Einfache lineare Regression und Residuenanalyse – Bericht

Yannic Lais, Marvin von Rappard, Luca Mazzotta

January 16, 2023

Abstract

In diesem Bericht wird das Vorgehen, einer einfachen linearen Regression beschrieben. Anhand der Wohnfläche wurde dabei der Preis verschiedener Immobilien mittels einfacher linearen Regression vorhergesagt. Anschliessend wurden die Resultate mittels Residuenanalyse verglichen und weiter untersucht. Weitere Informationen sind in folgendem Github Repository zu finden: https://github.com/marvinvr/fhnw-cml1

Contents

1 Immobilien Standorte **2**

2 Preisverteilung **3**
- 2.1 Preisverteilung von allen Daten . . . 3
- 2.2 Preisverteilung Daten ohne Ausreisser . . . 3
- 2.3 Untersuchung der Ausreisser . . . 4
- 2.4 Preisverteilung der Immobilientypen . . . 5

3 Abhängigkeit der Attribute **7**
- 3.1 Korrelationsmatrix . . . 7
- 3.2 Korrelation zwischen «Longitude» und «Zip» (0.94) . . . 8
- 3.3 Korrelation zwischen «WorkplaceDensityL» und «gde_population» (0.74) . . . 9
- 3.4 Korrelation zwischen «gde_area_settlement_percentage» und «gde_pop_per_km2» (0.81) . . . 9

4 Rückblick **10**

1 Einführung

Mittels linearen Regressionsmodellen wurde der Preis verschiedener Immobilien anhand der Fläche vorhergesagt. Die Funktionsweise einer einfachen linearen Regression wird dabei genauer beschrieben und erklärt.

2 Einfache Lineare Regression

2.1 Erklärung

Eine lineare Regression ist ein statistisches Modell, das verwendet wird, um die Beziehung zwischen einer abhängigen Variablen y und einer (einfache lineare Regression) oder mehreren unabhängigen Variablen x (multilinearen Regression) zu beschreiben. In der einfachsten Form besteht die Beziehung zwischen y und x in einer geraden Linie, die als "Regressionsgerade" bezeichnet wird. Für das verwendete Modell wurde als unabhängige Variable die Fläche der Immobilie und als abhängige Variable der Preis in CHF der Immobilie verwendet.

Die Regressionsgerade wird durch eine Gleichung der Form

$$y = \beta_0 + \beta_1 x \tag{1}$$

beschrieben, wobei β_0 der y-Achsenabschnitt und β_1 die Steigung der Geraden ist. Diese beiden Parameter werden durch die Methode der kleinsten Quadrate an die gegebenen Daten angepasst.
Um die Methode der kleinsten Quadrate anzuwenden, berechnet man zunächst den Fehler zwischen den tatsächlichen Datenwerten y_i und den vorhergesagten Werten $\hat{y_i}$ für jeden Datenpunkt i. Der Fehler entspricht der Differenz dieser Werte.
Die kleinsten Quadrate Methode sucht nach den Werten, welche die Summe der quadratischen Fehler (SSE) minimieren:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{2}$$

Es gibt analytische Lösungen, die diese Bedingung erfüllen. Sobald die Werte von β_0 und β_1 gefunden wurden, kann man die Regressionsgerade verwenden, um die abhängige Variable y anhand der unabhängigen Variable x vorherzusagen.

Wichtig zu erwähnen ist, dass die Annahme einer linearen Beziehung zwischen y und x einige Einschränkungen hat und es nicht immer die beste Wahl für die Modellierung von Daten ist, da wir die Abhängigkeit der abhängigen Variable mit nur einer unbhängigen Variable stark vereinfachen. Es ist also wichtig die Daten und die Beziehung zwischen abhängigen und unabhängigen Variablen vorher zu untersuchen und gegebenenfalls andere Modelle zu betrachten oder weitere unabhängige Variablen hinzuzuziehen.

3 Metriken

3.1 MSE: Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{D}(x_i - y_i)^2 \tag{3}$$

Der MSE zeigt uns den durchschnittlichen Quadratischen Vorhersagefehler. Vorteile des MSE sind, dass er die quadratische Abweichung verwendet, was bedeutet, dass grosse Abweichungen stärker gewichtet werden als kleine Abweichungen. Dies kann dazu beitragen, dass extreme Werte (wie Ausreisser) einen grösseren Einfluss auf die Genauigkeit der Vorhersage haben.

3.2 MAPE: Mean Absolute Percentage Error

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \tag{4}$$

Der MAPE zeigt uns den durchschnittlichen Vorhersagefehler in Prozent. Vorteile des MAPE sind, dass er die Grösse des Fehlers relativ zum Wert des Vorhersageobjekts misst und somit die Grösse des Fehlers in Relation zum Wert des Objekts misst. Seine Nachteile sind, dass er für negative Vorhersagen nicht geeignet ist.

3.3 MAE: Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{5}$$

Der MAE zeigt uns den durchschnittlichen Vorhersagefehler. Vorteile des MAE sind, dass er die Grösse des Fehlers misst und uns somit den Fehler relativ in Bezug zu der Grösse zurückgibt.

3.4 RMSE: Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} \tag{6}$$

Der RMSE zeigt uns den durchschnittlichen Quadratischen Vorhersagefehler. Vorteile des RMSE sind, dass er die quadratische Abweichung verwendet, was bedeutet, dass grosse Abweichungen stärker gewichtet werden als kleine Abweichungen. Dies kann dazu beitragen, dass die Ausreisser einen grösseren Einfluss auf die Genauigkeit der Vorhersage haben.

4 Modellierung

4.1 Vorgehen

Um ein einfaches Lineares Modell zu erstellen müssen keine Daten skaliert werden. Daher wurde mittels Jointplots einen ersten Überblick über die abhängige, sowie die unabhängige Variable verschaft.

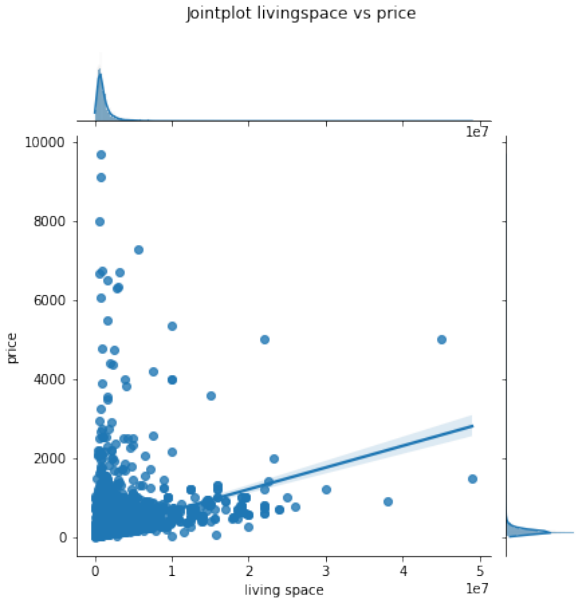

Figure 1: Datenverteilung, lineare Regression aller Immobilien.

Es ist keine klare lineare Abhängigkeit zwischen der Wohnfläche und dem Preis ersichtlich. Zudem sind die Daten der beiden Variablen rechtschief und nicht normal verteilt. Es wird ein erstes einfaches lineares Modell erstellt.

4.1.1 Lineare Regression aller Immobilientypen

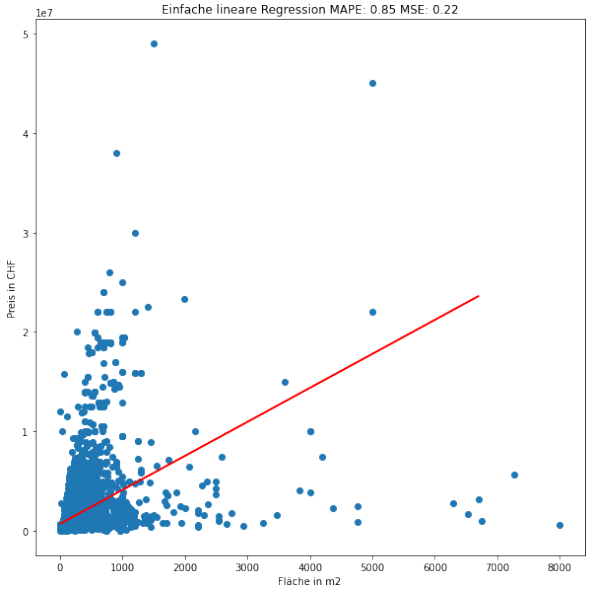

Figure 2: Datenverrteilung, lineare Regression aller Immobilien.

Es ist keine klare lineare Abhängigkeit zwischen der Wohnfläche und dem Preis ersichtlich. Daher ist auch der hohe MAPE und der tiefer MSE erklärbar. Bei beiden variablen ist eine rechtsschiefe Verteilung erkennbar. Daher werden beide Variablen mit der Quadratwurzel transformiert, um die lineare Abhängigkeit der Daten zu erhöhen. Bei einer Linksschiefen Verteilung müssten diese quadriert werden.

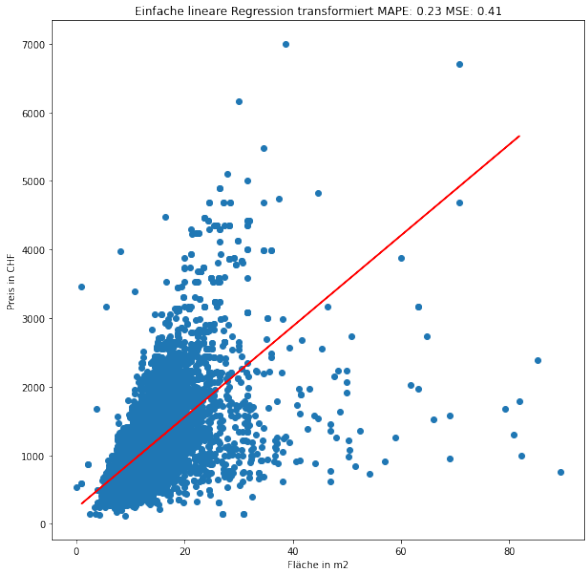

Figure 3: Lineare Regression aller Immobilien mit transformierten Variablen.

Mit dem jetzigen Modell ist eine deutlich bessere Fehlermetrik erreichbar. Dennoch sollte eine bessere Vorhersage möglich sein. Hierfür unterteilt man die Daten in unterschiedliche Kategorien wie Reihenhäuser, Wohnungen, Schlösser etc. Anschliessend werden mit diesen Datensätze erneut eine lineare Regression berechnet und die abhängige Variable Preis hervorgesagt. Dabei können einige besondere Beobachtungen gemacht werden.

4.1.2 Lineare Regression Attikawohnungen

So ist bei den Dachwohnungen bereits ein deutlicher linearer Zusammenhang sichtbar. Dennoch werden die Variablen transformiert, wodurch der lineare Zusammenhang noch besser sichtbar wird. So kann mit einem einfachen linearen Modell der Preis anhand der Wohnfläche mit einem MAPE 0,54 von und einem MSE von 0,35 bereits einigermassen vorhergesagt werden.

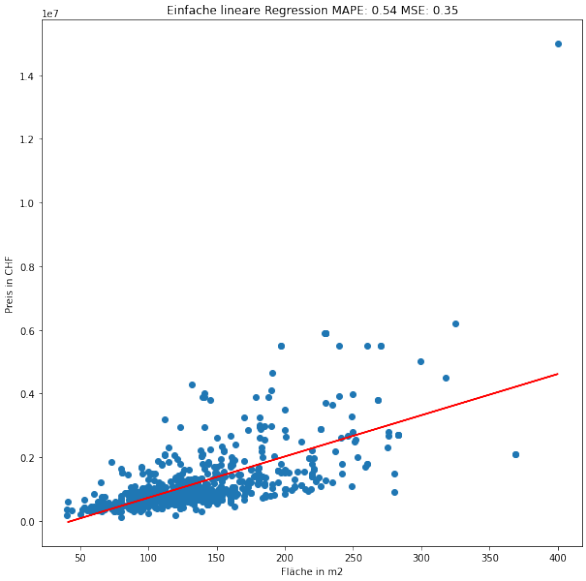

Figure 4: Lineare Regression der Dachgeschosswohnungen mit transformierten Daten

4.1.3 Lineare Regression Rusticos

Bei den Rustico kann eine weitere spannende Beobachtung gemacht werden. So sehen wir wie stark Ausreisser ein lineares Modell beeinflussen können. Es ist nur ein Datenpunkt, welcher die Regressionsgerade extrem verzerrt.

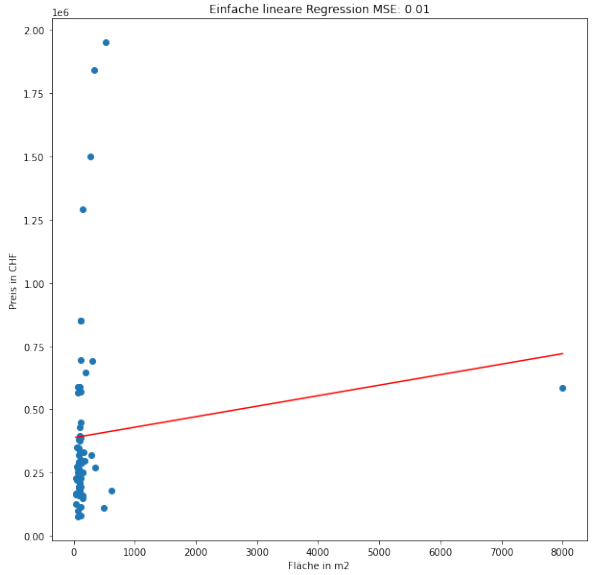

Figure 5: Lineare Regression der Rusticos.

Um die Auswirkung aufzuzeigen, werden diese beiden Punkte gelöscht und nochmals eine lineare Regression berechnet, wie zu sehen ist, ist nun ein deutlich besserer linearer Zusammenhang sichtbar.

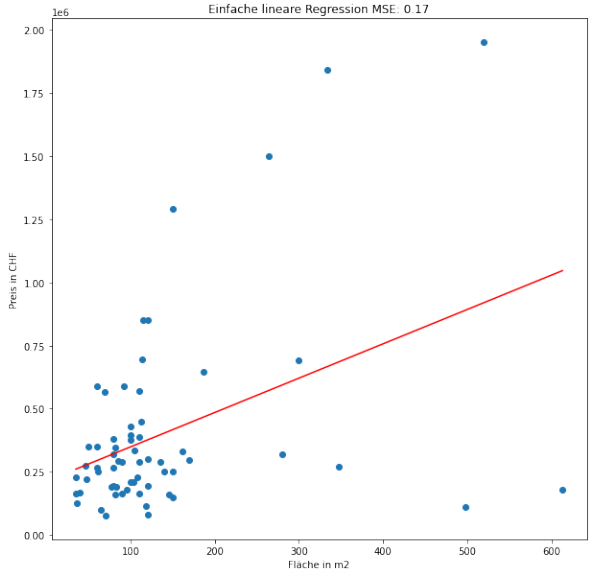

Figure 6: Lineare Regression der Rusticos ohne "Ausreisser".

4.2 Residuenanalyse

Die Residuenanalyse ist eine Methode in der Statistik, die dazu verwendet wird, die Qualität eines statistischen Modells zu bewerten. Sie wird häufig verwendet, um zu bestimmen, ob die Annahmen, die bei der Erstellung eines Modells gemacht wurden, gültig sind und ob das Modell eine gute Anpassung an die Daten bietet.

In einer Residuenanalyse untersucht man die Residuen, d.h die Abweichungen zwischen den tatsächlichen Daten und den von einem Modell vorhergesagten Daten. Wenn die Annahmen des Modells gültig sind und das Modell eine gute Anpassung an die Daten bietet, sollten die Residuen ungefähr normalverteilt sein mit einem Mittelwert von Null und konstanter Varianz.

Es ist wichtig zu beachten, dass die Residuenanalyse nur ein Teil der Überprüfung der Gültigkeit des Modells ist und es immer empfehlenswert ist, mehrere Methoden anzuwenden, um die Qualität des Modells zu bewerten.

Ein gutes Modell weist folgende Merkmale bei der Residuenanalyse auf.

- Residuen sollten um 0 Verteil sein
- Residuen sollten gleichmässig verteilt sein
- Residuen sollten Normalverteilt sein

zu sehen ist ein Beispiel einer Residuenanalyse für ein einfaches lineares Modell.

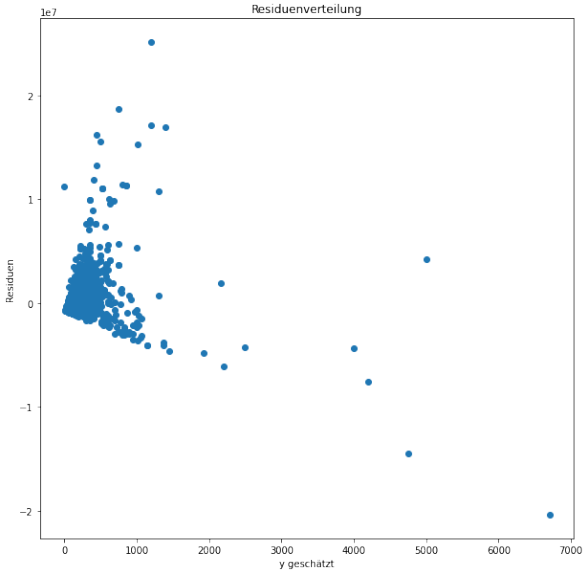

Figure 7: Test Nullverteilung der Residuen

Zu sehen sind die Residuen, welche um Null verteilt sind. Es ist zudem gut ersichtlich, dass das Modell mit zunehmndem Preis (y-geschätzt) einen höheren Fehler aufweist.
Nun werden die Residuen auf Ihre Verteilung getestet.

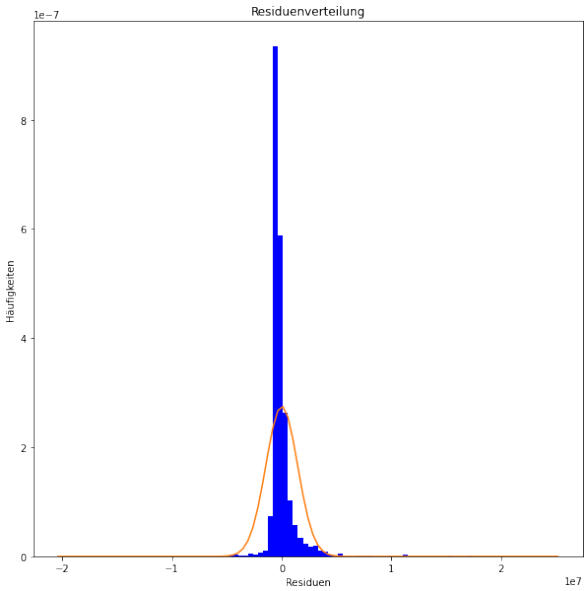

Figure 8: Residuentest auf eine Normalverteilung.

Die Residuen sind sehr gut Normalverteilt. Insgesamt performt das Modell anhand der Residuenanalyse gut und weist keinen Systematischen Fehler auf. Wichtig zu betonen ist, dass die Residuenanalyse aufgrund von Anschauungszwecken auf das Modell, der bereits unterteilten Daten erstellt wurde und somit nur Preise von Dachgeschosswohnungen schätzen kann.

5 Resultate

Für einige Kategorien wie bei den Dachgeschosswohnungen ist eine einfache lineare Regression durchaus ein Modell, mit welchem man einfach die abhängige Variable Preis mit der unabhängigen Variable Wohnfläche vorhersagen kann. Dennoch ist die Grenze dieses Modell rasch erreicht, da es die Abhängigkeiten mit nur einer variable doch stark vereinfacht.