Matthias Titze, s0563413

NHANES

National Health and Nutrition Survey

Durch die Untersuchung des „Food-Atlas“ konnte wenige Korrelationen zwischen Übergewicht (bzw. Adipositas) und sozioökonomischen Faktoren erkannt werden. Dies könnte aber auch in der Natur dieser Daten begründet liegen. Der Atlas bildet durch Bündelung von Erhebungen auf County-Ebene nur sehr grobe prozentuale Anteile für eine Vielzahl von Indikatoren. Über Haushalte und Individuen wird keine Aussage möglich. Dies spiegelt sich insbesondere in der Tatsache, dass die einzelnen Merkmale strukturell nicht verknüpft sind. Wir können zwangsläufig nicht wissen, ob etwa bei einem höheren Prozentsatz an einkommensschwachen Familien und gleichsam größeren Anteil an adipösen Patienten auch tatsächlich die Ärmeren stärker zu Übergewicht tendieren. Die Beziehung bleibt unbestimmt.

Lösungsansatz: Aus diesem Grund wird in diesem Abschnitt ein weiterer Datensatz untersucht: NHANES, eine großflächig angelegte Studie zur Ernährung und Gesundheit in den USA. Im Rahmen dieser Erhebung wurden über einen beachtlichen Zeitraum und im Abstand von zwei Jahren Teilnehmer klinisch untersucht. Dies erfolgte in speziellen Bussen oder Trucks, welche als mobile Kliniken durch die Staaten reisten. Da innerhalb dieser Erhebungen auch einige demographische und sozioökonomische Indikatoren erfasst wurden, verfügen wir so über einen Datensatz, über den sich Beziehungen realistisch erforschen lassen. Die Daten sind anonymisiert, aber über einen Patientencode verknüpft.

Jahrgänge und Umfang: Da NHANES eine Reihe von Studien umfasst, wählen wir die fünf Jahrgänge aus, die sich grob mit den schon im Atlas sondierten Zeiträumen decken: 2007-2015. Jeder dieser Abschnitte enthält rund 10.000 Untersuchungsdaten. Die über die Jahre durchlaufenden Patientennummern erleichtern das Zusammenfügen. Wir erhalten so einen umfassenden Satz von mehr als 48.000 Zeilen.

Bereinigen und kategorisieren der Daten

Die Daten sind weiterhin auf Aussagekraft, statistische Standards und fehlende Daten hin zu überprüfen. Die entsprechende Online Portal enthält die dafür notwendige Dokumentation, welche auch die Bedeutung der codierten Merkmale entschlüsselt. Im Rahmen dieser Arbeit wurden die Felder zum Teil umbenannt und ergänzt. Schließlich wurden stetige Daten zu diskreten oder kategorischen Daten konvertiert. In diesem Rahmen erfolgte auch eine Prüfung der Verteilung der Daten, mit dem Ziel widersprüchliche oder wenig representative Merkmale zu erfassen. Auf eine detaillierte Beschreibung dieses Prozesses soll hier verzichtet werden. Die vollständige Dokumentation liegt jedoch im Anhang vor. Dort finden sich ebenso kritische Überlegungen zur Auswahl der untersuchten Merkmale.

BMI und Einkommen

Die Untersuchung einer Korrelation von Body-Mass-Index und Einkommen ist der Ausgangspunkt für die Erforschung von sozioökonomischen Faktoren, welche in einem Zusammenhang mit der grassierenden Fettleibigkeit in den USA stehen könnten. Ähnlich wie zuvor mit den Daten aus dem Food Atlas, wird hier nun über den NHANES-Datensatz zunächst das Einkommen der Haushalte als Kernindikator herausgegriffen und in Beziehung zum BMI gesetzt.

Beide Datenfelder sind über die Identifikationsnummer des untersuchten Teilnehmers verknüpft und wurden zuvor kategorisiert (Klassenbildung). Somit lassen sich die absoluten Häufigkeiten der jeweiligen Merkmalsausprägungen zählen. Daraus ist eine Kontingenztabelle mit m Reihen für die BMI-Klassen und n Spalten für die Einkommensgruppen zu erstellen. Mittels dieser Tabelle lassen sich schließlich mit Chi-Quadrat-Test die Unabhängigkeit der Variablen prüfen.

Hypothesen

Als Hypothesenpaar für diese Untersuchung werden aufgestellt:

Nullhypothese H_0: Übergewicht und Fettleibigkeit korrelieren negativ mit Einkommen. Das heißt, mit sinkenden Einkommen steigt die Wahrscheinlichkeit für ungesundes Körpergewicht.

Alternativhypothese H_1: Übergewicht und Einkommen korrelieren nicht. Es besteht kein erkennbarer Zusammenhang zwischen Fettleibigkeit und diesem sozioökonomischen Indikator.

Selektion der Daten

Wir wählen zunächst die kategorisierten Tabellenspalten für BMI und Haushaltseinkommen aus.

BMI: Es stehen folgende Merkmalsausprägungen für den Body-Mass-Index zur Verfügung: {10-18; 18-25; 25-30; 30-35; 35-40; 40+; unbekannt}. Diese Ausprägungen sind geordnet aber nicht in streng gleichförmigen Intervallen. Sie repräsentieren die gesundheitlich relevanten Kategorien: Untergewicht, Normalgewicht, Übergewicht, Fettleibigkeit 1-3.

Einkommen: Die Einkommensklassen beschreiben Intervalle mit unterschiedlicher (ansteigender) Breite. Dazu kommt wiederum der nach oben offene Maximalwert "100+K" sowie die weniger präzisen Ausweichkategorien "<20K" und ">20K". Also: {000-005K; 005-010K; 010-015K; 015-020K; 020-025K; 025-035K; 035-045K; 045-055K; 055-065K; 065-075K; 075-100K; 100+K; <20K; >20K; unbekannt}.

Fehlende Daten: Von den 47873 Datenreihen enthalten einige die nicht aussagekräftigen "unbekannt" Einträge, welche ausgeschlossen werden. Damit reduziert sich der Datensatz auf 41800 Stichproben. Diese Menge scheint immer noch umfassend genug und deutlich weniger verzerrt als eine mit fiktiven Daten aufgefüllte Tabelle. Da die Einkommensklassen "<20K" und „>20K" die anderen Intervalle überlagern, müssen diese eventuell auch aus weiteren Test heraus genommen werden.

Kontingenztabelle BMI & Einkommensgruppen

Der selektierte Datensatz kann über Pivotieren in eine Kontingenztabelle umgewandelt werden. Als Reihen werden nun die BMI-Klassen und als Spalten die Einkommensgruppen abgebildet. Die jeweiligen Ausprägungen werden dabei zu Zählungen der Häufigkeiten aggregiert. Eine weitere Tabelle der relativen Häufigkeiten erhalten wir über Division der absoluten Häufigkeiten durch die Spaltensumme. Dieser Tabelle ist schließlich zu entnehmen, wie in jeder Einkommensgruppe, die BMI-Klassen prozentual verteilt sind.

Visuelle Analyse von BMI in Einkommensgruppen

Für die Untersuchung der Fettleibigkeit sind vor allem die BMI-Gruppen "25-30" (Übergewicht), "30-35" (Adipositas 1), 35-40" (Adipositas 2) und "40+" (Adipositas 3) interessant. Sie werden aus der Kontingenztabelle der prozentualen Verteilung gelesen und als gestapelte Balken für jede Einkommensgruppe dargestellt.

Das Resultat überrascht: Zunächst wird das Ausmaß des Problems deutlich - in den USA liegt der Anteil an Menschen mit krankhafter Fettleibigkeit im Schnitt bei knapp 30%. Werden übergewichtige Personen hinzu gezählt, erreicht der Anteil fast 50%. Eine Anstieg des prozentualen Anteils von übergewichtigen oder fettleibigen Menschen mit sinkendem Einkommen ist in keiner Form zu erkennen. Stattdessen finden wir in den untersten Einkommensklassen sogar einen Rückgang. Dies ist bei einem Jahreseinkommen von unter 5.000 vielleicht nicht ganz verwunderlich. In der Spanne von 10.000-75.000 scheint die prozentuale Verteilung von allen betrachteten BMI-Klassen jedoch eher gleichförmig zu verlaufen. Jenseits von kleineren Schwankungen ist kein Trend auszumachen. Erst ab 75.000 sehen wir einen minimalen und ab 100.000 einen deutlichen Rückgang der Fettleibigkeit.

Besonderheiten: sind in den übergreifenden Einkommensgruppen "<20K" und ">20K" zu finden. Hier scheint sich der vermutete Zusammenhang eher abzubilden. Bei jedoch gerade einmal 141 dargestellten

Proben in der viel breiteren Einkommensklasse "<20K", sollte dieser Beobachtung jedoch nicht zu viel Bedeutung beigemessen werden.

Chi2-Test für BMI und Einkommensgruppe

Um einen etwaigen Zusammenhang zwischen Fettleibigkeit und Einkommensklasse weiter mit statistischen Methoden zu erforschen, soll der Chi-Quadrat-Test zur Anwendung kommen. Dieser bietet sich zunächst an, da hier mit kategorischen Daten und einer Kontingenztabelle gearbeitet wurde. Der Test kann nur nachweisen ob die zwei Merkmale voneinander im Grundsatz statistisch unabhängig sind. Er basiert auf der Annahme, dass die zu erwartenden Werte innerhalb einer Kontingenztabelle über die Randsummen berechnet werden können. Wir betrachten nun die Differenzen zu den durch die Stichproben erhalten Werten und berechnen Chi-Quadrat nach folgender Formel:

$$\chi^2 = \sum_{i=1}^{j} \sum_{k=1}^{m} \frac{(h_{ij} - \tilde{h_{ij}})^2}{\tilde{h_{ij}}}$$

Die Hypothesen müssen also entsprechend der Aussagekraft des Tests präzisiert und umgestellt werden:
Nullhypothese H_0: Übergewicht und Fettleibigkeit sind stochastisch unabhängig.
Alternativhypothese H_1: Übergewicht bzw. Fettleibigkeit und Einkommensklasse sind voneinander abhängige Variablen.
Sigma: Gewählt wird ein Signifikanzniveau von 5% (p = 0.05).
Die Kontingenztabelle für die Merkmale, Fettleibigkeit und Einkommensgruppe wurde bereits für die visuelle Analyse erstellt. Aus den Randsummen der Kontingenztabelle werden die erwarteten Häufigkeiten nach der folgenden Formel ermittelt:

$$\tilde{h_{ij}} = \frac{h_{i*} h_{*j}}{n}$$

Somit lässt sich eine weitere Tabelle für die erwarteten Werte und schließlich eine dritte für die Differenz von Stichprobe und Erwartung aufsetzen. Diese beiden Tabellen werden schließlich genutzt um Chi-Quadrat entsprechend obiger Formel zu ermitteln. Wir erhalten **301,21**.

Kontingenzkoeffizient für BMI und Einkommensgruppe (nach Pearson)

Der berechnete Wert für Chi-Quadrat ist abhängig von der Anzahl der Merkmalsausprägungen und der Beobachtungen und somit noch nicht besonders aussagekräftig. Daher werden weitere Kenngrößen ermittelt bevor die Thesenprüfung erfolgt. Der korrigierte Kontingenzkoeffizient nach Pearson bringt den Wert von Chi-Quadrat in einen normalisierten Wertebereich von 0-1. Dafür wird zunächst der normierte Kontingenzkoeffizient nach folgender Formel ermittelt, wobei n=9752 die Anzahl unserer Merkmalsausprägungen ist.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Dieser Wert wird mit durch Korrekturfaktor dividiert, um den Wertebereich von 0-1 voll auszuschöpfen. In diesem Fall ist die minimale Tabellendimension k = 6 (die Zahl der BMI-Ausprägungen).

$$C_{korr} = \frac{C}{C_{max}}, \quad C_{max} = \sqrt{\frac{k-1}{k}} \quad \text{mit} \quad k = min(x, y)$$

Für die Auswertung des Chi-Quadrat-Test wird schließlich noch die Zahl der Freiheitsgrade DF benötigt. Sie ist für die reduzierte Kontingenztabelle 55. $DF = (k-1)(m-1)$ mit $k = 6$ und $m = 12$

Auswertung Chi-Quadrat Test

Mit 55 Freiheitsgraden überschreiten wir nach Chi-Quadrat-Tabelle bei einem Wert von 73.311 das gewählte Signifikanzniveau von 5%. Das heiß die Wahrscheinlichkeit, das der Wert für Chi-Quadrat einen Wert von 73,311 annimmt liegt gerade bei 5%. Bei dieser geringen Wahrscheinlichkeit müssen wir von einer signifikanten Beobachtung sprechen und die Nullhypothese verwerfen. Da der berechnete Wert für die Merkmale BMI und Einkommensgruppe bei 301,213 und somit weit über 73,311 liegt, ist dies der Fall:

Die Nullhypothese, dass die untersuchten Merkmale stochastisch unabhängig sind, ist zu verwerfen.

Dies scheint den Ergebnissen der visuellen Analyse entgegen zu laufen. Es ist jedoch zu beachten, dass aus dem Chi-Quadrat-Test nur ein vager Zusammenhang abzuleiten ist. Es wird keine Aussage über die Form der Abhängigkeiten getroffen. Eine Betrachtung des korrigierten Kontingenzkoeffizienten gibt weitere Auskunft: Der Wert liegt mit 0,095 dicht bei 0. Die 9,5% stehen für eine verhältnismäßig schwache Abhängigkeit von BMI und Einkommensgruppe.

Lineare Regression und p-Wert für BMI und Einkommensgruppe

Da die Hypothese einer Unabhängigkeit der von BMI und Einkommen verworfen wurde, ist die Alternativhypothese anzunehmen. Über die Art des Zusammenhangs dieser Variablen besteht jedoch weiterhin keine Klarheit. Daher wird in einem weiteren Schritt mittels linearer Regression und p-Wert auf eine lineare Korrelation getestet. In diesem Kontext ist ein neuen Hypothesenpaar aufzustellen:

Nullhypothese H_0: Es besteht keine lineare Korrelation zwischen BMI und Einkommensgruppe.

Alternativhypothese H_1: BMI und Einkommen antikorrelieren. Bei steigendem Einkommen sinkt der prozentuale Anteil an Testpersonen mit Übergewicht und somit der durchschnittliche Body-Mass-Index.

Alpha-Level: Das gewählte Signifikanzniveau bleibt bei 5% (p = 0.05).

Einschränkungen und modifizierte Datenselektion

Da für die lineare Regression und den damit verbundenen p-Test eigentlich stetige Daten oder zumindest Daten mit über Intervallskalen benötigt werden, ist diese Untersuchung nur eingeschränkt möglich. Einige der Probleme lassen sich jedoch durch eine leicht modifizierte Datenselektion beheben:

Der BMI liegt nicht nur als kategorischer Wert vor, sondern kann aus den ursprünglichen Tabellen als genauer Wert extrahiert werden. Für dieses Merkmal liegen somit die verlangten stetige Daten vor.

Das Einkommen der Haushalte wurde hingegen nur kategorisch erfasst und mit einem numerisch geordneten Code von 1-14 betitelt. Dieser Code lässt sich auf eine Achse abbilden. Es bleibt jedoch zu beachten, dass die Codes 13 und 14 für die aus oben beschriebenen Gründen auszuschließenden Klassen, "<020K" und ">020K", stehen. Reihen mit diesen Werten verwerfen wir auch hier. Schließlich bleibt die Einschränkung bestehen, das die codierten Einkommensintervalle bei höherem Einkommen an Breite gewinnen.

Eine mögliche Lösung ist die Annäherung an eine stetige Verteilung durch Zuweisung eines Zufallswertes innerhalb des codierten Intervalls. Alternativ kann auch der Mittelwert des Intervalls gewählt werden. Eine derartige Transformation der Daten birgt natürlich Gefahren. Diesen kann aber durch eine Gegenprüfung mit den originalen kategorischen Daten entgegen gewirkt werden. Wir führen zwei mit unterschiedlichen Mängeln behaftete Tests zu einer Aussage zusammen.

Auswertung Scatterplot und Box-Whisker-Plot

Der erstellte Scatterplot mit bietet schon einen Überblick über die Verteilung der Daten. Die Regressionslinien verlaufen um den mittleren BMI von 25 eher horizontal. Es scheint also auch hier wenig Korrelation zwischen BMI und Einkommen zu geben. Durch die Trennung der Datenpunkte nach Geschlecht ist jedoch zumindest bei den Frauen eine leichte Antikorrelation zu erahnen - die Linie fällt. Bei den Männer steigt die Kurve hingegen minimal an. Im Allgemeinen sind die Punkte aber für jede Einkommensklasse ähnlich (normal-)verteilt.
Dies bestätigt auch eine Betrachtung des Box-Whisker-Plots, der zur Überprüfung auf eine Modifizierung der Einkommensdaten (Randomisierung innerhalb der Intervalle) verzichtet. Interessant ist dort die leichte Erhöhung des oberen "Whiskers" in den Einkommensklassen von 5.000-10.000 Dollar. Die Formen der krankhaften Fettleibigkeit scheinen in diesen unteren Einkommensklassen etwas häufiger aufzutreten.

Auswertung Lineare Regression und p-Test

Alternativ zu Alpha-Fehler verwenden wir beim p-Test die Überschreitungswahrscheinlichkeit p. Diese wird im Rahmen der linearen Regression (Python SciPy-Bibliotheksfunktion) ermittelt. Der p-Wert für die lineare Korrelation von BMI und Einkommen ist mit 1.25e-19 erstaunlich niedrig und liegt deutlich unter der Grenze von 5%. Bei isolierter Betrachtung wäre also in Übereinstimmung mit dem Chi-Quadrat-Test die Nullthese abzulehnen. Es gibt demnach eine lineare Korrelation zwischen BMI und Einkommen. Gleichzeitig macht diese Aussage nach der visuellen Einschätzung der Daten intuitiv wenig Sinn. Der R-Quadrat-Wert ist mit 0.002 außerordentlich klein - gerade 0.2% der Abweichung wird somit durch das Regressionsmodell erklärt. Nehmen wir nun noch die Steigung der Regressionslinie hinzu, welche fast bei 0 liegt, dann muss festgestellt werden, dass eigentlich keine Aussage getroffen wurde.

> *Die Nullhypothese kann nicht direkt durch den p-Wert verworfen werden. Die Alternativhypothese einer Antikorrelation von BMI und Einkommen ist aber auch nicht anzunehmen.*

Dieser statistische Test kommt zu keiner entscheidenden Aussage. Der geringe p-Wert täuscht.

Auswertung Korrelationskoeffizient (nach Bravais-Pearson)

Zur abschließenden Betrachtung werden noch einige Korrelationskoeffizienten nach Bravais-Pearson ermittelt. Sie sind ein Maß für die Stärke des Zusammenhangs. die Berechnung erfolgt durch Division der Kovarianz der beiden betrachteten Variablen durch das Produkt ihrer Standardabweichungen.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Für alle Stichproben zusammen liegt der Wert bei -4,48%. Werden nur die Frauen betrachtet, dann liegt der Wert bei -9.98%. Demnach bestätigt sich die im Scatterplot ausgemachte leichte Antikorrelation. In weiterer Übereinstimmung läuft der positive Wert von 2,32% dem Trend entgegen. Insgesamt muss festgestellt werden, dass aller Werte sehr dicht bei 0 liegen. Der Zusammenhang ist nachweislich äußerst schwach.

Fazit: BMI und Einkommensgruppe

Für die Ausgangsthese der Untersuchung eines Zusammenhangs von BMI und Einkommen kann die These:

> „Übergewicht und Fettleibigkeit korrelieren negativ mit Einkommen. Das heißt, mit sinkenden Einkommen steigt die Wahrscheinlichkeit für ungesundes Körpergewicht."

nicht bestätigt werden. Auch wenn sowohl der Chi-Quadrat als auch der p-Test einen Zusammenhang nicht grundsätzlich ausschließen, zeigen weitere Untersuchungen doch offensichtlich, dass die Datenlage zu diffus ist, um eine Annahme der These zu rechtfertigen. Damit widerspricht diese Untersuchung der Mehrzahl der Studien, die Fettleibigkeit durch geringes Einkommen begünstigt sehen. Vermutlich müssen weitere sozioökonomische Faktoren hinzu gezogen, werden, um eine besseres Bild zu gewinnen. Dies soll im Folgenden in verkürzter Form durchgeführt werden.

Präzisierung des Einkommens durch Mitglieder im Haushalt

Wie bereits angedacht, muss das Haushaltseinkommen eventuell noch weiter präzisiert werden, indem die Anzahl der Mitglieder im jeweiligen Haushalt einbezogen werden. Die Vermutung wäre, dass Familien mit geringem noch zusätzlich benachteiligt werden, da vielleicht mehr Personen versorgt werden müssen. Tragen wir aber das berechnete Einkommen pro Kopf gegen den BMI in einem Scatterplot auf, so verändert sich der Eindruck nicht wesentlich. Die zuvor schon schwache Antikorrelation wird sogar noch weiter ausnivelliert.

Heatmap für Einkommen und Zahl der Haushaltsmitglieder

Eine Untersuchung des Zusammenhangs von Einkommen und Mitgliederzahl mittels Heatmap, macht deutlich, dass gering verdienende Familien keineswegs mehr Personen (Großfamilien) zu versorgen haben. Stattdessen überwiegen in den unteren Einkommensklassen Ein- und Zweipersonenhaushalte, während vier Personen bei den Großverdienern Vorrang erhalten. Dies erklärt, warum die Mitgliederzahl im Haushalt, die Antikorrelation von BMI und Einkommen eher schwächt: Das Einkommen wird etwas angeglichen, wenn wir es auf die Köpfe der Familie verteilen. Dieser Faktor kann also in den weiteren Modellen vernachlässigt werden.

BMI und Bildung

Ein gemeinhin mit dem Einkommen assoziierter sozioökonomischer Faktor ist die Bildung. Durch Einbezug dieses Feldes hoffen wir, die bisher diffusen Aussagen zu schärfen. In einem ersten Schritt soll geklärt werden, ob der NHANES-Datensatz die positive Korrelation von Bildung und Einkommensgruppe auch tatsächlich abbildet. Danach kann dann geprüft werden, welche Zusammenhänge zwischen Fettleibigkeit und Bildung bestehen.

Anmerkung: Die Methodik der statistischen Nachweise mit Hypothesenpaar wurde bereits ausführlich beleuchtet. In diesem Abschnitt wird auf diese Präzision weniger Wert gelegt, um einen breiteren Überblick zu gewinnen. Zudem werden einige Nachweise stark zusammengefasst, um den Leser nicht zu langweilen. zur genaueren Überprüfung verweisen wir auf den Anhang.

Selektion der Daten

Für diese Untersuchung werden weiterhin die nutzen Daten zum BMI und Einkommen verwertet. Hinzu kommen Tabellenspalten zur Ausbildung der Referenzperson im Haushalt. Diese Information liegt in Form von Bildungskategorien vor, die als geordneter Code (Bildungslevel) eine Ordinalskala bilden.

Bildungslevel: {0_unknown; 1_<9th-Grade; 2_<12th-Grade; 3_Highschool; 4_AA-Degree; 5_College+}

Schließlich ist auch das Geschlecht der Testpersonen für eine differenzierte Analyse relevant.

Chi-Quadrat Test für Einkommen und Bildung

Nullhypothese H_0: Einkommen und Bildung stochastisch unabhängig.
Alternativhypothese H_1: Einkommen und Bildung sind voneinander abhängige Variablen.
Sigma: Gewählt wird ein Signifikanzniveau von 5% (p = 0.05).

Der Chi-Quadrat-Test wird wiederum über die Kontingenztabelle mit den Merkmalen Bildungslevel und Einkommensklasse ermittelt. Der berechnete Wert ist **10.133**. Bei vorhandenen 44 Freiheitsgraden würde die Wahrscheinlichkeit, dass Chi-Quadrat einen Wert von 60,46 erreicht bei 5% liegen. Dies ist das gewählte Signifikanzniveau. Da der berechnete Wert weit darüber liegt, ist die Nullhypothese zu verwerfen.

Bildung und Einkommen sind stochastisch abhängig.

Korrelationkoeffizient und Plotanalyse

Über den Kontingenzkoeffizienten (Bravais-Pearson) lässt sich für Einkommen und Bildung klar eine positive Korrelation nachweisen. Der entsprechende Wert liegt bei 39,83% Prozent. Dieser Zusammenhang ist auch im Boxplot abzulesen: Das mittlere Einkommen steigt streng monoton mit dem Bildungslevel und auch die Quartile liegen durchweg höher. Der Anstieg scheint nicht linear sondern quadratisch bis exponentiell zu erfolgen. Da es sich bei den Bildungsleveln aber nicht um Werte auf einer Intervallskala handelt, ist die Herstellung einer echten mathematischen Funktion jedoch nicht möglich. Die ohne Bildungslevel erfassten Haushaltseinkommen (fehlende Daten) verteilen sich recht gut normal um den Mittelwert aller anderen Einkommen. Wir erwarten daher wenig statistische Verzerrung bei Ausschluss dieser Stichproben, die Zudem nur einen Anteil von 3,24% umfassen (siehe Kategorisierung der NHANES-Daten). Eine alternative Darstellung, die die positive Korrelation von Bildung und Einkommen offen legt, ist die "Heatmap". Diese kann aus der zuvor erstellten Kontingenztabelle abgeleitet werden. Sie zeigt die für jede Einkommensgruppe prozentual, aus welchen Bildungsklassen sich die Testpersonen speisen. Das diagonale Hitzeband bestätigt die Ergebnisse.

BMI Verteilung nach Bildungslevel

Da nun der triviale Zusammenhang zwischen Bildung und Einkommen nachgewiesen wurde, ist zumindest klar gestellt, dass die NHANES-Daten überhaupt eine Aussagekraft besitzen. Um so mehr ist demnach auch den der Intuition entgegen laufenden Erkenntnissen zu trauen. In diesem Schritt wird nun die Beziehung zwischen Bildungslevel und BMI beleuchtet. Als Hypothese anzunehmen wäre eine negative Korrelation zwischen diesen Variablen. Bei einem höheren Bildungslevel müsste demnach der Anteil and Übergewichtigen und Fettleibigen deutlich sinken.

Violinplot der BMI-Verteilung nach Bildungslevel

Um der Vermutung nachzugehen kann ein Violinplot genutzt werden. Dieser stellt die Dichtefunktion des BMI für jedes Bildungslevel dar. Für zusätzlichen Informationsgewinn wird zudem nach Geschlecht differenziert. Auch hier zeigt sich ein Bild, dass der zuvor erfolgten Gegenüberstellung von BMI und Einkommen entspricht: Der BMI liegt bei den Frauen zunächst etwas höher. Über die ersten vier Bildunglevel liegen die Median- und Quartilslinien in etwa auf gleicher Höhe. Dies würde implizieren, dass das Problem von Übergewicht und Fettleibigkeit als Breitenphänomen vollkommen losgelöst vom Bildungsstand ist. Das Bildungslevel 5 (College und Universität) weist jedoch einen starken Rückgang in den höheren BMI-Klassen auf. Insbesondere bei den Frauen liegen die Kennlinien dort deutlich niedriger und

erstmals sogar unter den Werten der männlichen Testpersonen. Wie schon beim Einkommen gilt für Frauen eine negative Korrelation von Bildung und BMI, die allerdings erst auf dem höchsten Level signifikant greift. Bei den Männern ist diese Tendenz weniger stark ausgeprägt. Der Rückgang betrifft vor allem die Bereiche krankhafter Fettleibigkeit (oberes Quartil).

Korrelationkoeffizient für BMI und Bildungslevel

Der Korrelationskoeffizient für BMI und Bildung (nach Bravais-Pearson) liegt bei -4,9% und bestätigt somit die aus den Violinplots abgelesene leichte Antikorrelation. Gerade im Vergleich zu den bei der Verknüpfung von Einkommen und Bildung erreichten 39,9% wir deutlich, wie dünn der Hinweis auf eine lineare Beziehung ist. Die Spur einer Korrelation speist sich vor allem aus dem deutlichen Abfall des starken Übergewichts im höchsten Bildungslevel bei gleichzeitig ausbleibender Gegentendenz in allen anderen Stufen.

BMI und Ethnische Gruppierung

Für das Einwanderungsland USA stellt die Zugehörigkeit zu einer ethnischen Gruppierung einen vielleicht entscheidenden sozioökonomischen Faktor dar. Daher soll dieser hier nicht vernachlässigt werden. Wenn die Tendenz zu Übergewicht und Fettleibigkeit in einigen dieser Bevölkerungsanteile auffällig aus dem Mittel fällt, dann ist weiterhin zu untersuchen, welchen Bildungszugang diese Gruppe hat und ob sich eventuell dementsprechend auch die Einkommensverteilung unterscheidet.

Selektion der Daten

Für die Betrachtung der der ethnischen Zugehörigkeit stehen im NHANES-Datensatz vier Klassen zur Verfügung: {white; black; mexican; hispanic; other} .

Die Unterscheidung zwischen Mexicans und Lateinamerikanern anderer Regionen scheint zunächst etwas sonderbar. Da beide Gruppen aber mit einem umfassenden Datenkörper vertreten sind und in den USA durchaus in einem anderen sozialen Kontext stehen, macht eine solche Teilung Sinn. Etwas problematisch ist hingegen die Gruppe "other". Sie umfasst viele marginale Minderheiten, sowie auch die größere Gruppe der Asiaten. Eine weitere Differenzierung wäre hier wünschenswert, wurde im Rahmen des NHANES-Programms aber erst in den letzten drei der insgesamt fünf verwendeten Jahrgänge vorgenommen. Daher musste mit Rücksicht auf die Homogenität der Daten eine Anpassung auf das gröbere Level hingenommen werden. Immerhin sind so alle definierten Merkmalsausprägungen mit einem repräsentativen Anteil and der Grundgesamtheit vertreten. Fehlende Daten gibt es nicht.

Plotanalyse zur BMI Verteilung in ethnischen Gruppen

Ein Box-Whisker-Plot, welcher die Verteilung des BMI in den verschiedenen ethnischen Gruppen darstellt, ist durchaus aufschlussreich: Die Tendenz zu einem erhöhten BMI ist bei Afroamerikanern und mexikanischen Einwanderern deutlich ausgeprägter. Insbesondere bei den Afroamerikanern gibt es eine markante Verschiebung des oberen Quartils hinein in den Bereich der krankhaften Fettleibigkeit (BMI über 30). In dieser Gruppierung erreicht der obere Whisker auch den sehr bedenklichen BMI-Wert von 50. Obwohl also die Medianwerte bei vier der fünf ethnischen Gruppen gleichauf im leichten Übergewicht bei 26 liegen, gibt es bei Afroamerikanern und auch Mexikanern einen entschieden größeren Anteil an gesundheitlich kritischer Fettleibigkeit. Das ganze Gegenteil findet sich in der Sammelgruppe "other". Hier liegt der Median im Normalgewicht und das obere Quartil rückt runter in das leichte Übergewicht. Dies ist vermutlich stark bedingt durch den Anteil an Asiaten in dieser Gruppe, was in einer Nebenuntersuchung kurz geprüft wurde (hier aber nicht dargestellt werden soll).

BMI Verteilung in Ethnischen Gruppen - weitere Differenzierung nach Geschlecht

Eine weiter Differenzierung nach Geschlecht wird Mittels eines geteilten Violinplots möglich. Hier überrascht die Rolle der weiblichen Testpersonen innerhalb der afroamerikanischen Gruppe. Während sich die Verteilung des BMI der männlichen Teilnehmer der Studie nicht allzu stark von Weißen, Mexikanern und Lateinamerikanern unterscheidet, wird der zuvor ausgemachte BMI-Anstieg in den krankhaften Bereich vor allem durch Frauen realisiert. Die obere Quartilsline nähert sich stark einem BMI, der einer Adipositas Grad II entspricht. In der ethnischen Sammelgruppe "other" kehrt sich diese Struktur hingegen um.
Da der Anteil an Frauen mit Adipositas I-III in der ethnischen Gruppe der Afroamerikaner so markant ausfällt, stellen wir diesen Aspekt noch einmal als gestapeltes Balkendiagramm dar. Hier werden für alle Ethnien nur die Frauen in den vier erhöhten BMI Klassen erfasst. Deutlich ist zu erkennen, dass mit Ausnahme der "others"-Gruppe der Gesamtanteil von Übergewichtigen (mit 52%) in etwa gleich groß ist. Bei den afroamerikanischen Frauen findet jedoch eine innerhalb des Übergewichts eine starke Verschiebung in den krankhaften Bereich (Adipositas I-III) statt.

Heatmap des Median BMI nach Einkommen und ethnischer Gruppe

Betrachtung des BMI in Hinblick auf die ethnische Gruppe, kann zusätzlich nach Einkommen differenziert werden. Dies wird durch die Bildung einer Heatmap ermöglicht, die den median des BMI für jede Ethnie und Einkommensklasse kartiert. Dies verschafft nur einen ersten Überblick, da die dargestellten Werte noch keine präzise Aussage über den eigentlichen Anteil an Übergewichtigen machen. Ein erhöhter median weist jedoch auch auf größere Anteile in den oberen BMI-Klassen hin. Interessant sind hier vor allem zwei Beobachtungen: Sowohl bei den Mexikanern als auch bei den Afroamerikanern liegen die Spitzen in den oberen Einkommensklassen. Die Antikorrelation von BMI und Haushaltseinkommen gilt hier gerade nicht. Nur bei den weißen Amerikanern gibt es einen minimalen Schwerpunkt in einem Intervall von 10-15 Tausend Dollar. Da der Anteil der weißen Bevölkerung aber in den USA mit 36,1% überwiegt, nivellieren sich die widersprüchlichen Tendenzen aus.

Heatmap des Median BMI nach Alter und ethnischer Gruppe

Wird die Beziehung von ethnischer Gruppenzugehörigkeit und BMI weiter nach Altersgruppe aufgespalten, dann erschein die Tendenz zum Übergewicht bei den weißen und mexikanischen Amerikanern eher ein Altersproblem. Bei Afroamerikanern hingegen findet sich der Hotspot in den jüngerem Jahren von 35-40. Dies kehrt die Beobachtung der zuvor erfolgten Aufteilung nach Einkommen geradezu um. Interessant wäre demnach eine Verknüpfung von Alter und Einkommen nach Ethnie. Eventuell gibt es in der afroamerikanischen Bevölkerung gleichsam eine Verschiebung der höheren Einkommen in jüngere Altersklassen. Auf einen so gearteten Test soll in dieser Arbeit jedoch verzichtet werden.

Vorhersagemodell - Supervised Machine Learning

Im Folgenden soll versucht werden, ein Vorhersagemodell (Klassifizierung) für Untersuchte Eigenschaft der BMI-Klasse zu erstellen. Dazu verwenden wir Supervised Machine Learning Algorithmen, d.h. die vorhandenen Daten werden in Trainingsdaten und Testdaten unterteilt. Mit den ersteren Trainieren wir den Prognose-Algorithmus. Anschließend lassen wir diesen über die Testdaten eine Vorhersagetreffen treffen und vergleichen das Resultat mit den für die Testdaten vorliegenden, realen Merkmalsausprägungen. So wird eine Einschätzung bezüglich der Qualität des Modells möglich.
Limitationen: Da die bisherigen Untersuchungen gezeigt haben, dass es für die Tendenz zu Übergewicht und Fettleibigkeit keine entscheidenden sozioökonomischen Faktoren gibt, muss vermutet werden, dass auch eine Machine-Learning-Algorithmus keine klaren Aussagen treffen kann. Dies ist insbesondere für den

Ansatz der Klassifizierung des Körpergewichts, einer Eigenschaft, die eher normal verteilt ist, anzunehmen: Auch wenn es Faktoren gäbe, die eine Tendenz zur Adipositas leicht begünstigen würden, wäre die so gesteigerte Wahrscheinlichkeit immer noch geringer als der Druck hin zum Erwartungswert der Normalverteilung. Klassifizierung als Vorhersagemodell ist somit eher nicht der treffende Ansatz. Günstiger wäre ein Modell, welches eine Aussage darüber treffen würde, in welchem Maß die Wahrscheinlichkeit durch entsprechende Einflussfaktoren beeinflusst wird. Trotz dieser grundsätzlichen Problematik soll, die Klassifizierung mit Naive Bayes getestet werden. Treffen wir dort auf überraschende Ergebnisse, so findet sich im Umkehrschluss in den Daten doch eine Kombination von Eigenschaften, die kritisch für die Ausbildung von Adipositas ist.

Vorherzusagen (y): BMI-Klasse {Untergewicht, Normal, Übergewicht, Adipositas 1-3}
Genutzte Merkmale (xi): {Geschlecht, Ethnie, Anzahl Mitglieder im Haushalt, Haushaltseinkommen, Bildung, Alter}

Klassifizierung mit Naive Bayes (Gauss)

Der Naive Bayes Klassifikator basiert auf Bayes Theorem. Er wird als "naive" bezeichnet, da er die Unabhängigkeit aller einbezogenen Merkmalspaare im Datensatz voraussetzt. V = (x_1, x_2, …, x_n) sei der Vektor der Merkmale und y die Eigenschaft, die durch den Vektor voraus gesagt werden soll. Dann gilt:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

und wenn x_1, x_2,...,x_n weiterhin als stochastisch unabhängig angenommen werden:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Diese Formel reduziert sich für die Klassifizierung weiter zu:

$$P(y \mid x_1, \ldots, x_n) = P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

‚da die errechneten Wahrscheinlichkeiten nur auf das Maximum ausgewertet werden. $P(x_i)$ ist eine Konstante, die in allen so verglichenen Werteberechnungen auftritt und somit ignoriert werden kann. Im Falle des hier verwendeten Gauss-Naive-Klassifikators aus der Python ScikitLearn Blibliothek, wird eine Normalverteilung jedes Merkmals angenommen. Dann ergibt sich die Wahrscheinlich dafür zu:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

Potenzielle Fehlerquellen: Schon vorab muss der gewählte Klassifikator kritisch betrachtet werden. Die Merkmale, Einkommen und Ausbildung, sind zum Beispiel keinesfalls unabhängig. Dies wurde in der Datenexploration nachgewiesen. Des Weiteren ist eigentlich nicht davon Auszugehen, das die zur Vorhersage genutzten Eigenschaften normal verteilt sind.

Daten Selektieren und Modell Trainieren

Aus dem NHANES-Datensatz werden die oben genannten Merkmale als Grundlage für die Klassifizierung extrahiert. Vorauszusagen ist die BMI-Gruppe. Alle genutzten Daten liegen als kategorisierte Eigenschaften vor, die hier in Form von Integer-Zahlen codiert wurden. Auf fehlende Daten wurde der Einfachheit halber mit Ausschluss reagiert. Die Anzahl der Zeilen schrumpft so auf 40.686, was für den angestrebten Rahmen genügen soll. Der so extrahierte Datensatz wir anschließend zufällig in 75% Trainings- und 25% Testdaten unterteilt. Mit den ersteren trainieren wir den den Naive-Bayes-Gauss Klassifikator aus der Python Bibliothek. Anschließend erfolgt die Vorhersage auf den Trainingsdaten.

Bewertung des Modells

Die erhaltenen Vorhersagen können in Beziehung zu den tatsächlich vorliegenden Daten für den BMI abgeglichen werden.

Wertebereiche: Schon bei der Überprüfung der Wertebereiche tritt Bedenkliches zu Tage: Während die tatsachlichen Werte, alle BMI-Klassen umfassen, liegen die vorher gesagten Ausprägungen nur in den Kategorien -1 (Untergewicht) bis 1 (leichteres Übergewicht). Der Klassifikator hat also versagt, Personen mit Adipositas 1-3 zu erkennen. Davon war jedoch von Beginn an auszugehen, da die betrachteten Faktoren nur einen leichten Einfluss auf die Tendenz zur Fettleibigkeit haben. Der Naive Bayes wählt bekanntlich, die wahrscheinlichste Klasse aus. Auch wenn also Adipositas durch einige Merkmale etwas wahrscheinlicher würde, so bleibt der Erwartungswert der Normalverteilung eben doch dominant.

Genauigkeit: Die Genauigkeit des Klassifikators ist mit 40,68 Prozent schwach.

Modifikation der BMI-Klassen

Da die drei oberen BMI-Klassen (Adipositas 1-3) im Einzelnen bei allen Testpersonen deutlich weniger häufig auftreten sind sie auch unwahrscheinlicher. Wir erhoffen uns demnach eine leicht Verbesserte Aussage, wenn wir diese Klassen in eine allgemeinere Gruppe, "Adipositas", verschmelzen. Der Trainings- und Vorhersageprozess muss dann wiederholt werden.

Bewertung des angepassten Modells

Auch nach Neueinteilung der BMI-Klassen bleibt das Problem erhalten: Es besteht keine Möglichkeit, krankhafte Fettleibigkeit vorher zu sagen. Dies ist vielleicht auch beruhigend, da somit klar wird, dass sozioökonomische Faktoren eben nicht allein über das Körpergewicht entscheiden, sondern maximal eine Tendenz fördern könnten. Der Naive Bayes Klasifikator ist jedoch nicht dafür geeignet eine derartige Verschiebung abzubilden. Die Genauigkeit ist mit 41,17 Prozent nur leicht verbessert und könnte schon damit erklärt werden, dass wir die Personen richtig vorher sagen, die ohnehin im Mittelfeld liegen.

Konfusionsmatrix

Um die Trefferquote genauer zu beleuchten, hilft eine Matrix, welche die realen und die vorher gesagten Merkmalsausprägungen gegenüber stellt. Wir können also erkennen, wie viele Treffer in jeder Klasse gefunden wurden und wohin sich die Alpha-Fehler bewegt haben. In normalisierter Form ist der prozentuale Trefferanteil erfasst. In unserem Modell wird somit eine erstaunlich präzise Aussage über die Testpersonen mit Untergewicht getroffen. Von 2114 Teilnehmern werden 1960 richtig eingeschätzt. Das ist eine Trefferquote von 93 Prozent. Dieser gute Wert verlangt eigentlich eine Folgeuntersuchung, die ergründet, ob es markante Einzelindikatoren wie Alter (Jugendliche), Ethnie (Asiaten) und Einkommen (extreme Armut) gibt. Dies lag aber nicht im Fokus dieser Arbeit. Die Vorhersage der Normalgewichtigen ist sehr ungenau - nur 32 Prozent werden richtig erfasst. Viel zu viele Personen wurden ins Untergewicht

verschoben (wo dann ein großer Beta-Fehler vorliegt). Die Vorhersage der leicht Übergewichtigen ist mit 52 Prozent etwas genauer. Immerhin wird dort der Schwerpunkt richtig gesetzt. Das Hauptproblem liegt bei der Kategorie "Adipös": Die Vorhersage dazu hat sich komplett in den über- und normalgewichtigen Teil verschoben. Die Trefferquote liegt so bei 0.

Fazit

Das Klassifikator-Modell ist sehr schwach. Bei guter Aussagekraft erwarten wir die Konfusionsmatrix eine deutlich ausgeprägte Hauptdiagonale. Diese ist jedoch zumindest schwach zu erkennen - die Verteilung der Schätzwerte ist keinesfalls zufällig. Dass bedeutet vielleicht, dass die angedachten Zusammenhänge nicht ganz abzustreiten sind, durch den Naive-Bayes aber auch schlecht nachgewiesen werden können. Bei der Anwendung von Klassifizierern bleibt as Problem bestehen, dass leicht gesteigerte Wahrscheinlichkeiten keinen Ausdruck finden. Gerade für diese Untersuchung wäre eine Berechnung solcher Werte interessant. Die Methodik für eine derartige Annäherung an das Thema lag zum Zeitpunkt der Arbeit nicht in Reichweite.