4 Vorgehen bei der Modellierung

In nachfolgenden Kapitel wird die Vorgehensweise bei der Erstellung der beiden hedonischen Modelle erläutert. Was hier als linearer Prozess dargestellt wird, war in Tat und Wahrheit ein iterativer Prozess, bei dem die verschiedenen Etappen nach neuen Erkenntnissen jeweils wiederholt wurden.

4.1 Deskriptive Analyse

In einem allerersten Schritt werden die erhobenen Daten, die als Basis für die Erstellung der hedonischen Modelle dienen, analysiert und auf allfällige Lücken, Fehler und Auffälligkeiten überprüft. Grundsätzlich können beim Modellieren nur Transaktionen verwendet werden, die keine Missings oder eindeutige Fehler bei den verwendeten Variablen aufweisen. Je nach Variablenumfang eines Modells werden deshalb in dieser Testphase auch unterschiedlich viele Transaktionen berücksichtigt. Neben Missings und Fehlern bei Variablen kann die Nichtverwendbarkeit einer Transaktion auch auf sogenannte No-Matchings zurückzuführen sein. Bei diesen Fällen konnte das Erhebungstool keinen erfolgreichen Adressabgleich vornehmen. In der Folge weist das Objekt keine Lagevariablen auf. Aus diesem Grund könnten die betroffenen Transaktionen nur in Modellen ohne Lagevariablen verwendet werden. Neben den No-Matchings werden bei der Modellierung auch Transaktionen mit einem Center-Community-Matching ausgeschlossen. Dies, weil die imputierten Lagevariablen, die vom Objekt stammen, das dem geographischen Mittelpunkt der Gemeinde am nächsten liegt, erheblich von den wirklichen Lagevariablen abweichen können.

Neben der Überprüfung der Verwendbarkeit der einzelnen Transaktionen ist es auch wichtig, sich ein gesamthaftes Bild bezüglich der Qualität der Daten zu machen. Dabei werden beispielsweise die Verteilungen der einzelnen Variablen angeschaut. Zudem werden die Korrelationen zwischen der abhängigen Variable (Preis) und den unabhängigen Variablen (Struktur-, Nutzungs-, Lagevariablen) bzw. zwischen den einzelnen unabhängigen Variablen studiert. Die aus diesen Analysen gewonnenen Erkenntnisse fliessen dann in die Modellierung und in die Interpretation der Resultate ein.

4.2 Variablentransformation und Modellierung

Nach der deskriptiven Analyse der Daten werden die vorhandenen Variablen auf verschiedene Art und Weise transformiert und aggregiert. So werden die linearen Variablen (VolumeOfBuilding, LandArea, NetLivingArea) beispielsweise logarithmiert oder quadriert. Bei kategoriellen Variablen bieten sich alternative Aggregationen der Kategorien an. Anschliessend werden sämtliche als sinnvoll erachteten Variablenkombinationen modelliert. Um sich ein Bild über den Einfluss einzelner Variablen machen zu können, werden zusätzlich die Verfahren der Vorwärtsselektion[19] und der Rückwärtselimination[20] beigezogen. Neben den einzelnen Variablen werden auch Interaktionen[21] zwischen verschiedenen Variablen getestet.

[19] Bei der Vorwärtsselektion werden die unabhängigen Variablen eine nach der anderen in das Modell aufgenommen. Dabei wird jeweils diejenige Variable ausgewählt, die die grösste partielle Korrelation mit der abhängigen Variable aufweist. Dieses Prozedere wird wiederholt, bis ein Abbruchkriterium erfüllt ist oder alle Variablen in das Modell aufgenommen wurden.

[20] Bei der Rückwärtselimination wird mit allen verfügbaren Variablen gestartet. Anschliessend werden die unabhängigen Variablen eine nach der anderen aus dem Modell entfernt. Dabei wird jeweils diejenige Variable aus dem Modell entfernt, die die kleinste partielle Korrelation mit der abhängigen Variable aufweist. Dieses Prozedere wird wiederholt, bis ein Abbruchkriterium erfüllt ist oder keine Variable mehr im Modell enthalten ist.

[21] Von Interaktionen wird gesprochen, wenn der Effekt einer unabhängigen Variable auf die abhängige Variable vom Wert einer anderen unabhängigen Variable abhängt. In einer Regressionsgleichung wird ein Interaktionseffekt als das Produkt von zwei oder mehr unabhängigen Variablen dargestellt: $\hat{y}$ = b0 + b1X1 + b2X2.

4.3 Erste Analyse der Modellvarianten

Aus den verschiedenen Variablenkombinationen resultieren für beide Objekttypen rasch einige hundert Modell-Varianten. Selbstverständlich wäre es nicht zielführend, diese Modelle alle von Hand zu analysieren und zu beurteilen. Aus diesem Grund werden die Modelle zuerst anhand verschiedener Kennzahlen beurteilt. Für sämtliche der unten aufgeführten und erläuterten Kennzahlen zur Modellgüte werden die jeweils ungefähr 10 besten Modelle ausgewählt.

R-Squared (R^2)
Das R^2 ist ein statistisches Maß dafür, wie nahe die beobachteten Daten an der geschätzten Regressionslinie liegen. Bei der Berechnung des R^2 wird die anhand des Modells erklärte Varianz durch die beobachtete Varianz dividiert.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Das R-Quadrat liegt immer zwischen 0 und 100%, wobei bei 100% das Modell die gesamte beobachtete Varianz erklärt. Das R^2 hat die Eigenschaft, dass es mit jeder Variable steigt, die zum Modell hinzugefügt wird. Würde man die verschiedenen Modelle nur aufgrund dieser Kennzahl beurteilen, wäre es demnach optimal, das Modell mit den meisten erklärenden Variablen zu wählen. Dies birgt die Gefahr der Überanpassung (Overfitting) des Modells.

Adjusted R-Squared (Adj. R^2)
Ein Gütemass, das neben der Modellanpassung auch das Problem der Überanpassung berücksichtigt, ist das Adjusted R-Squared. Es besteht aus dem R^2 sowie einem Strafterm, der mit jeder zum Modell hinzugefügten Variable weiter ansteigt. Aus diesem Grund nimmt das Adj. R^2 einen geringeren Wert als das R^2 an. Beim Hinzufügen einer neuen Variable steigt das Adj. R^2 nur dann an, wenn der zusätzliche Erklärungsgehalt den Anstieg des Strafterms ausgleichen kann. Im Strafterm werden die Anzahl Transaktionen (n) sowie die Anzahl Variablen (k) berücksichtigt.

$$Adj.R^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

Akaike Information Criterion (AIC)
Das Akaike Information Criterion vergleicht die verschiedenen Modellkandidaten anhand der Werte der log-Likelihood-Funktion ($\hat{L}$). Diese steigen, je grösser der Erklärungsgehalt des Modells ist. Daneben wird auch ein Strafterm berücksichtigt, der aus der Anzahl der im Modell enthaltenen Parameter (k) besteht. Das Modell mit dem kleinsten AIC wird bevorzugt.

$$AIC = 2k - 2\ln(\hat{L})$$

Bayesian Information Criterion (BIC)
Das Bayesian Information Criterion basiert ebenfalls auf der log-Likelihood-Funktion ($\hat{L}$) und unterscheidet sich nur beim Strafterm gegenüber dem Akaike Information Criterion. Das BIC ist neben der im Modell enthaltenen Parameter auch von der Anzahl Transaktionen (n) abhängig. Bereits ab einer Stichprobe von acht Observationen bestraft das BIC stärker als das AIC.

$$BIC = ln(n)k - 2\ln(\hat{L})$$

Mean Absolute Error (MAE)
Der Mean Absolute Error (MAE) einer Regression entspricht dem arithmetischen Mittel der absoluten Differenz zwischen den Vorhersagewerten ($\hat{y}$) und den Beobachtungswerten (y) des berücksichtigten Samples.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

4.4 Vertiefte Analyse der besten Modelle

Die aufgrund der im Kapitel 4.3 erwähnten Kennzahlen ausgewählten Modelle werden in einem nächsten Schritt genauer betrachtet. Dabei werden die Modelle nun einzeln und von Hand analysiert. Zuerst werden dabei die Koeffizienten der einzelnen Modellvariablen genauer studiert und plausibilisiert. Es stellt sich dabei die Frage, ob die Koeffizienten den theoretisch zu erwartenden Einflüssen entsprechen (z.B. hat Lärm im Modell wirklich einen negativen Einfluss auf den Preis?). Daneben wird auch die Signifikanz der einzelnen Koeffizienten angeschaut. Zudem wird anhand des Variance Inflation Factors (VIF)[22] überprüft, ob die verwendeten unabhängigen Variablen Multikollinearität[23] aufweisen.

Lineare Regressionen unterliegen verschiedenen Annahmen, die erfüllt sein sollten, damit die Methode ordnungsgemäss funktioniert. Dazu gehören neben der Abwesenheit von Multikollinearität, dem linearen Zusammenhang zwischen den metrischen unabhängigen Variablen und der abhängigen Variable auch die Normalverteilung und Homoskedastizität der Residuen[24]. Bei einer Verletzung einer der beiden letzten Annahmen sind die Schätzer zwar weiterhin konsistent, allerdings setzt beispielsweise der t-Test eine Normalverteilung und Homoskedastizität der Residuen voraus. Aus diesem Grund ist es wichtig, die beiden Hypothesen bei den ausgewählten Modellen zu überprüfen.

Die Normalverteilung der Residuen wird anhand eines QQ-Plots der Residuen beurteilt. Liegen die Residuen im Plot auf einer diagonalen Geraden, sind sie normalverteilt. Die Normalverteilungshypothese kann auch anhand eines Histogramms der Residuen mit darübergelegter Normalverteilungskurve überprüft werden. Neben den graphischen Check's wird der Jarque-Bera-Test[25] beigezogen.

Die Homoskedastizität der Residuen wird ebenfalls grafisch und mit Tests überprüft. Hierzu wird ein Scatterplot der Residuen versus die geschätzten Preisen verwendet. Wenn die Abbildung eine zufällige Streuung der Observationen ohne Clusterbildung und systematische Muster aufweist, kann Homoskedastizität angenommen werden. Da die Interpretation von Diagrammen immer auch eine subjektive Komponente aufweist, werden zusätzlich der Breusch-Pagan-Test[26] und eine Version des White Tests[27] zur Überprüfung der Homoskedastizitätsannahme beigezogen. Daneben werden auch sämtliche unabhängigen Variablen gegen die Residuen geplottet. Es ist nicht überraschend, wenn zu diesem Zeitpunkt der Modellierung nicht normalverteilte und heteroskedastische Residuen beobachtet werden können. Eine Massnahme zur Reduktion der Heteroskedastizität ist die Logarithmierung der abhängigen Variablen. Zudem haben Ausreisser einen grossen Einfluss auf die beiden zu Grunde liegenden Annahmen.

[22] Anhand des Variance-Inflation-Factor-Tests (VIF) kann die Multikollinearität zwischen verschiedenen Regressionsvariablen gemessen werden. Mathematisch gesehen ist der VIF für eine Regressionsmodellvariable gleich dem Verhältnis der gesamten Modellvarianz zur Varianz eines Modells, das nur diese eine unabhängige Variable enthält.

[23] Von Multikollinearität wird dann gesprochen, wenn es in einem Modell eine starke Korrelation zwischen zwei oder mehreren unabhängigen Variablen gibt. Bei starker Multikollinearität, wird die Schätzung der Regressionskoeffizienten unsicher, da die betroffenen Variablen zumindest teilweise denselben Teil der Varianz erklären.

[24] Die Residuen entsprechen dem Teil des Transaktionspreises, der durch das hedonische Modell nicht erklärt wird bzw. der Differenz zwischen dem beobachteten und dem geschätzten Transaktionspreis.

[25] Der Jarque-Bera-Test prüft anhand der Schiefe und Kurtosis ob eine Normalverteilung vorliegt. Vgl. Bera, A. K. and C. M. Jarque (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. In Economic Letters. Nr.3, p. 255–259.

[26] Breusch, T.S. and A. R. Pagan (1980). A simple test for heteroscedasticity and random coefficient variation. In: Journal of the Econometric Society, Econometrica, p. 817–838

[27] Wooldridge, J. M. (2013), Introductory Econometrics, a modern approach, 5th edition, South-Western, chap. 8.3, p. 280

4.5 Behandlung der Ausreisser

Bis zu diesem Zeitpunkt wurden nur Transaktionen mit Missings aus dem Datensatz entfernt. Daneben sind aber auch auffällige Transaktionen genauer anzuschauen. In einem nächsten Schritt werden die Daten deshalb auf Ausreisser überprüft und wenn nötig bereinigt. Wir unterscheiden dabei zwischen uni- und multivariaten Ausreissern. Bei univariaten Ausreissern handelt es sich um einen auffällig hohen bzw. niedrigen Wert in einer Variable (z.B. 200 Zimmer). Solche Fälle sind relativ einfach herauszufiltern. Komplizierter wird es hingegen bei der Detektion von multivariaten Ausreissern. Damit sind Transaktionen gemeint, die aufgrund der Kombination verschiedener Variablen auffallen. Dabei können die einzelnen Ausprägungen durchaus Sinn machen (z.B. 6 Zimmer; 30m^2 Nettowohnfläche).

In einem ersten Schritt kümmern wir uns um die multivariaten Ausreisser. Das Verfahren, das dabei angewendet wird, heisst Cook's Distance[28]. Die Outlier detection wird basierend auf dem hedonischen Modell durchgeführt. Cook's Distance gibt wieder, wie stark sich die Residuen verändern, falls eine bestimmte Observation aus der Schätzung des Regressionsmodells entfernt wird. Dadurch können Observationen ausfindig gemacht werden, die einen grossen Einfluss auf das Schätzmodell haben. Die Formel zur Berechnung der Cook's Distance für eine Transaktion lautet folgendermassen:

$$D_i = \frac{\{\beta - \beta_i\}^T X^T X \{\beta - \beta_i\}}{ps^2}$$

D_i Cook's Distance für Transaktion i
X Matrix ($n\ x\ p$) mit den Werten der unabhängigen Variablen
T Transposition einer Matrix bzw. eines Vektors
β Vektor ($p\ x\ 1$) der Least Squares Schätzer
β_i Vektor ($p\ x\ 1$) der Least Squares Schätzer ohne Berücksichtigung der Transaktion i im hedonischen Modell
p Anzahl der unabhängigen Variablen plus 1
s^2 Geschätzte Varianz unter Verwendung des ganzen Datensatzes

Das Resultat der Analyse ist ein Cook's Distance Wert für sämtliche Transaktionen des Datasets. Grundsätzlich gilt: Mit hohen Cook's Distance Werten steigt die Wahrscheinlichkeit, dass es sich um einen Outlier handelt. Beim Wohnimmobilienpreisindex werden sämtliche Transaktionen, die eine Cook's Distance grösser als «4/Anzahl Transaktionen» aufweisen, aus dem Sample gelöscht. Diese «cut-off»-Grenze ist weit verbreitet (z.B. offizieller Immobilienpreisindex Irland[29]) und wird unter anderem auch von Bollen und Jackman[30] propagiert.

Im Anschluss an die Eliminierung der multivariaten Ausreisser werden in einem zweiten Schritt gewisse Regeln zur Auffindung der allenfalls übrig gebliebenen univariaten Ausreisser angewendet. Dabei geht es vor allem darum, extreme Werte bei der abhängigen, wie auch bei den wichtigsten unabhängigen Variablen (VolumeOfBuilding, LandArea und NetLivingArea) zu eliminieren, die allenfalls die hedonischen Modelle verzerren könnten.

4.6 Auswahl des definitiven Modells

Nach der Ausreisserbehandlung werden die ausgewählten Modelltypen mit dem reduzierten Datensatz neu kalibriert. Wegen den bearbeiteten Stichproben können sich auch die Kennzahlen der einzelnen Modelle verändern. Durch die Entnahme von Ausreissern sollte sich die Schätzqualität verbessern. Es kann aber auch zu Veränderungen bei den Koeffizienten kommen. Deshalb müssen die neu kalibrierten Modelle nochmals analog zum Vorgehen in Kapitel 4.4. analysiert werden. Dazu gehört auch eine Überprüfung der der linearen Regression zu Grunde liegenden Hypothesen. Nach der Ausreisserbehandlung sollten sich die zu Grunde liegenden Annahmen für die lineare Regression verbessert haben. Aufgrund der vorliegenden Resultate wird anschliessend ein definitives Modell ausgewählt.

[28] Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. Technometrics Vol. 19 (1): p. 15–18

[29] O'Hanlon, N. (2011). Constructing a National House Price Index for Ireland. Journal of the Statistical and Social Inquiry Society of Ireland, Vol XL, Central Statistics Office and Centre for Policy Studies

[30] Bollen, K. A. and R. W. Jackman (1990). Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. Modern methods of data analysis, p. 257–291

4.7 Modellrevisionen

Auch wenn die ökonometrischen Modelle beim Ansatz des Hedonic Repricing eine gewisse Zeit lang stabil belassen werden können, gilt es zu berücksichtigen, dass sich die impliziten Preise der Qualitätseigenschaften mittel- bis langfristig verändern können. Um diesem Aspekt Rechnung zu tragen, wird das Hedonic Repricing-Modell in regelmässigen Abständen unter Berücksichtigung der neu zur Verfügung stehenden Daten kalibriert. Zusätzlich wird das BFS, parallel zur Anwendung der Methode des Hedonic Repricing einen zweiten Index mit der Rolling-Time-Dummy-Methode berechnen. Die Rolling-Time-Dummy-Methode ist eine Abwandlung bzw. Erweiterung der klassischen Time-Dummy-Methode, bei der neben den Struktur-, Nutzungs- und Lagevariablen auch periodenspezifische Dummy-Variablen in das hedonische Modell integriert werden. Das hedonische Modell wird dabei in jeder Periode neu berechnet. Als Basis dienen jeweils die Transaktionen aus der aktuellen und den drei vorangegangenen Perioden. Die Preisentwicklung kann direkt aus den periodenspezifischen Dummies abgeleitet werden. Der Rolling-Time-Dummy-Index wird nicht veröffentlicht und dient einzig als interner Benchmark sowie zur Überwachung der Entwicklung der impliziten Preise.

4.8 Gutachten

Im Nachgang zur Erstellung der Ausgangsmodelle mit den Daten der Jahre 2017 bis 2019 wurden die Resultate einem Gutachten unterzogen. Das Gutachten wurden von Mick Silver, einem emeritierten Professor für Economic Statistics an der Cardiff University und Senior Economist beim Internationalen Währungsfonds (IWF) durchgeführt. Aus Datenschutzgründen hatte der Experte keinen Zugang zu den Daten. Als Grundlage für das Gutachten wurden aggregierte Auswertungen sowie ein detaillierter Beschrieb der Vorgehensweise bereitgestellt. Ausserdem standen die Vertreter des BFS für alle Fragen zur Verfügung