4. Spezifische Fragen der linearen Regressionsanalyse

4.1 Datenskalierung

Falls abhängige und/oder erklärende Variablen in linearen Regressionsmodellen unterschiedlich skaliert sind (z.B. zur Verminderung der Anzahl an Nullen nach dem Komma bei Dezimalzahlen), verändern sich (mit OLS geschätzte) Regressionsparameter, Konfidenzintervalle und Teststatistiken, so dass alle geschätzten Effekte weiterhin aufrecht erhalten werden:

- Falls die abhängige Variable mit einer Konstanten c multipliziert wird, werden alle geschätzten Regressionsparameter, die geschätzte Standardabweichung des Störterms u, die geschätzten Standardabweichungen der geschätzten Regressionsparameter, die Unter- und Obergrenze der Konfidenzintervalle mit c sowie die Residualabweichungsquadratsumme SSR mit c^2 multipliziert. Die t- und F-Statistiken sowie das Bestimmtheitsmaß R^2 bleiben dagegen in diesem Fall gleich.
- Falls eine einzelne erklärende Variable mit einer Konstanten c multipliziert wird, werden der entsprechende geschätzte Steigungsparameter, die geschätzte Standardabweichung dieses geschätzten Steigungsparameters sowie die entsprechende Unter- und Obergrenze des Konfidenzintervalls durch c dividiert. Alle anderen Werte bleiben in diesem Fall dagegen gleich.

Beispiel: Erklärung von Geburtsgewichten

Mit Hilfe eines linearen Regressionsmodells soll der Effekt der durchschnittlichen Anzahl der von der Mutter während der Schwangerschaft täglich gerauchten Zigaretten (cigs) sowie des jährlichen Familieneinkommens (faminc) in 1000 Dollar auf das Geburtsgewicht des Kindes (bwght) in ounces (= 28,3495 Gramm) untersucht werden. Dabei haben sich mit STATA für n = 1388 Geburten folgende OLS-Schätzergebnisse gezeigt:

```
reg bwght cigs faminc
```

Source	SS	df	MS	Number of obs =	1388
Model	17126.2088	2	8563.10442	F(2, 1385) =	21.27
Residual	557485.511	1385	402.516614	Prob > F =	0.0000
				R-squared =	0.0298
Total	574611.72	1387	414.283864	Adj R-squared =	0.0284
				Root MSE =	20.063

bwght	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cigs	-.4634075	.0915768	-5.06	0.000	-.6430518	-.2837633
faminc	.0927647	.0291879	3.18	0.002	.0355075	.1500219
_cons	116.9741	1.048984	111.51	0.000	114.9164	119.0319

Beispiel: Erklärung von Geburtsgewichten (Fortsetzung)

Interpretation:

- Der geschätzte Steigungsparameter für cigs beträgt hier -0,4634. Dies bedeutet z.B. dass bei fünf zusätzlich von der Mutter während der Schwangerschaft täglich gerauchten Zigaretten eine Verminderung des Geburtsgewichts um $5 \cdot 0{,}4634 = 2{,}317$ ounces geschätzt wird. Da die entsprechende t-Statistik den Wert -5,06 besitzt, hat cigs einen statistisch hoch signifikanten negativen Effekt.
- Zudem gilt z.B. SSR = 557485,511 und $\hat{\sigma} = 20{,}063$

Im Folgenden wird nun zunächst das Geburtsgewicht des Kindes (bwghtlbs) nicht mehr in ounces, sondern in pounds (= 16 ounces) gemessen. Es gilt somit bwghtlbs = bwght/16. In einem nächsten Schritt wird das Geburtsgewicht des Kindes (bwght) zwar wieder in ounces, die durchschnittliche Anzahl der von der Mutter während der Schwangerschaft täglich gerauchten Zigaretten aber in Packungen mit 20 Zigaretten (packs) gemessen, d.h. packs = cigs/20.

In diesen Fällen haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt:

Beispiel: Erklärung von Geburtsgewichten (STATA-Output)

```
reg bwghtlbs cigs faminc
```

Source	SS	df	MS	Number of obs	1388
Model	66.8992533	2	33.4496266	F(2, 1385)	21.27
Residual	2177.67778	1385	1.57233052	Prob > F	0.0000
Total	2244.57703	1387	1.61829634	R-squared	0.0298
				Adj R-squared	0.0284
				Root MSE	1.2539

bwghtlbs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cigs	-.028963	.0057236	-5.06	0.000	-.0401907	-.0177352
faminc	.0057978	.0018242	3.18	0.002	.0022192	.0093764
_cons	7.310883	.0655615	111.51	0.000	7.182273	7.439494

```
reg bwght packs faminc
```

Source	SS	df	MS	Number of obs	1388
Model	17126.2088	2	8563.10442	F(2, 1385)	21.27
Residual	557485.511	1385	402.516614	Prob > F	0.0000
Total	574611.72	1387	414.283864	R-squared	0.0298
				Adj R-squared	0.0284
				Root MSE	20.063

bwght	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
packs	-9.268151	1.831536	-5.06	0.000	-12.86104	-5.675265
faminc	.0927647	.0291879	3.18	0.002	.0355075	.1500219
_cons	116.9741	1.048984	111.51	0.000	114.9164	119.0319

Standardisierung von Variablen:

- Zur Vermeidung von Problemen der Datenskalierung bzw. für den Fall, dass manche abhängige oder erklärende Variablen nicht leicht zu interpretieren sind (z.B. Testergebnisse bei der Erklärung von Löhnen), können diese Variablen in linearen Regressionsmodellen standardisiert werden
- Eine Variable wird standardisiert, indem man das arithmetische Mittel in der Stichprobe subtrahiert und dann durch die entsprechende Standardabweichung dividiert
- Die (abweichenden) standardisierten Regressionsparameter können in linearen Regressionsmodellen dadurch geschätzt werden, dass die standardisierten abhängigen Variablen auf die standardisierten erklärenden Variablen ohne Einbeziehung einer Konstante regressiert werden
- Ohne Standardisierung gibt der geschätzte Steigungsparameter bekanntlich entsprechend (2.18) die Veränderung des OLS-Regressionswertes an, falls die entsprechende erklärende Variable um eine Einheit steigt (und alle anderen erklärenden Variablen konstant gehalten werden). Mit Standardisierung gibt der geschätzte Steigungsparameter dagegen die Veränderung der Standardabweichung des OLS-Regressionswertes an, falls die entsprechende erklärende Variable um eine Standardabweichung steigt.
- Somit werden bei der Standardisierung der Variablen unterschiedliche Skalierungen irrelevant. Darüber hinaus kann die Relevanz einzelner erklärender Variablen besser verglichen werden.

4.2 Funktionale Form

Logarithmierte und quadrierte Variablen:

In Kapitel 2 wurde bereits diskutiert, dass lineare Regressionsmodelle durch die Einbeziehung von (natürlich) logarithmierten und quadrierten Variablen auch nichtlineare Zusammenhänge einbeziehen können

Übersicht zur Einbeziehung logarithmierter Variablen:

Lineares Regressions-modell	Abhängige Variable	Erklärende Variable	Interpretation des Steigungsparameters
Level-level	y	x_j	$\Delta\hat{y} = \hat{\beta}_j \Delta x_j$
Level-log	y	$\log x_j$	$\Delta\hat{y} \approx (\hat{\beta}_j/100)\%\Delta x_j$
Log-level	log y	x_j	$\%\Delta\hat{y} \approx (100\hat{\beta}_j)\Delta x_j$
Log-log	log y	$\log x_j$	$\%\Delta\hat{y} \approx \hat{\beta}_j \%\Delta x_j$

Die Einbeziehung von (natürlichen) Logarithmen erfolgt vor allem in log-level und log-log Ansätzen.

Interpretation der Steigungsparameter:

- Falls y auf x_j regressiert wird, gibt $\hat{\beta}_j$ entsprechend (2.18) die Veränderung des OLS-Regressionswertes $\hat{y}$ an, falls x_j um eine Einheit steigt (und alle anderen erklärenden Variablen konstant gehalten werden)
- Falls y auf logx_j regressiert wird, gibt $\hat{\beta}_j/100$ (näherungsweise) die Veränderung des OLS-Regressionswertes $\hat{y}$ an, falls x_j um 1% steigt (und alle anderen erklärenden Variablen konstant gehalten werden)
- Falls logy auf x_j regressiert wird, gibt $100\hat{\beta}_j$ (näherungsweise) die prozentuale Veränderung des OLS-Regressionswertes $\hat{y}$ an, falls x_j um eine Einheit steigt (und alle anderen erklärenden Variablen konstant gehalten werden). $100\beta_j$ wird in diesem Fall auch als Semi-Elastizität der entsprechenden erklärenden Variablen bezeichnet.
- Allerdings handelt es sich dabei lediglich um grobe Approximationen durch Ausnutzung des Zusammenhangs $\%\Delta y \approx 100\Delta \log y$. Diese Approximationen werden jedoch für große Veränderungen von y immer ungenauer. Genaue (geschätzte) prozentuale Veränderungen können aber berechnet werden.
- Falls logy auf logx_j regressiert wird, gibt $\hat{\beta}_j$ (für sehr kleine Veränderungen approximativ) die prozentuale Veränderung des OLS-Regressionswertes $\hat{y}$ an, falls x_j um 1% steigt (und alle anderen erklärenden Variablen konstant gehalten werden). β_j stellt in diesem Fall eine Elastizität der entsprechenden erklärenden Variablen dar.

Beispiel: Effekt von Luftverschmutzung auf Immobilienpreise

Mit Hilfe eines linearen Regressionsmodells wird nun mit einer Stichprobe von n = 506 Gemeinden der Effekt des Logarithmus der Stickoxide in der Luft (lognox) und der durchschnittlichen Anzahl an Räumen in Häusern (rooms) auf den Logarithmus des Medians der Immobilienpreise (logprice) untersucht. Mit STATA haben sich dabei folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,514):

```
reg logprice lognox rooms
```

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lognox	-.7176732	.0663397	-10.82	0.000	-.8480102	-.5873361
rooms	.3059183	.0190174	16.09	0.000	.268555	.3432816
_cons	9.233737	.1877406	49.18	0.000	8.864885	9.602589

Damit ergibt sich:

- Eine Erhöhung der Stickoxide in der Luft um 1% (d.h. %Δnox = 1), führt zu einer approximativ geschätzten Verminderung des Medians der Immobilienpreise um 0,718% (falls rooms konstant gehalten wird)
- Eine Erhöhung der durchschnittlichen Anzahl an Räumen in Häusern um eins (d.h. Δrooms = 1) führt zu einer approximativen geschätzten Erhöhung des Medians der Immobilienpreise um 0,306·100 = 30,6% (falls nox konstant gehalten wird)

Diskussion der Verwendung von logarithmierten Variablen:

- Logarithmierte Variablen erlauben häufig eine wünschenswerte Interpretation der geschätzten Regressionsparameter. Zudem können Probleme bei der unterschiedlichen Skalierung von Daten vermieden werden.
- Falls die abhängige Variable y ausschließlich positive Werte annimmt, werden bei der Logarithmierung logy häufig besser die klassischen linearen Modellannahmen approximiert (in diesem Fall können auch eher Probleme der Heteroskedastizität abgeschwächt werden)
- Mit logarithmierten Variablen können die Spannweite von Variablen vermindert und negative Auswirkungen von Ausreißerwerten eingedämmt werden
- Geldbeträge und Bevölkerungszahlen werden oft logarithmiert in linearen Regressionsmodellen untersucht. In Jahren gemessene Variablen (z.B. Bildungsjahre, Alter) erscheinen dagegen oft in ursprünglicher Form.
- Bei der Interpretation des Effektes von logarithmierten oder nicht-logarithmierten prozentualen Variablen (z.B. Arbeitslosenquote) muss sorgfältig zwischen prozentualen Veränderungen und Prozentpunktveränderungen unterschieden werden
- Logarithmierungen können allerdings nicht bei Variablen vorgenommen werden, die negative Werte annehmen. Zudem ist die Prognose der ursprünglichen abhängigen Variable schwieriger, falls diese logarithmiert eingeht.

Zur Einbeziehung quadrierter erklärender Variablen:

Damit können wachsende oder sinkende (partielle) marginale Effekte in linearen Regressionsmodellen untersucht werden

Zur Erinnerung:

Falls y auf x_j regressiert wird, gibt $\hat{\beta}_j$ entsprechend (2.18) die Veränderung des OLS-Regressionswertes $\hat{y}$ an, falls x_j um eine Einheit steigt (und alle anderen erklärenden Variablen konstant gehalten werden). Damit ist hier der (partielle) marginale Effekt konstant und hängt nicht von x_j ab.

Zusätzliche Einbeziehung einer quadrierten erklärenden Variablen x_1^2 (neben den k-1 erklärenden Variablen x_1, x_2,…, x_{k-1}):

$$(4.1) \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_{k-1} x_{k-2} + \beta_k x_{k-1} + u$$

In diesem Fall beschreibt β_1 nicht alleine die Veränderung von y bei einer Veränderung von x_1. Die OLS-Regressionsfunktion lautet:

$$(4.2) \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3 x_2 + \cdots + \hat{\beta}_{k-1} x_{k-2} + \hat{\beta}_k x_{k-1}$$

Falls x_2,…, x_{k-1} konstant gehalten werden, folgt daraus die Approximation:

$$(4.3) \quad \Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x_1)\Delta x_1 \quad \text{bzw.} \quad \frac{\Delta\hat{y}}{\Delta x_1} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x_1$$

Damit hängt der geschätzte (partielle) marginale Effekt von x_1 auf y auch von $\hat{\beta}_2$ sowie den Werten von x_1 ab. Häufig ist $\hat{\beta}_1$ positiv und $\hat{\beta}_2$ negativ.

Beispiel: Erklärung von Löhnen

Mit Hilfe eines linearen Regressionsmodells soll der Effekt der Berufserfahrung in Jahren (exper) und der quadrierten Berufserfahrung in Jahren (expersq) auf den Stundenlohn untersucht werden. Dabei haben sich mit STATA für n = 526 Personen folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,093):

```
reg wage exper expersq
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.2981001	.0409655	7.28	0.000	.2176229	.3785773
expersq	-.0061299	.0009025	-6.79	0.000	-.0079029	-.0043569
_cons	3.725406	.3459392	10.77	0.000	3.045805	4.405007

Damit ergibt sich ein geschätzter sinkender positiver Effekt von exper:

- Das erste zusätzliche Jahr an Berufserfahrung (von exper = 0) führt zu einer approximativ geschätzten Erhöhung des Stundenlohnes um 0,298
- Eine Steigerung der Berufserfahrung von ein auf zwei Jahre ergibt folgende approximativ geschätzte Erhöhung des Stundenlohnes:
  0,298 - 2·0,0061·1 = 0,286
- Eine Erhöhung der Berufserfahrung von zehn auf elf Jahre führt zu folgender approximativ geschätzten Steigerung des Stundenlohnes:
  0,298 - 2·0,0061·10 = 0,176

Falls in (4.2) $\hat{\beta}_1$ positiv und $\hat{\beta}_2$ negativ sind, ergibt sich immer ein positiver Wert von x_1, bei dem der geschätzte Effekt von x_1 auf y null ist. Vor diesem Punkt hat x_1 einen (mit x_1 sinkenden) positiven und nach diesem Punkt einen negativen geschätzten marginalen Effekt. Für diesen Wendepunkt gilt:

$$(4.4) \quad x_1^* = \left| \frac{\hat{\beta}_1}{2\hat{\beta}_2} \right|$$

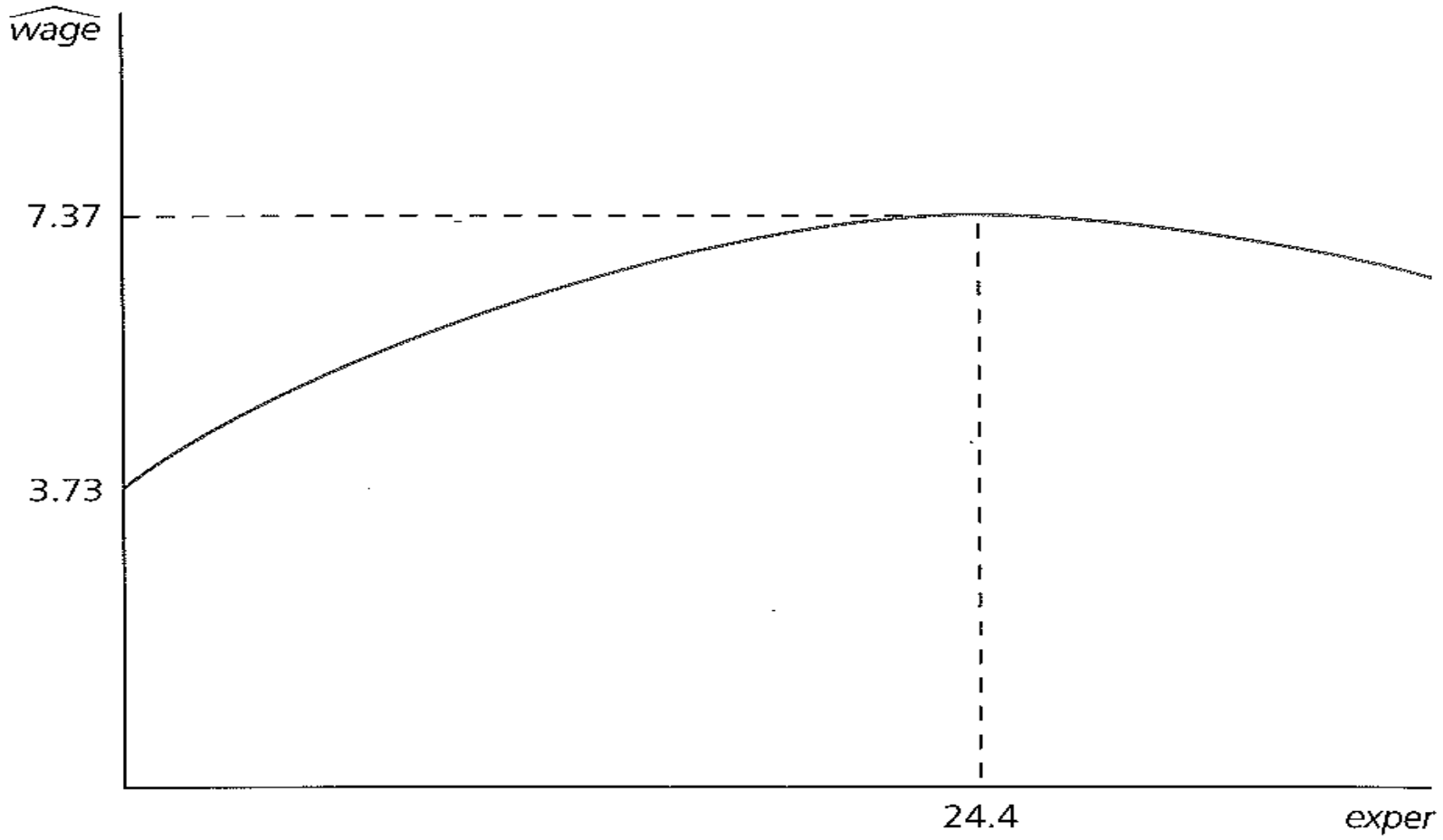
Beispiel: Erklärung von Löhnen

Im obigen Beispiel ergibt sich für den Wendepunkt des geschätzten Effektes von exper mit $\hat{\beta}_{exper} = 0{,}298$ und $\hat{\beta}_{expersq} = -0{,}0061$: $|0{,}298/[2(-0{,}0061)]| = 24{,}43$

Interpretation:

- Bei einer Berufserfahrung von mehr als 24 Jahren ergibt sich zunächst ein unerwarteter negativer geschätzter Effekt durch steigende Berufserfahrung
- Falls nur wenige Personen in der Stichprobe eine solch hohe Berufserfahrung besitzen würden, kann dieses Ergebnis ignoriert werden (da dann lediglich der positive geschätzte Effekt eine Rolle spielen würde)
- Allerdings haben 28% der Personen eine derart hohe Berufserfahrung. Eine Erklärung wären verzerrte Schätzungen, da wichtige Faktoren nicht einbezogen werden, oder aber tatsächliche negative Effekte bei hohem exper.

Beispiel: Erklärung von Löhnen (Fortsetzung)



Falls in (4.2) $\hat{\beta}_1$ negativ und $\hat{\beta}_2$ positiv sind, ergibt sich eine U-Form. Damit folgt nach dem Wendepunkt, dass x_1 einen (mit x_1 steigenden) positiven geschätzten marginalen Effekt hat. Die Einbeziehung von quadrierten und logarithmierten Variablen muss sorgfältig interpretiert werden.

Beispiel: Effekt von Luftverschmutzung auf Immobilienpreise

Mit einem linearen Regressionsmodell wird nun für n = 506 Gemeinden der Effekt des Logarithmus der Stickoxide in der Luft (lognox), des Logarithmus der gewichteten Entfernung zu fünf Beschäftigungszentren (logdist), der einfachen und quadrierten durchschnittlichen Anzahl an Räumen in Häusern (rooms, roomssq) und des Verhältnisses von Lehrern und Schülern in den Schulen (stratio) auf den Logarithmus des Medians der Immobilienpreise (logprice) untersucht. Es haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt:

```
reg logprice lognox logdist rooms roomssq stratio
```

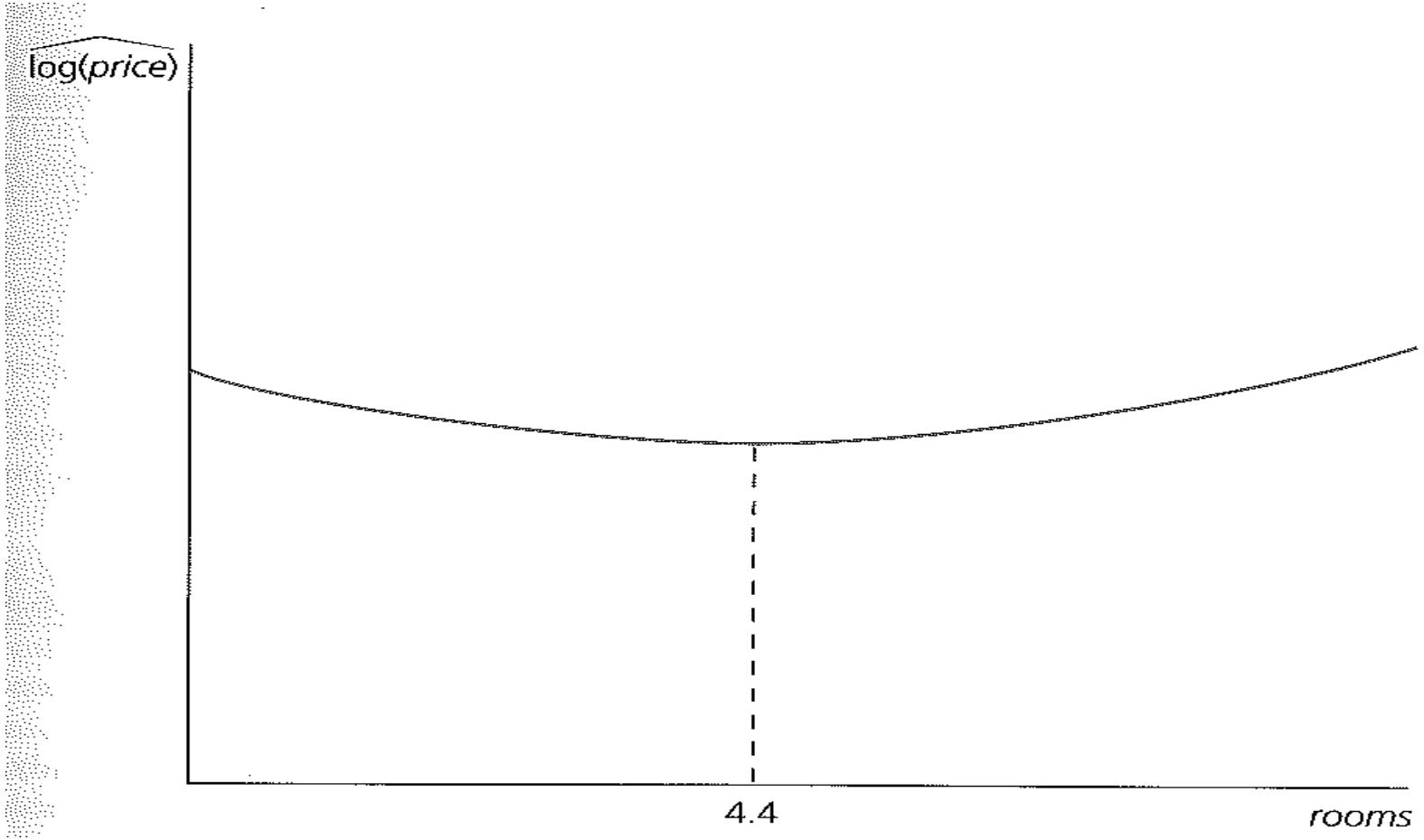
logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lognox	-.9016829	.114687	-7.86	0.000	-1.127011	-.676355
logdist	-.086782	.0432808	-2.01	0.045	-.1718165	-.0017474
rooms	-.5451124	.1654542	-3.29	0.001	-.8701835	-.2200412
roomssq	.0622612	.012805	4.86	0.000	.0371029	.0874194
stratio	-.0475902	.0058542	-8.13	0.000	-.0590921	-.0360884
_cons	13.38548	.5664733	23.63	0.000	12.27252	14.49844

Beispiel: Effekt von Luftverschmutzung auf Immobilienpreise (Fortsetzung)

Damit ergibt sich:

- Die geschätzten Parameter für rooms und roomssq sind statistisch stark signifikant von null verschieden
- Für den Wendepunkt des geschätzten Effektes von rooms ergibt sich mit $\hat{\beta}_{rooms} = -0{,}545$ und $\hat{\beta}_{roomssq} = 0{,}062$: $|-0{,}545/(2 \cdot 0{,}062)| = 4{,}40$. Da die geschätzten Parameter für rooms negativ und für roomssq positiv sind, ergibt sich das unplausible Ergebnis, dass bei kleinen Werten von rooms (d.h. kleiner als 4,4) diese Variable einen geschätzten negativen Effekt hat.
- Allerdings zeigt sich, dass in der Stichprobe nur bei fünf der 506 Gemeinden die durchschnittlichen Anzahl an Räumen in Häusern 4,4 oder kleiner ist. Aufgrund dieser kleinen Anzahl können die Werte von rooms vor dem Wendepunkt praktisch ignoriert werden.
- Wenn rooms > 4,4, ergeben sich mit steigenden Werten wachsende positive geschätzte marginale Effekte auf die prozentuale Veränderung von price:
  $\Delta \widehat{\log price} \approx [-0{,}545+2 \cdot (0{,}062)rooms]\Delta rooms$
  $\%\Delta \widehat{price} \approx 100 \cdot [-0{,}545+2 \cdot (0{,}062)rooms]\Delta rooms$
  Somit ergibt sich z.B. bei einer Erhöhung der durchschnittlichen Anzahl an Räumen in Häusern von fünf auf sechs eine approximative geschätzte prozentuale Erhöhung von price um $100[-0{,}545 + 0{,}124 \cdot 5] = 7{,}5\%$.

Beispiel: Effekt von Luftverschmutzung auf Immobilienpreise (Fortsetzung)



Interaktionsterme:

Diese Variablen erlauben, dass der (partielle) Effekt (bzw. die Elastizität oder Semi-Elastizität) einer erklärenden Variablen in linearen Regressionsmodellen von verschieden Werten einer anderen erklärenden Variablen abhängen kann

Zur Erinnerung:

Falls y auf x_j regressiert wird, gibt $\hat{\beta}_j$ entsprechend (2.18) die Veränderung des OLS-Regressionswertes $\hat{y}$ an, falls x_j um eine Einheit steigt (ceteris paribus)

Zusätzliche Einbeziehung eines Interaktionsterms für x_1 und x_2 (neben den k-1 erklärenden Variablen x_1, x_2,…, x_{k-1}):

$$(4.5) \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_3 + \cdots + \beta_{k-1} x_{k-2} + \beta_k x_{k-1} + u$$

Auch in diesem Fall beschreibt β_1 nicht alleine die Veränderung von y bei einer Veränderung von x_1. Die OLS-Regressionsfunktion lautet:

$$(4.6) \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_3 + \cdots + \hat{\beta}_{k-1} x_{k-2} + \hat{\beta}_k x_{k-1}$$

Falls x_2,…, x_{k-1} konstant gehalten werden, folgt daraus:

$$(4.7) \quad \Delta\hat{y} = (\hat{\beta}_1 + \hat{\beta}_3 x_2)\Delta x_1 \quad \text{bzw.} \quad \frac{\Delta\hat{y}}{\Delta x_1} = \hat{\beta}_1 + \hat{\beta}_3 x_2$$

Damit hängt der geschätzte (partielle) Effekt von x_1 auf y auch von $\hat{\beta}_3$ sowie von x_2 ab. Dabei sollten interessante Werte von x_2 betrachtet werden (z.B. arithmetisches Mittel in der Stichprobe). $\hat{\beta}_1$ alleine bildet lediglich den geschätzten (partiellen) Effekt von x_1 ab, wenn x_2 null ist.

Beispiel: Erklärung von Ergebnissen bei der Abschlussprüfung

Mit einem linearen Regressionsmodell soll für n = 680 Studierende der Effekt der relativen Häufigkeit des Besuchs einer Lehrveranstaltung in % (atndrte), der vorherigen College-Punktzahl (priGPA), der Punktzahl im College-Aufnahmetest (ACT), der quadrierten vorherigen College-Punktzahl (priGPAsq), der quadrierten Punktzahl im College-Aufnahmetest (ACTsq) sowie der Interaktion von priGPA und atndrte (priGPAatndrte) auf das standardisierte Ergebnis (im Vergleich zu den anderen Studierenden) bei der Abschlussprüfung (stndfnl) untersucht werden. Die Idee des Ansatzes ist, dass der Effekt von atndrte von unterschiedlichen Leistungen auf dem College abhängen könnte. Es haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,229):

```
reg stndfnl atndrte priGPA ACT priGPAsq ACTsq priGPAatndrte
------------------------------------------------------------------------------
      stndfnl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
      atndrte |  -.0067129   .0102321    -0.66   0.512    -.0268036    .0133777
       priGPA |  -1.628539   .4810025    -3.39   0.001    -2.572985   -.6840933
          ACT |  -.1280394    .098492    -1.30   0.194    -.3214279    .0653492
     priGPAsq |   .2959044   .1010495     2.93   0.004     .0974943    .4943146
        ACTsq |   .0045334   .0021764     2.08   0.038       .00026    .0088068
priGPAatndrte |   .0055859   .0043174     1.29   0.196    -.0028913    .0140631
        _cons |   2.050293   1.360319     1.51   0.132    -.6206865    4.721272
------------------------------------------------------------------------------
```

Beispiel: Erklärung von Ergebnissen bei der Abschlussprüfung (Fortsetzung)

Interpretation:

- Bei alleiniger Betrachtung des geschätzten Parameters von atndrte würde sich ein (insignifikanter) negativer Effekt ergeben. Allerdings wird hier der Effekt für priGPA = 0 gemessen, der irrelevant ist, da der minimale Wert in der Stichprobe priGPA = 0,86 beträgt.
- Bei einzelner Betrachtung der t-Statistiken für atndrte und des Interaktionsterms ergibt sich, dass beide Parameter nicht signifikant von null verschieden sind. Entscheidend ist jedoch, dass der p-Wert bei einem F-Test zur Überprüfung, dass beide Parameter gemeinsam null sind, 0,014 beträgt, so dass die entsprechende Nullhypothese z.B. schon bei einem Signifikanzniveau von 2% verworfen werden kann.
- Ein interessanter Wert zur Betrachtung des Effektes von atndrte liegt beim arithmetischen Mittel in der Stichprobe priGPA = 2,59 vor. Hier beträgt der geschätzte Effekt von atndrte: -0,0067 + 0,0056·2,59 = 0,0078.
- Die Frage, ob dieser Schätzwert 0,0078 signifikant von null verschieden ist, kann durch die Schätzung eines reparametrisierten linearen Regressionsmodells ermittelt werden, wobei priGPAatndrte durch (priGPA-2,59)·atndrte ersetzt wird. Der entsprechende t-Wert beträgt 3, so dass atndrte bei durchschnittlichem priGPA einen hochsignifikanten Effekt hat.

4.3 Zur Einbeziehung von erklärenden Variablen

→ In Kapitel 2.3 wurde der „omitted variable bias“ diskutiert, d.h. Verzerrungen bei der Vernachlässigung relevanter erklärender Variablen, wenn diese mit anderen erklärenden Variablen korreliert sind

Einbeziehung von zu vielen Kontrollvariablen („over controlling“):

- Vor dem Hintergrund des „omitted variable bias“ werden in empirischen Untersuchungen häufig alle verfügbaren und damit manchmal zu viele erklärende Variablen in lineare Regressionsmodelle einbezogen
- Dabei wird insbesondere die Erhöhung des Wertes des Bestimmtheitsmaßes R^2 überbetont
- Allerdings sollte grundsätzlich die ceteris paribus Betrachtung von Regressionsanlysen bedacht werden: In manchen Fällen macht es bei der Untersuchung des Effektes einer Variablen keinen Sinn, eine andere Kontrollvariable konstant zu halten (und deshalb in das lineare Regressionsmodell einzubeziehen), da genau die Variation dieser Kontrollvariablen durch die interessierende Variable relevant ist.
- Allerdings ist die Überlegung, ob eine (verfügbare) Kontrollvariable einbezogen werden soll, nicht immer eindeutig. Entscheidend ist dabei der Fokus der jeweiligen empirischen Untersuchung.

Beispiel: Erklärung der Anzahl von Verkehrstoten

Mit Hilfe eines linearen Regressionsmodells soll für verschiedene Bundesstaaten insbesondere der Effekt der Höhe der Biersteuer auf die Anzahl der Verkehrstoten untersucht werden:

- Die Überlegung dabei ist, dass eine höhere Steuer die Trunkenheit am Steuer und damit die Anzahl der Verkehrstoten senkt. Weitere denkbare erklärende Variablen sind die durchschnittlich gefahrenen Meilen, der Anteil der männlichen Bevölkerung, der Anteil der jungen Bevölkerung usw.
- Die Frage ist nun, ob auch der durchschnittliche pro-Kopf-Bierkonsum als Kontrollvariable einbezogen werden sollte. Diese Variable hätte sicherlich einen hohen Erklärungsgehalt für die Anzahl der Verkehrstoten, würde sicherlich den Wert des Bestimmtheitsmaßes R^2 erhöhen und ist sicherlich stark mit der Höhe der Biersteuer korreliert.
- Allerdings kann der Mechanismus des Effektes einer höheren Biersteuer auf die Anzahl der Verkehrstoten über einen geringeren Bierkonsum bei einer Einbeziehung des Bierkonsums als Kontrollvariable nicht mehr greifen. Diese Variable würde dann lediglich den eher uninteressanten indirekten Effekt bei konstantem Bierkonsum messen. Daher sollte die Bierkonsum-Variable hier nicht in die Untersuchung einbezogen werden.

Weitere Beispiele:

- Bei der Untersuchung des Effektes des Einsatzes von Pestiziden in landwirtschaftlichen Betrieben in Entwicklungsländern auf die Gesundheitsausgaben der bewirtschaftenden Landwirte ist es nicht sinnvoll, die Anzahl der Arztbesuche (die Bestandteil der Gesundheitsausgaben sind) als Kontrollvariable einzubeziehen
- Bei der Untersuchung des Effektes einzelner Häuserattribute wie Grundstücksgröße, Wohnflächengröße oder Anzahl an Schlafzimmern auf Häuserpreise macht es keinen Sinn, eine Beurteilung des Hauswertes vor dem Verkauf als Kontrollvariable einzubeziehen, selbst wenn hierdurch der Wert des Bestimmtheitsmaßes R^2 stark erhöht werden könnte

---

Zusätzliche Kontrollvariablen zur präziseren Parameterschätzung:

Zwar kann die Einbeziehung von Kontrollvariablen das Multikollinearitätsproblem verschärfen. Jedoch kann durch die Einbeziehung von zusätzlichen erklärenden Variablen die Varianz σ^2 des Fehlerterms u reduziert werden. Deshalb sollten relevante erklärende Variablen, die mit anderen erklärenden Variablen unkorreliert sind, immer einbezogen werden, weil dadurch die geschätzten Varianzen der geschätzten Parameter reduziert werden können.

4.4. Qualitative erklärende Variablen

→ Bisher wurde implizit auf quantitative (d.h. metrisch skalierte) abhängige und erklärende Variablen in linearen Regressionsmodellen fokussiert wie z.B. Löhne, Preise, Emissionen, Ausbildungszeit, Umsätze, Forschungs- und Entwicklungsausgaben. In empirischen Untersuchungen spielen aber häufig auch qualitative Faktoren eine wichtige Rolle wie z.B. Geschlecht, Hautfarbe, Besitz eines Produkts, Branchenzugehörigkeit, regionale Effekte usw.

Qualitative Variablen:

- Qualitative Informationen bei erklärenden Variablen können durch entsprechende binäre oder Dummy-Variablen eingefangen werden, die entweder den Wert null oder den Wert eins annehmen
- Dabei ist es für die Regressionsanalyse unerheblich, welche Ausprägung einer Dummy-Variablen (z.B. Geschlecht) den Wert null oder den Wert eins annimmt, wenngleich eine Festlegung getroffen werden muss
- Die Zuordnung der Zahlen null und eins bei binären Variablen ist letztlich willkürlich, führt aber zu linearen Regressionsmodellen, bei denen die entsprechenden Parameter sinnvoll zu interpretieren sind
- Die OLS-Schätzung und das Testen von Hypothesen erfolgt bei der Regressionsanalyse mit qualitativen erklärenden Variablen völlig analog zur ausschließlichen Einbeziehung von quantitativen Variablen

Einzelne binäre erklärende Variablen:

Einbeziehung von qualitativen Variablen mit zwei Ausprägungen

Ausgangspunkt ist zunächst ein multiples lineares Regressionsmodell entsprechend (2.9) mit ausschließlich quantitativen erklärenden Variablen. Nun wird (neben den jetzt k-1 quantitativen erklärenden Variablen $x_1, x_2,\ldots, x_{k-1}$) zusätzlich eine binäre erklärende Variable x_0 einbezogen, die die Werte null oder eins annimmt:

$$(4.8) \quad y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \cdots + \beta_k x_{k-1} + u$$

Mit $E(u|x_0,x_1,x_2,\ldots, x_{k-1}) = 0$ gilt:

$$E(y|x_0, x_1, x_2,\ldots, x_{k-1}) = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \cdots + \beta_k x_{k-1}$$

Daraus folgt:

$$(4.9) \quad \beta_1 = E(y|x_0 = 1, x_1, x_2,\ldots, x_{k-1}) - E(y|x_0 = 0, x_1, x_2,\ldots, x_{k-1})$$

β_1 ist also die Differenz im Erwartungswert von y zwischen $x_0 = 1$ und $x_0 = 0$, gegeben die gleichen Werte von $x_1, x_2,\ldots, x_{k-1}$ und u.

→ β_0 ist somit lediglich die Konstante für $x_0 = 0$. Für $x_0 = 1$ beträgt die Konstante $\beta_0 + \beta_1$, so dass β_1 die Differenz der Konstanten für $x_0 = 1$ und $x_0 = 0$ darstellt.

Beispiel: Erklärung von Löhnen

Mit Hilfe eines linearen Regressionsmodells soll der Effekt der Ausbildungszeit (educ) und des Geschlechts (female) auf den Stundenlohn (wage) untersucht werden. Die binäre Variable female nimmt den Wert eins an, wenn die Person weiblich ist und null (als Basisgruppe), wenn sie nicht weiblich (male) ist:

$$\text{wage} = \beta_0 + \beta_1 \text{female} + \beta_2 \text{educ} + u$$

Mit E(u|female, educ) = 0 ergibt sich:

$$\beta_1 = E(\text{wage}|\text{female} = 1, \text{educ}) - E(\text{wage}|\text{female} = 0, \text{educ}) \quad \text{bzw.}$$

$$\beta_1 = E(\text{wage}|\text{female}, \text{educ}) - E(\text{wage}|\text{male}, \text{educ})$$

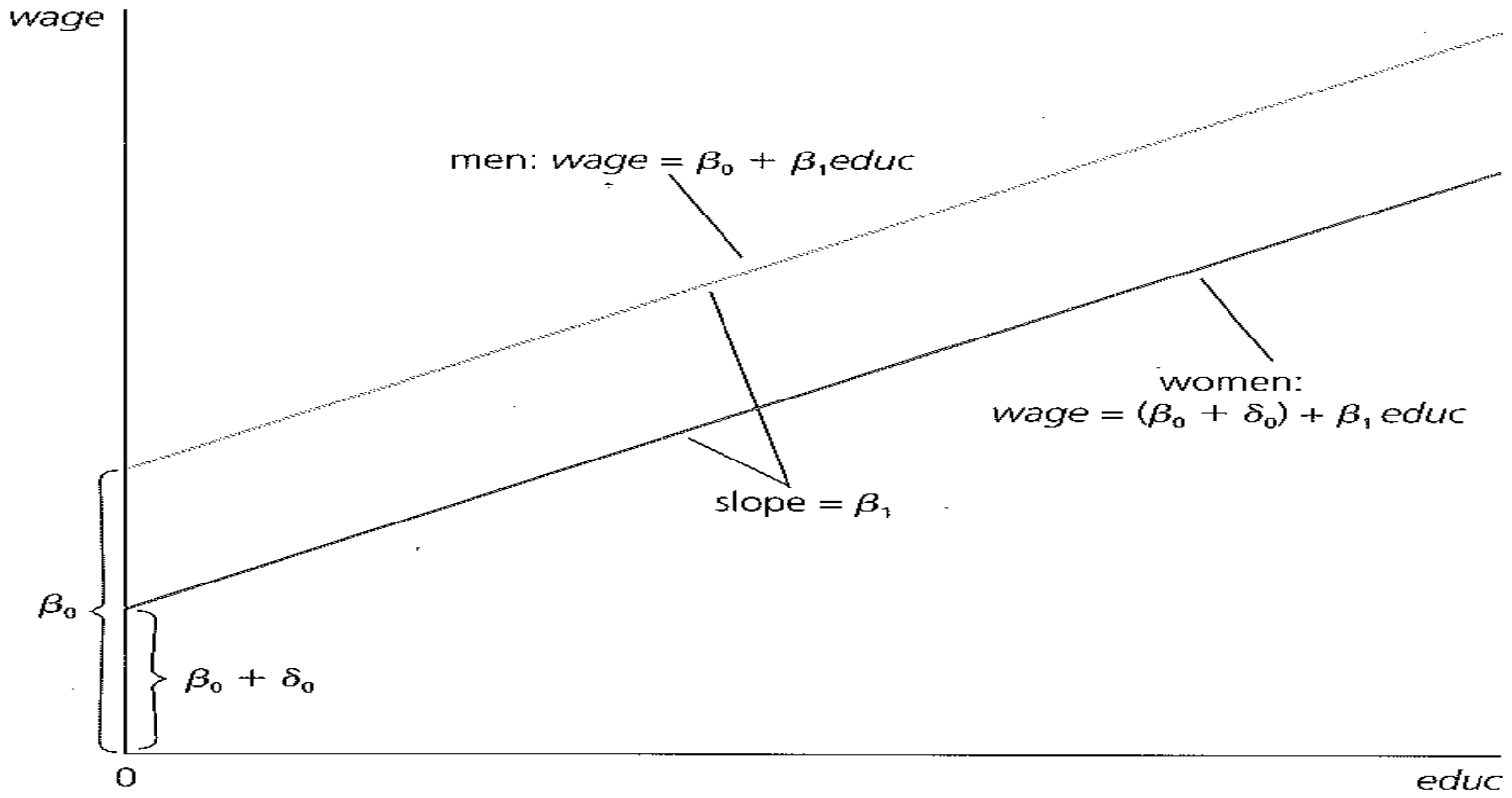
Interpretation:

- Mit $\beta_1 < 0$ haben Frauen bei gleicher Ausbildungszeit einen geringeren durchschnittlichen Stundenlohn als Männer, mit $\beta_1 > 0$ dagegen einen höheren durchschnittlichen Stundenlohn
- Der Parameter β_1 kann somit bei negativen Werten eine mögliche Diskriminierung von Frauen bei der Entlohnung offenbaren
- In diesem Regressionsmodell hängt ein möglicher Unterschied bei der Entlohnung zwischen Männern und Frauen nicht von der Ausbildungszeit ab

Beispiel: Erklärung von Löhnen (Fortsetzung)

Erwartungswerte von wage bei Frauen und Männern für $\beta_1 < 0$ (im Lehrbuch von Wooldridge $\delta_0 < 0$):

Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$.



Achtung:

Es dürfen für einen Faktor mit zwei Ausprägungen (z.B. Geschlecht) niemals zwei Dummy-Variablen (z.B. eine Variable, die den Wert eins annimmt für Frauen und eine Variable, die den Wert eins annimmt für Männer) gleichzeitig in ein lineares Regressionsmodell einbezogen werden, da dadurch eine perfekte Kollinearität vorliegen würde (einfache Form der „dummy variable trap“)

Beispiel 1: Erklärung von Löhnen

Mit Hilfe eines linearen Regressionsmodells wird der Effekt des Geschlechts (female), der Ausbildungszeit in Jahren (educ), der Berufserfahrung in Jahren (exper) und der Betriebszugehörigkeit in Jahren (tenure) auf den Stundenlohn (wage) untersucht. Dabei haben sich für n = 526 Personen mit STATA folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,364):

```
reg wage female educ exper tenure
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-1.810852	.2648252	-6.84	0.000	-2.331109	-1.290596
educ	.5715048	.0493373	11.58	0.000	.4745803	.6684293
exper	.0253959	.0115694	2.20	0.029	.0026674	.0481243
tenure	.1410051	.0211617	6.66	0.000	.0994323	.1825778
_cons	-1.567939	.7245511	-2.16	0.031	-2.99134	-.144538

Beispiel 1: Erklärung von Löhnen (Fortsetzung)

Interpretation:

- Aufgrund der t-Statistik von -6,84 bei female ist der durchschnittliche Stundenlohn von Frauen hochsignifikant geringer als bei Männern
- Die negative Konstante von -1,57, d.h. die Konstante für Männer, ist nicht sehr bedeutsam, da in der Stichprobe bei keiner Beobachtung bei educ, exper und tenure gleichzeitig der Wert null vorliegt
- Die OLS-Regressionsfunktionen für Frauen und Männer lauten damit:

$$\text{Frauen: } \hat{w}age = -1{,}57 - 1{,}81 + 0{,}572educ + 0{,}025exper + 0{,}141tenure$$

$$= -3{,}38 + 0{,}572educ + 0{,}025exper + 0{,}141tenure$$

$$\text{Männer: } \hat{w}age = -1{,}57 + 0{,}572educ + 0{,}025exper + 0{,}141tenure$$

Der geschätzte Stundenlohn ist somit bei Frauen (bei gleichen educ, exper und tenure) im Durchschnitt um 1,81 geringer.

- Diese Lohndifferenz kann somit nicht durch Unterschiede bei educ, exper und tenure erklärt werden, sondern bildet Geschlechtsunterschiede ab. Bei einer Regression von wage ausschließlich auf female ergibt sich ein geschätzter Parameter von -2,51. Dieser größere negative Wert zeigt sich, da nicht für Unterschiede in educ, exper und tenure kontrolliert wird.

Beispiel 2: Erklärung von Weiterbildungsmaßnahmen

Mit einem linearen Regressionsmodell wird für n = 105 Unternehmen der Effekt einer Unterstützungszahlung für Weiterbildungsmaßnahmen (grant), des Logarithmus der Umsätze (logsales) und des Logarithmus der Anzahl der Beschäftigten (logemploy) auf die Trainingsstunden pro Beschäftigten (hrsemp) untersucht. Grant nimmt den Wert eins an, wenn das Unternehmen öffentliche Unterstützungszahlungen erhalten hat. Dabei haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,237):

```
reg hrsemp grant logsales logemploy
```

hrsemp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grant	26.2545	5.591766	4.70	0.000	15.16194	37.34706
logsales	-.9845776	3.539904	-0.28	0.781	-8.006795	6.03764
logemploy	-6.069873	3.882894	-1.56	0.121	-13.77249	1.632744
_cons	46.66504	43.41211	1.07	0.285	-39.45291	132.783

Grant hat einen (zuvor erwarteten) hochsignifikant positiven Effekt auf hrsemp. Die geschätzten Trainingsstunden pro Beschäftigten sind damit (bei gleichen Umsätzen und Beschäftigungszahlen) bei Unternehmen, die Unterstützungszahlungen erhalten haben, im Durchschnitt 26,25 Stunden höher als bei Unternehmen ohne Unterstützungszahlungen.

Beispiel 3: Erklärung (des Logarithmus) von Löhnen

Mit Hilfe eines linearen Regressionsmodells wird für n = 526 Personen der Effekt des Geschlechts (female), der Ausbildungszeit in Jahren (educ), der Berufserfahrung in Jahren (exper), der quadrierten Berufserfahrung in Jahren (expersq), der Betriebszugehörigkeit in Jahren (tenure) und der quadrierten Betriebszugehörigkeit in Jahren (tenuresq) auf den Logarithmus des Stundenlohns (logwage) untersucht. Dabei haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,441):

```
reg logwage female educ exper expersq tenure tenuresq
```

logwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.296511	.0358055	-8.28	0.000	-.3668524	-.2261696
educ	.0801967	.0067573	11.87	0.000	.0669217	.0934716
exper	.0294324	.0049752	5.92	0.000	.0196584	.0392063
expersq	-.0005827	.0001073	-5.43	0.000	-.0007935	-.0003719
tenure	.0317139	.0068452	4.63	0.000	.0182663	.0451616
tenuresq	-.0005852	.0002347	-2.49	0.013	-.0010463	-.0001241
_cons	.4166909	.0989279	4.21	0.000	.2223425	.6110394

Damit ergibt sich, dass der geschätzte Stundenlohn bei Frauen (bei gleicher Ausbildungszeit, gleicher Berufserfahrung und gleicher Betriebszugehörigkeit) im Durchschnitt approximativ 100·0,297 = 29,7% geringer ist.

Binäre erklärende Variablen für multiple Kategorien:

Einbeziehung von qualitativen Variablen mit mehr als zwei Ausprägungen

Ausgangspunkt ist zunächst wieder ein multiples lineares Regressionsmodell entsprechend (2.9) mit ausschließlich quantitativen erklärenden Variablen. Nun wird zusätzlich eine qualitative (nominale oder ordinale) erklärende Variable (z.B. Branchen- oder regionale Zugehörigkeit) mit q verschiedenen Ausprägungen betrachtet, wobei im Gegensatz zu (4.8) $q > 2$. Für diesen Fall können (maximal) q-1 Dummy-Variablen $x_{01}, x_{02},\ldots, x_{0,q-1}$ (neben den jetzt k-q+1 quantitativen erklärenden Variablen $x_1, x_2,\ldots, x_{k-q+1}$) einbezogen werden:

$$(4.10)\quad y = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_{q-1} x_{0,q-1} + \beta_q x_1 + \beta_{q+1} x_2 + \cdots + \beta_k x_{k-q+1} + u$$

Die q-te Ausprägung der qualitativen Variablen (d.h. die Dummy-Variable x_{0q}) dient dabei als Basisgruppe. Das heißt, die geschätzten Regressionsparameter $\hat{\beta}_1, \hat{\beta}_2,\ldots, \hat{\beta}_{q-1}$ zeigen für die jeweilige Gruppe der qualitativen Variablen (d.h. für $x_{01}, x_{02},\ldots, x_{0,q-1}$) die geschätzte durchschnittliche Differenz in der abhängigen Variable y im Vergleich zur Basisgruppe, d.h. im Vergleich zu x_{0q}.

Achtung:

Es dürfen niemals alle q Dummy-Variablen $x_{01}, x_{02},\ldots, x_{0q}$ gleichzeitig einbezogen werden, da dadurch eine perfekte Kollinearität vorliegen würde (generelle Form der „dummy variable trap“). Viele ökonometrische Programmpakete wie z.B. STATA korrigieren aber einen solchen Fehler automatisch.

Beispiel 1: Erklärung von Einstiegsgehältern

Mit einem linearen Regressionsmodell wird für n = 136 juristische Fakultäten der Effekt eines Rankings der Fakultät (sowie anderer Faktoren) auf den Logarithmus des Medians des Einstiegsgehaltes (logsalary) für die Fakultäten untersucht. Bei der Rankingvariable werden sechs Kategorien einbezogen, d.h. top10, r11_25, r26_40, r41_60, r61_100 sowie Rankings jenseits von 100 als Basisgruppe. Dabei haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,911):

```
reg logsalary top10 r11_25 r26_40 r41_60 r61_100 ...
------------------------------------------------------------------------------
   logsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       top10 |   .6995647   .0534919    13.08   0.000     .5937058    .8054236
      r11_25 |   .5935445     .03944    15.05   0.000     .5154938    .6715951
      r26_40 |   .3750779   .0340812    11.01   0.000     .3076322    .4425236
      r41_60 |     .26282    .027962     9.40   0.000      .207484     .318156
     r61_100 |   .1315945   .0210418     6.25   0.000     .0899534    .1732357
        ⋮    |       ⋮          ⋮          ⋮      ⋮           ⋮           ⋮
------------------------------------------------------------------------------
```

Die geschätzten Regressionsparameter der Rankingvariablen sind hochsignifikant von null verschieden. Damit ergibt sich z.B., dass der geschätzte Median des Einstiegsgehaltes bei einem Ranking zwischen 61 und 100 durchschnittlich approximativ 13,2% höher ist im Vergleich zum Ranking jenseits von 100.

Beispiel 2: Erklärung (des Logarithmus) von Löhnen

Mit einem linearen Regressionsmodell wird für n = 526 Personen der Effekt der Ausbildungszeit in Jahren (educ), der einfachen und quadrierten Berufserfahrung in Jahren (exper, expersq), der einfachen und quadrierten Betriebszugehörigkeit in Jahren (tenure, tenuresq) auf den Logarithmus des Stundenlohns (logwage) untersucht. Neben diesen erklärenden Variablen wird zusätzlich auch eine kombinierte Familienstands- und Geschlechtsvariable betrachtet mit den Ausprägungen verheiratete Männer (marrmale), verheiratete Frauen (marrfem), unverheiratete Männer (singmale) und unverheiratete Frauen (singfem). Als Basisgruppe werden unverheiratete Männer betrachtet. Dabei haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt (R^2 = 0,461):

```
reg logwage marrmale marrfem singfem educ exper expersq tenure tenuresq
```

logwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marrmale	.2126756	.0553572	3.84	0.000	.103923	.3214283
marrfem	-.1982677	.0578355	-3.43	0.001	-.3118891	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.0920621
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenuresq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.321378	.100009	3.21	0.001	.1249041	.517852

Beispiel 2: Erklärung (des Logarithmus) von Löhnen (Fortsetzung)

Die OLS-Regressionsfunktion für unverheiratete Männer lautet somit:

$$\widehat{\log wage} = 0{,}321 + 0{,}079educ + 0{,}027exper - 0{,}00054expersq + 0{,}029tenure - 0{,}00053tenuresq$$

Für verheiratete Männer, verheiratete Frauen und nicht-verheiratete Frauen ergeben sich folgende OLS-Regressionsfunktionen:

Verheiratete Männer: $\widehat{\log wage} = 0{,}321 + 0{,}213 + 0{,}079educ + ... = 0{,}534 + 0{,}079educ + ...$

Verheiratete Frauen: $\widehat{\log wage} = 0{,}321 - 0{,}198 + 0{,}079educ + ... = 0{,}123 + 0{,}079educ + ...$

Unverheiratete Frauen: $\widehat{\log wage} = 0{,}321 - 0{,}110 + 0{,}079educ + ... = 0{,}211 + 0{,}079educ + ...$

Beispiel 2: Erklärung (des Logarithmus) von Löhnen (Fortsetzung)

Interpretation:

- Die geschätzten Regressionsparameter für marrmale, marrfem und singfem sind zu einem Signifikanzniveau von 5% von null verschieden
- Der geschätzte Stundenlohn ist somit bei verheirateten Männern bzw. verheirateten Frauen bzw. unverheirateten Frauen jeweils im Vergleich zu unverheirateten Männern (bei gleicher Ausbildungszeit, gleicher Berufserfahrung und gleicher Betriebszugehörigkeit) durchschnittlich approximativ um 21,3% höher bzw. 19,8% bzw. 11% geringer
- Mit den geschätzten Regressionsparametern können auch zwischen den Gruppen durchschnittliche Differenzen geschätzt werden: Zum Beispiel ist der geschätzte Stundenlohn bei verheirateten Männern im Vergleich zu verheirateten Frauen (bei gleicher Ausbildungszeit, gleicher Berufserfahrung und gleicher Betriebszugehörigkeit) durchschnittlich approximativ um 100[0,213-(-0,198)] = 41,1% höher.
- Zur Untersuchung, ob solche Differenzen signifikant von null verschieden sind, müssen neue OLS-Schätzungen mit neuen Basisgruppen durchgeführt werden (im vorherigen Fall müssen z.B. marrmale oder marrfem die Basisvariable darstellen)

Interaktionsterme mit binären erklärenden Variablen:

Interaktionsterme müssen sich nicht nur auf zwei quantitative erklärende Variablen beziehen, sondern können auch Dummy-Variablen einbeziehen

Zusätzliche Einbeziehung eines Interaktionsterms für zwei binäre erklärende Variablen x_{01} und x_{02} (neben den jetzt k-3 quantitativen erklärenden Variablen $x_1, x_2,\ldots, x_{k-3}$):

$$(4.11) \quad y = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_3 x_{01} x_{02} + \beta_4 x_1 + \beta_5 x_2 + \cdots + \beta_k x_{k-3} + u$$

Interpretation:

- Die Einbeziehung von Interaktionstermen für zwei binäre erklärende Variablen (neben der separaten Einbeziehung der Dummy-Variablen) ist eine Alternative zur Einbeziehung von drei binären erklärenden Variablen, wenn vier Kategorien untersucht werden sollen
- $\hat{\beta}_1$ (bzw. $\hat{\beta}_2$) zeigen für $x_{02} = 0$ (bzw. $x_{01} = 0$) die geschätzte durchschnittliche Differenz in der abhängigen Variable y zwischen $x_{01} = 1$ und $x_{01} = 0$ (bzw. zwischen $x_{02} = 1$ und $x_{02} = 0$)
- Für $x_{01} = 1$ und $x_{02} = 0$ (bzw. für $x_{01} = 0$ und $x_{02} = 1$) ergibt sich eine geschätzte Konstante von $\hat{\beta}_0 + \hat{\beta}_1$ (bzw. $\hat{\beta}_0 + \hat{\beta}_2$)
- Für $x_{01} = 1$ und $x_{02} = 1$ ergibt sich schließlich eine geschätzte Konstante von $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

Beispiel: Erklärung (des Logarithmus) von Löhnen

Mit Hilfe eines linearen Regressionsmodells wird der Effekt des Geschlechts (female), des Familienstands (married), der Interaktion von female und married (femmarried) (sowie der bereits zuvor betrachteten Faktoren) auf den Logarithmus des Stundenlohns (logwage) untersucht. Dabei haben sich mit STATA folgende OLS-Schätzergebnisse gezeigt (n = 526, R^2 = 0,461):

```
reg logwage female married femmarried educ ...
```

logwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
married	.2126756	.0553572	3.84	0.000	.103923	.3214283
femmarried	-.3005931	.071767	-4.19	0.000	-.4415838	-.1596024
⋮	⋮	⋮	⋮	⋮	⋮	⋮
_cons	.321378	.100009	3.21	0.001	.1249041	.517852

Interpretation:

- Die geschätzten Regressionsparameter für die anderen erklärenden Variablen sind zwingenderweise identisch mit den Werten zuvor
- Female = 0, married = 0 (bzw. female = 1, married = 1) korrespondieren mit unverheirateten Männern (bzw. verheirateten Frauen): Hier betragen die geschätzten Konstanten 0,321 (bzw. 0,321 – 0,110 + 0,213 - 0,301 = 0,123)

Zusätzliche Einbeziehung eines Interaktionsterms für eine binäre erklärende Variable x_0 und eine quantitative erklärende Variable x_1 (neben den jetzt k-2 quantitativen erklärenden Variablen x_1, x_2,…, x_{k-2}):

$$(4.12) \quad y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_0 x_1 + \beta_4 x_2 + \cdots + \beta_k x_{k-2} + u$$

Interpretation:

- Hier kann untersucht werden, inwieweit sich der (partielle) Effekt (bzw. die Elastizität oder Semi-Elastizität) der quantitativen erklärenden Variablen x_1 in linearen Regressionsmodellen bei den beiden Ausprägungen der binären erklärenden Variablen x_0 unterscheidet. Falls kein Unterschied vorliegt, gilt $\beta_3 = 0$.
- Falls $x_0 = 0$, gilt für die OLS-Regressionsfunktion:

  $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_1 + \hat{\beta}_4 x_2 + \cdots + \hat{\beta}_k x_{k-2}$$

  Die geschätzte Konstante lautet hier also $\hat{\beta}_0$ und der geschätzte (partielle) Effekt von x_1 beträgt $\hat{\beta}_2$.
- Falls $x_0 = 1$, gilt für die OLS-Regressionsfunktion:

  $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_1 + \hat{\beta}_4 x_2 + \cdots + \hat{\beta}_k x_{k-2}$$

  Die geschätzte Konstante lautet hier also $\hat{\beta}_0 + \hat{\beta}_1$ und der geschätzte (partielle) Effekt von x_1 beträgt $\hat{\beta}_2 + \hat{\beta}_3$.

Beispiel: Erklärung von Löhnen

Mit Hilfe eines linearen Regressionsmodells wird der Effekt des Geschlechts (female), der Ausbildungszeit in Jahren (educ) und der Interaktion von female und educ (femaleeduc) auf den Stundenlohn (wage) untersucht:

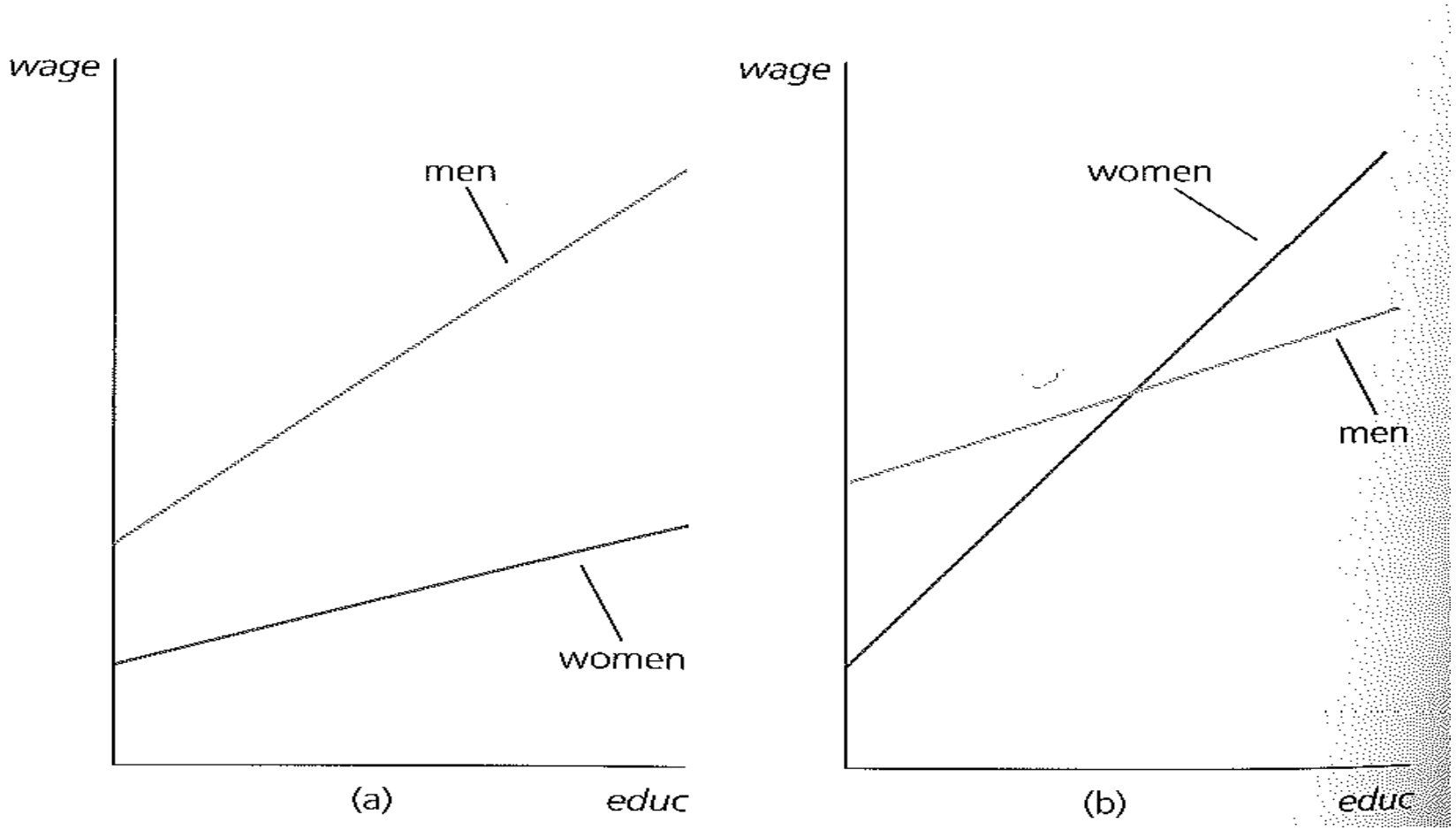
$$\text{wage} = \beta_0 + \beta_1 \text{female} + \beta_2 \text{educ} + \beta_3 \text{femaleeduc} + u$$

Daraus folgt:

- Die Nullhypothese zur Überprüfung, dass der Effekt von educ auf wage bei Männern und Frauen gleich ist, lautet H_0: $\beta_3 = 0$. Die Nullhypothese zur Überprüfung, dass die durchschnittlichen Stundenlöhne bei Frauen und Männer bei gleicher Ausbildungszeit identisch sind, lautet H_0: $\beta_1 = 0$, $\beta_3 = 0$.
- Mit $\beta_1 < 0$ und $\beta_3 < 0$ haben Frauen bei gleicher Ausbildungszeit einen geringeren durchschnittlichen Stundenlohn als Männer. Die Diskrepanz wächst mit zunehmender Ausbildungszeit, da der entsprechende Effekt von educ bei Frauen kleiner ist als bei Männern.
- Mit $\beta_1 < 0$ und $\beta_3 > 0$ haben Frauen bei einer geringen (gleichen) Ausbildungszeit einen geringeren durchschnittlichen Stundenlohn. Die Diskrepanz sinkt aber mit zunehmender Ausbildungszeit, da der entsprechende Effekt bei Frauen größer ist.

Beispiel: Erklärung von Löhnen (Fortsetzung)

Erwartungswerte von wage bei Frauen und Männern für (a) $\beta_1 < 0$ und $\beta_3 < 0$ und (b) $\beta_1 < 0$ und $\beta_3 > 0$:



4.5 Heteroskedastizität

In Kapitel 2 wurde für die Betrachtung der Varianz von OLS-Schätzern ausführlich die Annahme A5 der Homoskedastizität diskutiert:

- Falls $Var(u|x_1, x_2,\ldots, x_k) \neq \sigma^2$, liegt Heteroskedastizität vor
- Im Gegensatz z.B. zur Vernachlässigung relevanter erklärender Variablen, hat die Heteroskedastizität keinen Einfluss auf die Erwartungstreue oder Konsistenz von OLS-Schätzern
- Allerdings hat Heteroskedastizität einen Einfluss auf die (geschätzte) Varianz der mit OLS geschätzten Steigungsparameter in linearen Regressionsmodellen
- Entsprechend (2.33) ergibt sich bei Homoskedastizität, d.h. unter den in Kapitel 2 diskutierten Annahmen A1 bis A5 für die Varianz der geschätzten Steigungsparameter (wobei R_j^2 das Bestimmtheitsmaß bei einer Hilfsregression von x_j auf alle anderen erklärenden Variablen darstellt):

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2} = \frac{\sigma^2}{(1-R_j^2)SST_j} \quad \text{für } j = 1,\ldots, k$$

- Im einfachen linearen Regressionsmodell ergibt sich nach (2.34) speziell:

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \frac{1}{n}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ sowie } \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Damit ergibt sich in multiplen linearen Regressionsmodellen bei Homoskedastizität mit einer konsistenten Schätzung der Standardabweichung σ des Fehlerterms u nach (2.39) die geschätzte Standardabweichung der mit OLS geschätzten Steigungsparameter:

$$\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j)} = \frac{\hat{\sigma}}{\sqrt{(1-R_j^2)SST_j}} \quad \text{für } j = 1,\ldots, k$$

- Da die Varianz in (2.33) lediglich bei Homoskedastizität, nicht aber bei Heteroskedastizität gilt, ist (2.39) auch eine verzerrte Schätzung der Standardabweichung der OLS-Schätzer
- Damit sind die geschätzten Standardabweichungen in (2.39) bei Heteroskedastizität auch nicht mehr für die Konstruktion von Konfidenzintervallen und t-Statistiken gültig

- Das heißt, die entsprechenden t-Statistiken sind bei Heteroskedastizität (auch bei großen Stichprobenumfängen) nicht mehr t-verteilt. Ebenso sind entsprechende F-Statistiken bei Heteroskedastizität nicht mehr F-verteilt.
- Schließlich gilt bei Heteroskedastizität nicht mehr die wünschenswerte BLUE-Eigenschaft (bzw. Effizienz) von OLS-Schätzern sowie die Eigenschaft der asymptotischen Effizienz. Allerdings lassen sich bei Kenntnis der Form der Heteroskedastizität gegenüber den OLS-Schätzern effizientere Schätzungen ermitteln (siehe unten).

Falls $Var(u|x_1, x_2,\ldots, x_k) \neq \sigma^2$, d.h. falls die Varianz des Fehlerterms u nicht für alle Beobachtungen identisch ist, sondern von einzelnen Werten der erklärenden Variablen abhängt, finden die entsprechenden Varianzen σ_i^2 auch Eingang in die Varianzen der OLS-Schätzer. Speziell im einfachen linearen Regressionsmodell ergibt sich z.B. für die Varianz des mit OLS geschätzten Steigungsparameters:

$$(4.13) \quad Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

Falls $\sigma_i^2 = \sigma^2$, liegt Homoskedastizität vor und man gelangt zur entsprechenden Varianz in (2.34).

Alternative Schätzmethoden oder heteroskedastizitäts-robuste t-Statistiken:

- Da die OLS-Schätzungen auch bei Heteroskedastizität (unter den Annahmen A1 bis A4) erwartungstreu und konsistent sind, kann die Verwendung von OLS auch in diesem Fall weiterhin nützlich sein
- Für die Konstruktion von Konfidenzintervallen sowie die Durchführung von t- und F-Tests sollten bei Heteroskedastizität allerdings zumindest die geschätzten Standardabweichungen der OLS-Schätzer korrigiert werden

Ausgangspunkt dieser Korrekturen sind die tatsächlichen (unbekannten) Varianzen der OLS-Schätzer. Dabei werden die unbekannten Varianzen σ_i^2 durch die entsprechenden quadrierten Residuen $\hat{u}_i^2$ (die sich aus der ursprünglichen OLS-Schätzung ergeben) ersetzt. In Anlehnung an (4.13) ergibt sich somit für die geschätzte Varianz des mit OLS geschätzten Steigungsparameters im einfachen linearen Regressionsmodell:

$$(4.14) \quad \hat{\mathrm{Var}}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

Im multiplen linearen Regressionsmodell ergibt sich allgemein für die geschätzte Varianz der mit OLS geschätzten Steigungsparameter:

$$(4.15) \quad \widehat{\mathrm{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

Dabei bezeichnen $\hat{r}_{ij}$ das Residuum für Beobachtung i, das bei der Regression von x_j auf alle anderen erklärenden Variablen entsteht, und SSR_j die Residualabweichungsquadratsumme aus dieser Hilfsregression. Für die geschätzte Standardabweichung der mit OLS geschätzten Steigungsparameter ergibt sich entsprechend nach White (1980):

$$(4.16) \quad \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)} = \frac{\sqrt{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}}{SSR_j}$$

Auf der Grundlage von (4.16) sind verschiedene weitere asymptotisch äquivalente geschätzte Standardabweichungen entwickelt worden. Mit Hilfe dieser geschätzten Standardabweichungen können entsprechende heteroskedastizitäts-robuste Konfidenzintervalle und vor allem t-Statistiken konstruiert werden. Diese t-Statistiken werden mittlerweile fast standardmäßig verwendet und sind in ökonometrischen Programmpaketen wie z.B. STATA implementiert.

Beispiel: Erklärung (des Logarithmus) von Löhnen

Mit Hilfe eines linearen Regressionsmodells wird für n = 526 Personen erneut der Effekt der Ausbildungszeit in Jahren (educ), der Berufserfahrung in Jahren (exper), der quadrierten Berufserfahrung in Jahren (expersq), der Betriebszugehörigkeit in Jahren (tenure), der quadrierten Betriebszugehörigkeit in Jahren (tenuresq) sowie der drei kombinierten Familienstands- und Geschlechtsvariablen für verheiratete Männer (marrmale), verheiratete Frauen (marrfem) und unverheiratete Frauen (singfem) auf den Logarithmus des Stundenlohns (logwage) untersucht. Dabei wurde folgende OLS-Regressionsfunktion geschätzt, wobei jetzt neben den herkömmlichen auch die heteroskedastizitätsrobust geschätzten Standardabweichungen der geschätzten Parameter (eckige Klammern) ausgewiesen werden (R^2 = 0,461):

$$\widehat{\log wage} = 0{,}321 + 0{,}213\,marrmale - 0{,}198\,marrfem - 0{,}110\,singfem + 0{,}0789\,educ$$

(0,100) (0,055) (0,058) (0,056) (0,0067)

[0,109] [0,057] [0,058] [0,057] [0,0074]

$$+ 0{,}0268\,exper - 0{,}00054\,expersq + 0{,}0291\,tenure - 0{,}00053\,tenuresq$$

(0,0055) (0,00011) (0,0068) (0,00023)

[0,0051] [0,00011] [0,0069] [0,00024]

Beispiel: Erklärung (des Logarithmus) von Löhnen (STATA-Output)

Mit STATA haben sich folgende OLS-Schätzergebnisse mit heteroskedastizitäts-robust geschätzten Standardabweichungen der geschätzten Parameter gezeigt:

```
reg logwage marrmale marrfem singfem educ exper expersq tenure tenuresq, robust
Linear regression                                      Number of obs =     526
                                                       F(  8,   517) =   51.70
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4609
                                                       Root MSE      =  .39329
------------------------------------------------------------------------------
             |               Robust
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    marrmale |   .2126756   .0571419     3.72   0.000     .1004167    .3249345
     marrfem |  -.1982677     .05877    -3.37   0.001    -.3137251   -.0828103
     singfem |  -.1103502   .0571163    -1.93   0.054    -.2225587    .0018583
        educ |   .0789103   .0074147    10.64   0.000     .0643437    .0934769
       exper |   .0268006   .0051391     5.22   0.000     .0167044    .0368967
     expersq |  -.0005352   .0001063    -5.03   0.000    -.0007442   -.0003263
      tenure |   .0290875   .0069409     4.19   0.000     .0154516    .0427234
    tenuresq |  -.0005331   .0002437    -2.19   0.029    -.0010119   -.0000544
       _cons |    .321378    .109469     2.94   0.003     .1063193    .5364368
------------------------------------------------------------------------------
```

Test auf Heteroskedastizität:

Falls Heteroskedastizität vorliegt, gilt nicht mehr die BLUE-Eigenschaft (bzw. Effizienz) von OLS-Schätzern sowie die Eigenschaft der asymptotischen Effizienz, so dass es schon deshalb wünschenswert sein kann, Homoskedastizität statistisch zu testen.

Ein Standardansatz ist der Breusch-Pagan-Test (eine Alternative ist z.B. der White-Test). Die Nullhypothese für Homoskedastizität lautet:

$$(4.17) \quad H_0: \mathrm{Var}(u|x_1, x_2,\ldots,x_k) = \sigma^2 \quad \text{bzw.} \quad H_0: E(u^2|x_1, x_2,\ldots,x_k) = E(u^2) = \sigma^2$$

Falls H_0 nicht gilt, ist u^2 eine Funktion einer oder mehrerer erklärender Variablen. Bei der Betrachtung aller erklärenden Variablen und einer linearen Funktion ergibt sich in diesem Fall mit einem Störterm v mit (bedingtem) Erwartungswert null:

$$(4.18) \quad u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v$$

Die Nullhypothese für Homoskedastizität lautet dann:

$$(4.19) \quad H_0: \delta_1 = \delta_2 = \cdots = \delta_k = 0$$

Da die u_i unbekannt sind, werden diese durch die entsprechenden Schätzer ersetzt, d.h. den Residuen $\hat{u}_i$, so dass diese quadrierten Residuen auf die erklärenden Variablen regressiert werden:

$$(4.20) \quad \hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v$$

Ein hohes Bestimmtheitsmaß $R^2_{\hat{u}^2}$ bei dieser Hilfsregression spricht für die Gültigkeit der Alternativhypothese, d.h. für Heteroskedastizität. Eine Version einer Breusch-Pagan-Teststatistik (die keine Normalverteilung in u annimmt) lautet:

$$(4.21) \quad BP = nR^2_{\hat{u}^2}$$

Bei Gültigkeit der Nullhypothese (d.h. bei Homoskedastizität) ist BP asymptotisch χ^2-verteilt mit k Freiheitsgraden, d.h.:

$$(4.22) \quad BP \overset{a}{\sim} \chi^2_k$$

Damit wird die Nullhypothese der Homoskedastizität zugunsten der Alternativhypothese der Heteroskedastizität bei einem Signifikanzniveau α verworfen, falls (bei großem Stichprobenumfang n) für die Teststatistik gilt:

$$(4.23) \quad BP > \chi^2_{k;1-\alpha}$$

Die Testentscheidung kann natürlich auch mit Hilfe des entsprechenden p-Wertes getroffen werden.

Falls die Nullhypothese bei einem kleinen Signifikanzniveau verworfen wird, sollten zumindest heteroskedastizitäts-robust geschätzte Standardabweichungen bzw. robuste t-Statistiken verwendet werden. Alternativ können auch von OLS abweichende Schätzverfahren angewendet werden wie z.B. die gewichtete Methode der kleinsten Quadrate („WLS, weighted least squares“). Dazu sollte aber die genaue Form der Heteroskedastizität bekannt sein.

Beispiel 1: Erklärung von Häuserpreisen

Mit Hilfe eines linearen Regressionsmodells wird der Effekt der Grundstücksgröße in Quadratfuß (lotsize), der Wohnflächengröße in Quadratfuß (sqrft) und der Anzahl an Schlafzimmern (bdrms) auf Häuserpreise in 1000 Dollar (price) untersucht. Dabei wurde folgende OLS-Regressionsfunktion geschätzt:

$$\widehat{price} = -21{,}77 + 0{,}00207\,lotsize + 0{,}123\,sqrft + 13{,}85\,bdrms$$

$$(29{,}48)\quad (0{,}00064)\qquad (0{,}013)\qquad (9{,}01)$$

$$n = 88;\ R^2 = 0{,}672$$

Mit Hilfe des Breusch-Pagan-Tests wird nun bei einem Signifikanzniveau von 1% die Nullhypothese der Homoskedastizität überprüft:

- Zunächst werden die Residuen $\hat{u}_i$ berechnet. Bei der Hilfsregression von $\hat{u}^2$ auf lotsize, sqrft und bdrms ergibt sich ein Bestimmtheitsmaß in Höhe von $R^2_{\hat{u}^2} = 0{,}1601$.
- Für die entsprechende Breusch-Pagan-Teststatistik ergibt sich damit ein Wert von $BP = 88 \cdot 0{,}1601 = 14{,}09$
- Mit k = 3 lautet der Schrankenwert $\chi^2_{3;0,99} = 11{,}34$. Damit wird die Nullhypothese verworfen (der entsprechende p-Wert beträgt p = 0,0028)

Beispiel 1: Erklärung von Häuserpreisen (STATA-Output)

Mit STATA haben sich folgende OLS-Schätzergebnisse gezeigt:

```
reg price lotsize sqrft bdrms
```

Source	SS	df	MS		
Model	617130.702	3	205710.234	Number of obs	88
Residual	300723.806	84	3580.04531	F(3, 84)	57.46
Total	917854.508	87	10550.0518	Prob > F	0.0000
				R-squared	0.6724
				Adj R-squared	0.6607
				Root MSE	59.833

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lotsize	.0020677	.0006421	3.22	0.002	.0007908	.0033446
sqrft	.1227782	.0132374	9.28	0.000	.0964541	.1491022
bdrms	13.85252	9.010145	1.54	0.128	-4.06514	31.77018
_cons	-21.77031	29.47504	-0.74	0.462	-80.38466	36.84404

Testanweisung und Testergebnisse mit STATA (nur direkt nach Durchführung der OLS-Schätzung möglich):

```
estat hettest, rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: lotsize sqrft bdrms

         chi2(3)      =    14.09
         Prob > chi2  =   0.0028
```

Beispiel 2: Erklärung des Logarithmus von Häuserpreisen

In Kapitel 4.2 wurde erwähnt, dass durch die Logarithmierung insbesondere der abhängigen Variablen häufig Probleme der Heteroskedastizität abgeschwächt werden können. Aus diesem Grund wird nun mit Hilfe eines linearen Regressionsmodells mit denselben Daten der Effekt der logarithmierten Grundstücksgröße in Quadratfuß (loglotsize), der logarithmierten Wohnflächengröße in Quadratfuß (logsqrft) und der Anzahl an Schlafzimmern (bdrms) auf die logarithmierten Häuserpreise in 1000 Dollar (logprice) untersucht. Dabei wurde folgende OLS-Regressionsfunktion geschätzt ($R^2 = 0{,}643$):

$$\widehat{\log price} = -1{,}30 + 0{,}168\,\text{loglotsize} + 0{,}700\,\text{logsqrft} + 0{,}37\,\text{bdrms}$$

$$(0{,}65)\quad (0{,}038)\qquad (0{,}093)\qquad (0{,}028)$$

Breusch-Pagan-Test:

- Bei der Hilfsregression von $\hat{u}^2$ auf loglotsize, logsqrft und bdrms ergibt sich ein Bestimmtheitsmaß in Höhe von $R^2_{\hat{u}^2} = 0{,}048$. Damit ergibt sich für die Breusch-Pagan-Teststatistik ein Wert von $BP = 88 \cdot 0{,}048 = 4{,}22$.
- Der Schrankenwert lautet erneut $\chi^2_{3;0,99} = 11{,}34$, so dass jetzt die Nullhypothese der Homoskedastizität nicht verworfen werden kann (der entsprechende p-Wert beträgt $p = 0{,}238$).

Beispiel 2: Erklärung des Logarithmus von Häuserpreisen (STATA-Output)

Mit STATA haben sich folgende OLS-Schätzergebnisse gezeigt:

```
reg logprice loglotsize logsqrft bdrms
```

Source	SS	df	MS		
Model	5.15503927	3	1.71834642	Number of obs =	88
Residual	2.86256324	84	.034078134	F(3, 84) =	50.42
Total	8.01760251	87	.092156351	Prob > F =	0.0000
				R-squared =	0.6430
				Adj R-squared =	0.6302
				Root MSE =	.1846

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loglotsize	.1679668	.0382811	4.39	0.000	.0918406	.2440931
logsqrft	.7002321	.0928652	7.54	0.000	.5155594	.8849049
bdrms	.0369583	.0275313	1.34	0.183	-.0177907	.0917074
_cons	-1.297041	.6512835	-1.99	0.050	-2.592189	-.0018918

Testanweisung und Testergebnisse mit STATA (nur direkt nach Durchführung der OLS-Schätzung möglich):

```
estat hettest, rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: loglotsize logsqrft bdrms

         chi2(3)      =     4.22
         Prob > chi2  =   0.2383
```