$.

5.2.3 Tests und Konfidenzintervalle

Wir diskutieren hier die 2. und 3. Grundfragestellung (siehe Kapitel 3.1) im Kontext der einfachen linearen Regression. Dabei werden wir entscheidend mehr Schlussfolgerungen ziehen, als bloss eine best passende Regressionsgerade zu finden.

Der t-Test in Regression

Wir betrachten hier als Beispiel den folgenden Datensatz. Es wurden $n = 111$ Messungen gemacht von mittlerer täglicher Temperatur (x-Variable) und mittlerem täglichem Ozongehalt (Y-Variable). Die Daten und die Regressionsgerade $\hat{\beta}_0 + \hat{\beta}_1 x$ sind in Abbildung 5.2.3 ersichtlich. Die interessierend Frage in der Praxis lautet: hat die Temperatur einen Einfluss auf den Ozongehalt. Diese Frage kann man in ein Test-Problem übersetzen:

$$\begin{aligned} H_0 &: \ \beta_1 = 0, \\ H_A &: \ \beta_1 \neq 0. \end{aligned}$$

Es wird hier "per default" ein zwei-seitiger Test durchgeführt, nämlich der t-Test für die Steigung in der einfachen linearen Regression.

Wir machen hier die Annahme, dass

$$E_1, \ldots, E_n \text{ i.i.d. } \mathcal{N}(0, \sigma^2). \tag{5.3}$$

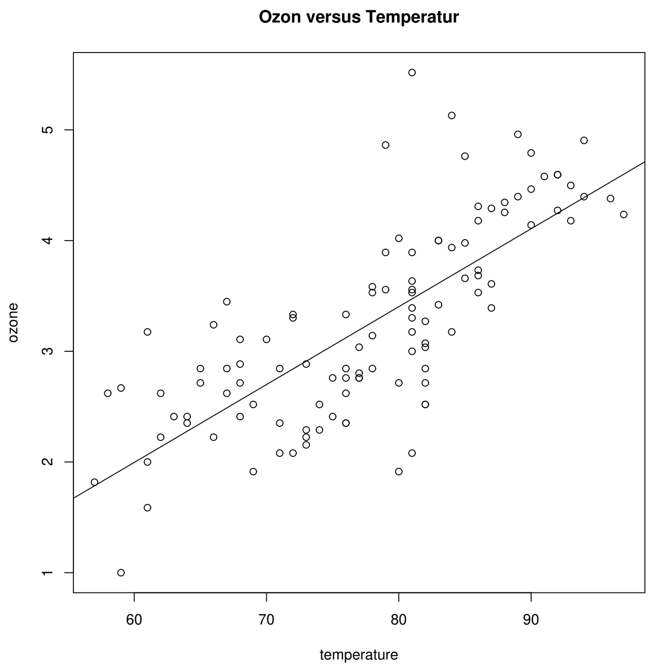

Abbildung 5.3: Streudiagramm und angepasste Regressionsgerade für den Ozon-Temperatur Datensatz.

Die Teststatistik ist

$$\frac{\hat{\beta}_1}{\widehat{s.e.}(\hat{\beta}_1)},$$

$$\widehat{s.e.}(\hat{\beta}_1) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Unter der Nullhypothese und der Annahme von normalverteilten Fehlern in (5.3) gilt:

$$T \sim t_{n-2} \text{ unter } H_0 :\ \beta_1 = 0,$$

und der P-Wert dieses zwei-seitigen t-Test kann dann analog wie in Kapitel 4.6.2 berechnet werden (mit $n-2$ anstelle von $n-1$ Freiheitsgraden), und er wird auch von statistischer Software geliefert.

Völlig analog erhält man auch einen Test für $H_0 :\ \beta_0 = 0$ bei zwei-seitiger Alternative $H_A :\ \beta_0 \neq 0$. Der entsprechende P-Wert, unter Annahme der Normalverteilung in (5.3) wird von statistischer Software geliefert.

Der Computer-Output vom `R` bei dem Anpassen einer einfachen linearen Regression für den Datensatz von Ozon als Funktion von Temperatur sieht wie folgt aus:

```
Call:
  lm(formula = ozone ~ temperature)

Residuals:
       Min       1Q   Median       3Q      Max
```

```
-1.49016 -0.42579  0.02521  0.36362  2.04439

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.225984   0.461408  -4.824 4.59e-06 ***
temperature   0.070363   0.005888  11.951  < 2e-16 ***
---

Residual standard error: 0.5885 on 109 degrees of freedom
Multiple R-Squared: 0.5672,Adjusted R-squared: 0.5632
F-statistic: 142.8 on 1 and 109 DF,  p-value: < 2.2e-16
```

Die zweite Kolonne bei "Coefficients" beschreibt die Punktschätzer $\hat{\beta}_i$ $(i = 0, 1)$; die dritte Kolonne die geschätzten Standardfehler $\widehat{s.e.}(\hat{\beta}_i)$ $(i = 0, 1)$; die vierte Kolonne die Teststatistik $\hat{\beta}_i/\widehat{s.e.}(\hat{\beta}_i)$ $(i = 0, 1)$, welche sich aus der zweiten dividiert durch die dritte Kolonne ergibt; die fünfte Kolonne bezeichnet den P-Wert für $H_0: \beta_i = 0$ und $H_A: \beta_i \neq 0$ $(i = 0, 1)$. Überdies ist die geschätzte Standardabweichung für den Fehler $\hat{\sigma}$ ersichtlich unter "Residual standard error"; die "degrees of freedom" sind gleich $n - 2$.

Konfidenzintervalle

Basierend auf der Normalverteilungsannahme erhält man die folgenden zwei-seitigen Konfidenzintervalle für β_i $(i = 0, 1)$ zum Niveau $1 - \alpha$:

$$\begin{aligned} &\hat{\beta}_0 \pm \widehat{s.e.}(\hat{\beta}_0) t_{n-2;1-\alpha/2} \quad \text{für } \beta_0, \\ &\hat{\beta}_1 \pm \widehat{s.e.}(\hat{\beta}_1) t_{n-2;1-\alpha/2} \quad \text{für } \beta_1. \end{aligned}$$

5.2.4 Das Bestimmtheitsmass R^2

Die Güte eines Regressionsmodells kann mit dem sogenannten Bestimmtheitsmass R^2 quantifiziert werden. Dazu betrachten wir eine Beziehungen zwischen verschiedenen Variations-Quellen: mit der Bezeichnung $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ für den Wert der angepassten Geraden beim Wert x_i ist

$$\underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{SS_Y} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SS_E} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}_{SS_R}. \tag{5.4}$$

Dabei beschreint SS_Y die totale Variation der Zielvariablen (ohne Einfluss der erklärenden Variablen x), SS_E die Variation des Fehlers (Residuen-Quadratsumme), und SS_R die Variation, welche durch die Regression erklärt wird (Einfluss der erklärenden Variablen x). Das Bestimmtheitsmass ist dann wie folgt definiert:

$$R^2 = \frac{SS_R}{SS_Y}, \tag{5.5}$$

und beschreibt den Anteil der totalen Variation, welche durch die Regression erklärt wird. Wegen 5.4 gilt, dass $0 \leq R^2 \leq 1$: falls R^2 nahe bei 1 ist, so erklärt das Regressionsmodell viel der totalen Variation und ist somit gut; falls $R^2 \approx 0$ taugt das Regressionsmodell nicht besonders viel. Die Realisation von R^2 ist im Computer-Ouput zu finden unter "Multiple R-squared".

Im Falle der einfachen linearen Regression gilt auch:

$$R^2 = \hat{\rho}_{XY}^2,$$

d.h. R^2 ist gleich der quadrierten empirischen Korrelation.

5.2.5 Allgemeines Vorgehen bei einfacher linearer Regression

Grob zusammengefasst kann bei einfacher linearer Regression folgendemassen vorgegangen werden.

1. Anpassen der Regressionsgeraden; d.h. Berechnung der Punktschätzer $\hat{\beta}_0,\ \hat{\beta}_1$.
2. Testen ob erklärende Variable x einen Einfluss auf die Zielvariable Y hat mittels t-Test für $H_0:\ \beta_1 = 0$ und $H_a:\ \beta_1 \neq 0$. Falls dieser Test nicht-signifikantes Ergebnis liefert, so ist das Problem "in der vorliegenden Form uninteressant".
3. Testen ob Regression durch Nullpunkt geht mittels t-Test für $H_0:\ \beta_0 = 0$ und $H_A:\ \beta_0 \neq 0$. Falls dieser Test nicht-signifikantes Ergebnis liefert, so benützt man das kleinere Modell mit Regression durch Nullpunkt.
4. Bei Interesse Angabe von Konfidenzintervallen für β_0 und β_1.
5. Angabe des Bestimmtheitsmass R^2. Dies ist in gewissem Sinne eine informellere (und zusätzliche) Quantifizierung als der statistische Test in Punkt 2.
6. Überprüfen der Modell-Voraussetzungen mittels Residuenanalyse. Dieser wichtige Schritt wird ausführlicher in Kapitel 5.2.6 beschrieben.

5.2.6 Residuenanalyse

Wir werden hier graphische Methoden beschreiben, basierend auf realisierten Residuen $r_i (i = 1, \ldots, n)$, welche zur Überprüfung der Modell-Voraussetzungen für die einfache lineare Regression eingesetzt werden können. Die Modell-Voraussetzungen sind, in prioritärer Reihenfolge, die folgenden.

1. $\mathcal{E}(E_i) = 0$.
   Somit gilt $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$, das heisst: es gibt keinen systematischen Fehler im Modell.
   Abweichungen von dieser Annahme könnte zum Beispiel durch einen nicht-linearer Zusammenhang zwischen x und Y verursacht sein.
2. $E_1, \ldots, E_n$ i.i.d.
   Abweichungen könnte z.B. eine nicht-konstante Varianz der Fehlers sein, d.h. $\mathrm{Var}(E_i) = \sigma_i^2$ mit verschiedenen σ_i^2 für $i = 1, \ldots, n$. Eine andere Abweichung könnte durch korrelierte/abhängige Fehler verursacht sein.
3. $E_1, \ldots, E_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$.
   Abweichungen könnte durch eine lang-schwänzige Fehlerverteilung verursacht sein.

Der Tukey-Anscombe Plot

Der wichtigste Plot in der Residuenanalyse ist der Plot der Residuen r_i gegen die angepassten Werte $\hat{y}_i$, der sogenannte Tukey-Anscombe Plot.

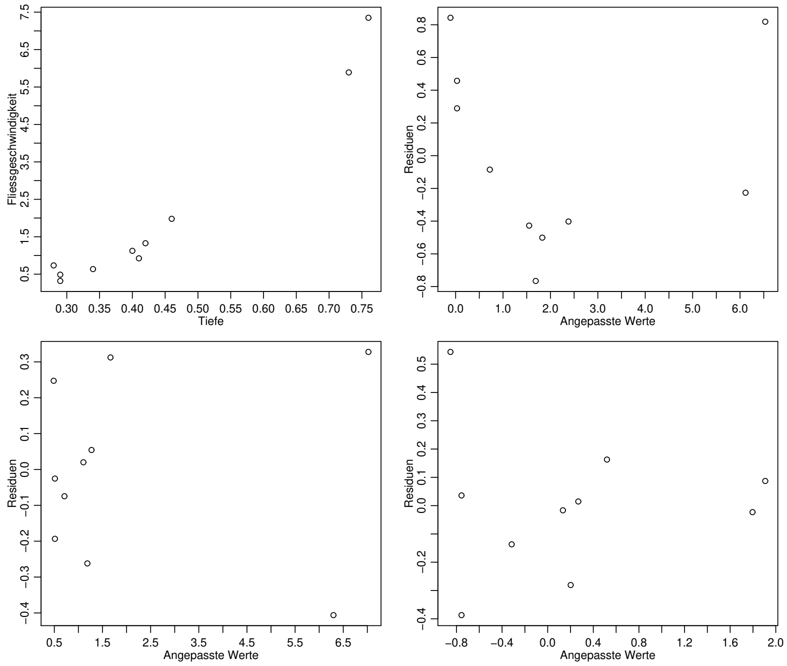

Abbildung 5.4: Streudiagramm von Tiefe und Fliessgeschwindigkeit (oben links), Tukey-Anscombe Plots für einfache lineare Regression (oben rechts), für quadratische Regression (siehe Kapitel 5.3.1) (unten links) und für einfache lineare Regression mit logarithmierten Variablen $\log(Y)$ und $\log(x)$ (unten rechts).

Im Idealfall: gleichmässige Streuung der Punkte um Null.
Abweichungen:
- kegelförmiges Anwachsen der Streuung mit $\hat{y}_i$
  evtl. kann man die Zielvariable logarithmieren (falls Y_i's positiv sind), d.h. man benutzt das neue Modell

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- Ausreisserpunkte
  evtl. können robuste Regressions-Verfahren verwendet werden (siehe Literatur)
- unregelmässige Struktur
  Indikation für nichtlinearen Zusammenhang
  evtl. Ziel und.oder erklärende Variablen transformieren (siehe auch das Beispiel in Abbildung 5.1.

Für den Ozon-Datensatz ist der Tukey-Anscombe Plot in Abbildung 5.5 gezeigt.

Nichtlineare Zusammenhänge können in der Praxis natürlich vorkommen: sie zeigen an, dass die Regressionsfunktion nicht korrekt ist. Abhilfe schaffen die Aufnahme zusätzlicher erklärender Variablen (z.B. quadratische Terme, siehe Kapitel 5.3.1) oder - wie bereits oben angedeutet - Transformationen der erklärenden und/oder der Ziel-Variablen. Ein einfaches

Beispiel ist in Abbildung 5.4 gezeigt, bei dem es um den Zusammenhang zwischen Tiefe und Fliessgeschwindigkeit von Bächen geht. Bei einfacher Regression zeigt der Tukey-Anscombe Plot eine klare nichtlineare Struktur, die verschwindet, wenn man entweder einen quadratischen Term dazunimmt (siehe Kapitel 5.3.1) oder wenn man beide Variablen logarithmiert (d.h. einen Potenzzusammenhang anpasst mit dem Modell

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i \ (i = 1, \ldots, n).$$

Mit so wenigen Daten kann man zwischen diesen beiden Modellen nicht unterscheiden. Die Nichtlinearität des Zusammenhangs ist natürlich auch im ursprünglichen Streudiagramm ersichtlich, wenn man genau hinschaut. Häufig sind aber Abweichungen von der Linearität im Tukey-Anscombe Plot besser zu sehen.

Plot bezüglich serieller Korrelation

Um die Unabhängigkeitsannahme der $E_1, \ldots, E_n$ zu überprüfen, kann der folgende Plot gemacht werden: plotte r_i gegen die Beobachtungsnummer i.

Im Idealfall: gleichmässige Streuung der Punkte um Null.
Abweichungen:
- langfristiges Zonen mit durchwegs positiven oder negativen Residuen
die Punktschätzungen sind immer noch OK, aber die Tests und Konfidenzintervalle stimmen nicht mehr evtl. Regression mit korrelierten Fehlern verwenden (siehe Literatur)

Für den Ozon-Datensatz ist der serielle Korrelations-Plot in Abbildung 5.5 gezeigt.

Der Normalplot

Mit dem Normalplot (siehe Kapitel 4.4.6) können wir die Normalverteilungsannahme in (5.3) überprüfen.

Im Idealfall: approximativ eine Gerade
Abweichungen:
- Abweichung von einer Geraden Evtl. robuste Regression benutzten (siehe Literatur)

Für den Ozon-Datensatz ist der Normalplot in Abbildung 5.5 gezeigt.

Das Auffinden eines guten Modells

Oftmals werden mehrere Modelle in einer Art "workflow-feedback" Prozeß betrachtet und angepasst. Man beginnt mit einem ersten Modell; dann, aufgrund von Residuenanalyse wird das Modell modifiziert. Das modifizierte Modell (immer noch als linear angenommen in evtl. transformierten Variablen) wird wiederum mit linearer Regression angepasst, und mit Residuenanalyse wird das neue Modell beurteilt. Dieses Vorgehen wird iteriert bis man ein "zufriedenstellendes" Modell gefunden und angepasst hat.

5.3 Multiple lineare Regression

Oftmals hat man mehrere erklärende Variablen $x_{i,1}, \ldots, x_{i,p-1}$ $(p > 2)$.

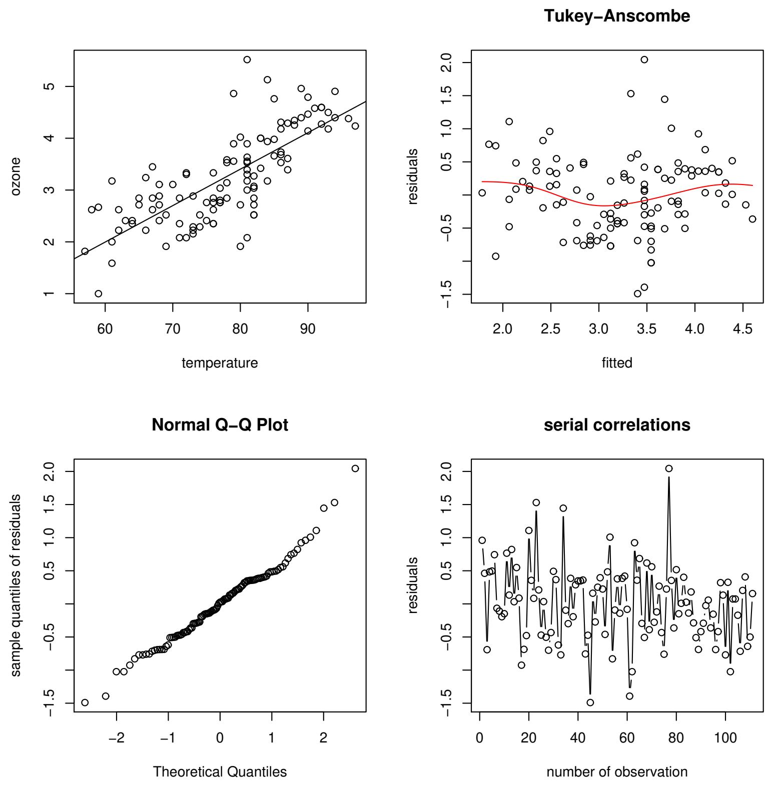

Abbildung 5.5: Ozon-Datensatz: Streudiagramm mit angepasster Regressiongerade (oben links); Tukey-Anscombe Plot (oben rechts); serieller Korrelations-Plot (unten links); Normalplot (unten rechts).

5.3.1 Das Modell der multiplen linearen Regression

Das Modell ist wie folgt:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j} + E_i,$$
$$E_1, \ldots, E_n \text{ i.i.d. }, \ \mathcal{E}(E_i) = 0, \ \mathrm{Var}(\mathcal{E}_i)) = \sigma^2.$$

Wie bei der einfachen linearen Regression nehmen wir an, dass die erklärenden Variablen deterministisch sind. Es ist oftmals nützlich, das obige Modell in Matrix-Schreibweise darzustellen:

$$\underset{n \times 1}{Y} = \underset{n \times p}{X} \times \underset{p \times 1}{\beta} + \underset{n \times 1}{E} \tag{5.6}$$

wobei X eine $(n \times p)$-Matrix ist mit Kolonnenvektoren $(1, 1, \ldots, 1)^T$, $(x_{1,1}, x_{2,1}, \ldots, x_{n,1})^T$ und letztendlich $(x_{1,p-1}, x_{2,p-1}, \ldots, x_{n,p-1})^T$.

Beispiele von multipler linearer Regression sind unter anderen:

Simple lineare Regression: $Y_i = \beta_0 + \beta_1 x_i + E_i \ (i = 1, \ldots n)$.

$$p = 2 \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Quadratische Regression: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \ (i = 1, \ldots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Zu beachten ist, dass die Funktion quadratisch ist in den x_i's, aber *linear* in den Koeffizienten β_j und deshalb ein Spezialfall des multiplen linearen Regressions Modells.

Regression mit transformierten erklärenden Variablen:
$Y_i = \beta_0 + \beta_1 \log(x_{i2}) + \beta_2 \sin(\pi x_{i3}) + E_i \ (i = 1, \ldots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & \log(x_{12}) & \sin(\pi x_{13}) \\ 1 & \log(x_{22}) & \sin(\pi x_{23}) \\ \vdots & \vdots & \vdots \\ 1 & \log(x_{n2}) & \sin(\pi x_{n3}) \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Wiederum, das Modell ist *linear* in den Koeffizienten β_j, aber nichtlinear in den x_{ij}'s.

5.3.2 Parameterschätzungen und t-Tests

Analog zur einfachen linearen Regression wird meist die Methode der Kleinsten Quadrate benutzt:

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1} \text{ sind Minimierer von } \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}))^2.$$

Die eindeutige Lösung dieser Optimierung ist explizit darstellbar falls $p < n$:

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

wobei $\hat{\beta}$ den $p \times 1$ Vektor $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})^T$ bezeichnet, und X, Y wie in (5.6).

Die Schätzung der Fehlervarianz ist

$$\frac{1}{n-p} \sum_{i=1}^{n} R_i^2, \quad R_i = Y_i - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{i,j}).$$

Unter der Annahme, dass die Fehler normalverteilt sind wie in (5.3), können auch ähnlich zur einfachen Regression t-Tests für die folgenden Hypothesen gemacht werden:

$$H_{0,j}: \ \beta_j = 0; \quad H_{A,j}: \ \beta_j \neq 0 \ (j = 0, \ldots, p-1).$$

Der wesentliche Unterschied besteht aber in der Interpretation der Parameter:

$$\beta_j \text{ misst den linearen Effekt}$$
$$\text{der } j\text{-ten erklärenden Variablen auf die Zielvariable } Y$$
$$\textbf{nach} \text{ Elimination der linearen Effekte}$$
$$\text{aller anderen Variablen auf } Y \ (j = 1, \ldots, p-1)$$

Insbesondere impliziert dies, dass man die Koeffizienten β_j nicht einfach durch einzelne, individuelle simple lineare Regressionen von Y auf die j-te erklärende erhalten kann.

Beispiel: Wir betrachten $p = 3$ und 2 erklärende Variablen. Wir nehmen an, dass die beiden erklärenden Variablen empirisch stark korreliert sind. Es kann dann durchaus geschehen, dass:
sowohl $H_{0,1}: \ \beta_1 = 0$ als auch $H_{0,2}: \ \beta_2 = 0$ werden nicht verworfen,
obschon mindestens einer der Koeffizienten β_1 oder β_2 ungleich Null ist.
Um den Trugschluss zu vermeiden, dass es keine Effekt der erklärenden Variable auf die Ziel-Variable gibt, muss man den sogenannten F-Test betrachten.

5.3.3 Der F-Test

Der (globale) F-Test quantifiziert die Frage, ob es mindestens eine erklärende Variable gibt, welche einen relevanten Effekt auf die Zielvariable (im Sinne der linear Regression). Die folgende Nullhypothese wird beim (globalen) F-Test betrachtet:

$$H_0: \ \beta_1 = \ldots = \beta_{p-1} = 0$$
$$H_A: \ \text{mindestens ein } \beta_j \neq 0 \ (j = 1, \ldots, p-1).$$

Der P-Wert des (globalen) F-Tests ist im Computer-Output gegeben unter "F-statistic".

5.3.4 Das Bestimmtheitsmass R^2

Das Bestimmtheitsmass R^2 ist in der multiplen linearen Regression über die Formel (5.5) definiert (mit Hilfe der Zerlegung in (5.4). Eine Interpretation im Sinne einer quadrierten Stichproben-Korrelation zwischen der Ziel-Variablen und den erklärenden Variablen lässt sich nicht mehr herstellen.

5.3.5 Residuenanalyse

Die Residuenanalyse geht völlig analog zu Kapitel 5.2.6. Das allgemeine Vorgehen bei multipler linearer Regression ist wie in Kapitel 5.2.5, unter Einbezug des F-Tests nach dem Schritt 1.

5.3.6 Strategie der Datenanalyse: ein abschliessendes Beispiel

Wir betrachten ein Beispiel wo die Asphalt-Qualität als Funktion von 6 erklärenden Variablen analysiert wird.

```
y = RUT  : log("rate of rutting") = log(change of rut depth in inches
                    per million wheel passes)
           ["rut":= 'Wagenspur", ausgefahrenes Geleise]
```

```
x1 = VISC : log(viscosity of asphalt)
x2 = ASPH : percentage of asphalt in surface course
x3 = BASE : percentage of asphalt in  base   course
x4 = RUN  : '0/1' indicator for two sets of runs.
x5 = FINES: 10* percentage of fines in surface course
x6 = VOIDS: percentage of voids in surface course
```

Die Daten sind in Abbildung 5.6 dargestellt. Die Zusammenhänge werden linearer, wenn

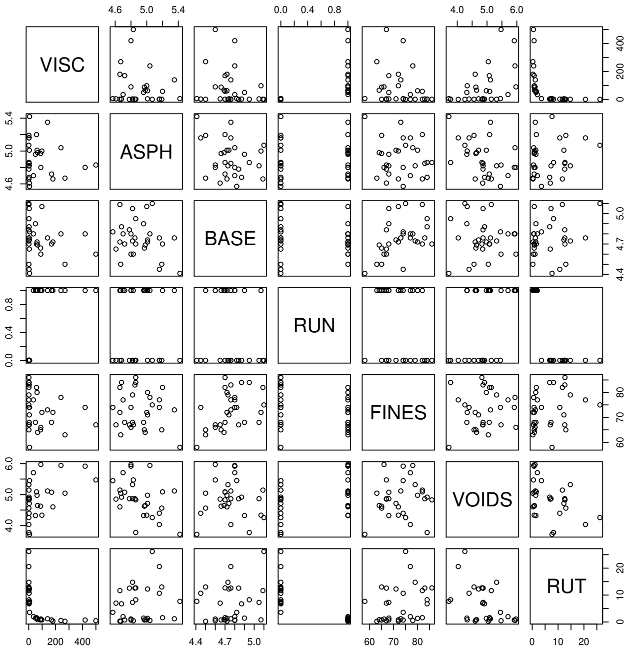

Abbildung 5.6: Paarweise Streudiagramme für den Asphalt-Datensatz. Die Zielvariable ist "RUT".

man die Zielvariable "RUT" logarithmiert und ebenfalls die erklärende Variable "VISC".

```
  y = LOGRUT  : log("rate of rutting") = log(change of rut depth in inches
                   per million wheel passes)
           ["rut":= 'Wagenspur", ausgefahrenes Geleise]
 x1 = LOGVISC : log(viscosity of asphalt)
 x2 = ASPH : percentage of asphalt in surface course
 x3 = BASE : percentage of asphalt in  base   course
 x4 = RUN  : '0/1' indicator for two sets of runs.
 x5 = FINES: 10* percentage of fines in surface course
 x6 = VOIDS: percentage of voids in surface course
```

Die transformierten Daten sind in Abbildung 5.7 dargestellt.

Mittels R wird ein multiples lineares Modell angepasst. Der Output sieht wie folgt aus:

```
Call:
lm(formula = LOGRUT ~ ., data = asphalt1)
```

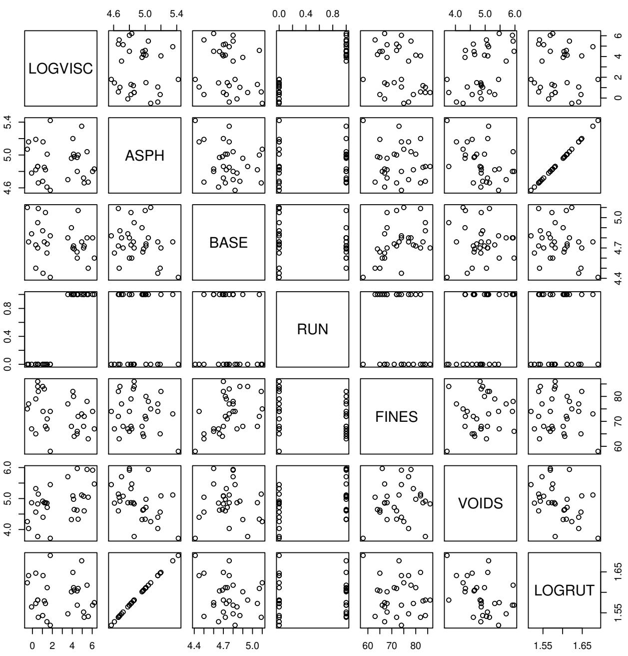

Abbildung 5.7: Paarweise Streudiagramme für den transformierten Asphalt-Datensatz. Die Zielvariable ist "LOGRUT", die log-transformierte ursprüngliche Variable "RUT". Die erklärende Variable "LOGVISC" ist ebenfalls die log-transformierte ursprüngliche Variable "VISC".

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.48348 -0.14374 -0.01198  0.15523  0.39652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.781239   2.459179  -2.351 0.027280 *
LOGVISC     -0.513325   0.073056  -7.027 2.90e-07 ***
ASPH         1.146898   0.265572   4.319 0.000235 ***
BASE         0.232809   0.326528   0.713 0.482731
RUN         -0.618893   0.294384  -2.102 0.046199 *
FINES        0.004343   0.007881   0.551 0.586700
VOIDS        0.316648   0.110329   2.870 0.008433 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2604 on 24 degrees of freedom
Multiple R-Squared: 0.9722,     Adjusted R-squared: 0.9653
F-statistic: 140.1 on 6 and 24 DF,  p-value: < 2.2e-16
```

Wir sehen, dass die Variablen "LOGVISC", "ASPH" und "VOID" signifikant oder sogar hoch-signifikant sind; die Variable "RUN" ist bloss schwach signifikant. Der F-Test ist hoch-signifikant, das Bestimmtheitsmass R^2 sehr nahe bei 1. Die degrees of freedom sind hier $n - p = 24$ mit $p = 7$, d.h. $n = 31$. Die Residuenanalyse ist mittels Tukey-Anscombe und Normalplot in Abbildung 5.8 zusammengefasst: die Normalverteilungsannahme für die Fehler ist eine vernünftige Approximation. Der Tukey-Anscombe Plot zeigt etwas systematische Variation was durch Nichtlinearität induziert sein könnte; das das R^2 aber bereits sehr nahe bei 1 liegt, so kann man trotzdem sagen, dass die multiple linear Regression sehr viel der totalen Variation erklären kann.

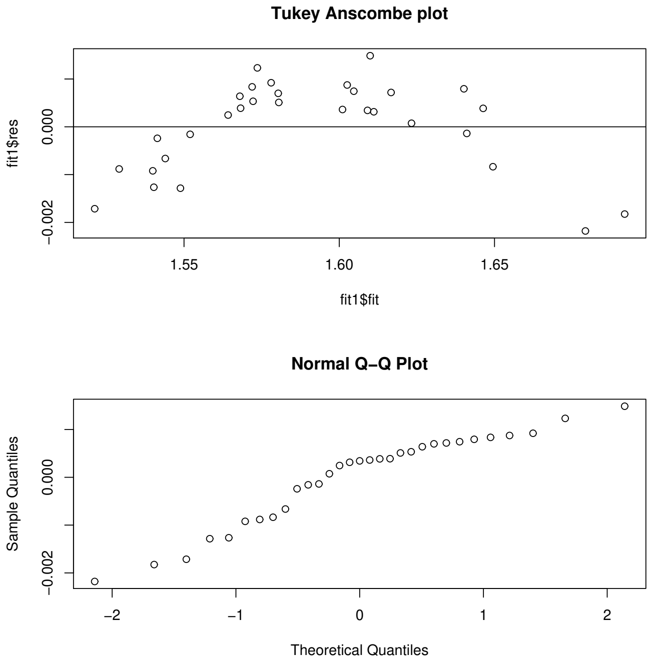

Abbildung 5.8: Tukey-Anscombe Plot (oben) und Normalplot (unten) beim Asphalt-Datensatz mit den transformierten Variablen "LOGRUT" und "LOGVISC".

Ohne log-Transformationen, d.h. das untransformierte Modell wie in Abbildung 5.6, ist das Bestimmtheitsmass $R^2 = 0.7278$, also wesentlich schlechter als im transformierten Modell.