Zur Ermittlung des Verkehrswerts bebauter Grundstücke in Kaiserslautern

Axel Krebs[1]
Betreuer: Prof. Dr. J. Franke

Inhaltsverzeichnis

Aufgabenstellung 1
Beschreibung des Datenmaterials 1
Herkunft der Daten 1
Beschreibung der Variablen 2
Einführung von Dummy-Variablen 7
Anwendung des automatischen Regressionsverfahrens 8
Beschreibung des Verfahrens 8
1. Modell (mit Originaldaten) 9
Voruntersuchung 9
Ergebnis der Auto-Regression 12
2. Modell (mit transformierten Daten) 15
Transformation 15
Voruntersuchung 17
Ergebnis der Auto-Regression 20
Vergleich 24
1. Beispiel 24
2. Beispiel 25
Vereinfachung der Modelle 26
Vereinfachtes Modell mit Original-Daten 26
Vereinfachtes Modell mit transformierten Daten 27
Schlußbemerkung 28
Literaturverzeichnis 29
Inhaltsverzeichnis 29

Literaturverzeichnis

[1] Angelika Schwarz: Abschlußbericht des Projektes "Bestimmungsfaktoren von Grundstückswerten"; Berlinforschung, Förderungsprogramm der Freien Universität;
[2] T. W. Anderson/S. L. Sclove: The Statistical Analysis of Data, Second Edition. Palo Alto, 1986.
[3] H.-W. Schaar: Vergleichswertverfahren für bebaute Grundstücke.

[1] Außer Herrn Prof. Franke möchte ich mich auch bei Frau Friederichs, Herrn v. Sachs und Herrn Scholl herzlich bedanken für zahlreiche klärende Gespräche.

Aufgabenstellung

Anhand des vom Gutachterausschuß der Stadt Kaiserslautern zur Verfügung gestellten Datenmaterials soll untersucht werden, welche Faktoren den Verkehrswert eines bebauten Grundstücks beeinflussen. Mit diesen Erkenntnissen soll eine möglichst einfache Formel ermittelt werden, die eine Schätzung für den Verkehrswert liefert, und die dabei die in der Vergangenheit erzielten Kaufpreise berücksichtigt.
Für die Lösung dieser Aufgabe bietet sich das Verfahren der multiplen linearen Regression an. Auf die theoretischen Grundlagen soll hier nicht näher eingegangen werden, man findet sie in jedem Buch über mathematische Statistik oder in [1].
Bei der Analyse der Daten wurde im großen und ganzen der Weg eingeschlagen, den Angelika Schwarz in [1] beschreibt. Ihre Ergebnisse lassen sich jedoch nicht direkt übertragen, da die dort betrachteten Grundstücke unbebaut waren.
Da bei der statistischen Auswertung großer Datenmengen ein immenser Rechenaufwand anfällt, ist es unverzichtbar, professionelle statistische Software einzusetzen. Es stand das Programm SPlus 2.0 (PC-Version für Windows) zur Verfügung. Sämtliche Berechnungen und alle Grafiken in diesem Bericht wurden in SPlus erstellt.

Beschreibung des Datenmaterials

Herkunft der Daten

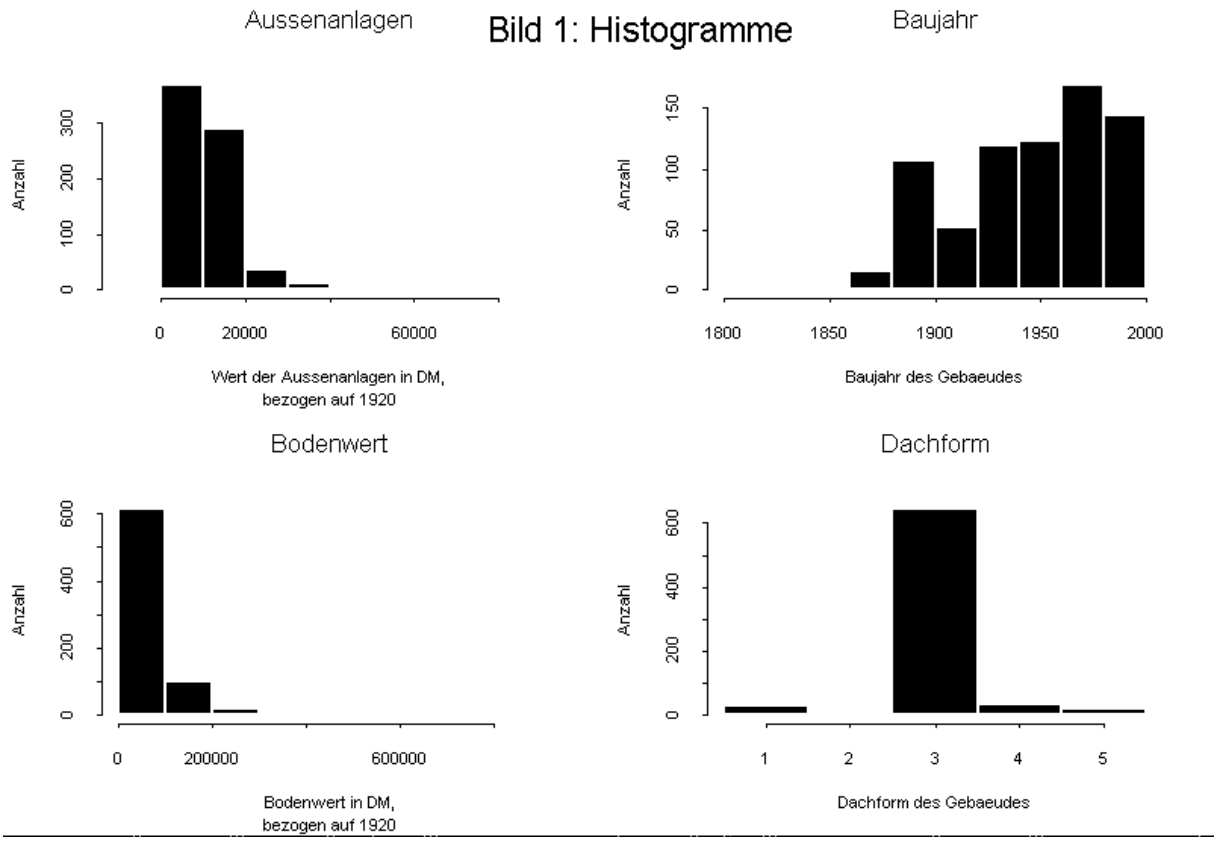
Der Gutachterausschuß der Stadt Kaiserslautern erfaßt die bei ihm eingehenden Kaufverträge seit 1990 per EDV. Auf diese Daten konnte zurückgegriffen werden. Um die Homogenität des Datenmaterials zu garantieren (nur dann machen statistische Untersuchungen dieser Art Sinn), wurden Verkaufsfälle von Grundstücken, die mit Ein- oder Zweifamilienhäusern bebaut sind, ausgewählt.
Datensätze, die mit dem Vermerk "zum Vergleich nicht geeignet" versehen waren, wurden aussortiert, ebenso einige nicht-repräsentative Verkaufsfälle in Vororten von Kaiserslautern. Nach diese ersten Auswahl standen 737 Datensätze zur Verfügung.
Aus den mehr als 50 vom Gutachterausschuß erfaßten Größen wurden in einer Vorauswahl 20 Variablen ausgewählt. Dabei wurde darauf geachtet, daß die Variablen zum einen als statistisch aussagekräftig gelten können (so wurden beispielsweise Variablen nicht aufgenommen, die reine Linearkombinationen der anderen Variablen sind, da sie in einem linearen Modell keine Verbesserung bringen können). Ein weiterer Gesichtspunkt war aber auch die Vollständigkeit der Daten: In vielen Fällen waren die Variablen nur bruchstückhaft eingegeben. Eine Nacherfassung hätte nicht nur den Rahmen dieses Projekts gesprengt, sondern hätte auch zu einem gewaltigen rechnerischen Mehraufwand geführt, der sich sicherlich nicht ausgezahlt hätte.
Da Datensätze, bei denen auch nur ein Wert nicht verfügbar ist, unbrauchbar sind, mußten 89 Datensätze entfernt werden. Es standen somit 648 Datensätze zur Verfügung, die aus Verkaufsfällen der Jahre 1990-93 stammen.

Beschreibung der Variablen

Im folgenden werden die Variablen, die für die weitere Untersuchung in Betracht kommen, in tabellarischer Form aufgelistet und beschrieben. Zu jeder Variablen werden einige Angaben gemacht, die für das zu findende Modell von großer Bedeutung sind. Es handelt sich um die Spannweite (Maximum und Minimum), Mittelwert, Median und Standardabweichung der Variablen.
Maximum und Minimum geben an, in welchem Bereich die Berechnungen durchgeführt wurden. Die Aussagen des zu berechnenden Modells (also die "Schätzformel" für den Verkehrswert) sind strenggenommen nur innerhalb dieser Grenzen gültig.
Mittelwert, Median und Standardabweichung können gemeinsam mit den ebenfalls aufgeführten Histogrammen (Häufigkeitsverteilungen) sowie der empirischen Schiefe und dem empirischen Exzeß zu Aussagen über die Verteilung der einzelnen Variablen herangezogen werden. Besonders interessant ist beispielsweise die Frage, ob die Daten einer Normalverteilung unterliegen.
Desweiteren wird angegeben, wie viele Datensätze in der Datenbank jeweils fehlen.

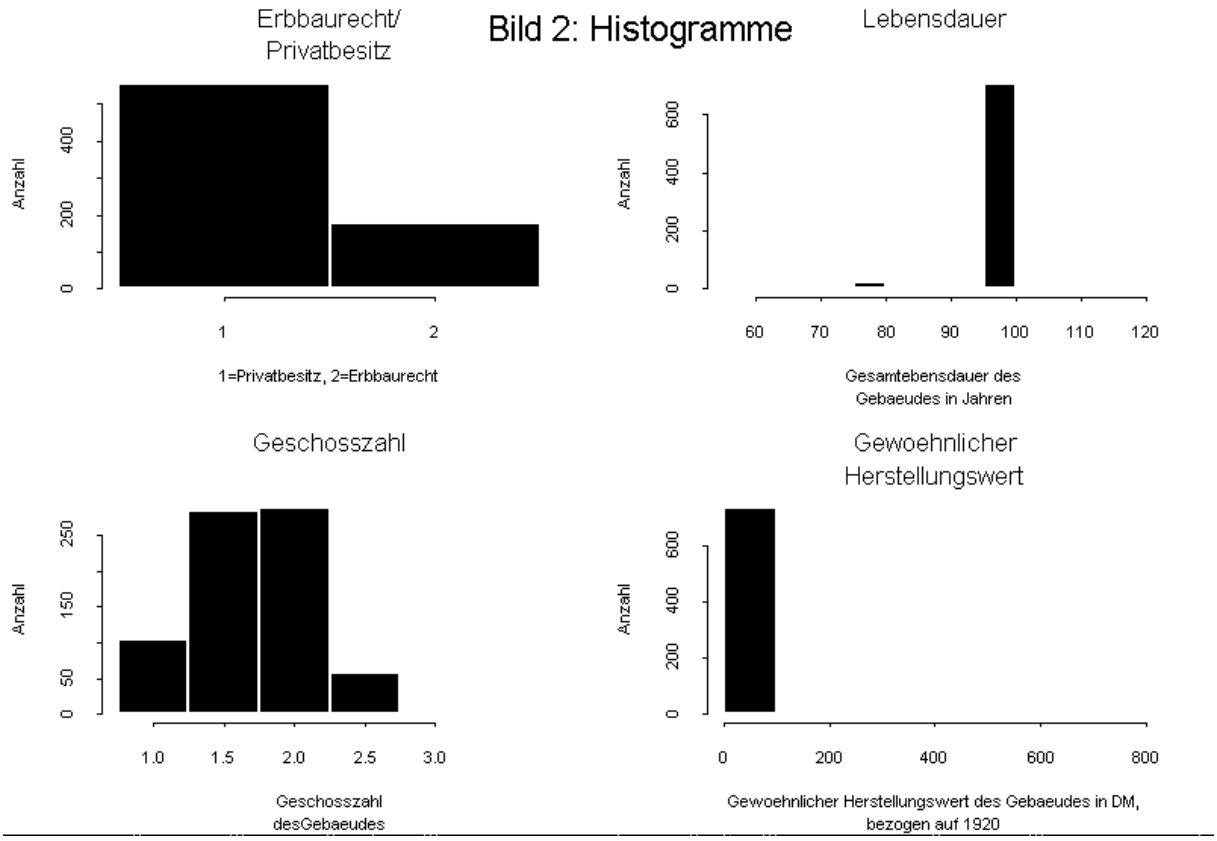
	Außenanlagen	Baujahr	Bodenwert	Dachform
Beschreibung	Wert der Außenanlagen des Grundstücks in DM	Baujahr des Gebäudes	Grundstückswert in DM, dividiert durch den Preisindex. Der Bodenwert ergibt sich aus Grundstücksfläche * Richtwert	Dachform des Gebäudes: 1 = Flachdach 2 = Pultdach 3 = Satteldach 4 = Walmdach 5 = andere Dachkonstruktion
Minimum	0	1810	1143	1
Maximum	88000	1994	886950	5
Mittelwert	11888,42	1947	64662,58	3,01
Median	10000	1956	46560	3
Standardabweichung	8345,76	33,96	63119,53	0,58
Fehlende Daten	20	3	0	0

Bild 1: Histogramme



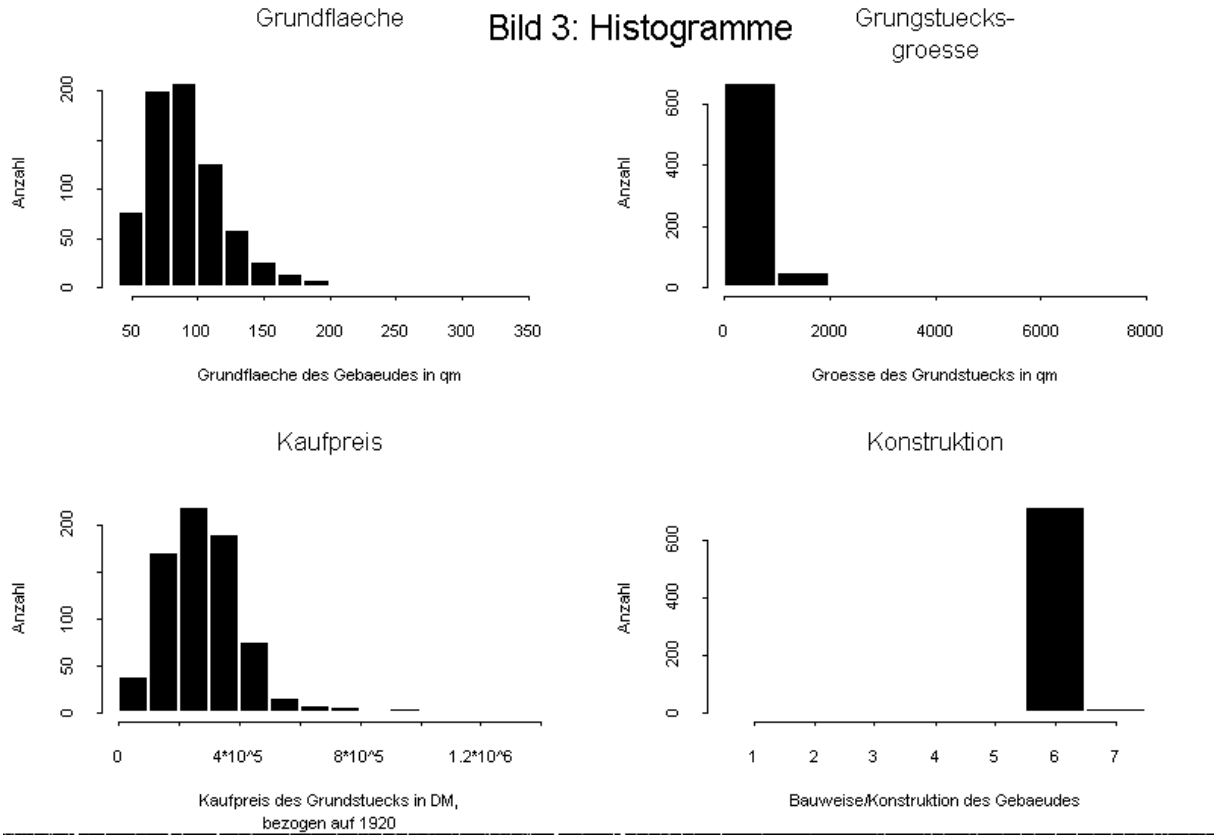
	Erbbau/privat	Ges.-Lebens-dauer	Geschoßzahl	Gew. Herst.wert
Beschreibung	1 = Privatbesitz 2 = Erbbaurecht	Gesamtlebensdauer des Gebäudes in Jahren	Geschoßzahl des Gebäudes (in 0.5-Schritten)	Gewöhnlicher Herstellungswert in DM,/m^3
Minimum	1	60	1	16
Maximum	2	120	3	730
Mittelwert	1,24	99,23	1,71	24,1
Median	1	100	1,5	23
Standardabweichung	0,43	4,44	0,43	26,2
Fehlende Daten	0	1	1	3

Bild 2: Histogramme



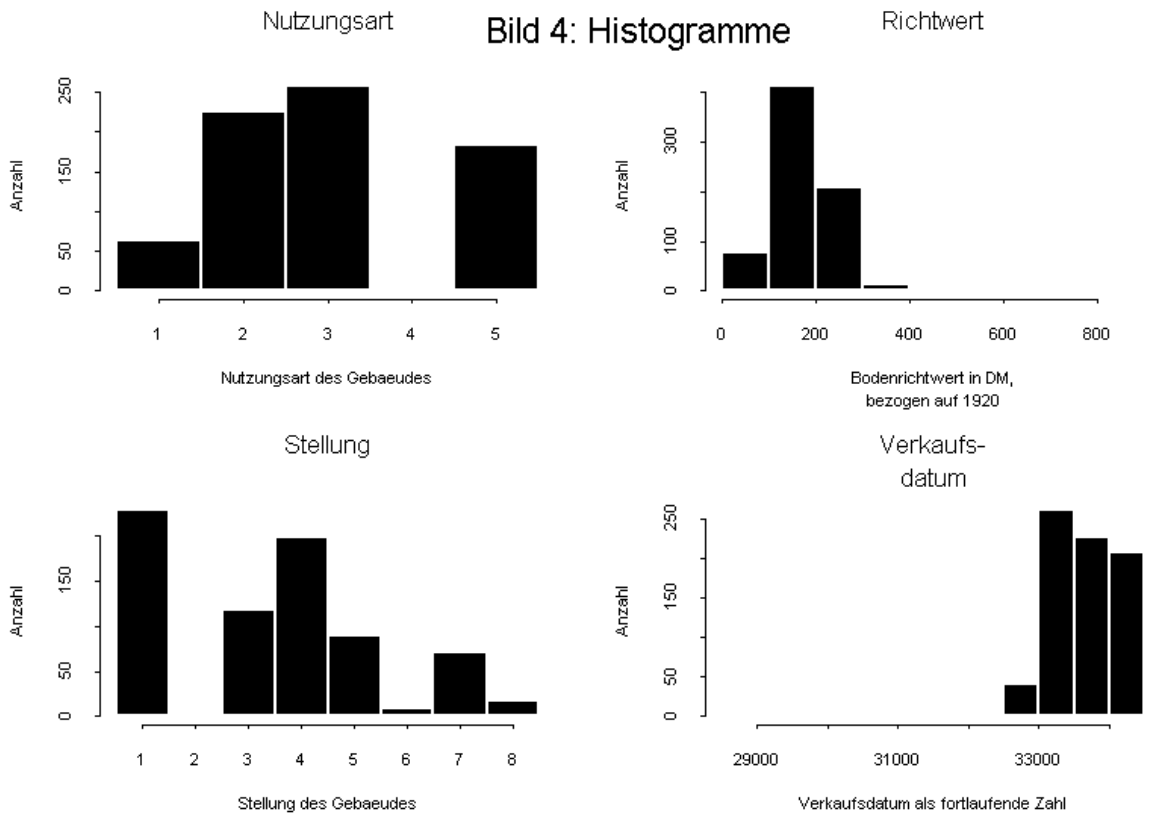
	Grundfläche	Grundstücks-größe	Kaufpreis	Konstruktion
Beschreibung	Grundfläche des Gebäudes	Größe des Grundstücks in m^2	Erzielter Kaufpreis in DM	Konstruktion/Bauweise des Gebäudes: 1 = Holzgebäude 2 = Fachwerkgebäude 3 = Gebäude in leichter Bauart 4 = Fertighaus in leichter Bauart 5 = Fertighaus in massiver Bauart 6 = Mauerwerksbauten 7 = Beton- oder Stahlbeton-Fertigteilbauten 8 = Stahlbeton-, Stahl- oder Stahlbeton-Skelettbau-ten
Minimum	45	80	11400	1
Maximum	350	7770	1344900	7
Mittelwert	95,58	541,89	292627,90	6,00
Median	88	409	279500	6
Standardabweichung	36,38	635,52	147980,79	0,30
Fehlende Daten	6	0	0	0

Bild 3: Histogramme



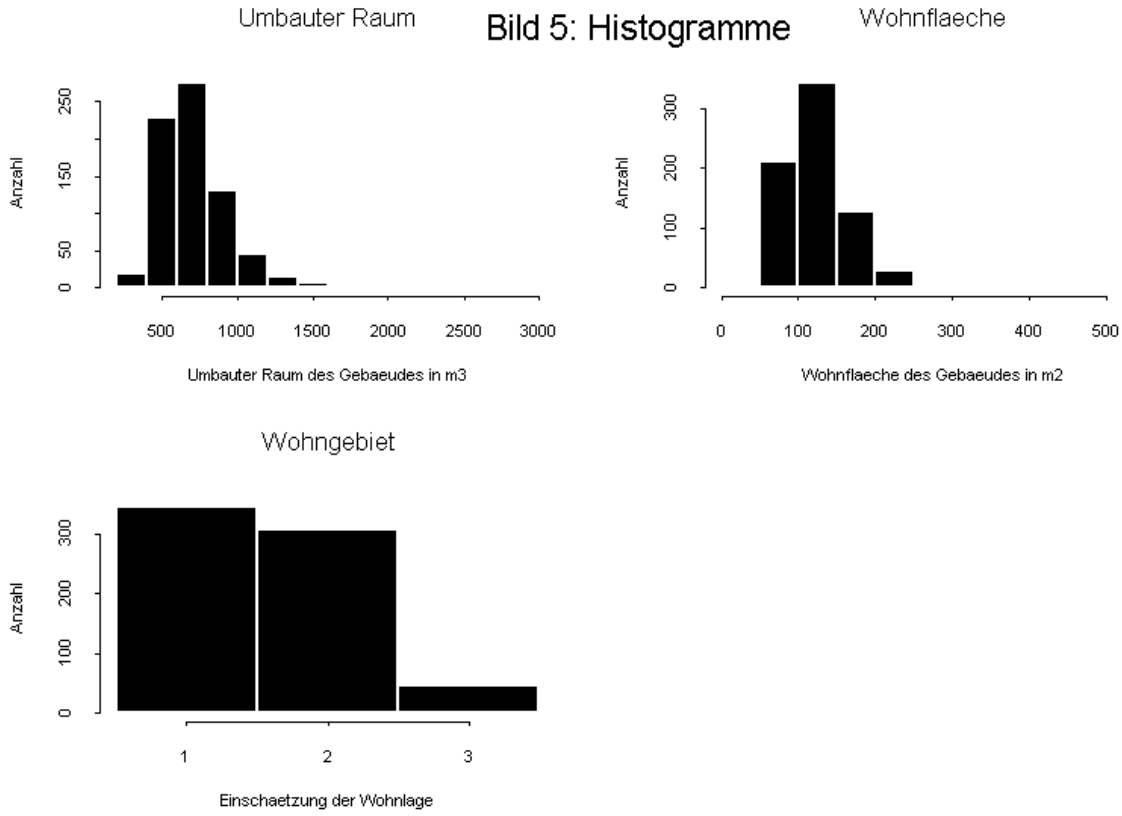
	Nutzungsart	Richtwert	Stellung	Tageszahl
Beschreibung	Art der Nutzung des Gebäudes: 1 = Kleinwohnhaus, Kleinsiedlungshaus 2 = Einfamilienhaus (normale Größe) 3 = größeres Einfamilienhaus 4 = Villa, Landhaus 5 = Zweifamilienhaus 6 = Mehrfamilienhaus 7 = Wochenendhaus 8 = Appartementhaus	Richtwert für den Bodenpreis in DM/m^2	Stellung des Gebäudes; 1 = einzeln stehend 2 = Doppelhaus 3 = Doppelhaushälfte 4 = Reihenmittelhaus 5 = Reihenendhaus 6 = Anbau 7 = freistehend (Nebengebäude) 8 = Eckhaus	Datum des Verkaufsfalls als fortlaufende Zahl. 0 = 1. Januar 1900
Minimum	1	25	1	28590
Maximum	5	850	8	34333
Mittelwert	3,02	180,1	3,44	33647,32
Median	3	180	4	33634
Standardabweichung	1,29	69,24	2,02	468,87
Fehlende Daten	0	22	0	1

Bild 4: Histogramme



	Umbauter Raum	Wohnfläche	Wohngebiet
Beschreibung	Umbauter Raum des Gebäudes in m^3	Wohnfläche des Gebäudes in m^2	Einschätzung der Wohnlage; 1 = sehr gute Wohnlage 2 = mittlere Wohnlage 3 = schlechte Wohnlage
Minimum	235	37	1
Maximum	3000	507	3
Mittelwert	719,31	127,80	1,57
Median	700	120	2
Standardabweichung	249,41	45,20	0,61
Fehlende Daten	5	11	39

Bild 5: Histogramme



Einführung von Dummy-Variablen

Betrachtet man sich die Variablen, so bemerkt man, daß einige qualitative Merkmale "quantifiziert" wurden in dem Sinn, daß jede mögliche Merkmalsausprägung mit einer Zahl kodiert wurde. Für ein Regressionsverfahren sind solche Angaben jedoch unbrauchbar, denn das Ergebnis der Regression hängt von der speziellen Codierung ab. Ändert man beispielsweise die Bedeutung der einzelnen Zahlen der Variablen "Nutzung", so erhält man ein völlig anderes Ergebnis.
Aus diesem Grund führt man sogenannte Dummy-Variablen ein, die nur die Werte 0 oder 1 annehmen, je nach dem, ob ein bestimmtes Merkmal auftritt oder nicht. Um auf diese Weise eine Variable mit *N* Merkmalsausprägungen zu kodieren, benötigen wir *N*-1 Dummy-Variablen, wenn wir alle Ausprägungen berücksichtigen wollen. In vielen Fällen ist dies jedoch nicht nötig. So ist etwa in der Variablen "Konstruktion" die Ausprägung 6 (Mauerwerksbau) 626 Mal vertreten. Die sieben anderen Klassen sind entsprechend dünn besetzt. Es wäre unsinnig, hier die theoretisch insgesamt nötigen sieben Dummies einzuführen. Stattdessen wird nur ein Dummy definiert, der den Wert 1 annimmt, wenn es sich bei dem betreffenden Gebäude um einen Mauerwerksbau handelt, und ansonsten 0.

Es ergeben sich folgende neuen Variablen:

Name	**Beschreibung**	**Anzahl "1"**
Dachform.3	= 1, falls Satteldach, 0 sonst	567
Erbbau.privat	= 1 falls Erbbaurecht, 0 sonst	155
Geschosszahl.1.0	= 1, falls Geschoßzahl 1,0, 0 sonst	85
Geschosszahl.1.5	= 1, falls Geschoßzahl 1,5, 0 sonst	256
Geschosszahl.2.0	= 1, falls Geschoßzahl 2,0, 0 sonst	258
Geschosszahl.2.5	= 1, falls Geschoßzahl 2,5, 0 sonst	44
Konstruktion	= 1, falls Mauerwerksbau, 0 sonst	626
Nutzung.1	= 1, falls Kleinwohnhaus, 0 sonst	57
Nutzung.2	= 1, falls Einfamilienhaus (normale Größe), 0 sonst	205
Nutzung.3	= 1, falls größeres Einfamilienhaus, 0 sonst	216
Stellung.1	= 1, falls einzeln stehend, 0 sonst	199
Stellung.3	= 1, falls Doppelhaushälfte, 0 sonst	109
Stellung.4	= 1, falls Reihenmittelhaus, 0 sonst	168
Stellung.5	= 1, falls Reihenendhaus, 0 sonst	75
Stellung.7	= 1, falls freistehend (Nebengebäude), 0 sonst	68
Wohngebiet.1	= 1, falls sehr gute Wohnlage, 0 sonst	317
Wohngebiet.2	= 1, falls mittlere Wohnlage, 0 sonst	386

Damit besteht das zu untersuchende Datenmaterial nun aus 28 unabhängigen und einer abhängigen Variablen.
Zusätzliche Variablen erhält man, wenn man mögliche Interaktionen betrachtet. Unter einer Interaktion versteht man die Beeinflussung der Zielvariablen durch gleichzeitiges Vorhandensein zweier Eigenschaften. Beispielsweise ist es durchaus vorstellbar, daß sich die Nutzungsart in den unterschiedlichen Wohnlagen unterschiedlich auf den Preis auswirkt.
Für die Lineare Regression ist es sinnvoll, für ein Paar von Dummy-Variablen, bei dem Interaktion vermutet wird, eine weitere Dummy-Variable einzuführen, die genau dann den Wert 1 annimmt, wenn beide in Frage kommenden Merkmale eintreten.
Zieht man sämtliche möglichen Interaktionen ins Kalkül, so erhöht sich die Zahl der Variablen um 136. Da dies jedoch zu aufwendig ist, betrachtet man vielmehr die Korrelationen der Interaktionen mit dem Kaufpreis. Diese vergleicht man mit den Korrelationen zwischen den ursprünglichen Variablen und dem Kaufpreis. Wenn die Interaktion einen größeren Korrelationskoeffizienten aufweist, als die Einzelvariablen, wird sie für die Regression in Betracht gezogen. Diese Berechnungen werden im Abschnitt "Anwendung des automatischen Regressionsverfahrens" für nichttransformierte und transformierte Daten getrennt durchgeführt.

Anwendung des automatischen Regressions-verfahrens

Beschreibung des Verfahrens

Ziel der Regression ist es, den Kaufpreis als gewichtete Summe der Einflußfaktoren zu modellieren. Die Aufgabe des Regressionsverfahrens ist es also, Gewichtsfaktoren so zu bestimmen, daß der tatsächliche Kaufpreis und die zu jedem Datensatz berechnete gewichtete Summe sich möglichst wenig unterscheiden.

Im Lauf der statistischen Voruntersuchung werden zunächst die Variablen daraufhin getestet, ob sie überhaupt Einfluß auf den Kaufpreis nehmen und deshalb für das Modell in Betracht gezogen werden müssen.

Für die kontinuierlichen Variablen werden dazu zunächst die paarweisen Korrelationen berechnet. Der Korrelationskoeffizient liegt zwischen -1 und +1. Werte nahe bei 0 bedeuten keine oder nur schwache Korrelation, große Werte deuten auf starke Korrelation hin. In [2] werden Korrelationen von mehr als 0,8 (bzw. weniger als -0,8) als stark bezeichnet. Ist ein Variablenpaar stark korreliert, so kann es problematisch sein, beide Variablen ins Modell aufzunehmen. Man muß sich für eine von beiden entscheiden. Dazu werden Modelle mit jeweils einer von beiden berechnet und das bessere der beiden Modelle ausgewählt.

Ein weiteres Kriterium für oder gegen die Aufnahme einer Variablen ins Modell ist das normierte "PRESS-Kriterium" (PRESS = Prediction sum of squares), im folgenden kurz PRESS-Statistik genannt. Sie gibt Auskunft darüber, wie gut die Vorhersagequalität einer Einfachregression der betreffenden Variablen auf die Zielvariable ist. Ist sie kleiner als 100, so ist die Vorhersagegenauigkeit des betreffenden Einfachmodells besser, als die Schätzung des Kaufpreises durch den Mittelwert.

Ist die PRESS-Statistik einer Variablen größer als 100, so entscheidet die t-Statistik, ob die Variable in das Modell aufgenommen wird oder nicht. Überschreitet die t-Statistik bzw. ihr Absolutbetrag einen bestimmten Schwellenwert, so kann man annehmen, daß der berechnete Regressionskoeffizient im entsprechenden Modell ungleich Null ist. Wie groß dieser Schwellenwert ist, hängt davon ab, wie groß man die statistische Sicherheit (d. h. das Signifikanzniveau, also die Wahrscheinlichkeit, mit der eine Fehlentscheidung getroffen wird) für diese Entscheidung wählt. Sofern nicht anders angegeben, wählen wir als Signifikanzniveau α = 10% und erhalten damit einen Schwellenwert von 1,282 für einen "einseitigen Test" (Test, ob der Regressionskoeffizient größer ist als 0), 1,645 für den "zweiseitigen Test" (Test, ob der Koeffizient von 0 verschieden ist).

Das automatische Regressionsverfahren, das für diese Untersuchung verwendet wurde, baut schrittweise ein Modell auf, das die PRESS-Statistik minimiert. Dazu geht es zunächst vom einfachsten denkbaren Modell aus: der Kaufpreis ist unabhängig von allen anderen Größen; die beste Vorhersage ist in diesem Fall der Mittelwert. Nach und nach wird dem Modell nun diejenige Variable hinzugefügt, die eine möglichst große Verbesserung der PRESS-Statistik bewirkt. Mit dem neuen Variablensatz wird eine Regression nach der Methode der kleinsten Quadrate durchgeführt. Hat man ein neues Modell gefunden, so überprüft man ob man es durch Weglassen einer Variablen verbessern kann. Dieser Prozeß wird solange durchgeführt, bis die PRESS-Statistik durch weiteres Hinzufügen oder Entfernen von Variablen nicht mehr verbessert werden kann.

Die von SPlus erstellte Zusammenfassung eines solchen Durchlaufs beinhaltet zahlreiche Informationen, die hier kurz erläutert werden sollen.

- Residuen: Es werden jeweils Minimum, Maximum, Median sowie erstes und drittes Quartil angegeben. Im Idealfall sind Mittelwert und Median beide Null, Minimum und Maximum haben beide die gleiche Größenordnung.
- Koeffizienten: Neben dem Wert des jeweiligen Koeffizienten werden seine Standardabweichung und seine t-Statistik angegeben. Wie bereits angegeben, sollte der Absolutbetrag der t-Statistik den Wert 1,645 nicht unterschreiten (0,05-Quantil der Standardnormalverteilung, durch die die t-Verteilung mit mehr als 200 Freiheitsgraden angenähert werden kann).
- Statistiken der Regression: Nicht nur die einzelnen Koeffizienten, sondern auch das Modell als gesamtes werden bewertet. Dazu wird die Standardabweichung der Residuen, die R^2-Statistik sowie die F-Statistik des Modells berechnet. Die R^2-Statistik gibt an, welcher Anteil der Schwankung der Zielvariablen durch das Regressionsmodell erklärt wird. Die F-Statistik verwendet man, um zu testen, ob mindestens einer der berechneten Koeffizienten von Null verschieden ist, ob also die Regression überhaupt einen Sinn macht. Wir akzeptieren das Modell auf dem Niveau α = 0,01, wenn die F-Statistik den Wert $F_{m,n}(0,01)$ (das 0,01-Quantil der F-Verteilung mit m und n Freiheitsgraden) überschreitet. Die Werte für m und n sind von Fall zu Fall verschieden. Deshalb unterscheiden sich die Schwellenwerte.

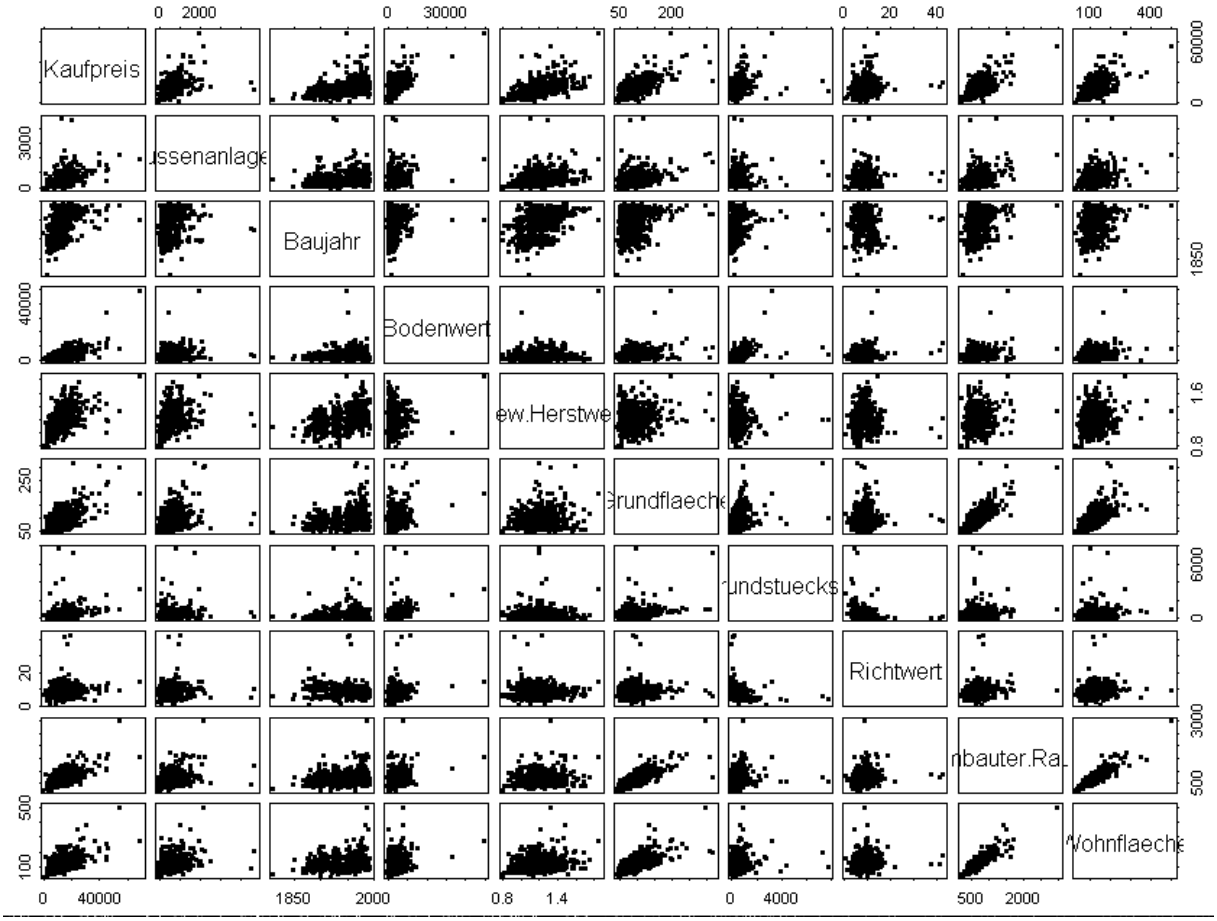
1. Modell (mit Originaldaten)

Während im nächsten Abschnitt transformierte Daten betrachtet werden, wurden die nun folgenden Berechnungen mit den Originaldaten durchgeführt. Auftretende DM-Beträge wurden normiert, indem sie durch den zum Verkaufszeitpunkt gültigen Preisindex dividiert wurden.

Voruntersuchung

Grafische Darstellung

Vor der eigentlichen Datenanalyse wollen wir uns einen optischen Eindruck der Daten verschaffen. Dazu dient die nachfolgende Abbildung. Hier sind die kuntinuierlichen Daten paarweise in Abhängigkeit voneinander aufgetragen. Generell kann man sagen, daß sich deutliche "Klumpen" bilden und daß es bei jedem Variablenpaar ausreißerverdächtige Datensätze gibt. Diese müßten in einer genaueren Untersuchung unter die Lupe genommen und gegebenenfalls entfernt werden.
Weiter fällt auf, daß manche Variablenpaare korreliert sind. Dies wird jeoch im nächsten Abschnitt noch näher zu untersuchen sein.



Korrelationen

Die paarweisen Korrelationen zwischen den einzelnen Variablen sind sehr gering. Das Maximum liegt bei 0,89 für die Variablen Wohnfläche/Umbauter Raum, was auch verständlich ist. Ebenfalls korreliert sind die Paare Grundfläche/Wohnfläche (0,7) sowie Grundfläche/Umbauter Raum (0,791). Es wird also nötig sein, zu untersuchen, welche der beiden Variablen Wohnfläche und Umbauter Raum in das Modell aufgenommen werden sollte.

	Kaufpreis	Aussen-anlagen	Baujahr	Boden-wert	Ges. Le-bens-dauer	Gew. Herstwert	Grund-flaeche	Grund-stuecks-groesse	Richtwert	Tageszahl	Umbauter Raum	Wohn-flaeche
Kaufpreis	1	0,47	0,476	0,579	0,162	0,492	0,596	0,294	0,196	-0,0425	0,625	0,588
Aussen-anlagen	0,47	1	0,215	0,229	0,0647	0,326	0,297	0,188	0,0673	-0,104	0,302	0,279
Baujahr	0,476	0,215	1	0,131	-7,18e-4	0,457	0,159	0,152	-0,0769	0,0427	0,113	0,164
Boden-wert	0,579	0,229	0,131	1	0,0678	0,108	0,385	0,457	0,175	-0,0352	0,331	0,304
Ges. Le-bens-dauer	0,162	0,0647	-7,18e-4	0,0678	1	0,156	0,0367	0,0233	0,0992	0,021	0,135	0,132
Gew. Herstwert	0,492	0,326	0,457	0,108	0,156	1	0,0878	0,0492	0,0317	-0,494	0,0932	0,128
Grund-flaeche	0,596	0,297	0,159	0,385	0,0367	0,0878	1	0,372	0,0951	-0,0114	0,791	0,7
Grund-stuecks-groesse	0,294	0,188	0,152	0,457	0,0233	0,0492	0,372	1	-0,225	-0,0491	0,162	0,161
Richtwert	0,196	0,0673	-0,0769	0,175	0,0992	0,0317	0,0951	-0,225	1	0,0845	0,209	0,165
Tages-zahl	-0,0425	-0,104	0,0427	-0,0352	0,021	-0,494	-0,0114	-0,0491	0,0845	1	0,00595	0,0123
Umbauter Raum	0,625	0,302	0,113	0,331	0,135	0,0932	0,791	0,162	0,209	0,00595	1	0,89
Wohn-flaeche	0,588	0,279	0,164	0,304	0,132	0,128	0,7	0,161	0,165	0,0123	0,89	1

PRESS-Statistik und t-Statistik für Einfach-Regression

Kontinuierliche Variablen

Die einzige Variable mit kontinuierlichen Werten, die eine PRESS-Statistik von mehr als 100 aufweist, ist die Variable Tageszahl. Sie liefert also das schlechteste Minimodell. Da außerdem auch die t-Statistik der Tageszahl im Minimodell sehr niedrig liegt (-1,081), scheidet diese Variable aus dem Verfahren aus. Ihre geringe Bedeutung läßt sich mit dem kurzen Beobachtungszeitraum begründen.
Alle anderen Variablen nehmen anscheinend größeren Einfluß auf den Kaufpreis. Ihre PRESS-Statistiken schwanken zwischen 61,07 und 97,66.
Es fällt auf, daß die t-Statistiken aller Variablen außer der Tageszahl sehr weit über dem Schwellenwert von 1,282 liegen. In der Tat könnte man das Niveau deutlich erhöhen. Bei 0,5% würde der Schwellenwert beispielsweise 2,576 betragen; auch bei diesem Niveau lägen alle t-Statistiken darüber.

Variable	PRESS-Statistik	t-Statistik
Aussenanlagen	78,13	13,54
Baujahr	77,61	13,75
Bodenwert	66,68	18,05
Ges.Lebensdauer	97,66	4,186
Gew.Herstwert	76,00	14,37
Grundflaeche	64,63	18,88
Grundstuecksgr	91,62	7,825
Richtwert	96,45	5,083
Tageszahl	100,1	-1,081
Umbauter.Raum	61,07	20,37
Wohnflaeche	65,57	18,50

Diskrete Variablen
Unter den diskreten Variablen finden sich immerhin vier, deren PRESS-Statistik über 100 liegt: Erbbau/privat, Geschoßzahl=1, Geschoßzahl=2,5 und Konstruktion. In allen Fällen liegt auch die t-Statistik so niedrig, daß wir diese Variablen nicht weiter zu betrachten brauchen.
Die Größenordnungen der PRESS- und der t-Statistiken lassen vermuten, daß die diskreten Variablen bei der Modellbildung eine geringe, aber nicht zu vernachlässigende Rolle spielen werden.

Variable	PRESS-Statistik	t-Statistik
Dachform.3	99,87	-1,69
Erbbau.privat	100,3	0,0395
Geschosszahl.1.0	100,1	1,03
Geschosszahl.1.5	99,50	-2,30
Geschosszahl.2.0	99,84	1,75
Geschosszahl.2.5	100,3	0,046
Konstruktion	100,3	0,191
Nutzung.1	89,80	-8,69
Nutzung.2	97,62	-4,22
Nutzung.3	85,68	10,5
Stellung.1	98,16	3,76
Stellung.3	99,23	-2,66
Stellung.4	97,98	-3,92
Stellung.5	99,45	2,36
Stellung.7	99,91	1,61
Wohngebiet.1	87,20	9,85
Wohngebiet.2	90,92	-8,17

Interaktionen
Da eine vollständige Liste aller berechneten Korrelationen recht unübersichtlich wäre, werden an dieser Stelle nur die Variablen angegeben, deren Interaktion eine größere Korrelation mit dem Kaufpreis aufweist als die beiden einzelnen Variablen. Von diesen wurden die PRESS-Statistiken berechnet. Für die Modellsuche wurden lediglich Variablen berücksichtigt, deren PRESS-Wert niedriger liegt als 100. In diesen Fällen liegt auch die t-Statistik unter dem Schwellenwert. Somit verbleiben 13 zu berücksichtigende Interaktionen.
In allen Fällen liegt die t-Statistik jenseits des Schwellenwerts für den zweiseitigen Test. Die PRESS-Statistiken hingegen liegen bis auf einen Fall sehr dicht bei 100. Von daher steht nicht zu erwarten, daß Interaktionen das endgültige Modell sehr stark beeinflussen werden.
Betrachtet man sich die folgende Liste, so bemerkt man, daß Variablen auftauchen, die im vorigen Abschnitt bereits entfernt wurden (Beispiel: die Variable Konstruktion wurde wegen den Werten der PRESS- und der t-Statistik nicht mehr berücksichtigt. Ihre Interaktion mit der Variablen Geschoßzahl.2.0 jedoch weist bei beiden Statistiken Werte auf, die es nahelegen, sie zu berücksichtigen).

Variablen	Korrelation zwischen Kaufpreis und...			PRESS-Statistik	t-Statistik
	Interaktion	1. Variable	2. Variable		
Geschosszahl.1.5:Dachform.3	-0,101	-0,0900	-0,0664	99,28	-2,58
Geschosszahl.2.0:Dachform.3	0,0849	0,0686	-0,0664	99,59	2,17
Geschosszahl.2.5:Erbbau.privat	0,0204	0,00180	0,00155	100,2	0,518
Konstruktion:Erbbau.privat	0,00817	0,00752	0,00155	100,3	0,208
Konstruktion:Geschosszahl.1.0	0,0426	0,00752	0,0403	100,1	1,08
Konstruktion:Geschosszahl.1.5	-0,0920	0,00752	-0,0900	99,46	-2,35
Konstruktion:Geschosszahl.2.0	0,0732	0,00752	0,0686	99,77	1,87
Nutzung.2:Konstruktion	-0,164	-0,164	0,00752	97,62	-4,22
Stellung.3:Erbbau.privat	-0,182	-0,104	0,00155	96,99	-4,70
Stellung.3:Geschosszahl.1.5	-0,163	-0,104	-0,0900	97,66	-4,19
Stellung.3:Konstruktion	-0,107	-0,104	0,00752	99,15	-2,74
Stellung.4:Geschosszahl.1.5	-0,163	-0,152	-0,0900	97,65	-4,19
Stellung.5:Erbbau.privat	0,101	0,0924	0,00155	99,29	2,57
Stellung.5:Konstruktion	0,0977	0,0924	0,00752	99,35	2,50
Stellung.7:Konstruktion	0,0638	0,0633	0,00752	99,90	1,63
Wohngebiet.1:Nutzung.3	0,422	0,362	0,382	82,45	11,8

Ausreißer

Da die Daten nicht zentriert (also transformiert) sind, enthalten sie eine sehr große Zahl an Datensätzen, die in einer oder mehreren Komponenten stark (d. h. um mehr als das Doppelte der Standardabweichung) vom Mittelwert abweichen. Daher macht es an dieser Stelle keinen Sinn, diese näher zu untersuchen. Bei der Analyse des Modells werden sich einige Datensätze angeben lassen, die sich sehr stark von der Mehrheit der anderen Datensätze unterscheiden. Diese werden dann aussortiert.

Ergebnis der Auto-Regression

Insgesamt stehen nun also 37 Variablen für die Modellbildung zur Verfügung: 11 kontinuierliche, 13 Dummy-Variablen und 13 Interaktionen.

1. Durchlauf: Ohne "Wohnfläche"

In einem ersten Durchlauf des automatischen Regressionsverfahrens wurde die Variable "umbauter Raum" ins Modell aufgenommen, die Variable "Wohnfläche" nicht. Das Verfahren liefert folgendes Ergebnis (man beachte, daß die Variablen Kaufpreis, Gew.Herstwert, Bodenwert und Aussenanlagen nicht in DM angegeben sind, sondern daß sie durch den Preisindex zum Zeitpunkt des Verkaufs dividiert wurden, daß sie also gewissermaßen normiert sind):

```
Residuals:
    Min     1Q Median    3Q   Max
 -15472 -1868  29.51 1806 14181

Coefficients:
                   Value  Std. Error     t value    Pr(>|t|)
  (Intercept) -93135.7124   9020.3803    -10.3250      0.0000
Umbauter.Raum     12.0046      0.5960     20.1424      0.0000
Gew.Herstwert  11953.5879    960.1574     12.4496      0.0000
    Bodenwert      0.7699      0.0431     17.8585      0.0000
      Baujahr     41.5489      4.8586      8.5516      0.0000
Aussenanlagen      1.9981      0.3323      6.0135      0.0000
   Stellung.7  -2200.2066    457.3153     -4.8111      0.0000
 Wohngebiet.1   1474.7888    305.9202      4.8208      0.0000
   Stellung.3  -1188.4079    369.1520     -3.2193      0.0014

Residual standard error: 3372 on 639 degrees of freedom
Multiple R-Squared: 0.7842
F-statistic: 290.3 on 8 and 639 degrees of freedom, the p-value is 0
```

Dieses Modell erklärt 78,42% der Varianz des Kaufpreises. Die Standardabweichung der Vorhersage für den Kaufpreis liegt bei 3372. Mit 290,3 liegt die F-Statistik deutlich über dem Schwellenwert 2,54.

2. Durchlauf: Ohne "umbauter Raum"

Das gleiche Verfahren, diesmal jedoch mit der Variablen "Wohnfläche" anstatt "umbauter Raum", liefert das folgende Ergebnis:

```
Residuals:
    Min     1Q Median    3Q   Max
 -15305 -1967 -57.36 1847 16393

Coefficients:
                   Value  Std. Error     t value    Pr(>|t|)
  (Intercept) -79903.4559   9337.4127     -8.5573      0.0000
 Grundflaeche     43.6058      5.9795      7.2926      0.0000
Gew.Herstwert  12249.3034    998.8290     12.2637      0.0000
    Bodenwert      0.7544      0.0453     16.6395      0.0000
      Baujahr     34.2854      5.0373      6.8063      0.0000
  Wohnflaeche     37.6052      4.3660      8.6131      0.0000
Aussenanlagen      2.1942      0.3425      6.4062      0.0000
   Stellung.7  -2405.2403    470.2186     -5.1152      0.0000
 Wohngebiet.1   1405.3675    314.7262      4.4654      0.0000

Residual standard error: 3490 on 639 degrees of freedom
Multiple R-Squared: 0.7688
F-statistic: 265.7 on 8 and 639 degrees of freedom, the p-value is 0
```

Dieses Modell erklärt 76,88% der Varianz des Kaufpreises. Die Standardabweichung der Vorhersage für den Kaufpreis liegt bei 3490. Die F-Statistik liegt bei 265,7, also abermals deutlich über dem Schwellenwert von 2,54.

Analyse

Da das Ergebnis des zweiten Durchlaufs geringfügig schlechter ist als das des ersten (es wird etwa 1% weniger der Varianz des Kaufpreises erklärt, die Standardabweichung der Vorhersage vom wahren Wert liegt um mehr als 100 höher), beschränken wir uns im folgenden auf das erste Modell, betrachten also nur die Variable "umbauter Raum" und nicht mehr "Wohnfläche".
Bei der Analyse des Ergebnisses fällt auf, daß es mehrere Datensätze gibt, die sich stark von der Mehrheit der anderen Datensätze unterscheiden. Diese Datensätze können das Ergebnis verfälschen, denn lineare Verfahren sind in dieser Hinsicht nicht besonders robust. Deshalb ist es wichtig, diese Ausreißer zu erkennen und aus der Gesamtheit des Datenmaterials herauszunehmen. Nachdem drei Datensätze grafisch als Ausreißer identifiziert und entfernt wurden (dazu wurde der Plot der Residuen gegen den Kaufpreis verwendet, der an dieser Stelle nicht gezeigt wird), wurde das automatische Regressionsverfahren nochmals gestartet. Das Ergebnis dieses zweiten Durchlaufs lautet:

```
Residuals:
    Min     1Q Median    3Q   Max
 -10546 -1823  46.77 1770 11460

Coefficients:
                  Value  Std. Error     t value    Pr(>|t|)
  (Intercept) -92534.3332   8502.5575    -10.8831      0.0000
Umbauter.Raum     11.9965      0.5614     21.3683      0.0000
Gew.Herstwert  10843.1682    914.3671     11.8587      0.0000
    Bodenwert      0.7377      0.0408     18.0928      0.0000
      Baujahr     41.8036      4.5801      9.1272      0.0000
Aussenanlagen      2.5554      0.3406      7.5028      0.0000
 Wohngebiet.1   1521.6773    289.0929      5.2636      0.0000
   Stellung.7  -2136.4619    431.4832     -4.9514      0.0000
   Stellung.3  -1161.7584    347.8135     -3.3402      0.0009

Residual standard error: 3175 on 635 degrees of freedom
Multiple R-Squared: 0.8005
F-statistic: 318.6 on 8 and 635 degrees of freedom, the p-value is 0
```

Dieses Modell erklärt 80,05% der Varianz des Kaufpreises, die Standardabweichung der Vorhersage des Kaufpreises beträgt 3175. Die F-Statistik liegt wieder sehr deutlich über dem Schwellenwert von 2,54. Durch das Entfernen der Ausreißer konnten also sämtliche Modellstatistiken verbessert werden.

Die von diesem Modell gelieferte Formel für den Kaufpreis lautet:

Kaufpreis	=	12 *	Umbauter. Raum	+	10800 *	Gew.Herstwert	+ 0,737 *	Bodenwert
		+ 41,8 *	Baujahr	+	2,56 *	Aussenanlagen	+1520 *	Wohngebiet.1
		- 2140 *	Stellung.7	-	1160 *	Stellung.3	- 92500	

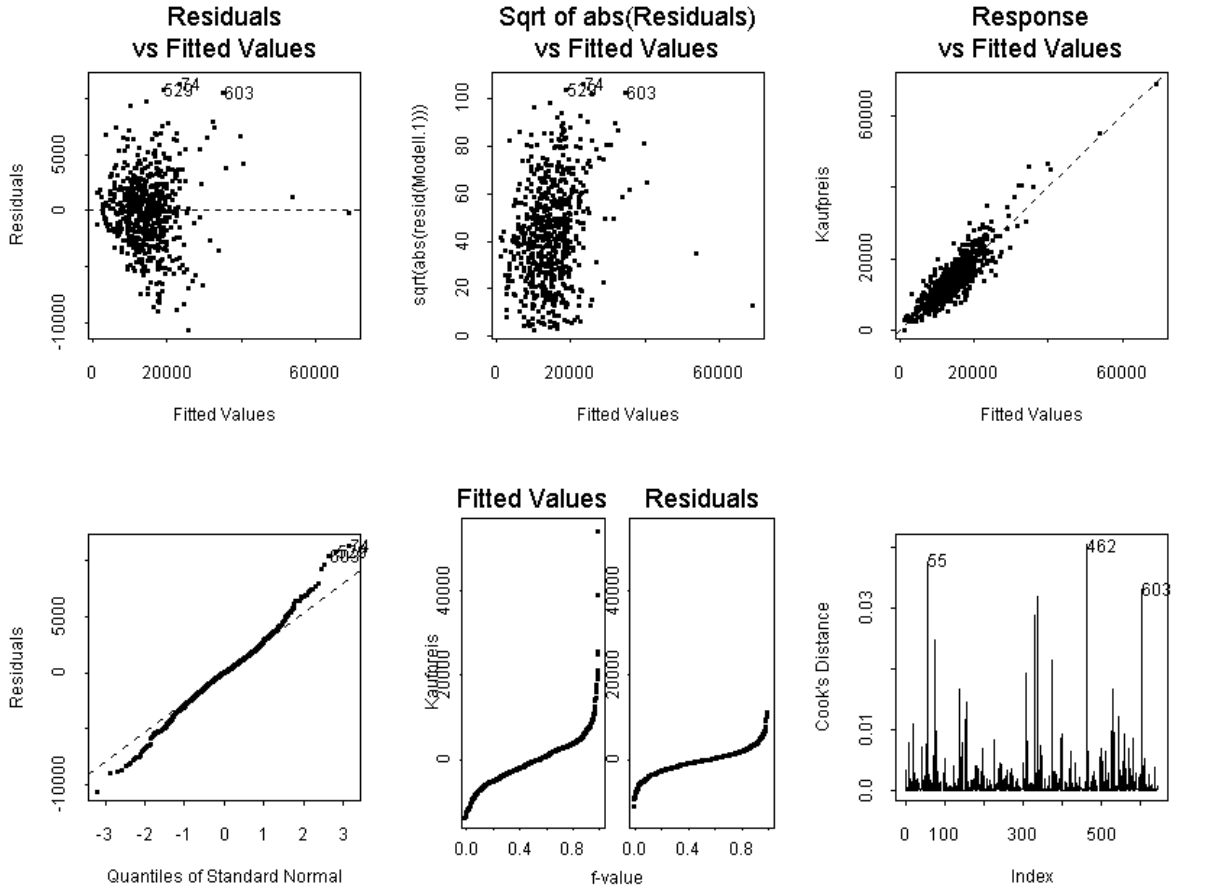
Der Einfluß, den die einzelnen Variablen auf den Kaufpreis nehmen, läßt sich nun nicht einfach an dem jeweiligen Koeffizienten ablesen. Vielmehr kommt es auch darauf an, welche Werte die jeweilige Variable annimmt. Zu diesem Zweck betrachten wir jeweils den minimalen und den maximalen sich aus unseren Daten ergebenden Summanden:

Variable	Minimaler Summand	Maximaler Summand
Umbauter.Raum	2820	36000
Gew.Herstwert	8697	20100
Bodenwert	44,70	36800
Baujahr	75600	83300
Aussenanlagen	0	11900
Wohngebiet.1	0	1520
Stellung.7	0	-2140
Stellung.3	0	-1160

Man sieht, daß Baujahr, gewöhnlicher Herstellungswert und umbauter Raum stets eine große Rolle spielen, während die anderen Variablen nicht ganz so viel zum Ergebnis beitragen.

In der folgenden Grafik werden nun einige diagnostische Plots des Modells gezeigt. Ausreißerverdächtige Datensätze werden markiert.

1. Die erste Grafik zeigt die Residuen in Abhängigkeit von den Vorhersagewerten. Hier sollte sich keine Struktur zeigen (weißes Rauschen). Es sieht jedoch so aus, als ob größere Vorhersagewerte auch mit größeren Residuen behaftet sind. Außerdem gewinnt man den Eindruck, daß die Abweichungen nach unten größer ausfallen als die Abweichungen nach oben.
2. In der zweiten Grafik wird die Quadratwurzel des Absolutwerts der Residuen gegen die gefitteten Werte dargestellt. Hier gilt das gleiche wie oben.
3. Die dritte Grafik sollte im Idealfall eine Gerade zeigen, denn hier wird der tatsächliche Kaufpreis gegen den gefitteten Wert aufgetragen. Hier macht sich wieder bemerkbar, daß die Daten nicht zentriert sind.
4. Daß die Modellannahme, die Residuen seien Standardnormalverteilt, nicht besonders gut erfüllt ist, zeigt sich in der vierten Grafik. Hier werden die Quantile der Standardnormalverteilung gegen die entsprechenden empirischen Quantile aufgetragen. Gerade an den Rändern zeigt sich eine deutliche Abweichung von der theoretisch zu erwartenden Geraden.
5. In der fünften Grafik werden im ersten Teil die Abweichungen der gefitteten Werte von ihrem Mittelwert und im zweiten Teil die Abweichung der Residuen von ihrem Mittelwert dargestellt. Es stellt sich auch hier wieder heraus, daß die Residuen nicht um 0 zentriert liegen. Ihre Spannweite ist jedoch geringer als die der Vorhersagewerte.
6. In der letzten Grafik kann man den Einfluß der einzelnen Datensätze auf das Ergebnis der Regression erkennen. Hier sieht man, daß sich alle Datensätze in etwa ähnlich verhalten, was für die Homogenität der Daten spricht.



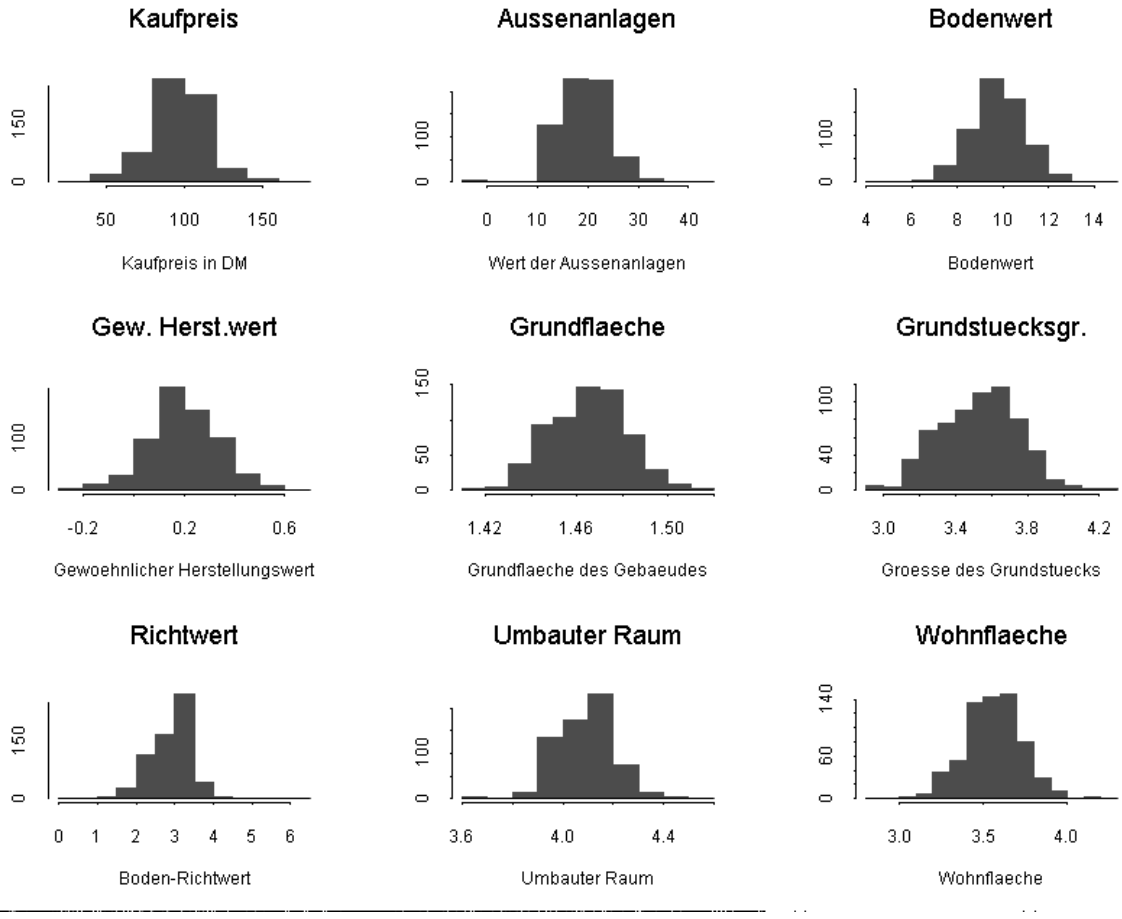
2. Modell (mit transformierten Daten)

Transformation

Die bisher verwendeten statistischen Verfahren sowie zahlreiche weiterführende theoretische Überlegungen sind speziell auf normalverteilte Zufallsgrößen zugeschnitten. Darauf wurde bisher keine Rücksicht genommen, obwohl man sehr leicht sieht, daß diese Voraussetzung nicht erfüllt ist (siehe etwa Histogramme: die Histogramme normalverteilter Zufallsgrößen bilden Gauß'sche Glockenkurven, sind also symmetrisch).
Nun sollen die kontinuierlichen Daten einer Transformation unterworfen werden, die diesen Mangel zumindest teilweise ausgleicht (für die diskreten Variablen ist dies nicht nötig, da die meisten sowieso nur die Werte 0 und 1 annehmen). Mit der sogenannten Box-Cox-Transformation werden die einzelnen Variablen "zentriert". Auf diese Weise wird ihre Varianz stabilisiert und sie verhalten sich ähnlich wie normalverteilte Daten.
Transformieren bedeutet, daß die Zahlenwerte der Daten durch andere Zahlen ersetzt werden. Diese neuen Zahlenwerte werden nach einer bestimmten Funktion, der Transformationsfunktion, berechnet. Im Falle der Box-Cox-Transformation hängt diese Funktion außer von den Daten selbst noch von einem Parameter ab, der so gewählt wird, daß die erhaltenen zentrierten Daten möglichst "gut" zentriert werden. Zu jeder Variablen erhält man auf diese Weise einen Transformationsparameter. Die folgende Tabelle enthält diese Parameter, soweit sie berechnet wurden:

Variable	Parameter	Bemerkung
Kaufpreis	0,380	
Aussenanlagen	0,309	Vor der Transformation wurde 1 addiert
Baujahr		keine Transformation durchgeführt
Bodenwert	0,0548	
Ges.Lebensdauer		keine Transformation durchgeführt
Gew.Herstwert	0,0903	
Grundflaeche	-0,644	
Grundstuecksgr	-0,194	
Richtwert	0,245	
Tageszahl		keine Transformation durchgeführt
Umbauter.Raum	-0,156	
Wohnflaeche	-0,131	

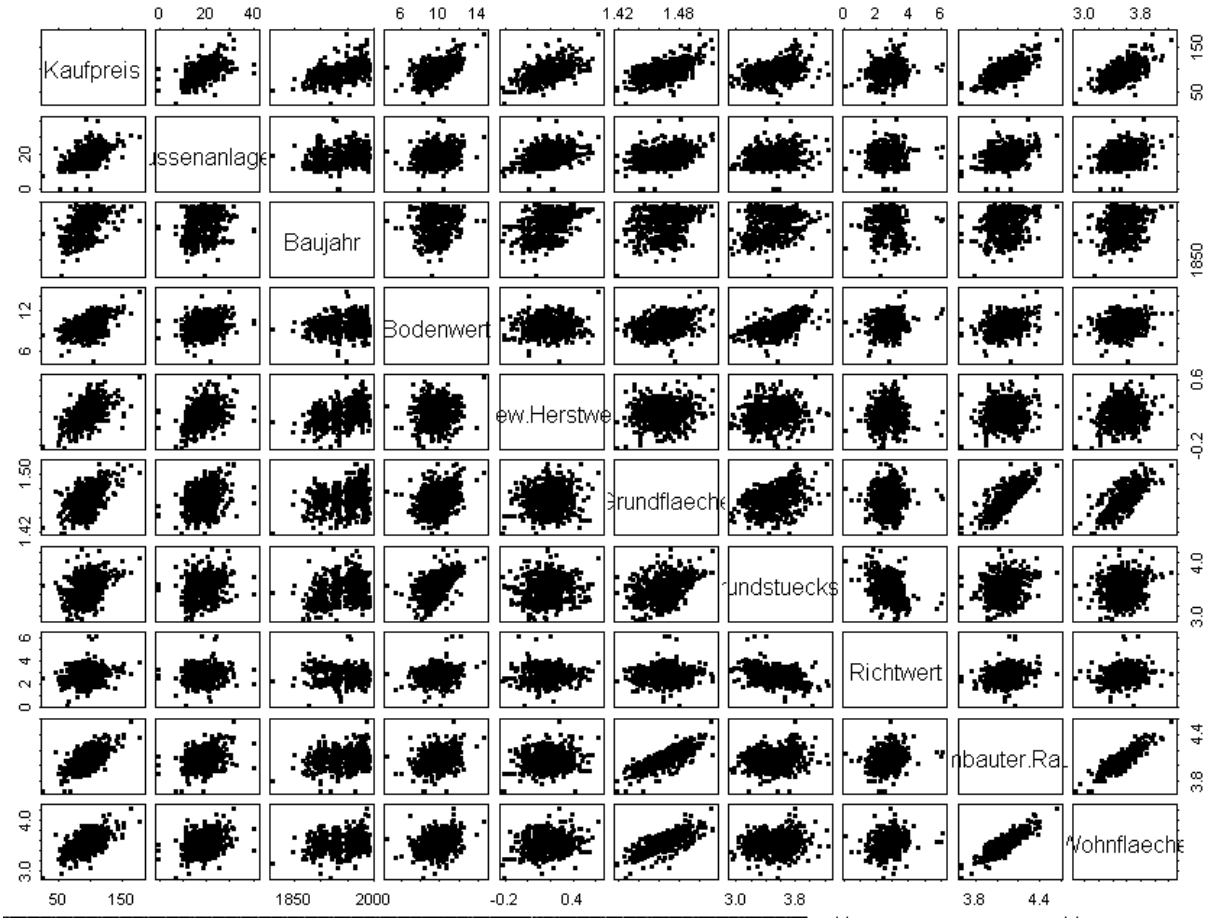
Um die Wirkung der Transformation zu veranschaulichen, sollen die Häufigkeitsverteilungen der transformierten Variablen abgebildet werden. Man erkennt, daß die Histogramme nach der Transformation einer Glockenkurve viel ähnlicher sehen.



Voruntersuchung

Grafische Darstellung

Abermals soll vor der eigentlichen Datenanalyse ein optischer Eindruck der Daten gewonnen werden. Wie schon bei den nichttransformierten Daten dient dazu der paarweise Plot der Daten gegeneinander. Der durch die Transformation erzielte Effekt wird bereits hier deutlich: die Datenpunte zeigen viel weniger Struktur, sie liegen gleichmäßig um ein Zentrum gestreut in der Ebene. Allerdings sind einige Datenpaare nach wie vor korreliert.



Korrelationen

Die Werte der paarweisen Korrelationen ändern sich durch die Transformation zwar, nicht aber die Größenordnungen. Nach wie vor sind die Korrelationen relativ niedrig, lediglich in drei Fällen liegen sie recht hoch. Der Grenzwert von 0,8 wird diesmal von zwei Variablenpaaren überschritten: Wohnfläche/Umbauter Raum (0,889) sowie Grundfläche/Umbauter Raum (0,818). Da sich gewissermaßen eine Dreiecksbeziehung zwischen den Korrelationen ergibt, werden wir vier verschiedene Modelle ausrechnen müssen: Ein Modell, das alle Variablen enthält, eines ohne Wohnfläche, eines ohne Grundfläche und eines ohne umbauten Raum.

	Kaufpreis	Aussen-anlagen	Baujahr	Boden-wert	Ges. Lebens-dauer	Gew. Herstwert	Grund-flaeche	Grund-stuecksgr	Richtwert	Tageszahl	Umbauter Raum	Wohn-flaeche
Kaufpreis	1	0,55	0,514	0,48	0,235	0,548	0,519	0,388	0,261	-0,0327	0,592	0,586
Aussen-anlagen	0,55	1	0,298	0,235	0,105	0,405	0,272	0,288	0,107	-0,0762	0,293	0,285
Baujahr	0,514	0,298	1	0,135	-7,18e-4	0,459	0,106	0,353	-0,0466	0,0427	0,108	0,173
Boden-wert	0,48	0,235	0,135	1	0,16	0,0843	0,401	0,505	0,18	-0,0302	0,362	0,324
Ges. Lebens-dauer	0,235	0,105	-7,18e-4	0,16	1	0,178	0,0702	-0,00651	0,114	0,021	0,194	0,197
Gew. Herstwert	0,548	0,405	0,459	0,0843	0,178	1	0,0586	0,121	0,0907	-0,492	0,11	0,164
Grund-flaeche	0,519	0,272	0,106	0,401	0,0702	0,0586	1	0,375	0,143	6,12e-4	0,818	0,736
Grund-stuecksgr	0,388	0,288	0,353	0,505	-0,00651	0,121	0,375	1	-0,339	-0,0595	0,229	0,2
Richtwert	0,261	0,107	-0,0466	0,18	0,114	0,0907	0,143	-0,339	1	0,0821	0,253	0,211
Tageszahl	-0,0327	-0,0762	0,0427	-0,0302	0,021	-0,492	6,12e-4	-0,0595	0,0821	1	0,0199	0,0178
Umbauter Raum	0,592	0,293	0,108	0,362	0,194	0,11	0,818	0,229	0,253	0,0199	1	0,889
Wohn-flaeche	0,586	0,285	0,173	0,324	0,197	0,164	0,736	0,2	0,211	0,0178	0,889	1

PRESS-Statistik und t-Statistik für Einfach-Regression

Kontinuierliche Variablen

Wie schon bei den nicht-transformierten Daten, liegt die PRESS-Statistik der Variablen Tageszahl auch nach der Transformation über 100. Abermals legt die t-Statistik es nahe, diese Variable nicht weiter zu berücksichtigen.

Sowohl die PRESS- als auch die t-Statistiken der übrigen Variablen haben in etwa die gleiche Größenordnung wie vor der Transformation.

Variable	PRESS-Statistik	t-Statistik
Aussenanlagen	69,97	16,74
Baujahr	73,76	15,25
Bodenwert	77,24	13,89
Ges.Lebensdauer	94,79	6,133
Gew.Herstwert	70,17	16,66
Grundflaeche	73,30	15,43
Grundstuecksgr	85,18	10,71
Richtwert	93,48	6,868
Tageszahl	100,2	-0,8323
Umbauter.Raum	65,10	18,69
Wohnflaeche	65,81	18,40

Diskrete Variablen
Die vier diskreten Variablen, die im Fall nicht-transformierter Daten bereits PRESS-Statistiken von über 100 aufwiesen, tun dies auch nach der Transformation. Außerdem kommt noch eine weitere Variable hinzu: Stellung.7. In allen fünf Fällen liegt die t-Statistik so niedrig, daß die Variablen aus dem Regressionsverfahren herausgenommen werden können.

Variable	PRESS-Statistik	t-Statistik
Dachform.3	99,89	-1,652
Erbbau.privat	100,3	-0,4113
Geschosszahl.1.0	100,3	0,3427
Geschosszahl.1.5	99,07	-2,840
Geschosszahl.2.0	99,28	2,591
Geschosszahl.2.5	100,3	0,3519
Konstruktion	100,3	-0,2787
Nutzung.1	83,04	-11,59
Nutzung.2	98,58	-3,358
Nutzung.3	85,13	10,73
Stellung.1	98,12	3,793
Stellung.3	99,07	-2,846
Stellung.4	98,50	-3,474
Stellung.5	99,30	2,565
Stellung.7	100,1	1,159
Wohngebiet.1	87,32	9,804
Wohngebiet.2	91,53	-7,872

Interaktionen

Bei den transformierten Daten ergeben sich 18 Interaktionen, die stärker mit dem Kaufpreis korreliert sind, als die Einzelvariablen (bei den Originaldaten waren es 17). Zwei davon können aufgrund der hohen PRESS-Statistik und der niedrigen t-Statistik vernachlässigt werden, eine weitere weist eine PRESS-Statistik auf die knapp über 100 liegt, ihre t-Statistik kann jedoch gerade noch akzeptiert werden.
Es bleiben also 16 Interaktionen übrig, die bei der Modellbildung zu berücksichtigen sind.

Variablen	Korrelation zwischen Kaufpreis und ...			PRESS-Statistik	t-Statistik
	Interaktion	1. Variable	2. Variable		
Geschosszahl.1.5:Dachform.3	-0,124	-0,111	-0,0649	98,77	-3,170
Geschosszahl.2.0:Dachform.3	0,114	0,101	-0,0649	99,00	2,925
Geschosszahl.2.5:Erbbau.privat	0,0260	0,0138	-0,0162	100,2	0,6616
Konstruktion:Dachform.3	-0,0738	-0,0110	-0,06487	99,76	-1,881
Konstruktion:Geschosszahl.1.0	0,0158	-0,0110	0,0135	100,3	0,4015
Nutzung.2:Konstruktion	-0,138	-0,131	-0,0110	98,41	-3,531
Stellung.1:Konstruktion	0,148	0,148	-0,0110	98,11	3,806
Stellung.3:Erbbau.privat	-0,228	-0,111	-0,0162	95,07	-5,965
Stellung.3:Geschosszahl.1.5	-0,209	-0,111	-0,111	95,93	-5,428
Stellung.3:Konstruktion	-0,117	-0,111	-0,0110	98,95	-2,982
Stellung.4:Geschosszahl.1.5	-0,179	-0,135	-0,111	97,10	-4,618
Stellung.4:Konstruktion	-0,139	-0,135	-0,0110	98,38	-3,563
Stellung.5:Erbbau.privat	0,111	0,100	-0,0162	99,08	2,830
Stellung.5:Konstruktion	0,103	0,100	-0,0110	99,23	2,644
Stellung.7:Geschosszahl.1.0	-0,0583	0,0456	0,0135	99,97	-1,484
Stellung.7:Geschosszahl.2.5	0,0505	0,0456	0,0138	100,1	1,286
Wohngebiet.1:Nutzung.3	0,418	0,360	0,389	82,78	11,70
Wohngebiet.2:Konstruktion	-0,297	-0,296	-0,0110	91,46	-7,905

Ausreißer

Bei einer normalverteilten Zufallsgröße erwartet man, daß der Anteil der Werte, die um mehr als die Standardabweichung vom Mittelwert abweichen, etwa 31,7% beträgt. Weicht der Anteil der Ausreißer in einer Stichprobe signifikant von dieser Zahl ab, so kann man davon ausgehen, daß die Annahme der Normalverteilung verletzt ist.
Um festzustellen, ob eine "signifikante" Abweichung vom erwarteten Wert vorliegt, werden 99,5%-Konfidenzintervalle zum Ausreißeranteil berechnet.
Lediglich beim Kaufpreis liegt der erwartete Wert von 0,317 nicht in dem berechneten Konfidenzintervall (untere Grenze: 0,218, obere Grenze: 0,298).

	Mittelwert	Median	Std.abw.	Ausreißer	KI unten	KI oben
Kaufpreis	96,6	97,4	18,1	0,258	0,218	0,298
Aussenanlagen	19,3	19,7	4,82	0,296	0,255	0,338
Bodenwert	9,8	9,73	1,19	0,318	0,275	0,361
Gew.Herstwert	0,194	0,189	0,138	0,293	0,252	0,335
Grundflaeche	1,46	1,47	0,0169	0,352	0,308	0,396
Grundstuecksgr	3,53	3,55	0,217	0,356	0,313	0,4
Richtwert	2,9	3,02	0,602	0,236	0,197	0,275
Umbauter.Raum	4,09	4,1	0,113	0,31	0,268	0,353
Wohnflaeche	3,56	3,56	0,173	0,29	0,249	0,332

Ergebnis der Auto-Regression

Für das automatische Regressionsverfahren standen 10 kontinuierliche und 12 diskrete Variablen sowie 15 Interaktions-Variablen zur Verfügung. Es sei an dieser Stelle darauf hingewiesen, daß es sich bei den Zahlenwerten, die in den nächsten Abschnitten auftauchen, um transformierte Zahlen handelt, daß es also keinen Sinn macht, von irgendwelchen Einheiten (z. B. DM) zu sprechen.

1. bis 3. Durchlauf: Kompletter Variablensatz, ohne "Wohnfläche" und ohne "Grundfläche"

Die drei Aufrufe liefern identische Ergebnisse, was daran liegt, daß die Variable "umbauter Raum" den größten Einfluß ausübt und die anderen beiden Variablen gewissermaßen überdeckt.
Das Ergebnis der Berechnung lautet:

```
Residuals:
    Min      1Q Median    3Q   Max
 -61.66 -5.024 0.1002 5.101 33.08

Coefficients:
                              Value Std. Error   t value  Pr(>|t|)
                (Intercept) -445.6694   27.9996   -15.9170    0.0000
              Umbauter.Raum   59.6047    3.5417    16.8296    0.0000
              Gew.Herstwert   38.5605    3.2011    12.0459    0.0000
                    Baujahr    0.1220    0.0131     9.2817    0.0000
                Bodenwert      3.2466    0.3352     9.6855    0.0000
              Aussenanlagen    0.6366    0.0860     7.4057    0.0000
                  Richtwert    3.3659    0.6375     5.2798    0.0000
                Stellung.1     2.6087    0.8550     3.0512    0.0024
Int.Wohngebiet2.Konstruktion   -2.3980    0.8309    -2.8859    0.0040

Residual standard error: 9.022 on 639 degrees of freedom
Multiple R-Squared: 0.7537
F-statistic: 244.4 on 8 and 639 degrees of freedom, the p-value is 0
```

Das gefundene Modell erklärt also 75,37% der Schwankung des Kaufpreises, die Standardabweichung bei der Vorhersage beträgt 9,022. Mit 244,4 überschreitet die F-Statistik den Schwellenwert von 2,54 deutlich.

4. Durchlauf: Ohne "umbauter Raum"

Das Ergebnis dieses Durchlaufs unterscheidet sich deutlich von den ersten drei. Zum einen ändert sich durch die Veränderung des Variablensatzes gewissermaßen die Wichtigkeit der einzelnen Variablen. Die Stellung des Gebäudes spielt in diesem Modell keine Rolle mehr, dafür taucht die Variable Wohngebiet auf.

```
Residuals:
    Min      1Q  Median    3Q   Max
 -61.07 -5.096 0.01551 5.276 34.79

Coefficients:
                   Value Std. Error   t value  Pr(>|t|)
    (Intercept) -543.1605   50.9497   -10.6607    0.0000
    Wohnflaeche   24.7745    3.2482     7.6271    0.0000
  Gew.Herstwert   33.7828    3.1954    10.5724    0.0000
      Bodenwert    3.2932    0.3415     9.6428    0.0000
        Baujahr    0.1245    0.0132     9.4241    0.0000
  Aussenanlagen    0.6876    0.0868     7.9228    0.0000
      Richtwert    3.4814    0.6257     5.5638    0.0000
   Grundflaeche  149.3140   33.5468     4.4509    0.0000
Ges.Lebensdauer    0.2995    0.0871     3.4403    0.0006
   Wohngebiet.2   -2.7653    0.8188    -3.3774    0.0008

Residual standard error: 9.166 on 638 degrees of freedom
Multiple R-Squared: 0.7462
F-statistic: 208.4 on 9 and 638 degrees of freedom, the p-value is 0
```

Dieses Modell erklärt nur 74,62% der Schwankung des Kaufpreises. Auch die Standardabweichung des Vorhersagewertes vom wahren Wert liegt mit 9,166 etwas höher. Mit 208,4 überschreitet die F-Statistik den Schwellenwert von 2,44 deutlich.

Analyse

Anhand des Residualplots wurden drei Datensätze als Ausreißer erkannt und entfernt. Danach wurde ein neues Modell berechnet. Das Ergebnis lautet:

```
Residuals:
    Min      1Q  Median    3Q   Max
 -25.69 -4.999 0.06239 5.102 26.89

Coefficients:
                              Value Std. Error   t value  Pr(>|t|)
                (Intercept) -442.8930   26.4467   -16.7467    0.0000
              Umbauter.Raum   59.8873    3.3456    17.9002    0.0000
              Gew.Herstwert   38.7844    3.0278    12.8093    0.0000
                    Baujahr    0.1210    0.0124     9.7512    0.0000
                Bodenwert      2.9028    0.3201     9.0687    0.0000
              Aussenanlagen    0.6705    0.0812     8.2534    0.0000
                  Richtwert    3.5875    0.6041     5.9386    0.0000
                 Stellung.1    2.8240    0.8087     3.4920    0.0005
Int.Wohngebiet2.Konstruktion   -2.5149    0.7870    -3.1955    0.0015

Residual standard error: 8.517 on 636 degrees of freedom
Multiple R-Squared: 0.7742
F-statistic: 272.6 on 8 and 636 degrees of freedom, the p-value is 0
```

Von diesem Modell werden also 77,42% der Schwankung des Kaufpreises erklärt. Die Standardabweichung der Vorhersage liegt bei 8,517. Mit 272,6 wird der Schwellenwert von 2,54 für die F-Statistik überschritten. Auch hier brachte das Weglassen der Ausreißer leichte Verbesserungen für alle Modellstatistiken.

Die von diesem Modell gelieferte Formel für den Kaufpreis lautet also (zur Verdeutlichung werden transformierte Variablen mit einem ' versehen):

$$\begin{aligned}\text{Kaufpreis}' = {} & 59{,}9 * \text{Umbauter.Raum}' + 38{,}8 * \text{Gew.Herstwert}' + 0{,}121 * \text{Baujahr} \\ & + 2{,}9 * \text{Bodenwert}' + 0{,}671 * \text{Aussenanlagen}' + 3{,}59 * \text{Richtwert}' \\ & + 2{,}82 * \text{Stellung.1} - 2{,}51 * \text{Int.Wohngebiet2.Konstruktion} \\ & - 443\end{aligned}$$

oder, wenn man die Transformation explizit durchführt:

$$\begin{aligned}\text{Kaufpreis} = {} & \big((-382 * \text{Umbauter.Raum}^{-0{,}156} + 430 * \text{Gew.Herstwert}^{0{,}0903} + 0{,}121 * \text{Baujahr} + 52{,}9 * \text{Bodenwert}^{0{,}0548} \\ & + 3{,}17 * \text{Aussenanlagen}^{0{,}309} + 14{,}7 * \text{Richtwert}^{0{,}245} + 2{,}82 * \text{Stellung.1} - 2{,}51 * \text{Int.Wohngebiet2.Konstruktion} - 562\big) \\ & * 0{,}380 + 1\big)^{2{,}63}\end{aligned}$$

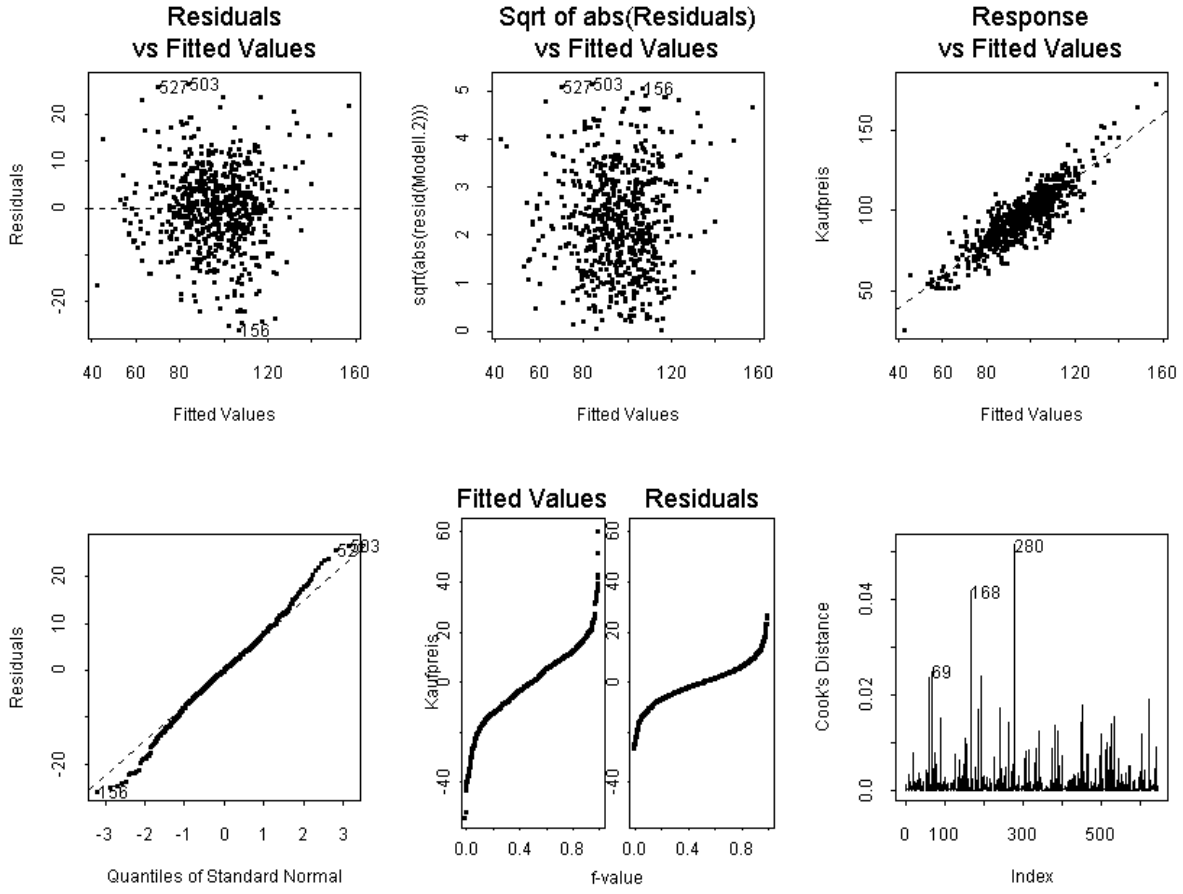
Auch hier wollen wir wieder die minimalen und die maximalen Summanden berechnet werden, um den Einfluß der Variablen auf den Kaufpreis abschätzen zu können:

Variable	Minimaler Summand	Maximaler Summand
Umbauter.Raum	220	274
Gew.Herstwert	-8,47	24,7
Baujahr	219	241
Bodenwert	13,4	42,9
Aussenanlagen	0	27,4
Richtwert	0,768	22,2
Stellung.1	0	2,82
Int.Wohngebiet2.Konstruktion	0	-2,51

Der umbaute Raum und das Baujahr tragen also sehr viel zum Kaufpreis bei, der gewöhnlicher Herstellungswert, der Bodenwert sowie der Wert der Außenanlagen tragen zwar etwas weniger bei, sind jedoch auch nicht zu vernachlässigen.

Im folgenden Bild werden nun wieder die bereits bei der Analyse des Modells mit nicht-transformierten Daten beschriebenen diagnostischen Plots des Modells gezeigt.

1. Die Residuen zeigen keine Abhängigkeit von den Vorhersagewerten.
2. Auch die Quadratwurzel des Absolutwerts der Residuen ist anscheinend unabhängig von den gefitteten Werten.
3. Zwischen Vorhersagewerten und wahrem Kaufpreis besteht der zu erwartende lineare Zusammenhang. Hier erkennt man ebenfalls kein Muster.
4. Daß die Modellannahme, die Residuen seien normalverteilt, auch nach der Transformation nicht besonders gut erfüllt ist, zeigt sich in der vierten Grafik. Nach wie vor fällt an den Rändern eine deutliche Abweichung von der theoretisch zu erwartenden Geraden auf.
5. Die Aussage der fünften Grafik ist wiederum sehr zufriedenstellend, denn sie zeigt, daß die Streuung der Residuen deutlich kleiner ist als die der Vorhersagewerte.
6. In der letzten Grafik schließlich werden die standardisierten Residuen der einzelnen Datensätze gezeigt. Gerade hier kann man besonders gut Ausreißer erkennen. Man sieht, daß sich alle Datensätze in etwa gleich verhalten.



Vergleich

In beiden Modellen wird der Kaufpreis durch acht Variablen bestimmt. In das aus den nicht-transformierten Daten berechnete Modell gehen fünf kontinuierliche Variablen ein. Die gleichen fünf Variablen gehen auch in das aus den transformierten Daten berechnete Modell ein.
Obwohl bei dem Modell mit nicht-transformierten Daten die Modellstatistiken geringfügig besser sind als beim Modell mit transformierten Daten, ist letzterem aufgrund des Verhaltens der Residuen aus der Sicht des Statistikers der Vorzug zu geben. Der zusätzliche Rechenaufwand, der bei einer Vorhersage durch die Transformation entsteht, wird jedoch unter Umständen die Entscheidung zugunsten des einfacheren Modells fallen lassen.
Anhand zweier praktischer Beispiele sollen die beiden Modelle ausgetestet und verglichen werden. Dazu wurde aus dem vorhandenen Datenmaterial zufällig Datensätze herausgegriffen. Die für die Schätzung relevanten Zahlen stehen in der Spalte "Wert". In der Spalte "Summand" steht die Zahl, die die entsprechende Größe zum Kaufpreis beiträgt (also Wert multipliziert mit Gewichtsfaktor, bei DM-Beträgen dividiert durch den Preisindex).
Die Berechnungen – einschließlich der Transformationen – werden jeweils mit den angegebenen (gerundeten) Zahlen mit einem Taschenrechner ausgeführt. Zusätzlich wird das von SPlus gelieferte Ergebnis angegeben. Diese Zahlen unterscheiden sich natürlich deutlich, denn die Rechengenauigkeit von SPlus übersteigt die eiens Taschenrechners um ein Vielfaches, und außerdem machen sich Rundungsfehler gerade bei den Transformationen sehr deutlich bemerkbar.

1. Beispiel

Berechnung ohne Transformation

Größe	Wert	Summand
Umbauter Raum in m^3	480	5760
Gewöhnlicher Herstellungswert in TDM	25	14500
Bodenwert in DM	15918	630
Baujahr	1936	80900
Wert der Außenanlagen in DM	15000	2060
Wohngebiet=1?	ja	1520
Stellung=7?	ja	-2140
Stellung=3?	nein	0
Konstante		-92500
Preisindex zum Zeitpunkt des Verkaufs	18,629	
Kaufpreis in DM	182000	
Summe		10730
geschätzter Kaufpreis in DM (Summe*Preisindex)		200000

Die etwas genauere Berechnung mit SPlus ergibt einen Schätzwert von 200830, sowie einen 95%-Konfidenzbereich zwischen 179280 und 222379. Der wahre Kaufpreis von 182000 DM liegt an der unteren Grenze gerade noch innerhalb dieses Vertrauensbereichs.

Berechnung mit Transformation

Größe	Wert	Transformiert	Summand
Umbauter Raum	480	3,96	237,20
Gewöhnlicher Herstellungswert	25	0,298	11,566
Baujahr	1936	1936	234,26
Bodenwert	15918	8,168	23,69
Außenanlagen	15000	22,35	15,00
Richtwert	140	2,609	9,366
Stellung=1?	nein	0	0
Wohngebiet=2 und Konstruktion=1?	nein	0	0
Konstante			-443,0
Preisindex zum Zeitpunkt des Verkaufs	18,629		
Kaufpreis in DM	182000	83,74	
geschätzter Kaufpreis in DM	207000	88,08	88,08

Die Berechnung in SPlus liefert hier als Ergebnis für den transformierten Kaufpreis 88,42, die obere und die untere Schranke des Konfidenzbereichs sind 86,29 bzw. 90,56. Umgerechnet ergeben diese Zahlen 208160 als Schätzwert, 195599 als untere Schranke und 221284 als obere Schranke. Der wahre Kaufpreis liegt also nicht im Konfidenzbereich.

2. Beispiel

Berechnung ohne Transformation

Größe	Wert	Summand
Umbauter Raum in m^3	845	10140
Gewöhnlicher Herstellungswert in TDM	24	13500
Bodenwert in DM	23175	887
Baujahr	1966	82200
Wert der Außenanlagen in DM	15000	1990
Wohngebiet=1?	ja	1520
Stellung=7?	nein	0
Stellung=3?	nein	0
Konstante		-92500
Preisindex zum Zeitpunkt des Verkaufs	19,258	
Kaufpreis in DM	355000	
Summe		17737
geschätzter Kaufpreis in DM (Summe*Preisindex)		342000

Die etwas genauere Berechnung mit SPlus ergibt einen Schätzwert von 340899, sowie einen 95%-Konfidenzbereich zwischen 328555 und 353242. Der wahre Kaufpreis von 355000 DM liegt nicht innerhalb dieses Vertrauensbereichs.

Berechnung mit Transformation

Größe	Wert	Transformiert	Summand
Umbauter Raum	845	4,17	249,8
Gewöhnlicher Herstellungswert	24	0,222	8,614
Baujahr	1966	1966	237,9
Bodenwert	23175	8,67	25,14
Außenanlagen	15000	22,09	14,82
Richtwert	150	2,67	9,585
Stellung=1?	nein	0	0
Wohngebiet=2 und Konstruktion=1?	nein	0	0
Konstante			-443,0
Preisindex zum Zeitpunkt des Verkaufs	19,258		
Kaufpreis in DM	355000	107,3	
geschätzter Kaufpreis in DM	319000	102,9	102,9

Die Berechnung in SPlus liefert hier als Ergebnis 105,79, als untere Schranke für das Konfidenzintervall 103,53 und als obere Schranke 108,07. Umgerechnet ergeben sich 340679 als Schätzwert, 322263 als untere Schranke und 359733 als obere Schranke. In diesem Fall liegt der wahre Kaufpreis also innerhalb des Berechneten Konfidenzintervalls.

Vereinfachung der Modelle

Bei derartig hochdimensionalen Problemen wie dem unseren kommt es häufig vor, daß die Zielgröße bereits durch wenige Variablen recht gut beschrieben wird. Beim Auto-Regressionsverfahren können aber noch weitere Variablen ins Modell aufgenommen werden, obwohl die Verbesserungen, die sich dadurch ergeben, nur sehr gering sind.
Es liegt also nahe, auszutesten, ob nicht ein kleineres Modell ein Ergebnis liefert, das nicht wesentlich schlechter ist als das ursprüngliche.
In den beiden berechneten Modellen haben sich folgende Größen als für den Kaufpreis relevant erwiesen:

- Umbauter Raum
- Gewöhnlicher Herstellungswert des Gebäudes
- Baujahr des Gebäudes
- Bodenwert
- Wert der Außenanlagen

Weitere Größen, die den Kaufpreis in geringerem Maß beeinflussen, sind:

- Wohngebiet
- Stellung des Gebäudes

Daher wollen wir – abermals sowohl mit den Originaldaten als auch mit den transformierten Daten – Modelle berechnen, die die als wichtig erkannten Variablen beinhalten.

Vereinfachtes Modell mit Original-Daten

Zunächst wurde ein lineares Modell berechnet, bei dem der Kaufpreis als gewichtete Summe von umbautem Raum, gewöhnlichem Herstellungswert, Baujahr, Bodenwert und Wert der Außenanlagen angenommen wird. Anhand der bereits erwähnten diagnostischen Plots wurden fünf Datensätze als Ausreißer identifiziert und entfernt. Danach wurde nochmals eine Regression durchgeführt. Das Ergebnis ist nicht wesentlich schlechter als das der schrittweisen Modellbildung:

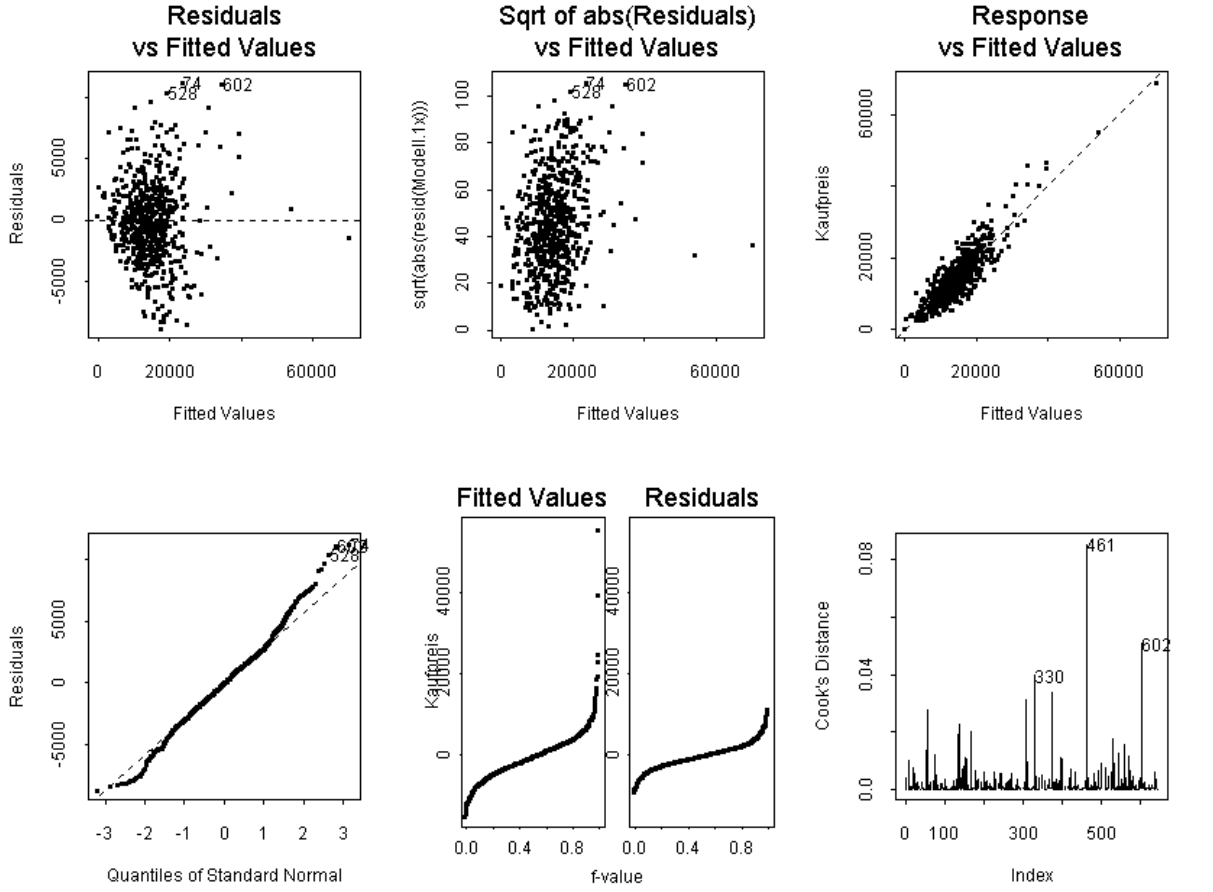
```
Residuals:
   Min     1Q Median    3Q   Max
 -8718 -2020 -79.75 1876 11337

Coefficients:
                    Value   Std. Error      t value     Pr(>|t|)
  (Intercept) -109954.5322    8066.3037     -13.6313       0.0000
Umbauter.Raum      12.4670       0.5750      21.6809       0.0000
Gew.Herstwert   10718.4727     911.8574      11.7545       0.0000
      Baujahr      50.8510       4.3652      11.6491       0.0000
    Bodenwert       0.7349       0.0416      17.6681       0.0000
Aussenanlagen       2.5235       0.3483       7.2446       0.0000

Residual standard error: 3268 on 637 degrees of freedom
Multiple R-Squared: 0.7881
F-statistic: 473.7 on 5 and 637 degrees of freedom, the p-value is 0
```

Die Modellstatistiken bewegen sich in den gleichen Größenordnungen wie zuvor (78,8% der Varianz des Kaufpreises werden erklärt, die Standardsabweichung der Vorhersage des Kaufpreises beträgt 3268, die F-Statistik 473,7). Die Koeffizienten dagegen haben sich deutlich verändert.

Die diagnostischen Plots sehen denen des ursprünglichen Modells sehr ähnlich: die Residuen zeigen eine schwache Struktur, und es macht sich wieder bemerkbar, daß die Daten nicht zentriert sind.



Das berechnete Modell liefert folgende Formel für den Kaufpreis:

Kaufpreis	=	12,5 *	Umbauter. Raum	+	10700 *	Gew.Herstwert	+ 50,9 *	Baujahr
	+	0,735 *	Bodenwert	+	2,52 *	Aussenanlagen	-	110000

Damit ergibt sich in unserem ersten Beispiel von vorhin ein Schätzwert für den Kaufpreis von 215000 DM. Der Konfidenzbereich liegt zwischen 203000 und 226000 DM. Für das zweite Beispiel ergibt sich ein Schätzwert von 323000, das Konfidenzintervall liegt bei 312000 bis 333000 DM.

Vereinfachtes Modell mit transformierten Daten

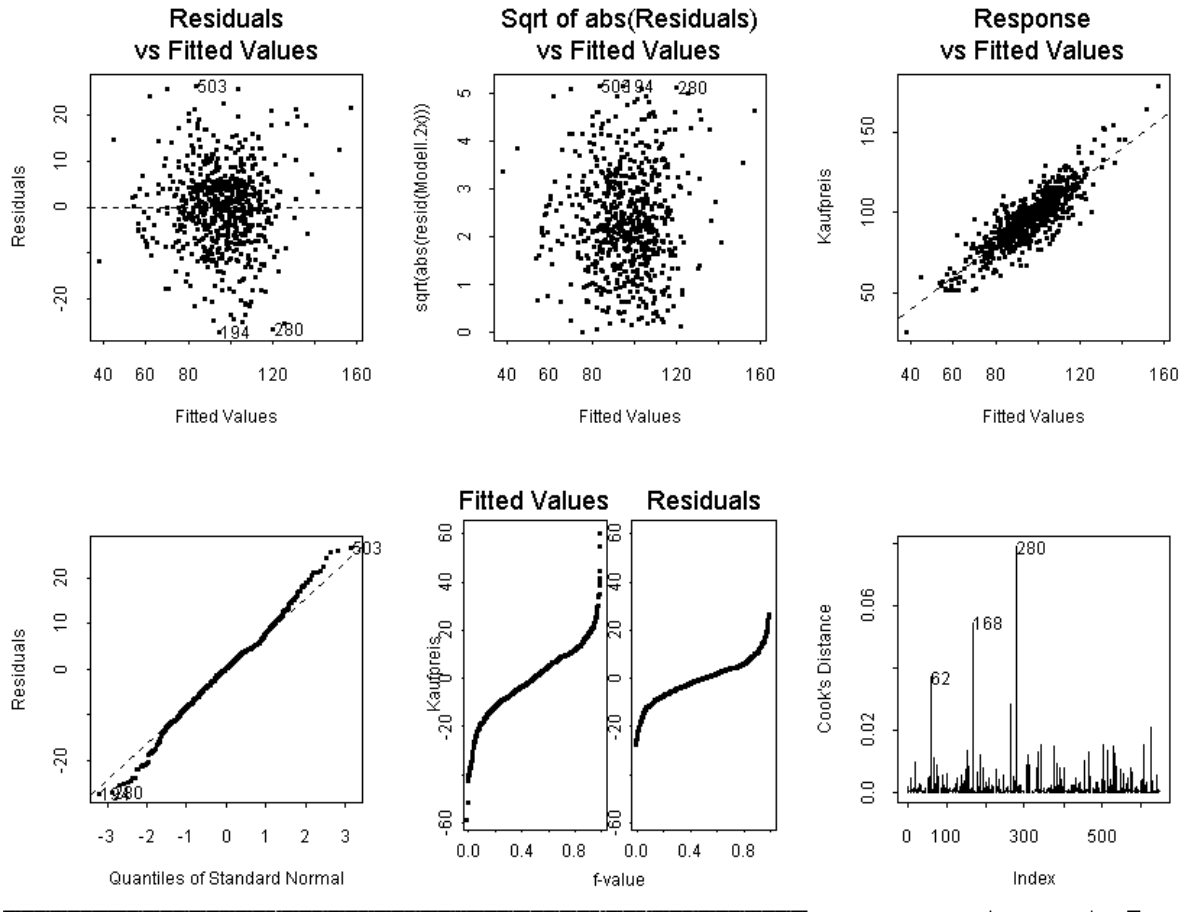
Analog zur Vorgehensweise im letzten Abschnitt wurde aus den transformierten Daten zunächst ein Modell berechnet, dann zwei Ausreißer entfernt, und schließlich nochmals das Modell berechnet.

```
Residuals:
    Min      1Q Median    3Q   Max
 -27.03 -5.647 0.2606 5.099 26.96

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
  (Intercept) -479.0824   25.9655   -18.4508    0.0000
Umbauter.Raum   63.2985    3.3986    18.6249    0.0000
Gew.Herstwert   38.6371    3.0061    12.8530    0.0000
      Baujahr    0.1351    0.0119    11.3387    0.0000
    Bodenwert    3.3249    0.3222    10.3205    0.0000
Aussenanlagen    0.7244    0.0839     8.6395    0.0000

Residual standard error: 8.88 on 640 degrees of freedom
Multiple R-Squared: 0.7544
F-statistic: 393.1 on 5 and 640 degrees of freedom, the p-value is 0
```

In diesem Fall haben sich sowohl die Modellstatistiken als auch die Koeffizienten deutlich verändert. Die diagnostischen Plots des Modells mit transformierten Daten sehen – wie schon bei der schrittweisen Regression – sehr viel besser aus als die des Modells mit nicht-transformierten Daten: die Residuen zeigen ein Verhalten, das dem von weißem Rauschen viel ähnlicher ist.



Die Formel zur Schätzung des Kaufpreises lautet:

Kaufpreis'	=	63,3 *	Umbauter. Raum'	+	38,6 *	Gew.Herstwert'	+ 0,135 *	Baujahr
	+	3,32 *	Bodenwert'	+	0,724 *	Aussenanlagen	- 479	

Danach erhält man als Schätzwert für den Kaufpreis in unserem ersten Beispiel 207000 DM bei einem Konfidenzbereich zwischen 195000 und 220000 DM. Im zweiten Beispiel ergibt sich ein Schätzwert von 314000 DM, das zugehörige Konfidenzband liegt zwischen 301000 und 328000 DM.

Schlußbemerkung

Da die vereinfachten Modelle doch deutlich schlechtere Ergebnisse erzielen als die vollständigen Modelle, und außerdem der eingesparte Rechenaufwand nur sehr gering ist, sollte man für Vorhersagen das komplette Modell heranziehen. In der Praxis dürfte es – zumindest in der näheren Zukunft – keinen großen Unterschied machen, ob man das auf den transformierten Daten beruhende oder das auf den Original-Daten beruhende Modell benutzt. Daher kann durchaus auch mit dem einfacheren Original-Daten-Modell gerechnet werden. Man sollte sich aber stets ins Gedächtnis rufen, daß dieses Verfahren weniger stabil ist als das aufwendigere.
Von Zeit zu Zeit sollte das Datenmaterial überprüft und das Modell aktualisiert werden. Zieht man Daten zur Modellfindung hinzu, die einen größeren Zeitraum abdecken, so ist es gut möglich, daß der Zeitaspekt eine Rolle spielt. Das hier gefundene Modell kann also nicht als endgültig gelten.