Seminar

Feinstaubkonzentration in der Luft

Thi Thanh Tu Phan (569723)

28. Februar 2022

Betreuende Dozentin: Alla Petukhina

Inhaltsverzeichnis

1. Motivation **2**

2. Regression **3**

2.1. Multiple lineare Regression 3

2.2. Lasso Regression 4

3. Datenanalyse **5**

4. Implementierung in R **6**

5. Fazit **9**

6. Literatur **9**

1. Motivation

Saubere Luft ist die Grundlage eines gesunden Lebens in der Stadt und auf dem Land. Besonders in den Innenstädten trägt das erhöhte Verkehrsaufkommen zur Belastung der Luftqualität bei. Insbesondere Feinstaub und Schadstoffe wie Stickstoffdioxid beeinträchtigen die Umwelt und die Gesundheit der Menschen. Feinstaub ist ein Teil des Schwebstaubs, die nicht sofort zu Boden sinken, sondern eine gewisse Zeit in der Atmosphäre verweilen. Je nach Korngröße werden Staubpartikel in verschiedene Klassen eingeteilt. Als Feinstaub (PM10) bezeichnet man Partikel mit einem aerodynamischen Durchmesser von weniger als 10 Mikrometer (µm). Je nach Größe und Eindringtiefe der Teilchen sind die gesundheitlichen Wirkungen von Feinstaub verschieden wie chronischer Husten, Lungeninfektion, Lungenkrebs, Herzerkrankungen. PM10 kann beim Menschen in die Nasenhöhle eindringen.

Abbildung 1: Tägliche Feinstaubkonzentration in Oslo, Norwegen

Obige Abbildung ist die tägliche Konzentration von Feinstaub in Oslo, Norwegen. Auf der Y-Achse wird die Konzentration in ug/m3 (mikrogramm pro kubikmeter) und auf der X-Achse das Messdatum angezeigt. Auffällig ist es, dass fast alle Tage im grünen Bereich liegen. Das heißt, die Konzentration von Feinstaub liegt unter 50 ug/m3. Außer am 17. und 21 Januar zeigen die Werte in den gelben Bereich. Die liegt bei über 50 ug/m3.

Laut EU-Feinstaubrichtlinie darf der Grenzwert von 50 Mikrogramm pro Kubikmeter Luft an einer Station nur an 35 Tagen im Jahr überschritten werden.

Um diese Änderung zu sehen, wurde eine Teilstichprobe von einer Studie in Oslo, Norwegen zwischen 2001 und August 2003 durchgeführt. Die Daten sind eine Teilstichprobe von 500 Beobachtungen aus einem Datensatz. Die Werte wurden an einer Straße mit hohem Verkehrsaufkommen und meteorologischen Variablen, in der auch die Luftverschmutzung vorkommt, von der norwegischen öffentlichen Straßenverwaltung erhoben. Gemessen wurden die stündlichen Werte des Logarithmus der Konzentration von Feinstaub (PM10) in abhängig von dem Logarithmus der *Anzahl der Autos pro Stunde*, *Temperaturmessung 2 Meter über Grund (Grad C)*, *die Temperaturdifferenz gemessen zwischen 25 und 2 Meter über Grund (Grad C)*, *Windgeschwindigkeit (Meter/Sekunde)*, *Windrichtung (Grad zwischen 0 und 360)*, *Tagesstunde* und *Tagesnummer ab dem 1. Oktober 2001.* Die Werte wurden in einem Datensatz, der "PM10 - Kopie.txt" heißt, zusammengefasst. Dabei soll die Fragestellung, inwiefern sich der Feinstaubkonzentration in der Luft durch Verkehrsaufkommen und Wettereinflüsse verändert, beantwortet werden. Dazu werden wir zunächst mit der statistischen Modellierung mittels linearer Regression und Lasso Regression bevor wir unseren Datensatz Analyse betreiben. Anschließen werden wir unsere Ergebnisse zusammentragen und auswerten. Für die Auswertung des Datensatzes werden wir RStudio, die integrierte Entwicklungsumgebung und grafische Benutzeroberfläche für die Statistik-Programmiersprache R, verwenden.

2. Regression

Regressionsanalysen sind statistische Analyseverfahren, die zum Ziel haben, Beziehungen zwischen abhängigen und unabhängigen Variablen zu modellieren. Sie werden insbesondere verwendet, wenn Zusammenhänge quantitativ zu beschreiben oder Werte der abhängigen Variablen zu prognostizieren sind.

2.1. Multiple lineare Regression

Die erste Methode, die zur Analyse verwendet werden soll, ist die multiple lineare Regression, welche zum Supervised Learning gehört. Ziel dieser ist es, einen linearen Zusammenhang zwischen mehreren Variablen zu modellieren. Es wird also ein Modell mit

einer abhängigen und mehreren unabhängigen Variablen aufgestellt. Die Modellgleichung multiplen linearen Modells lautet:

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon_i$$

mit
Y_i = Schätzer der abhängigen Variable
β_k = Regressionskoeffizient der Variable
X_k = Unabhängige Variable k
ϵ_i = Fehlerterm

Das multiple lineare Regressionsmodell lässt sich in Matrixschreibweise wie folgt formulieren:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

mit $\mathbf{Y} \in \mathbb{R}^{\mathrm{n}}, \mathbf{X} \in \mathbb{R}^{\mathrm{nxK}}, \beta \in \mathbb{R}^{\mathbb{K}}$ und $\epsilon \in \mathbb{R}^{\mathrm{n}}$.

Dabei wird angenommen, dass $E(\epsilon) = 0$ und $E(\epsilon_i^2) = o^2, i = 1, 2, ..., n$ und $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$. Außerdem gilt $rg(\mathbf{X}) = k$.

2.2. Lasso Regression

In manchen Situationen kommt es bei einer multiplen linearen Regression vor, dass die Prädiktorvariablen (erklärende Variablen) stark korreliert sind. So kann Multikollinearität zu einem Problem werden. Dies kann dazu führen, dass die Koeffizientenschätzungen des Modells unzuverlässig sind und eine hohe Varianz aufweisen. Das heißt, wenn das Modell auf einen neuen Datensatz angewendet wird, den es zuvor noch nicht gesehen hat, ist die Leistung wahrscheinlich schlecht. Die Grundidee der Lasso-Regression besteht darin, eine kleine Verzerrung einzuführen, damit die Varianz wesentlich verringert werden kann, was zu einer niedrigeren Gesamt-MSE führt. Hierfür ist der Term

$$arg\ min(\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|)$$

zu minimieren, wobei $L_1 = \sum_{j=1}^{p}|\beta_j|$ der Fehlerterm ist. Das zugehörige $\beta =: \hat{\beta}^{lasso}$ ist dann der Lasso Schätzer. Bei genügend großem λ werden so durch den Strafterm L_1 möglicherweise einige Parameter null und können aus dem Modell entfernt werden.

Lasso unterscheidet sich nur durch die über den Strafterm λ geregelten Penalty von einer linearen Regression. Wird $\lambda = 0$ gewählt, ist Lasso identisch zur linearen Regression. Je höher λ gewählt wird desto größer unterscheidet sich Lasso von der Linearen Regression.

3. Datenanalyse

Da die Variabke Tageszahl ab dem 1.10.2001 ledeglich nur als Information, wann die Stichprobe gemessen wird, kann man aus der Datensatz zur Untersuchung entfernen. Wir untersuchen nun die der Konzentration von Feinstaub (PM10) in abhängig von dem Logarithmus der *Anzahl der Autos pro Stunde*, *Temperaturmessung 2 Meter über Grund (Grad C)*, *die Temperaturdifferenz gemessen zwischen 25 und 2 Meter über Grund (Grad C)*, *Windgeschwindigkeit (Meter/Sekunde)*, *Windrichtung (Grad zwischen 0 und 360)*, *Tagesstunde*.
Wenn wir diese Siginfikanzentscheidungen nutzen wollen, um die Effekte in der Population auf diese Weise zu interpretieren, so müssen einige Voraussetzungen erfüllt sein, die zunächst noch geprüft werden müssten. Eine Voraussetzung für die Signifikanztests im Kontext der linearen Regression ist die Normalverteilung der Residuen (Als Residuum wird die Abweichung eines durch ein mathematisches Modell vorhergesagten Wertes vom tatsächlich beobachteten Wert bezeichnet.). Auch diese Annahme wird i.d.R. grafisch geprüft. Hierfür bietet sich ein Histogramm der Residuen an.Das nebenstehende Histogramm zeigt keine großen Verstöße gegen die Normalverteilungsannahme.

Eine grafische Prüfung der partiellen Linearität zwischen den einzelnen Prädiktorvariablen und dem Kriterium kann durch partielle Regressionsplots (engl. Partialplots) erfolgen. Dafür sagen wir in einem Zwischenschritt einen einzelnen Prädiktor durch alle übrigen Prädiktoren im Modell vorher und speichern die Residuen, die sich aus dieser Regression ergeben. Die hier dargestellten Plots zeigt die Korelation der einzelne erklärende Variable auf die Feibstaubkonzentration nach Pear-

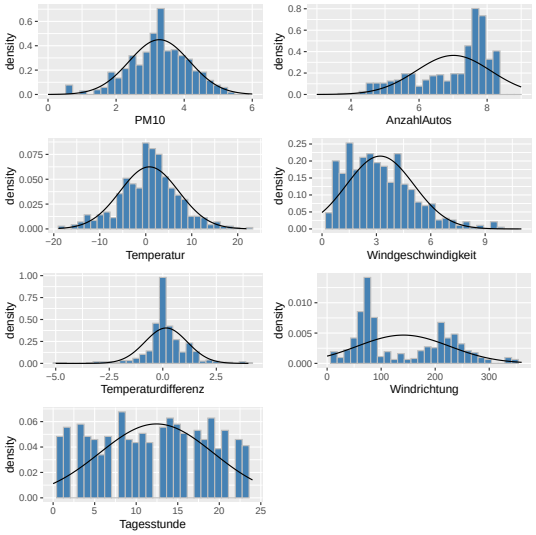

Abbildung 2: Normalverteilung der einzelnen Variablen

son. Der Korrelationskoeffizient R ist
das Maß für den Zusammenhang zwischen den beiden Variablen. der p-Wert. Überprüft, ob sich der Korrelationskoeffizient signifikant von Null unterscheidet. Hier sieht man eine positive Korrelation zwischen PM10-Konzentration und Anzahl der Autos sowie zwischen Windgeschwindigkeit und PM10 besteht. Sieht man sich Daten für Temperatur und PM10 sowie Windrichtung und PM10 und Tagesstunde und PM10 an, erkennt man fast keinen Zusammenhang.

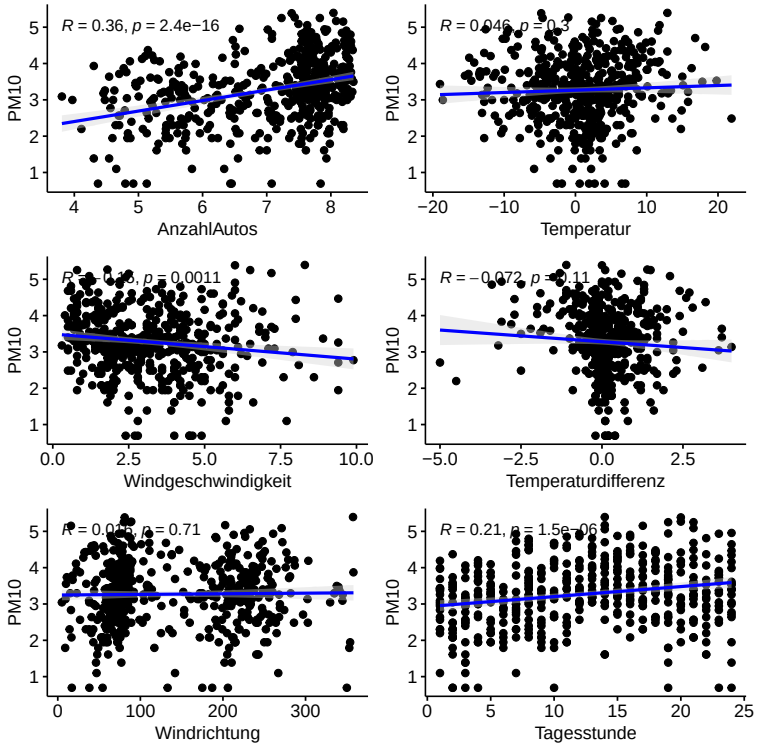

Abbildung 3: Korrelation der erklärende Variable auf PM10

4. Implementierung in R

Im Ergebnis der Regression, wird der Effekt jedes Faktors auf die abhängige Variable zu sehen sein. Dabei ist dieser Effekt jeweils für die Einflüsse der anderen im Modell enthaltenen Variablen kontrolliert und kann also unabhängig von deren Einfluss interpretiert werden. Der Effekt wird in Form des Regressionskoeffizienten angegeben, der in

der Höhe die Stärke und mit dem Vorzeichen die Richtung des Effekts beschreibt. Zusätzlich gibt der p-Wert an, ob dieser Effekt statistisch signifikant ist. Mit dem angepassten R-Quadrat-Wert wird die Güte des Modells beschrieben. Um das lineare Regressionsmodell anzupassen, wird der bestehende Algorithmus in der ersten Codezeile unten mit der Funktion lm() instanziiert. Die zweite Zeile gibt die Zusammenfassung des trainierten Modells an.

```
mlm      = lm(PM10 ~ AnzahlAutos+ Temperatur + Windgeschwindigkeit +
           Temperaturdifferenz + Windrichtung + Tagesstunde,
           data = dataset)
summary(mlm)
```

Output

```
Call:
lm(formula = PM10 ~ AnzahlAutos + Temperatur + Windgeschwindigkeit +
    Temperaturdifferenz + Windrichtung + Tagesstunde, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89425 -0.48252 -0.00869  0.50873  2.34489

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.337e+00  2.874e-01   4.653 4.21e-06 ***
AnzahlAutos          3.242e-01  4.357e-02   7.441 4.49e-13 ***
Temperatur          -9.339e-04  6.623e-03  -0.141    0.888
Windgeschwindigkeit -1.036e-01  2.081e-02  -4.980 8.81e-07 ***
Temperaturdifferenz  4.778e-03  4.218e-02   0.113    0.910
Windrichtung        -4.847e-05  4.567e-04  -0.106    0.916
Tagesstunde         -2.598e-05  6.463e-03  -0.004    0.997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8114 on 493 degrees of freedom
Multiple R-squared:  0.173,Adjusted R-squared:  0.1629
F-statistic: 17.19 on 6 and 493 DF,  p-value: < 2.2e-16
```

Die Teststatistik wird durch F-Test durchgeführt, der überprüft ob die Regressoren einen gemeinsamen Erklärungswert haben. Aus der Ausgabe durch die `summary()`-Funktion kann man erkennen, dass die Anzahl der Autos pro Stunde und die Windgeschwindigkeit (Meter/Sekunde) einen hohen signifikanten Einfluss auf die Konzentration von Feinstaub haben. Alle anderen Variablen wie *Temperaturmessung 2 Meter über Grund (Grad C), die Temperaturdifferenz gemessen zwischen 25 und 2 Meter über Grund (Grad C), Windrichtung (Grad zwischen 0 und 360)* und *Tagesstunde* gar keine Signifikanz auf unsere Zielvariable. Eine gute Erklärung für die Varianz der Zielvariable liefert dieses Modell jedoch nicht, was am niedrigen R^2 zu erkennen ist.

Nun wenden wir die Lasso Regression an, um unsere multiple linerare Regression zu verbessern. Die Lasso Regression lässt sich in R mit dem `glmnet`-Paket durchführen. Anders als bei der `lm()`-Funktion ist nun eine Matrix `x` und ein Vektor `y` an die Funktion `glmnet()` zu übergeben. Zuerst teilen wir den Datensatz in Test- und Trainingssatz auf. Hier wird der Datensatz zufällig aufgeteilt, wobei 70% den Trainings- und 30% den Testdatensatz bilden.

```
x        = model.matrix(PM10~.,dataset)[,-1]
y        = dataset$PM10
```

Der Schritt zum Erstellen eines Lasso Modells besteht darin, den optimalen Wert mithilfe des folgenden Codes zu finden. Dazu setzt man `alpha=1`.

```
grid      = 10^seq(2, -2, by = -.1)
lasso.mod = glmnet(x[train,],y[train],alpha=1,lambda=grid)
cv.out    = cv.glmnet(x[train,],y[train],alpha=1)
bestlam   = cv.out$lambda.min
[1] 0.02474522
```

Das beste λ wird dann zur Variablenselektion verwendet.

```
out        = glmnet(x,y,alpha=1,lambda=grid)
lasso.coef = predict(out,type="coefficients",s=bestlam)[1:7,]
lasso.coef
lasso.coef[lasso.coef!=0]
```

Als Ausgabe sehen wir, dass fast alle erklärende Variablen auf null gesetzt wurde. Somit hat dieser Prädikator kein Einfluss auf unsere Zielvariable. Anzahl der Autos pro Stunde und die Windgeschwindigkeit sind für die PM10-Konzentration veranwortlicht.

```
(Intercept)         AnzahlAutos      Temperatur       Windgeschwindigkeit
1.49488319          0.29360700      0.00000000           -0.08814452

Temperaturdifferenz        Windrichtung        Tagesstunde
0.00000000                 0.00000000          0.00000000
```

5. Fazit

Mit Hilfe der beiden Methoden Lineare Regression und Lasso Regression können wir unsere Fragestellung beantworten. Man kann feststellen, dass die Prädikatoren *Anzahl der Autos pro Stunde* und *Windgeschwindigkeit (Meter/Sekunde)* für die Feinstaunkonzentration verantwortlicht sind. *Temperaturmessung 2 Meter über Grund (Grad C)*, *die Temperaturdifferenz gemessen zwischen 25 und 2 Meter über Grund (Grad C)*,*Tagesstunde* und *Tagesnummer ab dem 1. Oktober 2001* haben fast keinen Einfluss darauf. Jedoch kann man durch die Multiple Lineare Regression erkennen, wie stark die einzelnen Prädikatoren sich auf unsere Zielvariable auswirken. Eine Erweiterung mit der Lasso Regression hift uns die irrelevante Variable aus der Modell zu entfernen.
Eine Regressiongleichung zur Vorhersage auf die Konzentration von Feinstaub lautet:
PM10 = 0.29360700*$x_1 - 0.08814452 * x_2 + 1.49488319 * x$
mit
$x_1 = AnzahlderAutosproStunde$
$x_2 = Windgeschwindigkeit$

6. Literatur

https://statistikguru.de/spss/produkt-moment-korrelation/ergebnisse-interpretieren.html
https://www.luft-reinheitsgebot.de/folgen/
http://lib.stat.cmu.edu/datasets/
https://pandar.netlify.app/post/regression-und-ausreisserdiagnostik/
https://www.kinews.de/ki-methoden/lasso-regression-einfach-gemacht/
https://github.com/TTTuPhan/Feinstaub-PM10.git **Skript in der Vorlesung**

[1] Statistical Learning - Vorlesung von Prof. Christina Erlwein-Sayer in SoSe2021.

[2] Statistik 3 Vorlesung von Prof. Christina Erlwein-Sayer in SoSe2021.

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Alle sinngemäß und wörtlich übernommenen Textstellen aus fremden Quellen wurden kenntlich gemacht.

Berlin, den 28.02.2022

Thi Thanh Tu Phan