

HUMBOLDT UNIVERSITÄT ZU BERLIN  
INSTITUT FÜR SOZIALWISSENSCHAFTEN  
BA SOZIALWISSENSCHAFTEN  
PS STATISTIK I - WS 2003/04

# Determinanten des Einkommens

## **Hausarbeit vorgelegt von:**

Dominik Hübner  
Matrikel: 178522  
dominickel@web.de

Daniel Kubiak  
Matrikel: 198892  
sir\_charlie@web.de

Norman Ludwig  
Matrikel: 194779  
opa\_joe@web.de

Andreas Westendorf  
Matrikel: 194710  
info@ungleichzeitig.de

# Inhaltsverzeichnis

1. Einleitung .....	3
2. Theoretischer Hintergrund .....	4
2.1 Theorien.....	4
2.1.1 Die Humankapitaltheorie .....	4
2.1.2 Die Klassentheorie .....	5
2.1.3 Soziales Kapital .....	6
2.1.4 Weitere Einflussfaktoren .....	6
2.2 Hypothesen.....	7
2.2.1 Zur Humankapitaltheorie .....	7
2.2.2 Zur Klassentheorie .....	7
2.2.3 Zum Sozialen Kapital .....	7
3. Mathematischer Hintergrund.....	8
3.1 Grundsätzliches zur linearen Regression.....	8
3.2 OLS- und Koeffizientenschätzung .....	9
3.3 Signifikanztests und Modellgüte .....	9
3.4 Prinzipien der Voraussetzungsprüfungen .....	11
3.4.1 Linearität.....	11
3.4.2 Homoskedastizität .....	12
3.4.3 Unabhängigkeit der Fehler.....	12
3.4.4. Vermeidung von Multikollinearität .....	13
4. Umsetzung .....	13
4.1 Datenbasis .....	13
4.2 Operationalisierung.....	13
4.2.1 abhängige Variable .....	13
4.2.2 Variablen der Humankapitaltheorie.....	14
4.2.3 Variablen der Klassentheorie .....	14
4.2.4 Variablen des Sozialen Kapitals.....	14
4.2.5 Variablen weiterer Einflussfaktoren.....	15
4.2.6. Interaktionsvariablen .....	15
4.3 Ergebnisse der Modelle.....	17
4.3.1 Modell zur Humankapitaltheorie.....	19
4.3.2 Modell zur Klassentheorie .....	19
4.3.3 Modell zum Sozialen Kapital .....	19
4.4 Auswertung der erwarteten Interaktionseffekte .....	20
4.5 Gesamtmodell.....	21
4.5.1 Auswertung der Voraussetzungsprüfungen .....	21
4.5.2 Ergebnisse .....	25
5. Fazit .....	26
6. Literatur .....	27
7. Anhang.....	28

# 1. Einleitung

Das Einkommen ist für die Sozialwissenschaften eine der zentralen Größen, die es in Bezug auf soziale Ungleichheit zu untersuchen gilt. Das Einkommen stellt den entscheidenden meßbaren Indikator der Dimension des individuellen Wohlstandes dar und ist deshalb für Sozialwissenschaftler eine der Standardquellen bei der Untersuchung sozialer Strukturen einer Gesellschaft.

Das klare Ergebnis, dass sich die Einkommenshöhen in einer Gesellschaft ungleich verteilen und sich hier eine Struktur sozialer Ungleichheit offenbart, besitzt aber bezüglich ihres Zustandekommens noch keine Aussagekraft. Welche Faktoren sind für Einkommensunterschiede verantwortlich? Was sind die Determinanten des Einkommens? Im Mittelpunkt der folgenden Untersuchung soll die Frage stehen, welche Determinanten den durchschnittlichen Nettostundenlohn als Basis des Einkommens im wesentlichen beeinflussen.

Das Wochenmagazin „Stern“ schrieb am 22. Februar 2001, dass Frauen im EU-Durchschnitt 16 % weniger verdienen als Männer. Daran zeigt sich bereits eine der sich auf das Einkommen auswirkenden Determinanten, nämlich das Geschlecht.

Auch gibt es statistische Untersuchungen darüber, dass die Bürger in den neuen Bundesländern auch bald vierzehn Jahre nach der Wiedervereinigung weniger Gehalt beziehen als ihre Kollegen in vergleichbaren oder gar den gleichen Berufen in den alten Bundesländern. Im Beamtenwesen spielt oft das Dienstalter eine große Rolle für die Entlohnung. Es gibt scheinbar viele Möglichkeiten, welche Faktoren das Einkommen bestimmen.

In den, der Hausarbeit zu Grunde liegenden Schriften, „Einführung in die Arbeitspolitik“ von Berndt Keller als auch in „Marxist Class Categories“ and Income Inequalities“ von Wright und Perrone, werden noch weitere Determinanten angegeben, die sich auf das Einkommen auswirken können. Dabei stützt sich Keller in seinem Text auf den Ansatz der neoklassischen Humankapitaltheorie, Wright und Perrone beziehen sich hingegen in ihrer Arbeit auf die Überlegungen der marxistischen Klassentheorie.

Mit dieser Hausarbeit soll nun, anhand von Theorien und Hypothesen mit Hilfe der linearen Regression aufgezeigt werden, welche Variablen Einfluß auf das Einkommen eines Menschen nehmen.

## 2. Theoretischer Hintergrund

Es werden die Annahmen zweier konkurrierender Theorien, die der Humankapital- und die der Klassentheorie, abgeleitet aus der Literatur von Becker und Mincer sowie von Keller untersucht und außerdem ein von uns selbst entwickelter Erklärungsansatz, der von der Theorie sozialen Kapitals ausgeht, vorgestellt. Weiterhin werden strukturelle Einflussfaktoren, wie Geschlecht und Regionszugehörigkeit untersucht.

Abschließend wird ein Gesamtmodell entwickelt, welches die aus den Einzelmodellen gewonnenen Erkenntnisse berücksichtigt und zusammenfaßt.

### 2.1 Theorien

#### 2.1.1 Die Humankapitaltheorie

Berndt Kellers Text „Einführung in die Arbeitspolitik“ beschreibt die *Humankapitaltheorie* nach Becker und Mincer. Diese älteste theoretische Weiterentwicklung des neoklassischen Basismodells versucht individuelle Unterschiede im Einkommen dadurch zu erklären, dass Individuen in unterschiedlichem Maß über humanes Kapital verfügen. Humankapital meint dabei das durch Bildungs- und Qualifikationsmaßnahmen aber auch durch Berufserfahrung erworbene Leistungspotential. Es ist somit für die Beschäftigungs- und Einkommenschancen wesentlich entscheidend. Bildung und Ausbildung erhöhen die persönliche Qualifikation und sind daher als Investition in das Humankapital zu betrachten, wobei solange in Bildung investiert wird, wie die zu erwartenden Einkommenszuwächse die Ausbildungskosten und etwaigen Lohnausfälle übersteigen. Einkommensungleichheit entsteht also durch ungleiche persönliche Investition in das Humankapital. So verdient bspw. jemand, der Abitur hat, einen Bachelor-Abschluss in Sozialwissenschaften erworben hat und danach ein Volontariat in einer Redaktion machte (18 Bildungsjahre), mehr Geld, als jemand der den Hauptschulabschluß mit einer anschließenden dreijährigen überbetrieblichen Ausbildung (13 Bildungsjahre) absolvierte und schon früher zu arbeiten begann. Diese Theorie ergibt eine u-förmige Alters-Einkommens-Kurve, da während der Ausbildungszeit wenig bis gar kein Geld verdient wird, dann aber mit dem Berufseinstieg das Gehalt in den Anfangsjahren rasant ansteigt. Im Laufe des Alters

stagniert das Einkommen und sinkt mit dem Eintritt ins Rentenalter und vorhergehendem möglichem Ruhestandsgeld wieder ab. Keller stellt jedoch auch fest, dass weitere Faktoren, wie Geschlecht, Religion, Rasse etc. das Einkommen beeinflussen können. Diese Tatsache schwächt die Theorie des Humankapitals ab. So könne beispielsweise davon ausgegangen werden, dass Frauen mit 18 Bildungsjahren weniger verdienen als Männer mit der gleichen Anzahl an Bildungsjahren.

### **2.1.2 Die Klassentheorie**

Wright und Perrone beziehen sich in ihrer Schrift „Marxist Class Categories and Income Inequalities“ auf die Klassentheorie von Karl Marx und stellen fest, dass diese viel zu sehr vernachlässigt wird, wenn es um Studien sozialer Schichtung und Einkommensungleichheit geht. Sie nehmen einen Zusammenhang zwischen Klassenposition und dem Einkommen insofern an, dass die Zugehörigkeit zu einer Klasse das zu erwartende Einkommen bestimmt. Unterschieden werden die Klassen nach dem Besitz von Produktionsmitteln bzw. dem Kauf oder Verkauf von Arbeitskraft. Darüber hinaus spielt laut Wright/Perrone in der gegenwärtigen kapitalistischen Produktionsweise die Aufsicht bzw. Macht über den Produktionsprozess eine entscheidende Rolle. Im 19. Jahrhundert war die Macht über den Produktionsablauf noch unmittelbar mit der über das fertige Produkt in den Händen der Kapitalistenklasse verbunden. In der modernen Gesellschaft sind diese beiden Positionen aber tendenziell voneinander getrennt. Durch diese Trennung habe sich mittlerweile eine Stellung im Produktionsprozess herausgebildet, die zwischen den Unternehmern und den Arbeitern steht. Wright/Perrone bezeichnen sie als Klasse der Manager, die zwar keine Produktionsmittel besitzen, aber den Produktionsprozess kontrollieren.

Gemäß dem Klassenmodell von Wright/Perrone ergeben sich somit drei zentrale Klassen: Die Klasse der Arbeiter, die ihre Arbeitskraft verkaufen, die Klasse der Manager, die darüber hinaus die Aufsicht über den Produktionsprozess haben sowie die Klasse der Kapitalisten, die den Produktionsprozess beaufsichtigen, Produktionsmittel besitzen und fremde Arbeitskraft einkaufen. Daneben gäbe es, so Wright/Perrone, noch die Klasse der Kleinunternehmer (bzw. Kleinbürger), die aber ein Überbleibsel einer früheren Phase der kapitalistischen Entwicklung darstelle und mit dem Fortschreiten dieser Entwicklung weniger wichtig werde (vgl. Wright/Perrone

43 Spalte 1). In dieser Untersuchung die Klasse der Kleinunternehmer zumindest anfangs dennoch weiter berücksichtigt werden. Auch soll mit der Unterscheidung nach Beamten eine weitere eigene Klasse eingeführt werden, da ein für das Einkommen bedeutender Unterschied zwischen Beamten und Arbeitern in der Unkündbarkeit von Beamten und dem automatischen Gehaltszuwachs nach Berufsjahren liegt. Außerdem sind staatliche Einrichtungen tendenziell nicht in den warenproduzierenden Kreislauf eingebunden. Von daher ergibt sich für Beamte ein anderer ökonomischer Status als für Arbeiter.

Die Klassenzugehörigkeit beeinflusst aber laut Wright/Perrone nicht nur die Höhe des Einkommens, sie wirkt sich auch vermittelt auf andere Ungleichheit erklärende Variablen wie Bildung aus und somit auch darauf wie stark sich diese wiederum auf das Einkommen auswirken. Gemäß dessen müßte sich eine höhere Bildung bei Managern stärker auf das Einkommen auswirken als bei Arbeitern, da Manager bestimmte durch die höhere Bildung angelernte (bzw. zumindest bescheinigte) Fähigkeiten in ihrer Stellung im Produktionsprozess benötigen, Arbeiter hingegen nicht.

Im Idealfall sollte das Einkommen innerhalb einer Klasse nicht von Kategorien wie „Rasse“ oder Geschlecht abhängen, alleinig die Klassenzugehörigkeit sollte ausschlaggebend sein.

### **2.1.3 Soziales Kapital**

Im selbstentwickelten Erklärungsansatz wird angenommen, dass vor allem die Ressource soziales Kapital bzw. individuelle soziale Netzwerke ausschlaggebend für die persönlichen Beschäftigungs- und Einkommenschancen sind. Die Qualität des sozialen Kapitals ist dabei von besonderer Bedeutung, so dass unbedingt eine mehrdimensionale Betrachtung wie nach sozialer Herkunft, sozialen Beziehungen oder sozialen Aktivitäten erfolgen soll.

### **2.1.4 Weitere Einflussfaktoren**

Zusätzlich zu den Einflussfaktoren der drei genannten Modelle sind auch strukturelle Variablen, wie Geschlecht und Regionszugehörigkeit als Einflussfaktoren in Betracht zu ziehen. Da angenommen wird, dass diese von grundsätzlicher Bedeutung für die

Erklärung von Einkommensungleichheit sind, soll die Untersuchung nicht getrennt, sondern als in die drei Modelle integriert erfolgen.

## **2.2 Hypothesen**

Entsprechend der drei theoretischen Modelle - das der Humankapitaltheorie, das der Klassentheorie und das des eigenentwickelten Erklärungsansatzes - werden gerichtete Hypothesen formuliert, die es zu untersuchen gilt:

### **2.2.1 Zur Humankapitaltheorie**

Gemäß der Humankapitaltheorie wird angenommen, dass je höher das persönliche Bildungsniveau ist, desto höher das Einkommen. Bedeutend ist ebenfalls die Berufserfahrung, also das um die Ausbildungszeiten bereinigte Alter, wobei angenommen wird, je älter desto höher der Verdienst. Ergänzt wird die Hypothese durch die Strukturvariablen Region und Geschlecht, unter der Annahme dass männliche oder aus den alten Bundesländern kommende Personen über ein höheres Einkommen verfügen.

Die Hypothese läßt sich dann folgendermaßen ausdrücken:

$$\text{Einkommen} = a + b_1 * \text{Ost/West} + b_2 * \text{Geschlecht} + b_3 * \text{Bildung} + b_4 * \text{Berufserfahrung}$$

### **2.2.2 Zur Klassentheorie**

Neben der wieder untersuchten Regionszugehörigkeit und dem Geschlecht mit der Vermutung männliche oder aus den alten Bundesländern kommende Personen verdienen mehr, wird angenommen, dass die Klassenzugehörigkeit über die Einkommenschancen entscheidet. Die Hypothese lautet demnach, dass je höher die Klassenposition, desto höher das Einkommen.

Als Gleichung der linearen Regression ausgedrückt:

$$\text{Einkommen} = a + b_1 * \text{Ost/West} + b_2 * \text{Geschlecht} + b_3 * \text{Klassenstatus}$$

### **2.2.3 Zum Sozialen Kapital**

Im selbstentwickelten Erklärungsmodell wird das Einkommen hinsichtlich der Beeinflussung durch soziales Kapital untersucht. Angenommen wird, dass individuell höheres soziales Kapital in höherem Einkommen resultiert. Soziales Kapital soll in dieser Untersuchung durch Beziehungen, familiärer Herkunft, außerberuflicher Aktivität und Freundeskreis zu definieren versucht werden. Auch wird wieder angenommen, dass Geschlecht und Regionszugehörigkeit einen Einfluß haben.

Die Hypothese lautet:

Einkommen =  $a + b_1 \cdot \text{Ost/West} + b_2 \cdot \text{Geschlecht} + b_3 \cdot \text{soziales Kapital}$

### 3. Mathematischer Hintergrund

#### 3.1 Grundsätzliches zur linearen Regression

Bei der Regressionsanalyse handelt es sich um ein Schätzverfahren, mit dem ein Merkmal, also eine Variable, durch ein oder mehrere andere Variablen bzw. deren Ausprägungen erklärt oder prognostiziert werden soll. Das zu erklärende Merkmal wird als "abhängige Variable" bezeichnet, die erklärenden Merkmale "unabhängige Variablen" oder Prädiktoren genannt. Eine abhängige Variable, wie zum Beispiel das Einkommen, soll also hinsichtlich des Einflusses unabhängiger, erklärender Variablen, wie bspw. Bildung oder Regionszugehörigkeit überprüft werden.

Die Regressionsanalyse erfordert, dass die abhängige Variable metrisch, d.h. mindestens intervallskaliert ist. Die unabhängigen Variablen müssen metrisch sein, ansonsten sind sie in binär kodierte Dummy-Variablen zu zerlegen. Bei der linearen Regression, der einfachsten Form der Regressionsanalyse, wird davon ausgegangen, dass eine lineare Beziehung zwischen abhängigen und unabhängigen Variablen besteht, die beobachteten Ausprägungen also als Gerade mit der Gleichung darstellbar sind:

$$y = b_1x_1 + b_2x_2 + \dots + b_nx_n + a$$

wobei:

y - Kriterium

$b_n$  - Regressionskoeffizient - gibt an, um wie viele Einheiten sich die abhängige Variable (AV) ändert, wenn sich die betreffende Variable (UV) um eine Einheit ändert

$x_n$  - Prädiktor

a - Zufallskomponente

Anhand dieser Gleichung soll sodann für jeden Fall der Stichprobe der Wert der abhängigen Variablen Y aus den jeweiligen Ausprägungen der, mit sog. Regressionskoeffizienten b multiplizierten, unabhängigen Variablen X vorhersagbar

sein. Die Einflüssen der unabhängigen Variablen werden durch eine Zufallskomponente  $a$  ergänzt, die Messfehler und nicht beobachtete Einflüsse auf  $Y$  berücksichtigt.

### 3.2 OLS- und Koeffizientenschätzung

Kann mit dem Regressionsmodell für jeden Einzelfall der Stichprobe exakt der beobachtete Wert der abhängigen Variablen vorhergesagt werden, so liegt ein perfektes Modell vor. Da es faktisch aber nicht gelingt, solche Modelle zu entwickeln, soll zumindest diejenige Regressionsgerade ermittelt werden, von der alle Werte insgesamt den kleinsten Abstand haben und die somit den Zusammenhang zwischen abhängiger und unabhängiger Variable am besten beschreibt. Die Regressionskoeffizienten müssen also aus den vorhandenen Daten so geschätzt werden, dass die Regressionsgerade möglichst nahe entlang den beobachteten Werten der abhängigen Variablen verläuft, d.h. die Abweichungen (Residuen) zwischen den aus der Regressionsgleichung vorhergesagten und den tatsächlich beobachteten Werten so gering wie möglich sind. Mathematisch ausgedrückt lautet die Grundforderung an das Ordinary Least Squares (OLS) genannte Schätzverfahren der optimalen Geraden:

$$\sum_{i=1}^n \underbrace{(y_i - \hat{y}_i)}_{e_i? \text{ Residuum}} \Rightarrow \min$$

mit:

$y_i$  - tatsächlicher Wert

$\hat{y}_i$  - geschätzter Wert

$n$  - Anzahl der untersuchten Fälle

### 3.3 Signifikanztests und Modellgüte

Um zu überprüfen, ob ein theoretisch angenommener bzw. in den Daten vorgefundener Zusammenhang nicht nur auf die untersuchte Stichprobe sondern auch auf die Grundgesamtheit zutrifft, muss das Regressionsmodell auf statistische Signifikanz hin überprüft werden. Als statistisch signifikant wird eine Aussage dann bezeichnet, wenn die Irrtumswahrscheinlichkeit unter 5% liegt, also mindestens 95% der Fälle mit dem theoretischen Modell erklärt werden können. Zur Durchführung der

Signifikanztests werden eine Null- ( $H_0$ ) und eine Alternativhypothese ( $H_1$ ) aufgestellt und auf ihre Gültigkeit hin überprüft. Null- und Alternativhypothese nehmen unterschiedliches über die Art des Zusammenhanges an und werden vor Durchführung der Signifikanztests sowohl für die Koeffizienten als auch für das Gesamtmodell aufgestellt.

Die einzelnen Regressionskoeffizienten des multivariaten Modells werden mittels T-Test - d.h. dem Vergleich der Mittelwertsdifferenzen - auf Signifikanz überprüft:

$$t_{df} = \frac{\bar{x} - \mathbf{m}}{\hat{\mathbf{s}}_{\bar{x}}}, \text{ mit dem geschätzten Standardfehler } \hat{\mathbf{s}}_{\bar{x}} = \sqrt{\frac{\mathbf{s}^2}{n}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}}$$

Als statistisch signifikant können die Regressionskoeffizienten dann angesehen werden, wenn der Betrag des t-Wertes größer 1.96 (95%-Wahrscheinlichkeit) bzw. größer 2.58 (99%-Wahrscheinlichkeit) ist.

Um die Qualität des gesamten Regressionsmodells zu überprüfen, wird ein F-Test

nach der Formel  $F = \frac{MSS / k - 1}{RSS / n - k}$  durchgeführt, wobei MSS (Modal Sum of Squares)

die mit Hilfe des Modells erklärbaren und RSS (Residual Sum of Squares) die Fehlervariation, also die nicht erklärten Quadrierungen der Abweichungen bezeichnen. Die Anzahl der Fälle wird durch n und die Anzahl der unabhängigen Variablen durch k dargestellt.

Ergibt sich ein F-Wert größer als 1.96, kann wieder mit 95%iger-Wahrscheinlichkeit von einem Zusammenhang zwischen abhängiger Variable und erklärenden, unabhängigen Variablen ausgegangen werden.

Die Modellgüte ganz allgemein wird bspw. durch den Korrelationskoeffizienten  $R^2$ , beschrieben, der den Anteil erklärter Variation im Verhältnis zu gesamter Variation

ausdrückt  $\left( R^2 = \frac{MSS}{RSS} \right)$  und Werte zwischen 0 und maximal 1 annimmt, wobei ein  $R^2$ -

Wert zwischen 0.2 und 0.5 (entspr. 20% bis 50% erklärter Varianz) in dieser Untersuchung als das Modell genügend erklärend betrachtet wird. Weitere die Modellgüte beschreibende Maßzahlen sind das irrelevante Regressoren

berücksichtigende ‚adjusted R<sup>2</sup>‘ oder das ‚Root MSE‘ - dem, in der gleichen Einheit wie die abhängige Variable ausgedrückten Standardfehler der Abweichungen.

### 3.4 Prinzipien der Voraussetzungsprüfungen

Für die Schätzung der Regressionskoeffizienten mit OLS müssen bestimmte Bedingungen erfüllt sein, die so genannten *Gauss-Markow-Bedingungen*. Denn nur dann stellt das OLS-Verfahren den *best linear unbiased estimator* (BLUE) dar, bei dem die Koeffizienten bestmöglich, unverzerrt (unbiased) und effizient (efficient) geschätzt werden. Bei den *Gauss-Markow-Bedingungen* handelt es sich im Einzelnen um:

#### 3.4.1 Linearität

Mit der Bedingung der Linearität ist gemeint, dass sich die abhängige Variable als eine Linearkombination aus unabhängigen Variablen darstellen lässt. Dies bedeutet, dass sich abhängige und unabhängige Variable(n) proportional zueinander verhalten, so dass die Steigung der Regressionsgeraden konstant bleibt. Daraus folgt, dass die Fehler, d.h. die nicht beobachtbaren Einflüsse auf die abhängige Variable, einen Erwartungswert von Null aufweisen sollen ( $E(e_i) = 0$ ). Diese Bedingung wird verletzt, wenn sich der Zusammenhang zwischen abhängiger und unabhängiger Variable nichtlinear gestaltet, wenn einzelne Ausreißer das Regressionsergebnis unverhältnismäßig stark beeinflussen oder wenn Einflussfaktoren übersehen wurden, die mit den unabhängigen Variablen korrelieren.

Eine nichtlineare Beziehung der Variablen lässt sich u. U. abhängig von ihrer Form korrigieren. Im Falle eines logarithmischen Zusammenhangs wird die betreffende unabhängige Variable logarithmiert in der Regressionsgleichung verwendet, bei hyperbolischen Zusammenhängen hilft eine Quadrierung der Variablen. Wenn ein parabelförmiger Verlauf vorliegt, sollte die Variable zusätzlich in ihrer quadrierten Form mit aufgenommen werden.

Multivariate Ausreißer ergeben sich, wenn einzelne Fälle eine ungewöhnliche Kombination der Variablen aufweisen, bezogen auf unsere Untersuchung beispielsweise eine Person mit sehr hohem Einkommen, geringer Bildung und keiner Berufserfahrung. Diese Fälle zeichnen sich durch eine hohe Hebelwirkung, d.h. Leverage, auf das Regressionsergebnis aus, weisen aber nur einen kleinen Fehler auf. Es gilt, diese herauszufinden und zu überprüfen, ob ein Ausschluss der

jeweiligen Beobachtung gerechtfertigt ist (z. B. falls ein Messfehler erkennbar ist) oder ob sich das Modell als ungenügend erweist und korrigiert werden sollte.

Als „übersehene Einflussfaktoren“ gelten Variablen, die die abhängige Variable beeinflussen und mit einer der unabhängigen Variablen korrelieren. Eine Regressionsgleichung kann zwar beispielsweise durch einen Plot der Residuen gegen nicht im Modell enthaltene Variablen auf mögliche übersehene Einflussfaktoren hin überprüft werden, jedoch kann das Problem trotzdem bestehen. Statt einer statischen Anwendung von möglichen Hilfsmitteln sollte dieses Problem in erster Linie theoretisch bearbeitet werden, d.h. ein theoretisches Modell entwickeln, das die wichtigen Einflussfaktoren mit einschließt. Durch die Aufnahme sämtlicher wichtiger Faktoren kann es aber auch leicht zu erhöhter Multikollinearität zwischen unabhängigen Variablen kommen.

### 3.4.2 Homoskedastizität

Als Homoskedastizität wird das Vorliegen einer gleichmäßigen, d.h. homoskedastischen Streuung der Fehler um die Regressionsgeraden bezeichnet:  $VAR(\mathbf{e}_i) = \mathbf{d}^2$ . Eine Verletzung dieser Annahme nennt man dementsprechend Heteroskedastizität. Diese führt aber nicht, wie bei einer Verletzung der Linearitätsbedingung, zu einer Verzerrung der Koeffizienten, jedoch sind die Regressionskoeffizienten nicht mehr effizient und weichen deshalb umso wahrscheinlicher vom wahren Wert in der Grundgesamtheit ab. Mögliche Ursachen für Heteroskedastizität sind beispielsweise nicht-symmetrische Verteilungen der abhängigen Variable. In einem solchen Falle hilft es, die abhängige Variable zu transformieren und somit eine symmetrisch verteilte Variable zu erhalten. Bei einer rechtsschiefen Verteilung lässt sich dies durch die Logarithmisierung der Variable herstellen. In STATA gibt es aber auch die Möglichkeit, über den *boxcox*-Befehl eine Normalisierung der Verteilung durchzuführen, die über eine *Maximum-Likelihood-Schätzung* funktioniert.

### 3.4.3 Unabhängigkeit der Fehler

Diese Bedingung fordert, dass die Fehler untereinander unkorreliert sein sollten:  $COV(\mathbf{e}_i, \mathbf{e}_j) = 0; i \neq j$ . Eine Verletzung dieser Annahme wird Autokorrelation genannt und führt zu ineffizienten Schätzungen der Regressionskoeffizienten. Nachzuweisen ist dies über den Durbin-Watson-Test, der in STATA allerdings nur für Zeitreihen definiert ist.

### 3.4.4. Vermeidung von Multikollinearität

Neben dem Beachten den *Gauss-Markov-Bedingungen* gilt es noch zu vermeiden, dass die unabhängigen Variablen untereinander korrelieren:  $COV(x_i, x_j) = 0; i \neq j$ . Anzeichen hierfür sind nicht oder nur sehr knapp signifikante Koeffizienten, bei gleichzeitig signifikantem F-Test und hohem  $R^2$ . In STATA lässt sich mögliche Multikollinearität mit dem *vif*-Befehl überprüfen. Mögliche Lösungen sind die Respezifikation des Modells, das Zusammenfassen von Variablen, das Herausnehmen von Variablen aus dem Modell, die Erhebung von mehr Daten, oder die Anwendung einer anderen Schätzmethode.

## 4. Umsetzung

### 4.1 Datenbasis

Die dieser Untersuchung zugrunde liegenden Daten stammen aus der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (ALLBUS) aus dem Jahr 2002, erhoben von der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V.(GESIS). Diese Umfrage dient der Erhebung von Daten über Einstellungen, Verhaltensweisen und Sozialstruktur in Deutschland. Seit 1980 wird in der Regel alle zwei Jahre ein repräsentativer Bevölkerungsquerschnitt mit einem teils stetig replikativen, teils variablen Fragenprogramm befragt. Die Daten werden vom ALLBUS Datenservice unmittelbar nach ihrer Aufbereitung und Dokumentation für wissenschaftliche Anwendungen zur Verfügung gestellt.

### 4.2 Operationalisierung

#### 4.2.1 abhängige Variable

Da das Einkommen in Form des Nettostundenlohns untersucht werden soll, wurde die Variable ‚wage‘, generiert, die sich aus dem monatlichen Nettoeinkommen (v359) dividiert durch die vierfache Wochenarbeitszeit (v214) ergibt und auf Vollzeitbeschäftigte unter 65 Jahre beschränkt ist.

#### **4.2.2 Variablen der Humankapitaltheorie**

Um entsprechend dem Humankapitalmodell eine Bildungsvariable „bbild“ zu operationalisieren, die Schul- und Berufsbildungszeiten berücksichtigt, wurden die Variablen „bschubi“ (aus v187, den Angaben zur Schulbildung) und „berubi“ (aus v188-v199, Angaben zur beruflichen Qualifikation), generiert und addiert. Da auch die Berufserfahrung zur persönlichen Qualifikation und somit zur Steigerung des Humankapitals beiträgt, wurde die Variable „bererf“ generiert, indem vom Alter (v185) die Bildungsjahre (bbild), die Dauer der Arbeitslosigkeit (v218) und 6 Jahre, nämlich die der ersten sechs Lebensjahre vor der Einschulung, subtrahiert wurden.

#### **4.2.3 Variablen der Klassentheorie**

Entsprechend der Klassentheorie, wurde eine Variable „class“ anhand von jetziger beruflicher Stellung (v202) und dem Kriterium der beruflichen Aufsichtsfunktion (v216) operationalisiert und als Dummy-Variable in den Ausprägungen Kapitalist (kap), Manager (man), Kleinunternehmer (kunt), Arbeiter (arb) und Beamter (bea) sowie Missings als „kmis“, generiert. Die Variable Kapitalist „kap“ umfasst demnach freiberufliche Akademiker mit zwei und mehr Mitarbeitern und Selbständige mit mindestens zwei Mitarbeitern. Die Variable Manager „man“ beinhaltet Beamte und Angestellte mit beruflicher Aufsichtsfunktion; in die Variable der Kleinunternehmer „kunt“ fallen selbständige Landwirte sowie Selbständige oder freiberufliche Akademiker. Die Vergleichsvariable Arbeiter bilden Angestellte und Arbeiter ohne berufliche Aufsichtsfunktion und in der Variable für Beamte „bea“ werden Beamte die ebenfalls keine berufliche Aufsichtsfunktion besitzen zusammengefasst. Um alle Fälle der Stichprobe in die Regressionsanalyse, auch die derjenigen, die keine Angaben machten, mit einbeziehen zu können, wurde für die Missings eine eigene Variable „kmis“ generiert.

#### **4.2.4 Variablen des Sozialen Kapitals**

Um den Einfluss des selbstentwickelten Erklärungsansatzes, Soziales Kapital genannt, auf das Einkommen untersuchen zu können, wurde eine Variable „sozkap“ generiert. Diese Sozialkapitalvariable vereint Angaben zu Beziehungen, über die man an den Job gelangt sein könnte („bezi“ aus v603) mit sozialer Herkunft, die über Schulabschluss des Vaters und der Mutter definiert wurde („sozher“ aus v320 und

v318) und außerberuflichen Aktivitäten („aktiv“ aus v80 und v587-592), von denen angenommen wurde, dass sie zu Jobkontakten verhelfen könnten sowie Angaben zu Freunden, die zur Jobfindung nützlich sein könnten („freund“ aus v581-v583) in einer Dummy -Variable mit zwei Ausprägungen, nämlich viel oder wenig sozialem Kapital („vielsozkap“ und „wenigsozkap“). Die, die einzelnen Dimensionen sozialen Kapitals darstellenden Teilvariablen „bezi“, „sozher“, „aktiv“ und „freund“, wurden ebenfalls als Dummyvariablen kodiert, wobei die binäre Zuordnung jeweils nach persönlichem Ermessen vorgenommen wurde.

Die Definition der ‚sozkap‘-Variable bzw. ihrer einzelnen Dimensionen (soziale Beziehungen und soziale Herkunft bspw.) war durch die im ALLBUS-Datensatz vorgegebenen Fragestellungen stark eingeschränkt, da soziales Kapital und dessen Determinanten nicht direkt erfragt wurden sondern nur aus mehreren ähnlichen Fragestellungen indirekt zusammengesetzt und generiert werden konnten.

#### **4.2.5 Variablen weiterer Einflussfaktoren**

Um Strukturvariablen wie Geschlecht oder Regionszugehörigkeit in die Betrachtung mit einbeziehen zu können, wurde die Variable „sex“ (männliche Befragte aus v182) und die Variable „west“ (Befragte mit Wohnsitz in den alten Bundesländern aus v3) generiert.

#### **4.2.6. Interaktionsvariablen**

Es ist davon auszugehen, dass es bei der Vielzahl von Variablen in unserer Regressionsgleichung zu Interaktionseffekten zwischen unabhängigen Variablen kommt. Man kann vermuten, dass es vor allem bei der gemeinsamen Verwendung der Klassen- und Humankapitalvariablen zu solchen Interaktionen kommt. Hierbei verdienen vor allem Hypothesen, die sich aus der Kombination einer individuellen Ausprägung (Bildung oder Berufserfahrung) und einem gesellschaftlichen Strukturmerkmal (Klasse) ergeben, eine nähere Betrachtung.

Eine Hypothese (und damit auf Interaktionseffekte zu untersuchen) in diesem Zusammenhang wäre beispielsweise, dass sich ein Bildungsplus bei einem Angehörigen der Managerklasse stärker auf das Einkommen auswirken wird als bei einem der Arbeiterklasse, da ein Manager im Gegensatz zum Arbeiter eine kontrollierende Stellung im Produktionsprozess einnimmt, in der er die erworbene

Bildung produktiv umsetzen kann. Dies müsste auch für einen Kapitalisten zutreffen, da dieser ebenfalls eine Position im Produktionsprozess inne hat, in der er ein Mehr an Bildung positiv verwerten kann. Zwar gewinnt ein Kapitalist in der Theorie sein Einkommen aus seinem Besitz an Produktionsmitteln (insofern wäre ein Bildungsvorsprung nicht produktiv umsetzbar), jedoch trifft dies eher auf Großkapitalisten zu und nicht auf die Mehrzahl der Beobachtungen, die wir der Klasse der Kapitalisten zugeordnet haben. So ist beispielsweise bei Selbstständigen mit mehr als zwei Mitarbeitern noch davon auszugehen, dass diese eine aktiv mitarbeitende Rolle und Funktionen des Managements im Produktionsprozess übernehmen. Deshalb wurden zwei Interaktionsvariablen generiert: „*lbilman*“ (*Bildung\*Manager*) und „*lbilkap*“ (*Bildung\*Kapitalist*).

Analog zu diesen Hypothesen können auch Interaktionseffekte zwischen der Klassenzugehörigkeit und der Berufserfahrung vermutet werden. Hier muss die Annahme untersucht werden, ob andere Klassen als die Arbeiterklasse ein Mehr an Einkommen pro Berufsjahr gewinnen. Bei Managern wäre ein Argument hierfür, dass sie in ihrer beruflichen Tätigkeit zusätzliches Wissen erwerben, dass sie in höheren Berufsjahren produktiv umsetzen und beruflich aufsteigen können. Beides sollte für Arbeiter nicht in diesem Maße zutreffen. Bei Kapitalisten ist mit jedem Berufsjahr auch ein verwertbarer Wissenserwerb als bei Arbeitern zu erwarten, daneben besitzen sie die Möglichkeit, ihren Besitz an Produktionsmitteln (womit das Unternehmen gemeint ist) und somit ihren Gewinn zu vergrößern. Hier muss aber noch zusätzlich die Beamtenklasse betrachtet werden. Aufgrund der praktischen Unkündbarkeit im Beamtenstatus und der allgemein steigenden Entlohnung pro Berufsjahr sollte es auch hier zu einem Interaktionseffekt zwischen Berufserfahrung und Einkommen kommen. Deshalb wurden hier drei Interaktionsvariablen gebildet: „*lbererfman*“ (*Berufserfahrung\*Manager*), „*lbererfkap*“ (*Berufserfahrung\*Kapitalist*) und „*lbererfbea*“ (*Berufserfahrung\*Beamter*).

Neben den vermuteten Interaktionseffekten in Bezug auf die Klassenzugehörigkeit sind auch Effekte für die Strukturvariablen Geschlecht und Regionszugehörigkeit in Verbindung mit den Humankapitaltheorievariablen Bildung und Berufserfahrung zu erwarten. Es ist anzunehmen, dass sich bei Männern ein Mehr an Bildung positiver auf das Einkommen auswirkt als bei Frauen. So ist zwar die Studienquote bei Frauen gleich, wenn nicht höher als bei Männern (analog dazu bei höheren Schulabschlüssen), jedoch sind die Studierendenzahlen in Bezug auf einzelne Studiengängen stark geschlechtsspezifisch verteilt. Insofern spielt hierbei nicht die

Quantität, sondern die Qualität (im Sinne von Inhalt) der Bildung eine Rolle. Absolventen von männlich dominierten Fachrichtungen wie beispielsweise Ingenieursfächer verdienen erheblich mehr als Absolventen weiblich dominierter Studiengänge wie Germanistik. Die Hypothese lautet daher, dass Männer statistisch gesehen ihr Mehr an Bildung profitabler umsetzen können als Frauen dies können. Hier wurde eine weitere Interaktionsvariable generiert: „*lbilsex*“ (*Bildung\*Geschlecht*). Des Weiteren ist zu erwarten, dass Männer pro Beschäftigungsjahr einen signifikant höheren Einkommenszuwachs haben als Frauen, da diese aufgrund diskriminierender Entscheidungen nicht dieselben Aufstiegschancen haben wie Männer. Daraus ergibt sich eine weitere Interaktionsvariable: „*lbererfsex*“ (*Berufserfahrung\*Geschlecht*).

Ein ähnlicher Zusammenhang kann in Bezug auf die Regionszugehörigkeit vermutet werden. Personen aus Westdeutschland sollten ihre Bildung besser in Einkommen umsetzen können als ostdeutsche Personen, da deren Bildung nach der Wende „entwertet“ wurde. Ostdeutsche Qualifikationen waren nicht auf dem nötigen technologischen Stand (z. B. Computerkenntnisse), manche Berufsbilder/-richtungen nicht oder nicht in der notwendigen Ausprägung vorhanden (z. B. Betriebswirtschaft). Diese Annahme wurde durch die Interaktionsvariable „*lbilwest*“ (*Bildung\*Regionszugehörigkeit*) abgebildet.

Weiterhin kann angenommen werden, dass eine westdeutsche Person pro Beschäftigungsjahr mehr Einkommenszuwachs erhält als eine ostdeutsche Person. Durch die Wiedervereinigung wurden viele Personen in Ostdeutschland aus ihren bisherigen Stellungen herausgerissen und fanden in geringerem Maße zurück in den Arbeitsmarkt. Oft war dies auch verbunden mit der Annahme von geringer qualifizierten und schlechter bezahlten Arbeitsstellen. Zudem liegt die Vermutung nahe, dass aufgrund der oben angeführten „Entwertung“ der ostdeutschen Qualifikationen es in geringerem Maße zu einem kontinuierlicher beruflicher Aufstieg kam. Die letzte Interaktionsvariable lautet daher: „*lbererfwest*“ (*Berufserfahrung\*Regionszugehörigkeit*).

### **4.3 Ergebnisse der Modelle**

Die Modelle basieren auf einer einheitlichen Fallzahl von 1061 Observations (1060 nach dem Entfernen einer Beobachtung). Alle einzelnen Modelle beinhalten die immer sehr signifikanten Strukturvariablen Geschlecht und Regionszugehörigkeit und

sind schließlich nach Durchführung der Voraussetzungsprüfungen zu einem Endmodell vereint.

Die folgende Übersicht enthält Angaben zu Regressionskoeffizienten und erklärter Varianz ( $R^2$ ) :

	Humankapital		Klasse	Soziales Kapital		Gesamtmodell	+ Interaktion Bildung Kapital	+ Interaktion BerErf Beamte	+ Interaktion BerErf West	+ quadr. Berufserfahrung	Gesamtmodell mit logarithmiertem Einkommen + Interaktionseffekte	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	wage	wage	wage	wage	wage	wage	wage	wage	wage	wage	lwage	lwage
sex	1.916	1.704	2.000	1.821	1.821	1.784	1.782	0.610	0.617	0.602	0.067	0.062
	(5.18)**	(4.70)**	(5.25)**	(4.61)**	(4.61)**	(4.92)**	(4.99)**	(8.38)**	(8.50)**	(8.25)**	(12.05)**	(10.93)**
west	2.253	2.345	1.774	2.068	2.068	2.146	2.175	1.789	1.757	1.716	0.185	0.183
	(6.03)**	(6.42)**	(4.61)**	(5.18)**	(5.18)**	(5.86)**	(6.04)**	(5.03)**	(4.95)**	(4.83)**	(6.61)**	(6.62)**
bbild	0.819	0.884				0.725	0.615	2.163	2.855	2.919	0.260	0.309
	(12.34)**	(13.52)**				(10.12)**	(8.43)**	(6.02)**	(6.38)**	(6.51)**	(9.20)**	(8.86)**
bererf		0.114				0.100	0.100	0.093	0.033	0.146	0.040	0.034
		(7.21)**				(6.21)**	(6.36)**	(5.84)**	(1.18)	(2.32)*	(8.75)**	(6.88)**
kap			4.587			2.965	-0.187	-0.156	-0.316	-0.271	0.095	-0.061
			(5.76)**			(3.85)**	(0.20)	(0.17)	(0.34)	(0.29)	(1.60)	(0.85)
man			2.119			0.995	1.146	1.158	1.114	1.104	0.137	0.142
			(5.02)**			(2.40)*	(2.80)**	(2.84)**	(2.74)**	(2.72)**	(4.32)**	(4.48)**
bea			4.950			2.620	3.008	6.135	5.885	5.748	0.239	0.412
			(4.68)**			(2.53)*	(2.95)**	(4.13)**	(3.96)**	(3.87)**	(3.01)**	(3.56)**
kmis			-2.186			-1.051	-1.167	-1.200	-1.175	-1.077	-0.159	-0.166
			(4.13)**			(2.05)*	(2.31)*	(2.38)*	(2.34)*	(2.14)*	(4.02)**	(4.23)**
bererf 2										-0.003	-0.001	-0.001
										(2.01)*	(5.84)**	(5.56)**
lbilkap							1.586	1.583	1.610	1.585		0.075
							(5.88)**	(5.89)**	(6.00)**	(5.91)**		(3.60)**
lbererf bea								0.284	0.266	0.252		0.014
								(2.88)**	(2.70)**	(2.56)*		(1.88)
lbererf west									0.086	0.086		0.006
									(2.58)*	(2.57)*		(2.38)*
sozka p				0.665								
				(1.72)								
bezi					-1.252							
					(2.65)**							
sozhe r					1.148							
					(2.32)*							
aktiv					0.586							
					(1.51)							
freund					1.045							
					(1.74)							
Constant	6.107	7.061	6.180	6.460	5.672	6.720	6.732	6.679	6.244	5.602	1.691	1.665
	(16.06)**	(17.91)**	(14.89)**	(13.88)**	(8.16)**	(16.16)**	(16.45)**	(16.36)**	(14.16)**	(10.31)**	(41.58)**	(39.33)**
Obs.	1061	1061	1061	1061	1061	1061	1061	1061	1061	1061	1060	1060
R <sup>2</sup>	0.17	0.21	0.14	0.05	0.06	0.23	0.25	0.26	0.26	0.26	0.39	0.40

Absolute value of t statistics in parentheses

\* significant at 5%; \*\* significant at 1%

### **4.3.1 Modell zur Humankapitaltheorie**

Das Regressionsmodell zur Humankapitaltheorie wurde in zwei Schritten untersucht. Zuerst der alleinige Einfluss von Bildung (1) und dann Bildung und Berufserfahrung zusammen (2). Es zeigt sich, dass beide Determinanten der Humankapitaltheorie sehr signifikant sind und zusammen mit den Strukturvariablen 21% der Varianz des Nettostundenlohnes erklären können. Die Hypothese der Humankapitaltheorie von Berndt&Keller kann deshalb als bestätigt betrachtet werden.

Nach diesem Modell steigt pro Bildungsjahr der Nettostundenlohn um 0,88 € und mit jedem zusätzlichen Jahr an Berufserfahrung um 0,11 €.

### **4.3.2 Modell zur Klassentheorie**

Bei der Regressionsanalyse des Modells zur Klassentheorie nach Wright&Perrone (3) stellte sich heraus, dass die gebildeten Klassen Kapitalisten, Manager und Beamte tatsächlich einen statistisch signifikanten Einfluss auf den Nettostundenverdienst haben, nicht jedoch die Klasse der Kleinunternehmer, die deswegen auch nicht in das Regressionsmodell miteinbezogen wurde. Als Vergleichsgruppe wurde die Klasse der Arbeiter herangezogen.

Ein der Klasse ‚Kapitalist‘ Angehöriger hat so beispielsweise einen Mehrverdienst von 4,59 € pro Stunde und auch ein als Manager Eingestufte verdient immer noch 2,12 € mehr als die Vergleichsgröße der Arbeiter. Beamte verdienen pro Stunde 4,95 € mehr als Arbeiter. Das Modell der Klassentheorie kann 14% der Stichprobenvarianz erklären.

### **4.3.3 Modell zum Sozialen Kapital**

Das selbstentwickelte Modell des sozialen Kapitals wurde ebenfalls in zwei Schritten untersucht [(4) und (5)] da sich herausstellte, dass die aus sozialen Beziehungen, sozialer Herkunft, außerberufliche Aktivität und Freundeskreis generierte Variable ‚Soziales Kapital‘, keinen statistisch signifikanten Einfluss auf den Nettostundenlohn hat und selbst in Verbindung mit den Strukturvariablen Geschlecht und Regionszugehörigkeit nur 5% der Nettostundenlohn-Varianz erklären kann.

Um zu untersuchen, ob die einzelnen Determinanten der Variable ‚soziales Kapital‘, soziale Beziehungen, soziale Herkunft, ausserberufliche Aktivität oder Freundeskreis, für sich genommen in vielleicht eindeutigerem Zusammenhang mit

dem Stundenlohn stehen, wurde die Regression im Modell (5) noch einmal durchgeführt. Es zeigt sich, dass die soziale Herkunft durchaus statistisch signifikant den Verdienst beeinflusst, dies jedoch nur für einen sehr geringen Teil der Stichprobe, so dass das selbstentwickelte Modell des sozialen Kapitals mit seinen Variablen als nicht erklärend verworfen wird und im folgenden Gesamt- wie auch im Endmodell nicht mehr miteinbezogen wird.

Ob die Generierung der Variablen unzureichend und fehlerhaft ist oder die vorhandene Datenbasis (ALLBUS 2002) zur Erstellung solch eines Modells ungeeignet ist und deshalb keine erklärende Aussage gewonnen werden konnte, bedarf einer gesonderten Überprüfung.

#### **4.4 Auswertung der erwarteten Interaktionseffekte**

Entgegen unserer Erwartungen erweisen sich die meisten der vermuteten Interaktionseffekte als nicht signifikant. Die Interaktionsvariablen „*ibilman*“ (Bildung\*Manager), „*ibilsex*“ (Bildung\*Geschlecht) und „*ibilwest*“ (Bildung\*Regionszugehörigkeit), die auf mögliche Effekte in Bezug auf Bildung hin generiert wurden, erweisen sich als nicht signifikant und werden daher im endgültigen Modell nicht berücksichtigt. Hingegen konnten wir bei der Interaktionsvariable „*ibilkap*“ (Bildung\*Kapitalist) einen sehr signifikanten Interaktionseffekt zwischen Bildung und Zugehörigkeit zur Kapitalistenklasse feststellen. Auch hat sie großen Einfluss auf den Nettostundenlohn: Pro Bildungsjahr steigt der Nettostundenlohn bei Mitgliedern der Kapitalistenklasse um 1,59 € mehr als bei Mitgliedern der Arbeiterklasse. Jedoch führt die Aufnahme dieser Interaktionsvariable dazu, dass die Variable Kapitalist nicht mehr signifikant ist. Wir behalten sie aber dennoch bei, da sie integraler Bestandteil der Klassentheorie ist. Allerdings werden wir aufgrund dessen zwei Gesamtmodelle errechnen, eines ohne und eines inklusive der Interaktionsvariablen. Bei den Interaktionsvariablen in Bezug auf Berufserfahrung erweisen die beiden Variablen „*lbererfbeat*“ (Berufserfahrung\*Beamter) als sehr und „*lbererfwest*“ (Berufserfahrung\*Regionszugehörigkeit) als signifikant, „*lbererfman*“ (Berufserfahrung\*Manager), „*lbererfkapt*“ (Berufserfahrung\*Kapitalist) und „*lbererfsex*“ (Berufserfahrung\*Geschlecht) jedoch nicht. Jedoch haben beide geringeren Einfluss auf das Einkommen. Gemäß der Variable „*lbererfbeat*“ bekommen Mitglieder der Beamtenklasse pro Berufsjahr 0,28 € mehr Nettostundenlohn als Mitglieder der Arbeiterklasse. Bei „*lbererfwest*“ ist der Effekt noch geringer: Pro Berufsjahr

bekommen Personen, die in Westdeutschland arbeiten, 0,09 € mehr an Nettostundenlohn als Personen, die in Ostdeutschland arbeiten (bezogen auf die Arbeiterklasse).

Insgesamt sind also drei Interaktionseffekte nachweisbar gewesen. Diese werden in das Gesamtmodell übernommen.

## 4.5 Gesamtmodell

Für das Gesamtmodell wurden nun das Humankapitalmodell und Klassenmodell kombiniert. Daraus ergibt sich folgende Regressionsgleichung:

$$\text{Einkommen} = a + b_1 * \text{Ost/West} + b_2 * \text{Geschlecht} + b_3 * \text{Bildung} + b_4 * \text{Berufserfahrung} + b_5 * \text{Klassenstatus}$$

Im erweiterten Gesamtmodell wurden noch die Interaktionsvariablen mit einbezogen:

$$\text{Einkommen} = a + b_1 * \text{Ost/West} + b_2 * \text{Geschlecht} + b_3 * \text{Bildung} + b_4 * \text{Berufserfahrung} + b_5 * \text{Klassenstatus} + b_6 * \text{Interakt\_Bildung\&Kapitalist} + b_7 * \text{Interakt\_Berufserfahrung\&Beamter} + b_8 * \text{Interakt\_Berufserfahrung\&Ost/West}$$

### 4.5.1 Auswertung der Voraussetzungsprüfungen

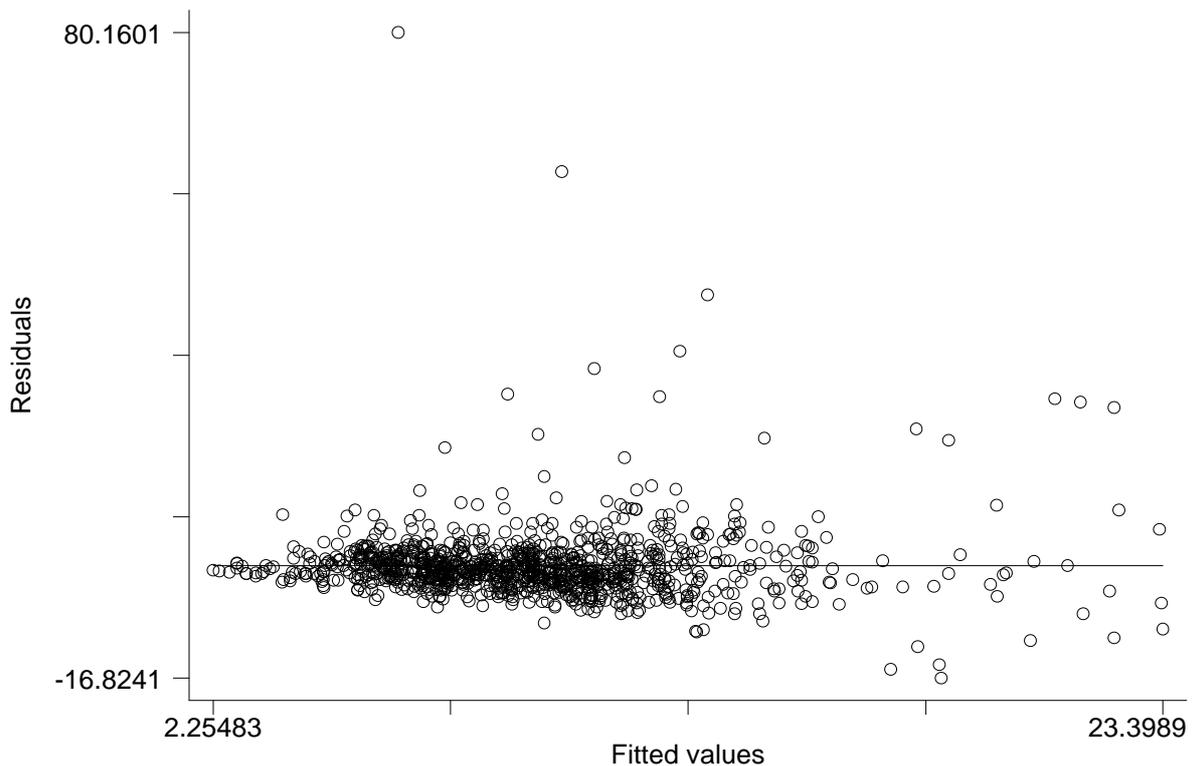
Dieses Gesamtmodell wurde schließlich der Prüfung der Voraussetzungsbedingungen (siehe Abschnitt 3.4) unterzogen.

#### Linearität

Es zeigt sich, dass die Berufserfahrung nicht linear zum Einkommen verläuft. Dieser Effekt konnte schon dadurch abgemildert werden, dass die Berufserfahrung nicht alleinig über das Alter berechnet wird, sondern Zeiten der Arbeitslosigkeit ebenfalls noch von der Berufserfahrung abgezogen wurden. Der nicht lineare Verlauf weist darauf hin, dass mit zunehmender Berufserfahrung die Einkommenszuwächse abflachen, unter Umständen sogar negativ werden. Um diesen Kurvenverlauf zu korrigieren wurde die Berufserfahrung quadriert und als zusätzlicher quadratischer Term in die Regressionsgleichung mit aufgenommen: „bererf2“

#### Homoskedastizität

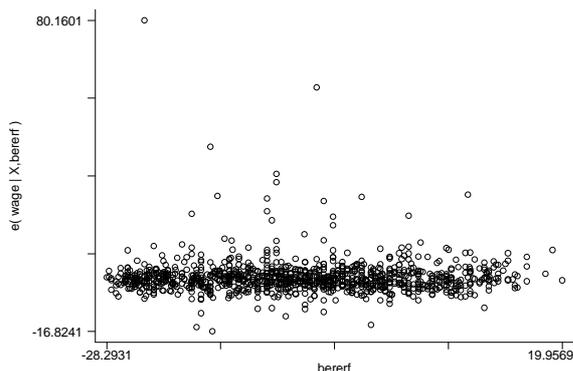
Bei Betrachtung der Voraussetzung der Homoskedastizität fällt auf, dass sich die Fehler in unserem Modell nicht gleichmäßig, sondern mit zunehmendem Einkommen weiter auseinander streuen:



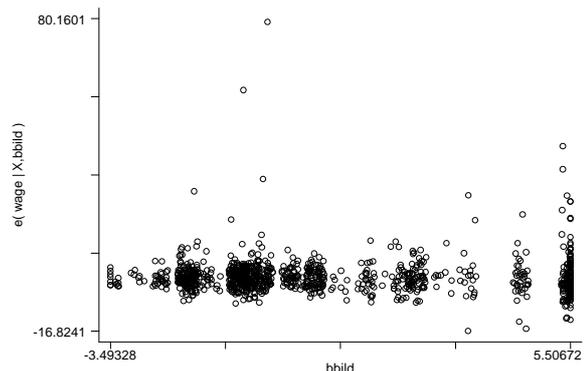
Auch der *hettest*-Befehl bei STATA für den Cook-Weisberg-Test gibt aus, dass die Irrtumswahrscheinlichkeit der Hypothese H1: Streuung der Fehler ist nicht konstant, gegen null geht. Der Chi-Quadrat-Wert als Maß für Heteroskedastizität liegt bei sehr signifikanten 137.68.

Eine mögliche Ursache für die vorliegende Heteroskedastizität könnte im schon oben festgestellten Problem der unterschiedlichen „Qualität“ von als gleichwertig gezählten Bildungsabschlüssen liegen. Während Schulabschlüsse noch stark standardisiert und vereinheitlicht sind, so dass mit einem gewissen Schulabschluss auch immer gleich oder zumindest ähnliche Qualifikationen zu erwarten sind, trifft dies im ausdifferenzierten höheren Bildungsbereich sicher nicht mehr zu: Studium ist nicht gleich Studium, Diplom ist nicht gleich Diplom. Diese Annahme lässt sich durch die Residuendiagramme stützen:

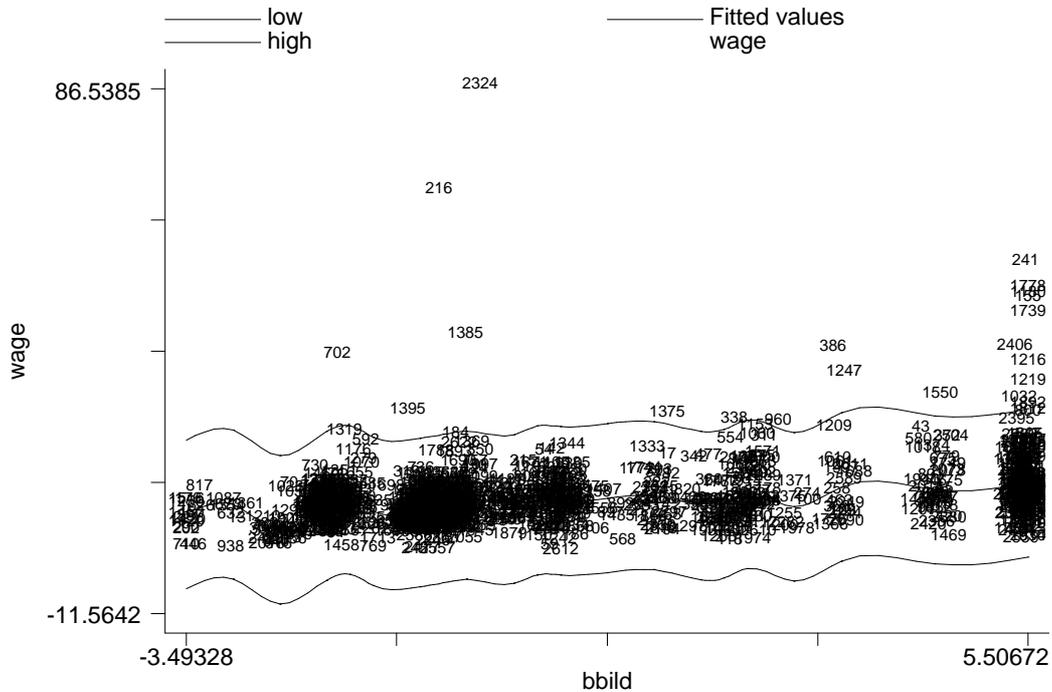
Residuen gegen Beschäftigungsdauer



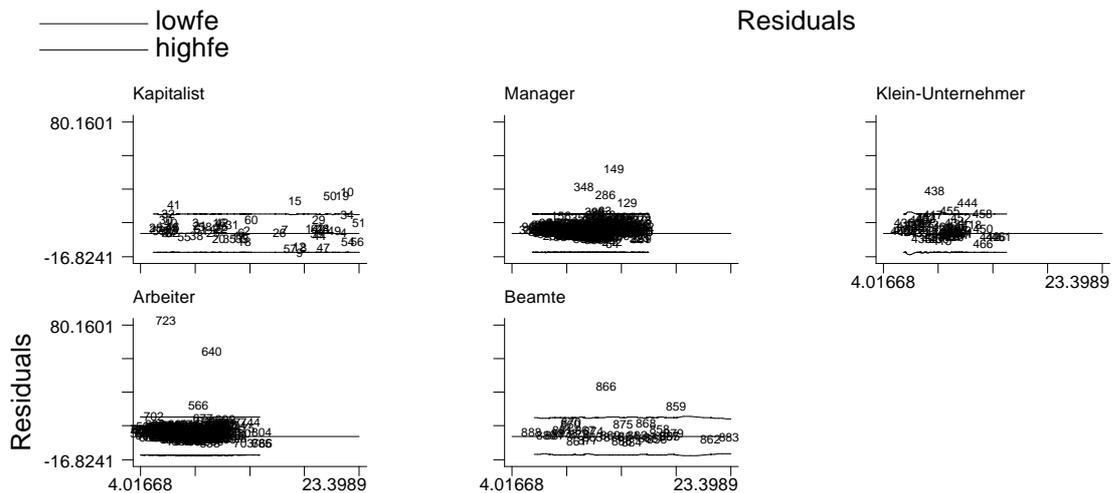
Residuen gegen Bildung



Auch der Cook-Weisberg-Test zeigt, dass Bildung einen hohen heteroskedastischen Einfluss auf die Fehlerverteilung hat: der Chi-Quadrat-Wert für Beschäftigungsdauer liegt bei 42.39, für Bildung liegt er bei 80.39 und ist damit fast doppelt so groß. Heteroskedastizität kann aber auch durch Ausreißer hervorgerufen werden. Deshalb werden von uns alle Beobachtungen jenseits des 99%-Bereichs näher untersucht, d.h. diejenigen, die am stärksten vom vorhergesagten Wert abweichen:



Die nähere Betrachtung wird nach Klassen getrennt vorgenommen:

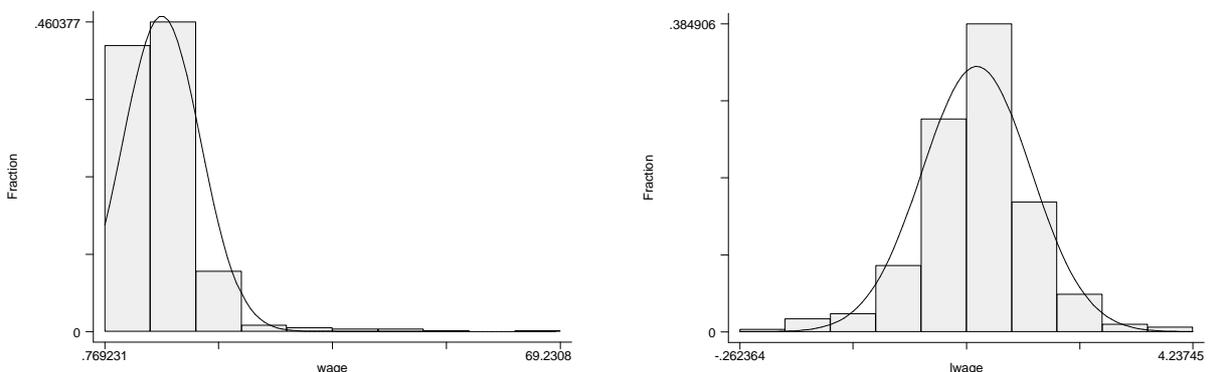


Fitted values  
Graphs by class

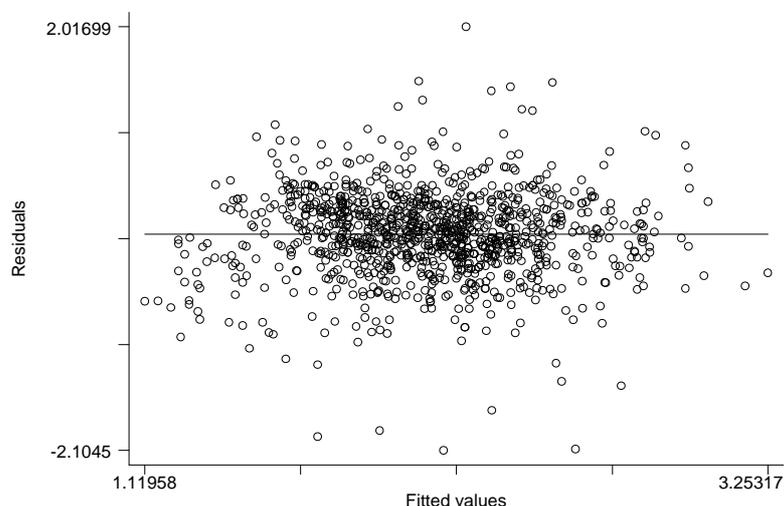
Bei der näheren Untersuchung erweist sich nur eine Beobachtung als für uns nicht nachvollziehbar. Es handelt sich um ein ostdeutsches männliches Mitglied der Arbeiterklasse ( $v2==2496$ ), das nach 12 Jahren Schulbildung und 4 Jahren Berufsbildung bereits fast 70 € in der Stunde verdient. Alle weiteren von uns untersuchten Fälle sind im do-file aufgeführt.

Obwohl wir nur einen Fall entfernt haben, zieht dies starke Veränderungen der Chi-Quadrat-Werte des Cook-Weisberg-Tests nach sich. Die Hypothese H1 (Heteroskedastizität) erweist sich in diesem Test als genauso sehr signifikant wie vorher, der Chi-Quadrat-Wert ist sogar auf 419.44 gestiegen. Bei Bildung zeigt sich der gleiche Effekt, hier hat der Chi-Quadrat-Wert auf 197.82 zugelegt. Jedoch zeigt sich bei der Berufserfahrung ein anderes Bild. Hier trifft die Hypothese H1 (Heteroskedastizität) nicht mehr zu ( $\text{Prob} > \text{chi}^2 = 0.4264$ ), auch ist der Chi-Quadrat-Wert auf 0.63 gesunken. Insgesamt liegt nach wie vor eine stark heteroskedastische Verteilung der Fehler vor.

Deshalb wurde von uns noch die Verteilung der abhängigen Variable untersucht und festgestellt, dass das Einkommen stark rechtschief verteilt ist. Diese Verletzung lässt sich durch die Logarithmisierung des Einkommens beheben:



Durch die Logarithmierung hat die Heteroskedastizität stark abgenommen:



Das Modell weist nun nur noch einen Chi-Quadrat-Wert von 0.04 auf. Für Bildung beträgt er 17.81, für Berufserfahrung 7.97. Jedoch hat die Logarithmisierung auch dazu geführt, dass manche unserer Variablen nicht mehr signifikant sind. Die betrifft die Interaktionsvariable „*lbererf*bea“ (Berufserfahrung\*Beamter) (t-test: 1.88) und vor allem die Variable „*kap*“ (t-test: 1,60).

### Multikollinearität

Diese Nicht-Signifikanz wichtiger Variablen kann Folge von Multikollinearität sein. Dies hat zur Folge, dass sich die Konfidenzintervalle verbreitern und wichtige Variablen als insignifikant ausgegeben werden. Der durchschnittliche Multikollinearitätswert (vif-Wert bei STATA) liegt bei unserem Gesamtmodell inklusive aller Interaktionsvariablen jedoch nur bei 4.26, für die Variable „*kap*“ beträgt er 1.70. In erster Linie weisen die Variablen „*bererf*“ und „*bererf2*“ einen starken multikollinearen Zusammenhang auf, was aber bei ihrer Herkunft zwangsläufig der Fall ist und von uns bewusst in Kauf genommen wird. Werden die Interaktionsvariablen aus unserem Modell entfernt, reduziert sich der Multikollinearitätswert auf 4.18. Wenn die quadrierte Berufserfahrung noch aus dem Modell entfernt wird, resultiert dies in einen durchschnittlichen vif-Wert von 1.15.

Es ließ sich kein multikollinearer Zusammenhang nachweisen (außer dem zwischen „*bererf*“ und „*bererf2*“), so dass die Kapitalisten-Variable nach wie vor insignifikant bleibt. Da sie aber integraler Bestandteil des Klassenmodells ist und mit 1.60 (im Modell ohne Interaktionsvariablen) nicht weit von der Signifikanzgrenze entfernt liegt, wird die Variable beibehalten.

## **4.5.2 Ergebnisse**

### Gesamtmodell ohne Interaktionsvariablen

Das Gesamtmodell ohne Interaktionsvariablen liefert eine Erklärungsrate von 39% der Varianz des Nettostundenlohns ( $R^2=0,39$ ). Es fällt jedoch schwer, bei einem logarithmierten Einkommen noch genaue Aussagen über die Zusammenhänge abzugeben, es sind nur noch relative Aussagen möglich. Der größte Einfluss liegt nach wie vor bei der Bildung vor. So erhöht sich der Nettostundenlohn laut der Regressionsgleichung um 26% pro Bildungsjahr. Aber auch die Klassentheorie weist zum großen Teil einen sehr signifikanten Einfluss nach: Der Nettostundenlohn von Mitglieder der Managerklasse liegt um fast 14% über dem von Arbeitern. Bei Beamten ist der Einfluss noch stärker: sie verdienen netto 24% mehr in der Stunde. Dieser starke prozentuale Unterschied ist für uns leider nicht nachzuvollziehen.

Genauso wenig wie die Feststellung, dass Kapitalisten in der Stunde nur 10% mehr verdienen als Arbeiter. Ein möglicher Erklärungsansatz hierfür wäre eine hohe Arbeitszeit bei Kapitalisten. Dies dürfte in erster Linie auf kleinere Unternehmer zutreffen, die ihre eigene Arbeitskraft noch stark in den Produktionsprozess einbringen. Unter Umständen müsste man die Kapitalistenklasse noch stärker zuspitzen und solche Fälle in Gänze in die Kleinkapitalistenklasse einordnen.

#### Gesamtmodell mit Interaktionsvariablen

Wenn man die (zumindest ursprünglich) signifikanten Interaktionsvariablen noch mit in die Regressionsgleichung aufnimmt, erreicht man eine Erklärungsrate von 40% der Varianz des Nettostundenlohns. Dieses eine Prozent an Zuwachs bei der Erklärungsrate geht aber auf Kosten des Modells, da vor allem durch die Aufnahme Interaktionsvariable „*lbilkap*“ die wichtige Kapitalisten-Variable gänzlich unsignifikant wird.

### **5. Fazit**

Ausgehend von der anfänglichen Fragestellung, die Determinanten des Nettostundenverdienstes zu bestimmen, zeigt sich, dass vor allem die Strukturvariablen Geschlecht und Regionszugehörigkeit, die Variablen Bildung und Berufserfahrung entsprechend der Humankapitaltheorie und auch Variablen des klassentheoretischen Ansatzes als den Verdienst beeinflussend wirken. Jedoch hat keines der untersuchten einzelnen Modelle oder einzelnen Variablen eine ausreichend hohe Aussagekraft, so dass nur mit der Zusammenfassung aller zu einem Gesamtmodell die Einkommensunterschiede annähernd gut erklärt werden können. Auch muss davon ausgegangen werden, dass es noch weitere, vielleicht auch weit besser erklärende Determinanten des Einkommens gibt, die in dieser Untersuchung keine Beachtung gefunden haben.

## 6. Literatur

Bortz, Jürgen: Statistik für Sozialwissenschaftler / Jürgen Bortz. – 5., vollst. überarb. Auflage. Berlin; Heidelberg; New York: Springer, 1999.

Keller, Berndt: Einführung in die Arbeitspolitik: Arbeitsbeziehungen und Arbeitsmarkt in sozialwissenschaftlicher Perspektive. München: Oldenbourg 1999

Kohler, Ulrich/Kreuter, Frauke: Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung. München, Wien: Oldenbourg 2001

Wright, E.O./Perrone, L.: Marxist Class Categories and Income Inequalities. In: American Sociological Review, 1977, Vol. 42

# 7.Anhang

## Do-File

```
capture log close
log using "tryout.log", replace
set more off
use "allb02.dta", clear

*****
*GRUNDLEGENDE *
*VARIABLENBILDUNG*
*****

*Variable für Nettostundenlohn der unter 65jährigen Vollzeitbeschäftigten
gen wage=v361/((v214*52)/12) if v185<=65 & v200==1

*Geschlecht, binär
gen sex =.
recode sex . = 1 if v182 ==1
recode sex . = 0 if v182 ==2

*Ost-West-Variable binär generieren
gen west =.
recode west . =1 if v3 ==1
recode west . =0 if v3 ==2

*****
*HUMANKAPITALTHEORIE*
*Variablenbildung *
*****

* Bildung in Jahren messen
*Schulbildung in Jahren
generate bschubi =.
recode bschubi . =9 if v187==2
recode bschubi . =10 if v187==3
recode bschubi . =11 if v187==4
recode bschubi . =13 if v187==5

* Berufsbildung in Jahren, nur höchste Bildungsstufe zählt; 5 Jahre falls Hochschulbildung
generate berubi=.
recode berubi . =5 if v197==1
recode berubi . =4 if v196==1
recode berubi . =3 if v195==1
recode berubi . =2.5 if v193==1
recode berubi . =2 if v194==1
recode berubi . =1.75 if v191==1
recode berubi . =1.5 if v190==1
recode berubi . =1 if v189==1
recode berubi . =0.5 if v192==1
recode berubi . =0.5 if v188==1
recode berubi . =0 if v199==1

*generieren von Gesamtbildung
gen berbild=bschubi+berubi

*Generieren einer Variable, die Arbeitslosigkeit in Jahren misst
gen alk=.
replace alk=v218/52
replace alk=0 if v218==.

*Generieren d. Berufserfahrungsvariable: Alter-Bildungsdauer-6-Arbeitslosigkeit
gen gesbererf=v185-berbild-6-alk

*zentrieren von Gesamtbildung, um Konstante (_cons) besser lesen zu können
sum berbild
gen bbild=berbild-r(mean)

*zentrieren von Berufserfahrung, um Konstante (_cons) besser lesen zu können
sum gesbererf
gen bererf=gesbererf-r(mean)
*zentrieren von quadrierter Berufserfahrung, um Konstante (_cons) besser lesen zu können
gen bererf2=gesbererf*gesbererf
```

```

sum bererf2
replace bererf2=bererf2-r(mean)

*****
*KLASSENTHEORIE *
*Variablenbildung*
*****

*Generieren der Klassenvariablen
gen class=.
replace class=1 if (v202==16 | v202==17 | v202==22 | v202==23 | v202==24) & wage!=.
replace class=2 if v202>=40 & v202<=54 & v216==1 & wage!=.
*Hab mal bei der Manager Klasse die tatsächlichen Arbeiterrollen komplett rausgenommen
*da auch Vorarbeiter mit Aufsichtsfunktion nicht zur Managerklasse gehören
*Damit Managerklasse wieder signifikant
replace class=3 if (v202==10 | v202==11 | v202==12 | v202==13 | v202==15 | v202==21) & wage!=.
replace class=4 if v202>=50 & v202<=64 & v216 == 2 & wage!=.
replace class=5 if v202>=40 & v202<=43 & v216 == 2 & wage!=.
*Habe vorhandenes Einkommen als Bedingung eingebaut
*da nur diejenigen, die auch ein Einkommen angegeben haben, Klassen zugeordnet werden sollen.

label define class 1 "Kapitalist" 2 "Manager" 3 "Klein-Unternehmer" 4 "Arbeiter" 5 "Beamte"
label value class class

*Bilden von Klassendummies, inklusive Missingvariable kmis
gen kap=0
replace kap =1 if class==1

gen man=0
replace man =1 if class==2

gen kunt=0
replace kunt =1 if class==3

gen arb=0
replace arb =1 if class==4

gen bea=0
replace bea=1 if class==5

*Missingvariable
gen kmis=0
replace kmis=1 if class==. & wage!=.

*****
*SOZIALES KAPITAL/NETZWERK*
*Variablenbildung *
*****

*DHU: Habe mal bei dem Modell keine Missings zugelassen, dadurch Berechnung möglich

*BEZIEHUNGEN, über die man an den Job gelangte
gen bezi=0
replace bezi=1 if v603==2 | v603==3 | v603==4 | v603==5 | v603==8 | v603==10
replace bezi=0 if v603==6 | v603==7 | v603==9 | v603==11

*soziale Herkunft, die über Schulabschluss des Vaters und der Mutter definiert ist
gen sozher=0
replace sozher=0 if v320==1 | v320==2 | v320==3 | v318==1 | v318==2 | v318==3
replace sozher=1 if v320==4 | v320==5 | v318==4 | v318==5

*ausserberufl. Aktivität, die zu Jobkontakten verhilft
gen aktiv=0
replace aktiv=0 if v80>=4 | v587>=3 | v588>=3 | v589>=4 | v590>=3 | v591>=3 |v592>=3
replace aktiv=1 if v80<=3 | v587<=2 | v588<=2 | v589<=3 | v590<=2 | v591<=2 |v592<=2

*Freunde, die zur Jobfindung nützlich sein könnten
gen freund=0
replace freund=0 if v581<=2 | v582<=2 | v583<=2
replace freund=1 if v581>=3 | v582>=3 | v583>=3

*Zusammenfassen der einzelnen Elemente zum vorläufigen (pre) sozialen Kapital
gen presozkap=bezi+sozher+aktiv+freund

*Umformen des sozialen Kapitals in binär kodierte Variable
gen sozkap=.

```

```

replace sozkap=1 if presozkap==2 | presozkap==3 | presozkap==4
replace sozkap=0 if presozkap=1 | presozkap==0

/*
gen vielsozkap=.
replace vielsozkap=.
replace vielsozkap =1 if sozkap==2 | sozkap==3 | sozkap==4
replace vielsozkap =0 if sozkap==1 | sozkap==0

gen wenigsozkap=.
replace wenigsozkap=.
replace wenigsozkap =1 if sozkap==0 | sozkap==1
replace wenigsozkap =0 if sozkap==2 | sozkap==3 | sozkap==4
*/

***** R E G R E S S I O N 1.STUFE *****

keep if bbild!=. & bererf!=. & kap!=. & man!=. & bea!=. & kunt!=. & sozkap!=. & sex!=. &
west!=.

*****
*Regression für erstes Modell*
*****
reg wage bbild

reg wage bererf

reg wage sex

reg wage west

reg wage bbild sex

reg wage bbild west

reg wage bbild sex west

reg wage bererf sex

reg wage bererf west

reg wage bererf sex west

*Nur mit Bildung (+sex&west)
reg wage bbild sex west, beta
outreg using tablesix, replace
predict wage_bbild
graph wage_bbild wage bbild, s(.) c(1.) pen(29)
translate @Graph wage_bbild.wmf, replace

*Mit Bildung und Berufserfahrung (+sex&west)
reg wage bbild bererf sex west, beta
outreg using tablesix, append
predict wage_bererf
graph wage_bererf wage bererf, s(.) c(1.) pen(29)
translate @Graph wage_bererf.wmf, replace

*****
*Regression für zweites Modell*
*****

*Mit Kleinunternehmerklasse und Beamtenklasse
reg wage kap man kunt bea sex west

*Mit Kleinunternehmerklasse und Beamtenklasse und Klassenmissings
reg wage kap man kunt bea kmis sex west

*Mit Beamtenklasse und Klassenmissings, ohne Kleinunternehmerklasse
reg wage kap man bea kmis sex west
outreg using tablesix, append

*****
*Regression für drittes Modell*
*****
reg wage sozkap
outreg using tablesix, append

/*
Soziales Kapital nicht signifikant, zu viele Missings

```

da zu viele nicht erwerbstaetig bzw. nicht Zusatzfragebogen beantwortet (von dort stammen Variablen).

Deshalb verzichten wir auf das soziale Kapital in unserem Endmodell

\*/

\*\*\*\*\*

\*Regression für vorläufiges Gesamtmodell\*

\*\*\*\*\*

\*Kleinunternehmer (kunt) rausgenommen, da nicht signifikant, Klasmisings (kmis) eingebaut

\*soziales Kapital auch rausgenommen (s.o.)

reg wage bbild bererf kap man bea kmis sex west

outreg using tablesix, append

\*\*\*\*\* V O R A U S S E T Z U N G S P R Ü F U N G \*\*\*\*\*

\*\*\*\* A D D I T I V I T Ä T \*\*\*\*

\*\* Interaktionsvariablen im Zshg. mit Bildung \*\*

gen Ibilkap = bbild\*kap

reg wage bbild bererf kap man bea kmis sex west Ibilkap

\*ja ist signifikant, bleibt drin

outreg using tablesix, append

gen Ibilman = bbild\*man

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibilman

\*nein ist nicht signifikant, also wieder rausnehmen

gen Ibilsex = bbild\*sex

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibilsex

\*nein ist nicht signifikant, also wieder rausnehmen

gen Ibilwest = bbild\*west

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibilwest

\*nein ist nicht signifikant, also wieder rausnehmen

\*\* Interaktionsvariablen im Zshg. mit Berufserfahrung \*\*

gen Ibererfkap = bererf\*kap

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibererfkap

\*nein ist nicht signifikant, also wieder rausnehmen

gen Ibererfman = bererf\*man

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibererfman

\*nein ist nicht signifikant, also wieder rausnehmen

gen Ibererfbea = bererf\*bea

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibererfbea

\*ja ist signifikant, bleibt drin

outreg using tablesix, append

gen Ibererfsex = bererf\*sex

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibererfbea Ibererfsex

\*nein ist nicht signifikant, also wieder rausnehmen

gen Ibererfwest = bererf\*west

reg wage bbild bererf kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest

\*ja ist signifikant, bleibt drin

outreg using tablesix, append

\*\*\*\* L I N E A R I T Ä T \*\*\*\*

/\*

Die Berufserfahrung verläuft nicht linear zum Einkommen!

Deshalb Verwendung einer quadrierten Funktion:  $E\text{inkommen} = a + b_1 \cdot \text{bererf} + b_2 \cdot \text{bererf}^2$

bererf2 wurde bereits oben bei Variablenbildung erstellt, bisher nur nicht verwendet.

\*/

reg wage bbild bererf bererf2 kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest

outreg using tablesix, append

\*\*\*\* H O M O S K E D A S T I Z I T Ä T \*\*\*\*

reg wage bbild bererf bererf2 kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest, beta

rvfplot, yline(0)

translate @Graph residuals\_before\_homo.wmf, replace

hettest

```

**** theoretische Untersuchung ****
rvpplot bererf
translate @Graph homo2_Resid_bererf.wmf, replace
rvpplot bbild, ji(2)
translate @Graph homo2_Resid_bbild.wmf, replace
hetttest bererf
hetttest bbild

**** Untersuchung auf Ausreisser ****
predict ydach if e(sample)
predict fehler if e(sample), res
predict se if e(sample),stdr

gen low= ydach - 2.58*se
gen high = ydach + 2.58*se
graph low ydach high wage bbild, s(iii[_n]) c(sss.) ji(2) pen(4342)
translate @Graph homo3_high_low.wmf, replace
gen lowfe = -2.58*se
gen highfe = 2.58*se

sort class
by class: graph lowfe fehler highfe ydach, yline(0) s(i[_n]i) c(s.s) ji(2) pen(434)
graph lowfe fehler highfe ydach, yline(0) s(i[_n]i) c(s.s) ji(2) pen(434)by(class)
translate @Graph homo4_class_high_Low.wmf, replace

gen abfehler = abs(fehler)

** Kapitalisten **
list v2 wage v361 v214 v202 v203 berbild gesbererf class sex west if abfehler > 2.58*se &
kap==1
/*8 (v2==165), 25 (1493), 33 (1522), 37 (2743), 49 (1598), 50 (1288), 63 (2583) und 82 (1926)
liegen außerhalb des 99%Bereiches
Dies sind folgende Fälle:
08: Arzt mit mehr als 9 Mitarbeitern, der mehr als 12000 Euro verdient
25: selbstständiger Möbeltischler mit 2-9 Mitarbeitern, der nur 850 Euro im Monat verdient
(wäre eher Kleinkapitalist)
33: selbstständiger Handelsgeschäftsführer mit 10-49 Mitarbeitern der 6000 Euro verdient
37: selbstständige Hotelinhaberin mit 2-9 Mitarbeitern, die nur 563 Euro im Monat verdient und
dafür 72 Stunden arbeitet
49: Hauptgeschäftsführer mit 2-9 Mitarbeitern, der nur 500 Euro im Monat verdient und dafür 70
Stunden verdient
50: Rechtsvertreter mit mehr 10-49 Mitarbeitern, der 15000 Euro im Monat verdient
63: Ärztin in Ostdeutschland mit 2-9 Mitarbeitern, die 10000 Euro im Monat verdient
Davon könnte man #8 entfernen. Lassen wir aber auch drin.*/
*nur mal testweise eingebaut: drop if v2==1493 | v2==2743 | v2==1598*

** Manager **
list v2 wage v361 v214 v202 v203 berbild gesbererf class sex west if abfehler > 2.58*se &
man==1
*Keine ungewöhnlichen Fälle, bis auf vielleicht den Versicherungsvertreter 445 (257) der 15200
Euro im Monat verdient

** Arbeiter **
list v2 wage v361 v214 v216 v202 v203 berbild gesbererf class sex west if abfehler > 2.58*se &
arb==1, nol
list v2 wage v361 v214 v216 v202 v203 berbild gesbererf class sex west if abfehler > 2.58*se &
arb==1
*Fall 943(2496): Arbeiter mit nur 12 Jahren Bildung und 4 Jahren Berufserfahrung verdient
15000 Euro bei 40 Stunden Arbeit.*
*Fall 862(232): Arbeiter verdient in 50 Stunden 15000 Euro, aber immerhin schon über 22 Jahre
Berufserfahrung.*
*Fall 943(2496) entfernen wir, da kaum erklärbar (15000 Euro Einkommen, niedrige Bildung etc.)
drop if v2==2496

** Beamte **
list v2 wage v361 v214 v216 v202 v203 berbild gesbererf class sex west if abfehler > 2.58*se &
bea==1, nol
list v2 wage v361 v214 v216 v202 v203 berbild gesbererf class sex west if abfehler > 2.58*se &
bea==1
*Keine ungewöhnlichen Fälle, nur seit wann verdienen Gymnasiallehrer (1245, v2==1883) soviel
(8750 Euro)?

**** Untersuchung auf Leverage ****
reg wage bbild bererf bererf2 kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest
lvr2plot, s([v361]) ji(2)
translate @Graph homo5_leverage.wmf, replace
lvr2plot, s([_n]) ji(2)
predict lev, lev
list v2 wage v361 v214 v216 v202 v203 berbild gesbererf class sex west if wage==. & class!=.

```

```

list v2 wage v361 v214 v216 v202 v203 berbild gesbererf class sex west lev if lev>0.1 & lev!=.
& wage!=.
*Wir entfernen keine Fälle*

*****
* Zusammenfassung: Heteroskedastizität *
*****
reg wage bbild bererf bererf2 kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest, beta
rvfplot, yline(0)
translate @Graph residuals_after_homol.wmf, replace
hetttest
hetttest bbild
hetttest bererf

*****
* Die Normalverteilung der Fehler *
*****
graph fehler, bin(10) normal
translate @Graph fehler_NV.wmf, replace
*Fehler rechtsschief verteilt
sum fehler, detail

*****
* Normalverteilung der abhängigen Variable (Einkommen) *
*****
graph wage, bin(10) normal
translate @Graph NV_AV_wage.wmf, replace
*Einkommen rechtsschief verteilt, deshalb wird es logarithmiert
gen lwage=ln(wage)
reg lwage bbild bererf bererf2 kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest, beta
graph lwage, bin(10) normal
translate @Graph NV_AV_logwage.wmf, replace

rvfplot, yline(0)
translate @Graph residuals_after_homo2Log.wmf, replace
hetttest
hetttest bbild
hetttest bererf

*****
* Multikollinearität der Variablen *
*****

reg lwage bbild bererf bererf2 kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest
vif
reg lwage bbild bererf bererf2 kap man bea kmis sex west Ibererfbea Ibererfwest
vif
reg lwage bbild bererf bererf2 kap man bea kmis sex west
vif
reg lwage bbild bererf kap man bea kmis sex west
vif

*****
* ZUM ABCHLUSS: REGRESSION FÜRS ENDGÜLTIGE GESAMTMODELL *
*****
reg lwage bbild bererf bererf2 kap man bea kmis sex west
outreg using tablesix, append

reg lwage bbild bererf bererf2 kap man bea kmis sex west Ibilkap Ibererfbea Ibererfwest
outreg using tablesix, append

*log close
set more on
exit

```