# Breusch-Pagan-Test I

auf Heteroskedastie in den Störgrößen

- Ein weiterer Test auf Heteroskedastie in den Störgrößen ist der **Breusch-Pagan-Test**.
- Im Gegensatz zum Goldfeld-Quandt-Test ist es nicht erforderlich, eine (einzelne) Quelle der Heteroskedastizität anzugeben bzw. zu vermuten.
- Vielmehr lässt sich mit dem Breusch-Pagan-Test eine konstante Störgrößenvarianz $\sigma^2 \equiv \sigma_i^2$ gegen eine recht allgemeine Abhängigkeit der Störgrößenvarianzen von $Q$ Variablen $z_{1i}, z_{2i}, \ldots, z_{Qi}$, $i \in \{1, \ldots, n\}$, in der Form

$$\sigma_i^2 = h(\gamma_0 + \gamma_1 \cdot z_{1i} + \ldots + \gamma_Q \cdot z_{Qi}) \tag{1}$$

  mit einer Funktion $h$, an die nur moderate Bedingungen gestellt werden müssen, abgrenzen.
- Im Breusch-Pagan-Test entspricht der Fall einer konstanten Störgrößenvarianz der Nullhypothese

$$H_0 : \gamma_1 = \ldots = \gamma_Q = 0 \qquad \Longleftrightarrow \qquad \sigma_i^2 \equiv h(\gamma_0)$$

  im allgemeineren „Varianz-Modell“ aus Formel (1).

# Breusch-Pagan-Test II

auf Heteroskedastie in den Störgrößen

- Häufig werden als Variablen $z_{1i}, z_{2i}, \ldots, z_{Qi}$ gerade wieder die Regressoren des ursprünglichen Regressionsmodells eingesetzt, es gilt dann also

$$Q = K \qquad \text{und} \qquad z_{ji} = x_{ji} \quad \text{für} \quad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, K\}\ .$$

- Durch die Freiheit bei der Auswahl der Einflussvariablen $z_{1i}, z_{2i}, \ldots, z_{Qi}$ sind aber auch zahlreiche Varianten möglich, zum Beispiel
  - die Verwendung nicht nur der Regressoren des ursprünglichen Modells, sondern auch Potenzen hiervon und/oder Produkte verschiedener Regressoren oder
  - die Verwendung der aus der ursprünglichen Modellschätzung gewonnenen $\widehat{y}_i$.
- Unter dem Namen „Breusch-Pagan-Test“ (BP-Test) werden üblicherweise zwei unterschiedliche Versionen subsumiert, nämlich
  - der ursprüngliche Test von Breusch und Pagan (Econometrica, 1979), der unabhängig auch von Cook und Weisberg (Biometrika, 1983) vorgeschlagen wurde, sowie
  - eine „robuste“ Modifikation von Koenker (Journal of Econometrics, 1981), die geeigneter ist, wenn die Störgrößen nicht normalverteilt sind.

# Breusch-Pagan-Test III

auf Heteroskedastie in den Störgrößen

- Beide Versionen des BP-Tests sind als „Score-Test“ konzipiert, die Teststatistik lässt sich jedoch jeweils leicht auf Basis von (OLS-)Schätzergebnissen einer (linearen) Hilfsregression berechnen.
- Sind $\widehat{u}_i$ die Residuen aus der Schätzung des auf heteroskedastische Störgrößen zu untersuchenden linearen Modells und RSS die *Residual Sum of Squares* (mit $RSS = \sum_{i=1}^{n} \widehat{u}_i^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}$), so benötigt man als *abhängige* Variable der Hilfsregression die gemäß

$$w_i := \frac{n}{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}\widehat{u}_i^2 = \frac{n}{\mathrm{RSS}}\widehat{u}_i^2 \quad \text{für} \quad i \in \{1, \ldots, n\}$$

„standardisierten“ **quadrierten** Residuen $w_i$.

# Breusch-Pagan-Test IV

auf Heteroskedastie in den Störgrößen

- Für beide Versionen des BP-Tests ist dann die Hilfsregression

$$w_i = \gamma_0 + \gamma_1 \cdot z_{1i} + \ldots + \gamma_Q \cdot z_{Qi} + e_i, \quad i \in \{1, \ldots, n\},$$

  (per OLS-/KQ-Methode) zu schätzen.
- Im ursprünglichen BP-Test erhält man die unter der Nullhypothese näherungsweise $\chi^2(Q)$-verteilte Teststatistik dann als die **Hälfte der „Explained Sum of Squares" der Hilfsregression**, mit der Bezeichnung $\widehat{e}_i$ für die Residuen der **Hilfs**regression und der Abkürzung $\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$ also zum Beispiel unter Verwendung von ESS = TSS − RSS durch

$$\chi^2 = \frac{1}{2} \cdot \left( \left( \sum_{i=1}^{n} (w_i - \overline{w})^2 \right) - \left( \sum_{i=1}^{n} \widehat{e}_i^2 \right) \right) .$$

# Breusch-Pagan-Test V

auf Heteroskedastie in den Störgrößen

- In der robusteren Version von Koenker erhält man die unter der Nullhypothese ebenfalls näherungsweise $\chi^2(Q)$-verteilte Teststatistik als *n*-**faches multiples Bestimmtheitsmaß der Hilfsregression**, es gilt also

$$\chi^2 = n \cdot R_H^2$$

  mit der Bezeichnung $R_H^2$ für das Bestimmtheitsmaß der Hilfsregression.
- Offensichtlich kann **(nur)** bei Verwendung der Version von Koenker auf die Standardisierung der quadrierten Residuen der ursprünglichen Modellschätzung verzichtet werden und die Hilfsregression auch direkt mit der abhängigen Variablen $\widehat{u}_i^2$ durchgeführt werden, da dies das Bestimmtheitsmaß nicht ändert (wohl aber die ESS!).

# Zusammenfassung: Breusch-Pagan-Test („Original“)

auf Heteroskedastizität der Störgrößen

| | |
|---|---|
| Anwendungs-voraussetzungen | approx.: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ mit $\mathrm{E}(\mathbf{u}) = \mathbf{0}$, $\mathrm{V}(\mathbf{u}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, $\mathbf{X}$ deterministisch mit vollem Spaltenrang $K+1$, Realisation $\mathbf{y} = (y_1, \ldots, y_n)'$ beobachtet, $Q$ Einflussvariablen $z_{1i}, \ldots, z_{Qi}$, $\sigma_i^2 = h(\gamma_0 + \gamma_1 \cdot z_{1i} + \ldots + \gamma_Q \cdot z_{Qi})$ |
| Nullhypothese<br>Gegenhypothese | $H_0 : \gamma_1 = \ldots = \gamma_Q = 0 \iff \sigma_i^2 \equiv h(\gamma_0)$<br>$H_1 : \gamma_q \neq 0$ für mindestens ein $q \in \{1, \ldots, Q\}$ |
| Teststatistik<br><br>Verteilung ($H_0$) | $\chi^2 = \frac{1}{2} \cdot \left( \left( \sum_{i=1}^{n} (w_i - \overline{w})^2 \right) - \left( \sum_{i=1}^{n} \widehat{e}_i^2 \right) \right)$<br>$\chi^2$ ist approx. $\chi^2(Q)$-verteilt, falls $\sigma_i^2 \equiv h(\gamma_0)$ konstant. |
| Benötigte Größen | $\widehat{\mathbf{u}} = (\widehat{u}_1, \ldots, \widehat{u}_n)' = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, $w_i = \frac{n}{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}} \widehat{u}_i^2$,<br>$\widehat{e}_i$ die Residuen der Hilfsregression<br>$w_i = \gamma_0 + \gamma_1 \cdot z_{1i} + \ldots + \gamma_Q \cdot z_{Qi} + e_i$ |
| Kritischer Bereich zum Niveau $\alpha$ | $(\chi^2_{Q;1-\alpha}, \infty)$ |
| $p$-Wert | $1 - F_{\chi^2(Q)}(\chi^2)$ |

# Zusammenfassung: Breusch-Pagan-Test („Koenker“)

auf Heteroskedastizität der Störgrößen

| | |
|---|---|
| Anwendungs-voraussetzungen | approx.: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ mit $\mathrm{E}(\mathbf{u}) = \mathbf{0}$, $\mathrm{V}(\mathbf{u}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, $\mathbf{X}$ deterministisch mit vollem Spaltenrang $K+1$, Realisation $\mathbf{y} = (y_1, \ldots, y_n)'$ beobachtet, $Q$ Einflussvariablen $z_{1i}, \ldots, z_{Qi}$, $\sigma_i^2 = h(\gamma_0 + \gamma_1 \cdot z_{1i} + \ldots + \gamma_Q \cdot z_{Qi})$ |
| Nullhypothese<br>Gegenhypothese | $H_0 : \gamma_1 = \ldots = \gamma_Q = 0 \iff \sigma_i^2 \equiv h(\gamma_0)$<br>$H_1 : \gamma_q \neq 0$ für mindestens ein $q \in \{1, \ldots, Q\}$ |
| Teststatistik<br>Verteilung ($H_0$) | $\chi^2 = n \cdot R_H^2$<br>$\chi^2$ ist approx. $\chi^2(Q)$-verteilt, falls $\sigma_i^2 \equiv h(\gamma_0)$ konstant. |
| Benötigte Größen | $\widehat{\mathbf{u}} = (\widehat{u}_1, \ldots, \widehat{u}_n)' = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, $R_H^2$ das Bestimmtheitsmaß der Hilfsregression $\widehat{u}_i^2 = \gamma_0 + \gamma_1 \cdot z_{1i} + \ldots + \gamma_Q \cdot z_{Qi} + e_i$ |
| Kritischer Bereich zum Niveau $\alpha$ | $(\chi^2_{Q;1-\alpha}, \infty)$ |
| $p$-Wert | $1 - F_{\chi^2(Q)}(\chi^2)$ |

# White-Test

auf Heteroskedastie in den Störgrößen

- White hat in seiner Arbeit von 1980 (Econometrica) nicht nur heteroskedastie-konsistente Schätzverfahren, sondern auch einen Test auf Heteroskedastie in den Störgrößen vorgeschlagen.
- Es zeigt sich, dass der **White-Test** auf heteroskedatische Störgrößen ein Spezialfall der „Koenker"-Version des Breusch-Pagan-Tests ist.
- Konkret erhält man den White-Test bei der Durchführung eines Breusch-Pagan-Tests nach Koenker, wenn man als Einflussvariablen $z_{qi}$ für die Varianz der Störgrößen gerade
  - alle Regressoren, zusätzlich
  - alle quadrierten Regressoren sowie zusätzlich
  - alle gemischten Produkte von Regressoren

  des ursprünglichen Modells wählt.
- In einem Modell mit 2 Regressoren wäre also die Hilfsregression

$$\widehat{u}_i^2 = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{1i}^2 + \gamma_4 x_{2i}^2 + \gamma_5 x_{1i} x_{2i} + e_i$$

  durchzuführen.

# Beispiel: Breusch-Pagan-Test/White-Test I

- Im Folgenden werden zwei Varianten des Breusch-Pagan-Test am bereits mehrfach verwendeten „Lohnhöhen“-Beispiel illustriert.
- Ausgehend von den quadrierten Residuen $\widehat{u}_i^2$ der ursprünglichen Regression der `Lohnhöhe` auf die beiden Regressoren `Ausbildung` und `Alter` (sowie ein Absolutglied) werden für die „Original“-Version des Breusch-Pagan-Tests zunächst die standardisierten quadrierten Residuen $w_i = \frac{n}{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}\widehat{u}_i^2$ berechnet:

```
> uhat <- residuals(lm(Lohnhöhe~Ausbildung+Alter))
> w    <- uhat^2/mean(uhat^2)
```

Als Summe der quadrierten Abweichungen vom arithmetischen Mittel $\sum_{i=1}^{n}(w_i - \overline{w})^2$ der $w_i$ (also als TSS der folgenden Hilfsregression!) erhält man:

```
> sum((w-mean(w))^2)

[1] 72.66564
```

# Beispiel: Breusch-Pagan-Test/White-Test II

- Werden als Einflussvariablen für die Varianz der Störgrößen die beiden ursprünglichen Regressoren `Ausbildung` und `Alter` gewählt, ist dann die Hilfsregression

$$w_i = \gamma_0 + \gamma_1 \texttt{Ausbildung}_i + \gamma_2 \texttt{Alter}_i + e_i$$

zu schätzen und die zugehörige RSS zu bestimmen, man erhält

```
> sum(residuals(lm(w~Ausbildung+Alter))^2)
[1] 45.76786
```

und damit (gerundet) die Teststatistik

$$\chi^2 = \frac{1}{2} \cdot \left( \left( \sum_{i=1}^{n} (w_i - \overline{w})^2 \right) - \left( \sum_{i=1}^{n} \widehat{e}_i^2 \right) \right) = \frac{1}{2}(72.666 - 45.768) = 13.449 \ .$$

- Ein Vergleich zum kritischen Wert $\chi^2_{2;0.95} = 5.991$ bei einem Test zum Niveau $\alpha = 0.05$ erlaubt die Ablehnung der Nullhypothese und damit den Schluss auf das Vorliegen von Heteroskedastie in den Störgrößen.

# Beispiel: Breusch-Pagan-Test/White-Test III

- Wird in der beschriebenen Situation ein White-Test durchgeführt, so muss eine der Hilfsregressionen

$$\begin{aligned}\widehat{u}_i^2 = \gamma_0 &+ \gamma_1 \cdot \texttt{Ausbildung}_i + \gamma_2 \cdot \texttt{Alter}_i + \gamma_3 \cdot \texttt{Ausbildung}_i^2 \\ &+ \gamma_4 \cdot \texttt{Alter}_i^2 + \gamma_5 \cdot \texttt{Ausbildung}_i \cdot \texttt{Alter}_i + e_i\end{aligned}$$

  oder

$$\begin{aligned}w_i = \gamma_0 &+ \gamma_1 \cdot \texttt{Ausbildung}_i + \gamma_2 \cdot \texttt{Alter}_i + \gamma_3 \cdot \texttt{Ausbildung}_i^2 \\ &+ \gamma_4 \cdot \texttt{Alter}_i^2 + \gamma_5 \cdot \texttt{Ausbildung}_i \cdot \texttt{Alter}_i + e_i\end{aligned}$$

  durchgeführt werden.
- In der Statistik-Software **R** müssen diese „Rechenoperationen“ von Regressoren bei der Modellformulierung in den Befehl „`I()`“ eingeschlossen werden, da „^“ und „*“ bei der Notation von Modellgleichungen andere Bedeutungen haben!

# Beispiel: Breusch-Pagan-Test/White-Test IV

- Man erhält als OLS-Schätzergebnis:

```
Call:
lm(formula = uhat^2 ~ Ausbildung + Alter + I(Ausbildung^2) +
    I(Alter^2) + I(Ausbildung * Alter))

Residuals:
    Min      1Q  Median      3Q     Max
-104762  -17524   -9639   29687   78007

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            5778.593 125459.783   0.046   0.9639
Ausbildung            -5788.874  23416.039  -0.247   0.8083
Alter                    -6.682   6568.457  -0.001   0.9992
I(Ausbildung^2)       -6319.607   2139.021  -2.954   0.0105 *
I(Alter^2)              -58.640     92.777  -0.632   0.5375
I(Ausbildung * Alter)  1826.589    549.299   3.325   0.0050 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58820 on 14 degrees of freedom
Multiple R-squared:  0.7093,        Adjusted R-squared:  0.6055
F-statistic: 6.831 on 5 and 14 DF,  p-value: 0.002013
```

# Beispiel: Breusch-Pagan-Test/White-Test V

- Unter Verwendung des Bestimmtheitsmaßes dieser Hilfsregression ergibt sich $\chi^2 = n \cdot R_H^2 = 20 \cdot 0.7093 = 14.186 > \chi^2_{2;0.95} = 5.991$, also wird auch hier zum Niveau $\alpha = 0.05$ signifikante Heteroskedastie in den Störgrößen festgestellt.
- Schneller: mit dem Befehl `bptest()` im Paket `lmtest`:
  - „Original"-Breusch-Pagan-Test (1. Beispiel):

```
> bptest(lm(Lohnhöhe~Ausbildung+Alter),studentize=FALSE)

        Breusch-Pagan test

data:  lm(Lohnhöhe ~ Ausbildung + Alter)
BP = 13.4489, df = 2, p-value = 0.001201
```

  - „White"- bzw. „Koenker"-Variante (2. Beispiel):

```
> bptest(lm(Lohnhöhe~Ausbildung+Alter),
+ ~Ausbildung+Alter+I(Ausbildung^2)+I(Alter^2)+I(Ausbildung*Alter))

        studentized Breusch-Pagan test

data:  lm(Lohnhöhe ~ Ausbildung + Alter)
BP = 14.1857, df = 5, p-value = 0.01447
```

# Inhaltsverzeichnis

(Ausschnitt)

# Nichtlinearität in den Regressoren I

- Eine Variable $y$ hängt linear von einer Variablen $x$ ab, wenn der Differenzenquotient bzw. die Ableitung bzgl. dieser Variablen konstant ist, wenn also

$$\frac{\Delta y}{\Delta x} = c \qquad \text{bzw.} \qquad \frac{\partial y}{\partial x} = c$$

  für eine Konstante $c \in \mathbb{R}$ gilt.

- Im bisher betrachteten linearen Regressionsmodell

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + u_i, \qquad i \in \{1, \ldots, n\},$$

  hängt $y$ also linear von jedem Regressor $x_k$ ($k \in \{1, \ldots, K\}$) ab, denn es gilt

$$\frac{\Delta y}{\Delta x_k} = \beta_k \qquad \text{bzw.} \qquad \frac{\partial y}{\partial x_k} = \beta_k \ .$$

- Die hier als „marginaler Effekt“ einer Änderung von $x_k$ auf $y$ interpretierbare (partielle) Ableitung ist also kostant und damit insbesondere unabhängig von $x_k$ (sowie unabhängig von anderen Variablen).

# Nichtlinearität in den Regressoren II

- Bereits im White-Test verwendet: „Regressionsfunktion“

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} \ ,$$

  die zwar linear in den Regressionsparametern $\beta_0, \ldots, \beta_5$, aber **nichtlinear in den Regressoren** $x_1$ und $x_2$ ist.

- Der marginale Effekt einer Änderung von $x_1$ auf $y$ beträgt hier beispielsweise (abhängig vom Wert der Regressoren $x_1$ und $x_2$!)

$$\frac{\partial y}{\partial x_1} = \beta_1 + 2\beta_3 x_1 + \beta_5 x_2 \ .$$

- Allgemein betrachten wir nun Regressionsmodelle, die sich in der Form

$$g(y_i) = \beta_0 + \beta_1 h_1(x_{1i}, \ldots, x_{Ki}) + \ldots + \beta_M h_M(x_{1i}, \ldots, x_{Ki}) + u_i, \quad i \in \{1, \ldots, n\},$$

  mit $M$ Transformationen $h_1, \ldots, h_M$ der $K$ Regressoren und (ggf.) einer Transformation $g$ der abhängigen Variablen darstellen lassen.

# Nichtlinearität in den Regressoren III

- Unter den üblichen Annahmen an die Störgrößen $u_i$ und unter der Voraussetzung, dass die Transformationen $h_1, \ldots, h_M$ zu einer „neuen“ Regressormatrix

$$\widetilde{\mathbf{X}} := \begin{pmatrix} 1 & h_1(x_{11}, \ldots, x_{K1}) & \cdots & h_M(x_{11}, \ldots, x_{K1}) \\ 1 & h_1(x_{12}, \ldots, x_{K2}) & \cdots & h_M(x_{12}, \ldots, x_{K2}) \\ \vdots & \vdots & & \vdots \\ 1 & h_1(x_{1n}, \ldots, x_{Kn}) & \cdots & h_M(x_{1n}, \ldots, x_{Kn}) \end{pmatrix}$$

  mit vollem Spaltenrang $M + 1$ führen, bleiben die bisher besprochenen Eigenschaften der OLS-/KQ-Schätzung dieses Modells bestehen.
- Bezeichnet $\widetilde{\mathbf{y}} := (g(y_1), \ldots, g(y_n))'$ den transformierten (bzw. – falls $g(y) = y$ für alle $y \in \mathbb{R}$ gilt – untransformierten) Vektor der abhängigen Variable, erhält man beispielsweise den KQ-Schätzer als

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{y}} \, .$$

# Nichtlinearität in den Regressoren IV

- Weitere Beispiele für Modelle mit Regressionsfunktionen, die nichtlinear in den (ursprünglichen) Regressoren $x_k$ sind:
  1. $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + u_i$,
  2. $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + u_i$,
  3. $y_i = \beta_0 + \beta_1 \ln(x_{1i}) + u_i$,
  4. $\ln(y_i) = \beta_0 + \beta_1 x_{1i} + u_i$,
  5. $\ln(y_i) = \beta_0 + \beta_1 \ln(x_{1i}) + \beta_2 \ln(x_{2i}) + u_i$.

**Wichtig!**

Unabhängig von der konkreten Form der Regressionsfunktion muss (wie auch bisher!) die Korrektheit der Spezifikation der Regressionsfunkion gewährleistet sein, um die Ergebnisse der Schätzung überhaupt sinnvoll verwerten zu können!

- Im Folgenden werden zunächst Regressionsfunktionen untersucht, die nur von einer unabhängigen Variablen $x_1$ abhängen (wie in den Beispielen 1 – 4).

# Polynomiale Modelle I

in nur einer Variablen $x_1$

- Die Modelle aus (1) bzw. (2),

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + u_i \quad \text{bzw.} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + u_i,$$

sind Beispiele für **polynomiale Modelle** (in einer Variablen) der Form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \ldots + \beta_r x_{1i}^r + u_i$$

zu *vorgegebenem* Grad $r \in \{2, 3, \ldots\}$ des Polynoms.

- In polynomialen Modellen (in einer Variablen) sind die marginalen Effekte einer Änderung von $x_1$ auf $y$ gegeben durch

$$\frac{\partial y}{\partial x_1} = \beta_1 + 2\beta_2 x_1 + \ldots + r\beta_r x_1^{r-1}$$

und damit insbesondere nicht konstant, sondern abhängig vom Regressor $x_1$.

# Polynomiale Modelle II

in nur einer Variablen $x_1$

- Konfidenzintervalle für die marginalen Effekte an einem vorgegebenen Wert $x_1$ des Regressors können dann als Konfidenzintervalle für Linearkombinationen $\mathbf{a}'\boldsymbol{\beta}$ bestimmt werden, wenn der Vektor $\mathbf{a} \in \mathbb{R}^{r+1}$ (abhängig von $x_1$) entsprechend gewählt wird, im polynomialen Modell mit Polynomgrad $r$ also als

$$\mathbf{a} = \begin{bmatrix} 0 & 1 & 2x_1 & \dots & rx_1^{r-1} \end{bmatrix}' .$$

- Bei einer sehr großen Wahl von $r$ besteht die Gefahr des „Overfittings“: Sind bei einer „Punktwolke“ aus $n$ Beobachtungen $(x_{1i}, y_i)$ alle $x_i$ unterschiedlich, so kann die Punktwolke durch ein Polynom vom Grad $r = n - 1$ perfekt „interpoliert“ werden!
- In der Praxis finden sich häufig polynomiale Modelle mit $r = 2$ oder $r = 3$.

# Polynomiale Modelle III

in nur einer Variablen $x_1$

- Gelegentlich wird – unter der Annahme, dass die wahre Regressionsfunktion ein Polynom von unbekanntem Grad ist – zunächst ein Modell mit „großem“ $r$ geschätzt und dann sukzessive mit Hilfe von $t$-Tests überprüft, ob $\beta_r$ signifikant von Null verschieden ist, um ggf. den Grad $r$ des Polynoms in der Regressionsfunktion um 1 zu reduzieren.
- Die Nullhypothese eines linearen Zusammenhangs gegen die Alternative eines polynomialen Zusammenhangs (mit Polynomgrad $r \geq 2$) kann offensichtlich durch einen $F$-Test mit

$$H_0 : \beta_2 = \ldots = \beta_r = 0$$

überprüft werden.

- Natürlich können Tests bzw. Konfidenzintervalle auch unter der Annahme heteroskedastischer Störgrößen durchgeführt werden, wenn die entsprechende heteroskedastie-konsistente Schätzung $\widehat{\mathrm{V}}_{\mathrm{hc}}(\widehat{\boldsymbol{\beta}})$ der Varianz-Kovarianzmatrix $\mathrm{V}(\widehat{\boldsymbol{\beta}})$ und die dafür geeigneten Darstellungen der jeweiligen Tests verwendet werden.

# (Semi-)logarithmische Modelle I

in nur einer Variablen $x_1$

- **Log-Transformationen von $x_{1i}$ in $\ln(x_{1i})$ und/oder $y_i$ in $\ln(y_i)$** bieten sich dann an, wenn anstelle der Annahme eines konstanten Effekts $\Delta y = \beta_1 \Delta x_1$ von absoluten Änderungen $\Delta x_1$ auf absolute Änderungen $\Delta y$ eher dann ein konstanter Effekt $\beta_1$ erwartet wird, wenn *relative, prozentuale* Änderungen bei der Ursache $(\frac{\Delta x_1}{x_1})$ und/oder bei der abhängigen Variablen $(\frac{\Delta y}{y})$ betrachtet werden.
- Grundlage dafür ist $\frac{\partial \ln(x)}{\partial x} = \frac{1}{x}$ bzw.

$$\ln(x + \Delta x) - \ln(x) = \ln\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}, \text{ wenn } |\Delta x| \ll |x|.$$

- Abhängig davon, ob nur die unabhängige Variable, nur die abhängige Variable oder beide Variablen transformiert werden, sind die folgenden Spezifikationen möglich:

# (Semi-)logarithmische Modelle II

in nur einer Variablen $x_1$

1. Linear-log-Spezifikation:

$$y_i = \beta_0 + \beta_1 \ln(x_{1i}) + u_i.$$

Konstanter Effekt $\beta_1$ der relativen Änderung von $x_1$ auf eine absolute Änderung von $y$, bzw. abnehmender marginaler Effekt bei steigendem $x$:

$$\Delta y \approx \beta_1 \frac{\Delta x_1}{x_1} \quad \text{bzw.} \quad \frac{\partial y}{\partial x_1} = \frac{\beta_1}{x_1}$$

Bsp.: $x_{1i}$ Düngemitteleinsatz, $y_i$ Ernteertrag (auf Feld $i$).

- Eine (relative) Erhöhung des Düngemitteleinsatzes um 1% erhöht den (absoluten) Ernteertrag (etwa) um $0.01 \cdot \beta_1$.
- Eine (absolute) Erhöhung des Düngemitteleinsatzes um einen Betrag $\Delta x_1$ hat dort mehr Wirkung, wo noch nicht so viel Dünger eingebracht wurde („abnehmende Grenzerträge“).

# (Semi-)logarithmische Modelle III

in nur einer Variablen $x_1$

2. Log-linear-Spezifikation:

$$\ln(y_i) = \beta_0 + \beta_1 x_{1i} + u_i.$$

Konstanter Effekt $\beta_1$ der absoluten Änderung von $x_1$ auf eine relative Änderung von $y$, bzw. steigender marginaler Effekt bei steigendem $y$:

$$\frac{\Delta y}{y} \approx \beta_1 \Delta x_1 \quad \text{bzw.} \quad \frac{\partial y}{\partial x_1} = \beta_1 y$$

Bsp.: $x_{1i}$ Berufserfahrung von BWL-Absolventen (in Jahren), $y_i$ Einkommen.

- Ein Jahr zusätzliche Berufserfahrung erhöht danach das mittlere Einkommen um etwa $100\beta_1\%$.
- Eine (absolute) Erhöhung der Berufserfahrung hat also einen höheren (absoluten) Effekt auf das Einkommen dort, wo das Einkommen ohnehin bereits ein höheres Niveau hatte.

# (Semi-)logarithmische Modelle IV

in nur einer Variablen $x_1$

3. Log-log-Spezifikation:

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_{1i}) + u_i.$$

Konstanter Effekt $\beta_1$ (=Elastizität) der relativen Änderung von $x_1$ auf eine relative Änderung von $y$:

$$\frac{\Delta y}{y} \approx \beta_1 \frac{\Delta x_1}{x_1} \quad \text{bzw.} \quad \frac{\partial y}{\partial x_1} \frac{x_1}{y} = \beta_1$$

Bsp.: $x_{1i}$ Kapitaleinsatz pro Arbeitskraft, $y_i$ Output pro Arbeitskraft.

- Erhöhung des per-capita-Kapitaleinsatzes um 1% führt zur Erhöhung des per-capita-Output um $\beta_1$% (Cobb-Douglas-Produktionsfunktion).
- Modellierung von „konstanten Skalenerträgen“.

# (Semi-)logarithmische Modelle V

in nur einer Variablen $x_1$

**Anmerkungen zu Log-transformierten abhängigen Variablen ($\ln(y)$)**

- Insbesondere Log-log-Spezifikationen können bei der sog. „Linearisierung“ von Regressionsmodellen entstehen, die zunächst nichtlinear (auch!) in den Regressionsparametern sind, zum Beispiel erhält man aus dem Modell (hier: mit mehreren Regressoren)

$$y_i = \beta_0 \cdot x_{1i}^{\beta_1} \cdot x_{2i}^{\beta_2} \cdot e^{u_i}, \qquad i \in \{1, \ldots, n\},$$

durch Logarithmieren auf beiden Seiten mit

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_{1i}) + \beta_2 \ln(x_{2i}) + u_i, \qquad i \in \{1, \ldots, n\}.$$

ein „linearisiertes“ Modell.

# (Semi-)logarithmische Modelle VI

in nur einer Variablen $x_1$

- Bei der Prognose von $y_0$ gegeben $\mathbf{x_0}$ bzw. der Bestimmung von $\widehat{y}_i$ auf Basis von Modellen mit log-tranformierter abhängiger Variablen $\ln(y)$ ist zu beachten, dass wegen $\mathrm{E}\left(e^{u_i}\right) \neq e^{\mathrm{E}(u_i)}$ trotz der Annahme $\mathrm{E}(u_i) \equiv 0$ im Allgemeinen $\mathrm{E}\left(e^{u_i}\right) \neq 1 = e^0$ gilt. Für $u_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ gilt insbesondere $\mathrm{E}\left(e^{u_i}\right) = e^{\frac{\sigma^2}{2}}$, damit erhält man für $\ln(y_i) = h(x_{1i}) + u_i$ mit $u_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

$$\begin{aligned} \mathrm{E}(y_i) &= \mathrm{E}\left(e^{\ln(y_i)}\right) = \mathrm{E}\left(e^{h(x_{1i})+u_i}\right) = \mathrm{E}\left(e^{h(x_{1i})} \cdot e^{u_i}\right) \\ &= e^{h(x_{1i})} \cdot \mathrm{E}\left(e^{u_i}\right) = e^{h(x_{1i})} \cdot e^{\frac{\sigma^2}{2}} > e^{h(x_{1i})} . \end{aligned}$$

- Wenn die abhängige Variable $y$ in $\ln(y)$ transformiert wird, kann man das Bestimmtheitsmaß für die geschätzte Regression nicht sinnvoll mit dem Bestimmtheitsmaß einer Regressionsgleichung für $y$ vergleichen!
*(Anteil der erklärten Varianz der* $\ln(y_i)$ *vs. Anteil der erklärten Varianz der* $y_i$*)*

# Beispiel zur Nichtlinearität in einer Variablen I

- Im Folgenden soll am Beispiel der Abhängigkeit der Milchleistung von Kühen von der zugeführten Futtermenge die Schätzung einiger in den Regressoren nichtlinearer Modelle illustriert werden.
- Es liege hierzu folgender Datensatz vom Umfang $n = 12$ zu Grunde:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Milchleistung (Liter/Jahr) $y_i$ | 6525 | 8437 | 8019 | 8255 | 5335 | 7236 |
| Futtermenge (Zentner/Jahr) $x_{1i}$ | 10 | 30 | 20 | 33 | 5 | 22 |

| $i$ | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|
| Milchleistung (Liter/Jahr) $y_i$ | 5821 | 7531 | 8320 | 4336 | 7225 | 8112 |
| Futtermenge (Zentner/Jahr) $x_{1i}$ | 8 | 14 | 25 | 1 | 17 | 28 |

(vgl. *von Auer, Ludwig: Ökonometrie – Eine Einführung, 6. Aufl., Tabelle 14.1*)

- Es wird nacheinander die Gültigkeit einer linearen, quadratischen, kubischen, linear-log-, log-linear- bzw. log-log-Spezifikation unterstellt und das zugehörige Modell geschätzt (unter Homoskedastieannahme).

# Beispiel zur Nichtlinearität in einer Variablen II

- Lineares Modell: $\texttt{Milch}_i = \beta_1 + \texttt{Futter}_i + u_i$

```
Call:
lm(formula = Milch ~ Futter)

Residuals:
   Min     1Q Median     3Q    Max
-768.2 -275.0 -115.6  353.4  880.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4985.27     312.84  15.935 1.95e-08 ***
Futter        118.91      15.39   7.725 1.60e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 527.9 on 10 degrees of freedom
Multiple R-squared:  0.8565,        Adjusted R-squared:  0.8421
F-statistic: 59.68 on 1 and 10 DF,  p-value: 1.597e-05
```

# Beispiel zur Nichtlinearität in einer Variablen III

- Quadratisches Modell: $\texttt{Milch}_i = \beta_1 + \texttt{Futter}_i + \texttt{Futter}_i^2 + u_i$

```
Call:
lm(formula = Milch ~ Futter + I(Futter^2))

Residuals:
    Min      1Q  Median      3Q     Max
-699.14 -135.47   -2.44  179.63  490.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4109.445    290.487  14.147 1.87e-07 ***
Futter       271.393     38.626   7.026 6.14e-05 ***
I(Futter^2)   -4.432      1.087  -4.076  0.00277 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 329.9 on 9 degrees of freedom
Multiple R-squared:  0.9496,        Adjusted R-squared:  0.9384
F-statistic: 84.74 on 2 and 9 DF,  p-value: 1.452e-06
```

# Beispiel zur Nichtlinearität in einer Variablen IV

- Kubisches Modell: $\texttt{Milch}_i = \beta_1 + \texttt{Futter}_i + \texttt{Futter}_i^2 + \texttt{Futter}_i^3 + u_i$

```
Call:
lm(formula = Milch ~ Futter + I(Futter^2) + I(Futter^3))

Residuals:
    Min      1Q  Median      3Q     Max
-641.92 -117.82    5.13  202.86  447.31

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 3954.93841  389.73064  10.148 7.61e-06 ***
Futter       327.00926   97.73076   3.346   0.0101 *
I(Futter^2)   -8.50791    6.63147  -1.283   0.2354
I(Futter^3)    0.07951    0.12747   0.624   0.5502
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 341.7 on 8 degrees of freedom
Multiple R-squared:  0.9519,        Adjusted R-squared:  0.9339
F-statistic: 52.79 on 3 and 8 DF,  p-value: 1.29e-05
```

# Beispiel zur Nichtlinearität in einer Variablen V

- Linear-log-Modell: $\texttt{Milch}_i = \beta_1 + \ln(\texttt{Futter}_i) + u_i$

```
Call:
lm(formula = Milch ~ log(Futter))

Residuals:
    Min      1Q  Median      3Q     Max
-635.74 -287.21   33.02  373.09  517.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3818.3      358.2  10.660 8.82e-07 ***
log(Futter)   1268.8      130.1   9.754 2.00e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 429.8 on 10 degrees of freedom
Multiple R-squared:  0.9049,        Adjusted R-squared:  0.8954
F-statistic: 95.14 on 1 and 10 DF,  p-value: 1.996e-06
```

# Beispiel zur Nichtlinearität in einer Variablen VI

- Log-linear-Modell: $\ln(\texttt{Milch}_i) = \beta_1 + \texttt{Futter}_i + u_i$

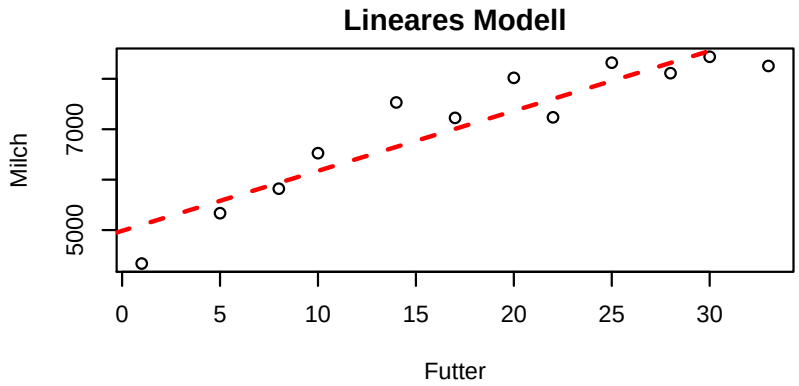
```
Call:
lm(formula = log(Milch) ~ Futter)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16721 -0.03642 -0.01678  0.05692  0.14677

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.523601   0.055220 154.358  < 2e-16 ***
Futter      0.018315   0.002717   6.741  5.1e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09318 on 10 degrees of freedom
Multiple R-squared:  0.8196,        Adjusted R-squared:  0.8016
F-statistic: 45.44 on 1 and 10 DF,  p-value: 5.098e-05
```

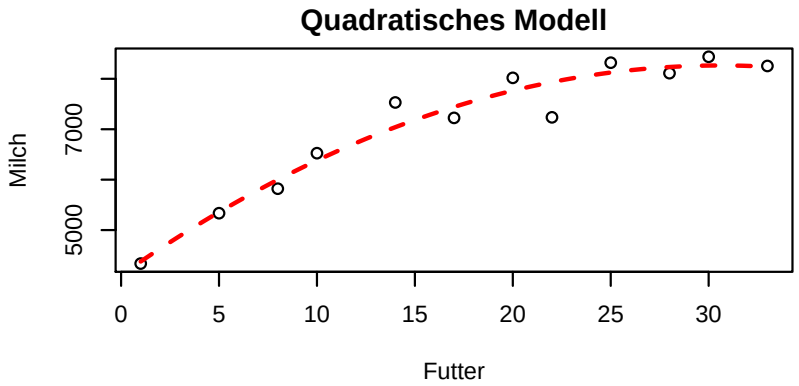
# Beispiel zur Nichtlinearität in einer Variablen VII

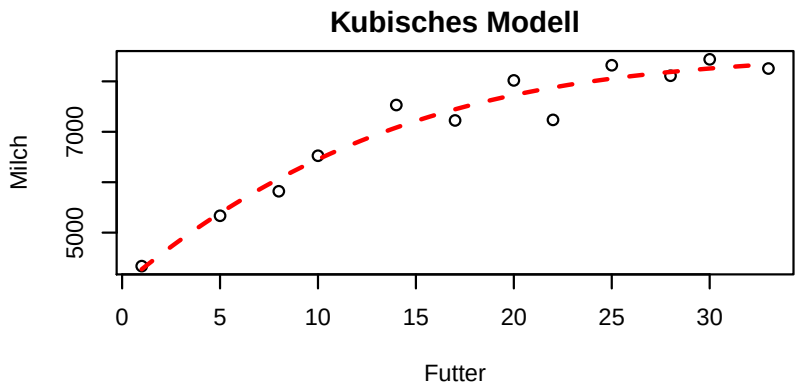
- Log-log-Modell: $\ln(\texttt{Milch}_i) = \beta_1 + \ln(\texttt{Futter}_i) + u_i$

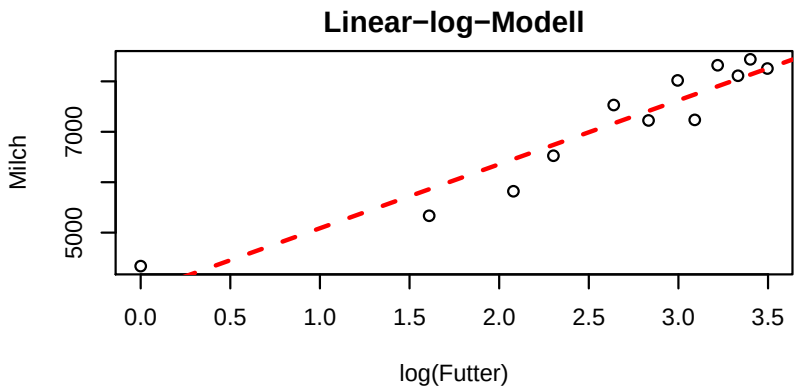
```
Call:
lm(formula = log(Milch) ~ log(Futter))

Residuals:
      Min         1Q     Median         3Q        Max
-0.076867 -0.028385 -0.004122  0.049235  0.066730

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.32264    0.04468  186.29  < 2e-16 ***
log(Futter)  0.20364    0.01622   12.55 1.91e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0536 on 10 degrees of freedom
Multiple R-squared:  0.9403,        Adjusted R-squared:  0.9343
F-statistic: 157.5 on 1 and 10 DF,  p-value: 1.912e-07
```

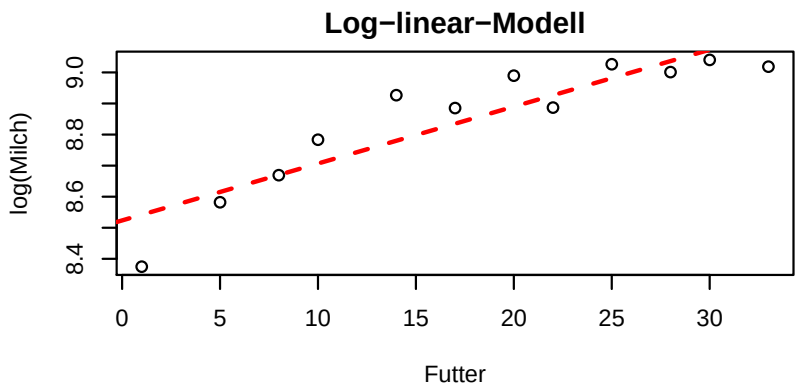
# Geschätzte Regressions-/Prognosefunktionen I

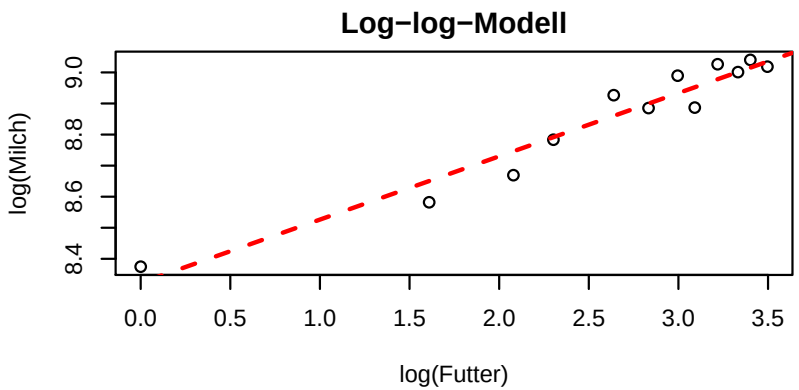


# Geschätzte Regressions-/Prognosefunktionen II