Universität Bremen

Abhängigkeit des Einkommens von soziodemografischen Merkmalen

Jan Klüver und Chris Michel Peters

betreut von
Prof. Dickhaus und Nico Steffens

5. Juni 2018

Inhaltsverzeichnis

1 Einführung

In dieser Ausarbeitung wird der Einfluss von soziodemographischen Faktoren auf das Einkommen in Deutschland untersucht. Als Grundlage der Untersuchung dient ein CAMPUS-File Datensatz der Verdienststrukutrerhebung aus dem Jahre 2010. Dementsprechend unterliegt dieser Datensatz einigen vereinfachenden Maßnahmen, wie z.B. der Rundung von erhobenen soziodemographischen Merkmalen (z.B. wöchentliche Arbeitszeit), der Deckelung des Bruttomonatseinkommens auf 7000 Euro usw. Die Interpretation der Ergebnisse unserer Untersuchung im soziodemographischen Kontext (soweit mit unserem Fachwissen überhaupt möglich) ist somit nicht auf die Wirklichkeit, d.h. der Grundgesamtheit aller Einkommen in Deutschland übertragbar. Wir werden uns daher im Folgenden auf die mathematische Analyse des Datensatzes beschränken. Natürlich wurde in der Analyse versucht, möglichst sinnvolle (im Sinne der Soziodemographie) Fragestellungen und Antworten zu finden.

Die Hauptfragestellung der Untersuchung ist folgende: " Ist der Bruttostundenlohn (BSL) einer betrieblich angestellten Person abhängig vom Geschlecht? " . Anschließende Fragestellungen sind

1. Welche Ursache hat die Abhängigkeit des BSL vom Geschlecht?
   Dazu wird die Region in der der Betrieb liegt, die Ausbildung, das Alter, die Berufsgruppe, die Stellung im Beruf der Person usw. mit in Beziehung zum BSL gesetzt.

2. Welche weiteren Abhängigkeiten des BSL lassen sich ermitteln?

Annmerkung: Wir verwenden den Bruttolohn anstatt des Nettolohns, um möglichst viele nicht im Datensatz erhobenen Einflüsse politischer und gesellschaftlicher Natur zu unterbinden, die nicht direkt mit dem Betrieb in Verbindung stehen (unterschieliche Steuerklassen usw.).

Um diese Fragen zu beantworten, führen wir hauptsächlich lineare Regressionen und deren zugehörige Tests durch. Die theoretischen Grundlagen hierfür werden in Kapitel 2 erörtert und später dann in Kapitel 3, Datenanalyse, angewendet. Für die Datenanalyse wurde die Programmiersprache R benutzt. Wichtige Teile des Codes werden per E-Mail verschikt. Anschließend werden wir in der Diskussion kurz unsere Ergebnisse der Untersuchung interpretieren und vorstellen.

2 Theoretischer Hintergrund

2.1 Modell

Bei der multiplen linearen Regression (ANCOVA) wird der messbare Raum $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ betrachtet. Es wird allgemein der Zusammenhang zwischen y_i, einer Störung ε_i und

$x_{i,1}, \ldots, x_{i,k}$ durch eine Funktion $f(x_{i,1}, \ldots, x_{i,k})$ modelliert, wobei die Störung additiv zugrunde liegt, also

$$y_i = f(x_{i,1}, \ldots, x_{i,k}) + \varepsilon_i$$

für alle $1 \leq i \leq n$. Die Funktion f wird als Linearkombination dargestellt mit

$$f(x_{i,1}, \ldots, x_{i,k}) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}.$$

Von Interesse ist der Vektor $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ mit seinen Parametern, wobei der Parameter β_0 als Intercept bezeichnet wird. Die einzelnen y_i sind eine Realisierung von reellwertigen stochastisch unabhängigen Zufallsvariablen Y_i, damit ergibt sich nun folgende Schreibweise

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i$$

für alle $1 \leq i \leq n$. Hierbei ist $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ der Response-Vektor, für $p = k+1$ ist

$$X = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

die Design-Matrix, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ sind die Fehlerterme und $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^p$ ist der Parametervektor. Für die Matrixschreibweise ergibt sich also

$$Y = X\beta + \varepsilon.$$

2.2 Modellannahmen

Das Modell wird komplettiert durch Annahmen über die Fehlerterme und die Designmatrix.

1) Die Designmatrix hat vollen Rang, so dass $X^\top X \in \mathbb{R}^{p \times p}$ positiv definit und invertierbar ist.

2) Die Fehlerterme sind unabhängige identisch verteilte Zufallsvariablen, für die gilt $\mathbb{E}(\varepsilon_1) = 0$ und $0 < \sigma^2 = \mathrm{Var}(\varepsilon_1) < \infty$. Da die Fehlerterme identisch verteilt sind, haben alle Fehlerterme den gleichen Erwartungswert und die gleiche Varianz. Bei konstanter Varianz werden die Fehler auch homoskedastisch genannt. Neben der Homoskedastizität wird angenommen, dass die Fehlerterme unkorreliert sind, das heißt $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ für $i \neq j$. Dadurch kann die Kovarianzmatrix bestimmt werden $\mathrm{Cov}(\varepsilon) = \mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 E_n$, hierbei ist E_n die $(n \times n)$-Einheitsmatrix.

3) Zusätzlich kann angenommen werden, dass die Fehlerterme normalverteilt sind, also $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$

Aus dem Model und den Modellannahmen können bereits einfache Folgerungen geschlossen werden, die Aussagen über den Erwartungswert, die Varianz und die Kovarianz liefern

$$\begin{aligned} \forall 1 \leq i \leq n: \ \mathbb{E}(Y_i) &= \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} \\ \forall 1 \leq i \leq n: \ \mathrm{Var}(Y_i) &= \sigma^2 \\ \forall 1 \leq i \neq j \leq n: \ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) &= 0 \end{aligned}$$

wird zusätzlich angenommen, dass die Fehlerterme normalverteilt sind, so gilt

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 E_n).$$

2.3 Schätzungen

Die Schätzungen für die Parameter β_i und die Varianz σ^2 werden bezeichnet mit $\hat{\beta}_i$ und $\hat{\sigma}^2$. Ist $\hat{\beta}_i$ der Schätzer des Parametervektors, so erhält man für den Erwartungswert $\widehat{\mathbb{E}(Y)} = X\hat{\beta}$, wobei die einzelnen Komponenten definiert sind durch

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k}$$

für alle $1 \leq i \leq n$. Die Residuen sind die Abweichung der Werte der Responsvariable und den Schätzwerten der Erwartungswerte, also $\hat{\varepsilon}_i = y_i - \hat{y}_i$ für $1 \leq i \leq n$.

2.3.1 Methode der kleinsten Quadrate

Ist β der Parametervektor, dann ist der kleinste Quadrate (KQ)-Schätzer gegeben durch

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \tag{1}$$

Damit folgt nun für die Responsevariable

$$\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1} X^\top Y$$

Hierfür beschreibt $H := X(X^\top X)^{-1} X^\top$ eine $(n \times n)$-Matrix und heißt Prädiktionsmatrix. Mit dieser Prädiktionsmatrix werden die Prädiktionswerte der linearen Regression bestimmt, welche zur Berechnung der Residuen beitragen. Wird für den Schätzer $\hat{\beta}$ der Erwartungswert und die Kovarianz betrachtet, so erhält man

$$\mathbb{E}(\hat{\beta}) = \beta \text{ und } \mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}.$$

2.4 Hypothesentests

Im vorherigen Abschnitt wurde der KQ-Schätzer eingeführt, der eine Schätzung $\hat{\beta}$ für den Parametervektor β liefert. In der Praxis möchte man anhand dieser Parameter Aussagen über das Bruttomonatseinkommen treffen, indem überprüft wird, von welchen Faktoren das Einkommen abhängt und ob bestimmte Faktoren einen großen, kleinen oder auch keinen Einfluss auf das Bruttomonatseinkommen haben. Um zu testen welchen Einfluss ein Faktor hat, verwendet man den Hypothesentest, zuerst werden allerdings zwei Verteilungen eingeführt die zur Bestimmung des kritischen Niveaus beitragen.

2.4.1 χ^2- und F-Verteilung

Sind $X_i \sim \mathcal{N}(\mu_i, 1)$ mit $i = 1, \dots n$ und unabhängig, dann heißt

$$V = \sum_{i=1}^{n} X_i^2$$

nichtzentral χ^2-verteilt mit Nichtzentralitätsparameter $\theta = \sum_{i=1}^{n} \mu_i$ oder auch $\chi_n^2(\theta)$-verteilt. Hierbei bedeutet nichtzentral, im Gegensatz zu den zentrierten Verteilungen, dass die Erwartungswerte der normalverteilten Zufallsvariablen nicht verschwinden.

Um Varianzen zu vergleichen, werden Quotienten betrachtet, die zur F-Verteilung führen. Sind also $V \sim \chi_n^2(\theta)$ und $W \sim \chi_m^2$, sowie V und W unabhängig, dann heißt

$$F := \frac{V/n}{W/m}$$

die $F_{n,m}(\theta)$-Verteilung, wobei θ der Nichtzentralitätsparameter ist.

2.4.2 Allgemeine Einführung des Hypothesentests

Wie bereits erwähnt werden Hypothesentests verwendet, um Hypothesen bezüglich bestimmter Faktoren zu überprüfen, dabei wird der Parameterraum disjunkt in zwei Hypothesen aufgeteilt. Sei nun $\{\mathbb{P}_\theta : \theta \in \Theta\}$ ein statistisches Modell, wobei Θ der Parameterraum und $\mathbb{P}_\theta$ das Wahrscheinlichkeitsmaß für alle $\theta \in \Theta$. Da der Parameterraum disjunkt aufgeteilt wird, gilt $\Theta = \Theta_0 \cup \Theta_1$ und $\Theta_0 \cap \Theta_1 = \emptyset$ für $\Theta_0, \Theta_1 \in \Theta$. Eine häufige Frage für das Bruttomonatseinkommen ist, ob ein Faktor einen Einfluss hat, hierbei würde Θ_0 dafür stehen, dass der Faktor keinen Einfluss hat, Θ_1 hingegen hat einen Einfluss auf das Bruttomonatseinkommen. Hierfür schreibt man $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_1$. Dabei wird H_0 Nullhypothese genannt und H_1 wird als Alternative bezeichnet. Bei einem Hypothesentest wird H_0 generell als die Hypothese gewählt, welche widerlegt werden soll.

Es gibt eine Entscheidungsregel, ob die Nullhypothese angenommen oder verworfen wird, welche Test genannt wird. Ein Test δ ist eine messbare Funktion der Daten X mit Werten in $[0, 1]$. Für $\delta(X) = 0$ wird die Nullhypothese akzeptiert und bei $\delta(X) = 1$ wird die Nullhypothese verworfen. Der Bereich $\{x : \delta(x) = 1\}$ wird als kritischer Bereich bezeichnet. Ist $T(X)$ eine Statistik und gilt $\delta(x) = \mathbb{1}_{\{T(X) \geq c\}}$, so ist c der kritische Wert. Natürlich ist es sinnvoll $\delta(x) = p \in (0, 1)$ zuzulassen, dies ist möglich über einen randomisierten Test, der über eine Bernoulliverteilung definiert ist, darauf soll nun aber nicht weiter eingegangen werden.

In den Tests werden stets zwei Hypothesen betrachtet, bei Entscheidungen für jeweils eine der beiden Hypothesen können Fehler auftreten, die in der unteren Tabelle dargestellt werden:

Einen Fehler 1. Art bekommt man also, wenn die Nullhypothese wahr ist, diese allerdings verworfen wird. Einen Fehler 2. Art bekommt man, wenn die Nullhypothese nicht wahr ist, aber angenommen wird. Die Nullhypothese ist so gewählt, dass diese abgelehnt werden soll, deswegen ist der Fehler 1. Art für die Fragestellung wichtiger als der

	H_0 wahr	H_1 wahr
H_0 wird akzeptiert	kein Fehler	Fehler 2. Art
H_1 wird akzeptiert	Fehler 1. Art	kein Fehler

Fehler 2. Art. Es wird ein Niveau α vorgegeben und der Test wird so gewählt, dass der Fehler 1. Art höchstens α ist. Das Niveau α wird als Signifikanzniveau bezeichnet und wird im Folgenden stets bei $\alpha = 0,05$ liegen, sodass man ein 95%-Konfidenzintervall bekommt. Unterschiedliche Tests werden anhand des Fehlers 2. Art verglichen und führen zur Gütefunktion. Ist δ ein Test, so ist die Gütefunktion $G_\delta : \Theta \to [0,1]$ definiert durch

$$G_\delta(\theta) = \mathbb{E}_\theta(\delta(X))$$

Ist $\theta \in \Theta_1$, so ist das die Güte des Tests für die Alternative. Ist andersrum $\theta \in \Theta_0$, so ist das die Wahrscheinlichkeit für einen Fehler 1. Art für den wahren Wert θ. Gilt nun für den Test δ

$$\sup_{\theta \in \Theta_0} G_\delta(\theta) \leq \alpha,$$

dann ist das Signifikanzniveau des Tests α. Ist $\sup_{\theta\in\Theta_0} G_\delta(\theta) = \alpha$, so spricht man von einem Level-α-Test.

2.4.3 Hypothesentests für lineare Regression

Für Hypothesentests bei linearen Regressionen wird ausgegangen von W_0 als Unterraum von W und von Interesse ist weiterhin die Nullhypothese. Bei einer einfachen linearen Regression $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ für $i = 1, \ldots, n$ und der Normalverteilungsannahme der Fehlerterme, so stellt sich die Frage, ob der Parameter β_1 einen Einfluss auf das Bruttomonatseinkommen hat. Mit dem Testproblem

$$H_0 : \beta_1 = 0 \quad \text{gegen} \quad H_1 : \beta_1 \neq 0$$

wird nun festgestellt, ob der Parameter einen linearen Einfluss auf das Bruttomonatseinkommen hat. Wird die Nullhypothese verworfen, so hat der Parameter β_1 einen linearen Einfluss zum gegebenen Signifikanzniveau.

Für die Bestimmung des kritischen Niveaus c werden die Verteilungen aus (2.4.1) benötigt. Mit der Normalverteilungsannahme führt dies nun zu folgendem Satz

Satz 1. *Sei $\xi_0 := P_{W_0}\xi$ die Projektion mit $\dim(W) = r$ und $\dim(W_0) = q$, wobei $r > q$. Dann ist das linearen Modell $V_n(Y)$[1] nichtzentral $F_{r-q,n-r}$ verteilt mit*

$$\delta^2 = \frac{||\xi - \xi_0||^2}{\sigma^2}.$$

Insbesondere gilt mit der Nullhypothese $H_0 : \xi \in W_0$, dass $V_n \sim F_{r-q,n-r}$.

[1] Aus dem Likelihood-Quotienten-Test aus *Mathematische Statistik*

Dieser Satz führt nun zum folgenden Test mit der Teststatistik $V_n(Y)$. Mit $F_{1-\alpha,r-q,n-r}$ wird das $(1-\alpha)$-Quantil der $F_{r-q,n-r}$-Verteilung beschrieben. Nach Satz (1) ist mit

$$\delta(Y) := \mathbb{1}_{\{V_n(Y) \geq F_{1-\alpha,r-q,n-r}\}}$$

ein Level-α-Test für $H_0 : \xi \in W_0$ gegen $H_1 : \xi \notin W_0$ beschrieben. Dieser Test heißt F-Test.

2.4.4 Globaler F-Test

Das Resultat des letzten Abschnittes war der F-Test. Bisher wurde jedoch nur für einen Parameter bestimmt, ob dieser einen Einfluss auf das lineare Modell, also auf das Bruttomonatseinkommen hat. Mit dem globalen F-Test kann nun eine Gesamtaussage über die Parameter getroffen werden. Dabei lautet die Nullhypothese

$$H_0 : \beta_1 = \cdots = \beta_k = 0.$$

Würde die Nullhypothese akzeptiert werden hätte das lineare Modell keinen linearen Einfluss, das heißt $Y_i = \beta_0 + \varepsilon_i$. Zum Testen dieser Hypothese wird der F-Test angewendet, wobei hier nun Quadratsummen betrachtet werden, anstatt die abstrakte Betrachtung aus Satz 1. Setze

$$SQT := \sum_{i=1}^{n}(Y_i - \bar{Y})^2,\ SQR := \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \text{ und } SQE := \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \qquad (2)$$

Für diese Quadratsummen gilt $SQT = SQE + SQR$. Mit

$$MQR := \frac{SQR}{k},\ MQE := \frac{SQE}{n-k-1}$$

wird nun der F-Wert bestimmt

$$F = \frac{MQR}{MQE}.$$

Die Nullhypothese wird verworfen, falls $F > F_{\{1-\alpha,k,n-k-1\}}$.

2.4.5 Partieller F-Test

Im Gegensatz zum globalen F-Test wird beim partiellen F-Test nur ein Teilmodell betrachtet. Die Nullhypothese lautet hier für $m < k$

$$H_m : \beta_{m+1} = \cdots = \beta_k = 0.$$

Unter H_m wird also nur ein Teilmodell betrachtet, der Unterschied hier liegt bei der Schätzung der Parameter, da hier nicht alle Parameter, sondern nur eine Auswahl von Parametern betrachtet wird. Definiere analog zu (2)

$$SQE_k = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,\ SQE_m = \sum_{i=1}^{n}(Y_i - \tilde{Y}_i)^2.$$

Dabei stehen die Indizes für das jeweilige Modell und $\tilde{Y}_i$ kommt aus (1), dem KQ-Schätzer. Ebenfalls wie in (2.4.4) lässt sich der F-Wert berechnen mit $MQE_k = SQE_k/(n-k-1)$ und

$$MQD = \frac{SQE_m - SQE_k}{k-m} = \frac{SQR_k - SQR_m}{k-m}.$$

Der F-Wert ist nun $F_m = MQD/MQE_k$ und ist $F_m > F_{\{1-\alpha,k-m,n-k-1\}}$, so wird die H_m Hypothese verworfen.

3 Datenanalyse

In diesem Teil wird die in Kapitel 2 vorgestellte Theorie auf den bereits erwähnten Datensatz angewendet. Wir werden uns zuerst der Frage nach der Abhängigkeit des BSL vom Geschlecht widmen. Als Einstieg führen wir eine einfache ANOVA durch, um einen signifikanten Einkommensmittelwertunterschied zwischen Mann und Frau festzustellen. Sei Y_i der BSL der i-ten Person im Datensatz (ef$21 / ef$19 im Datensatz) und GE_i das Geschlecht der i-ten Person (weitere Bezeichnungen siehe Tabelle). Das Signifikanzniveau sei $\alpha = 0.05$.

Bezeichnung / Faktor	Abkürzung / Variablenname
BSL	Y
Geschlecht	GE
Region	RE
Ausbildung	AB
Wirtschaftszweig	WZ
Stellung im Beruf	SB
Art des Arbeitsvertrages	AV
Beruf nach ISCO	IS
Alter (in 5 Jahresschritten)	A
Unternehmenszugehörigkeit	UZ

3.1 Einfache ANOVA

Eine einfache ANOVA entspricht einer einfachen linearen Regression. Wir haben also das Model $Y = \beta_0 + \beta_1 \mathbb{1}_{\{GE=2\}} + \varepsilon$, wobei $\mathbb{1}_{\{GE=2\}}$ als Indikatorfunktion den Faktor Geschlecht modeliert (Dummyvariable). Ist eine Person weiblich, so gilt $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1$. Andernfalls wird das Einkommen nur über den geschätzten Intercept bestimmt.
Das Ziel ist es, mittels eines globalen F-Tests festzustellen, ob das Geschlecht einen signifikanten Einfluss auf das Einkommen hat. Die Nullhypothese lautet in diesem Fall $H_0 : \beta_0 = \beta_1$. Die Nullhypothese kann aber verworfen werden, welches folgende Analyse zeigt:

```
               Df  Sum Sq Mean Sq F value Pr(>F)
GE              1   87829   87829    1098 <2e-16
Residuals   25972 2078007      80
```

In der Tabelle ist vor allem der F-Wert $F_{25972,1} = 1098$ und der daraus ermittelte P-Wert von Bedeutung. Letzterer liegt deutlich unter dem Signifikanzniveau α. Es ist damit, unter der Annahme der Nullhypothese, sehr unwahrscheinlich einen solchen Datensatz aus einer Grundgesamtheit zu erhalten.

Es besteht nach Obigem ein Einkommensunterschied bzgl. des Geschlechtes. Nun ist es von Interesse, festzustellen, wie groß der Einkommensunterschied ist und welche Gründe dies hat. Zum Beispiel könnten alle Personen eines Geschlechtes in Führungspositionen angestellt und alle anderen Personen gering bezahlte Praktikanten sein. Einen Hinweis, wie gut ein Modell die gegebenen Daten beschreibt, gibt einem der adjustierte R^2_{adj}-Wert. In unserer einfachen Regression ergibt sich ein sehr kleiner Wert $R^2_{adj} = 0,041$. Das Modell mit seinem einzigen Faktor Geschlecht erklärt damit die Variabilität des BSL eher schlecht und ungenügend.

3.2 Lineare Regression mit mehreren Faktoren

Die unten aufgeführten Modelle sind so modeliert, dass die von diesen Modellen erzielten Ergebnisse die in am Ende von 3.1 aufgeworfenen Fragen möglichst präzise beantworten.

Nr.	Modelname	Model
1	Region und Geschlecht	$Y_i = \beta_0 + \beta_1 GE_i + \beta_2 RE_i + \beta_3 GE_i RE_i$
2	Ausbildung und Geschlecht	$Y_i = \beta_0 + \beta_1 GE_i + \beta_2 AB_i + \beta_3 GE_i AB_i$
3	Arbeitsvertrag und Geschlecht	$Y_i = \beta_0 + \beta_1 GE_i + \beta_2 AA_i + \beta_3 GE_i AA_i$
4	Beruf (ISCO) und Geschlecht	$Y_i = \beta_0 + \beta_1 GE_i + \beta_2 IS_i + \beta_3 GE_i IS_i$

Es sei angemerkt, dass die Faktoren in obigen Modellen stets (bis auf Y_i) kategorialer Natur sind. Diese werden somit als Dummyvariablen aufgefasst (s. 3.1).Es müssen also mehr Parameter geschätzt werden als in der Tablle angeben sind. Die einzelnen Modelle werden nun genauer untersucht. Neben der Residualanalyse, um die Annahmen der Modelle zu verifizieren, werden die Ergebnisse der Parameterschätzungen in Tabellenform präsentiert. Die Residualanalysen befinden sich zu größten Teilen im Anhang. Wir haben hauptsächlich graphische Analysemittel, wie QQ-Plots und Varianz-Plots verwendet. Als Varianz-Plot wird ein Plot bezeichnet, der die studentisierten Residuen gegen die Prädiktionswerte plottet, der Hinweise auf Heteroskedastizität geben kann.

3.2.1 Modell Nr.1

Dieses Model modelliert die Wechselwirkung zwischen Geschlecht und der geographischen Lage (alte bzw. neue Bundesländer) des Betriebes, in welchem die Person angestellt ist. Die Schätzung des Intercept in Höhe von 19,80 Euro gibt das durchschnittliche Einkommen einer männlichen Person, die im Westen angestellt ist, an. Ist die Person

aber weiblich, so verdient diese ca. 5 Euro weniger und nochmal 1,51 Euro weniger, wenn sie zudem im Osten lebt. Ein Mann im Osten verdient im Mittel ca. 13,20 Euro. Damit sind die Löhne von Mann und Frau im Osten in etwa gleich.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.80102    0.08092  244.69   <2e-16
GE2         -5.06969    0.12528  -40.47   <2e-16
RE2         -6.63379    0.16087  -41.24   <2e-16
GE2:RE2      5.12792    0.25593   20.04   <2e-16
```

Alle Parameter haben einen signifikanten Einfluss auf die Responsevariable. Hier betrachtet man einen partiellen F-Test, wie im Theorieteil beschrieben. Die Residualanlyse ergibt, dass die Modellannahmen erfüllt sind. Es liegen also homoskedastische, normalverteilte Residuen vor, aber der R^2_{adj}-Wert liegt bei $0,1$. Der geringe Zuwachs der zu erklärenden Variabilität durch den geographischen Faktor könnte deshalb so gering ausfallen, weil Geschlecht und Region stark unkorelliert sind (dies ist eigentlich auch eine Modellannahme, die aber vernachlässigbar ist) und der BSL vermutlich eher von persönlichen Merkmalen beeinflusst wird. Damit wären die nächsten drei Modelle motiviert.

3.2.2 Modell Nr.2

Dieses Modell, welches die Ausbildung mit in den Fokus setzt, hat einen R^2_{adj}-Wert in Höhe von $0,3$. Es sei angemerkt, dass duch Hinzunahme von Faktoren der R^2-Wert nicht sinken kann und somit der R^2-Wert bei zu vielen Faktoren keine zuverlässigen Aussagen über die Passgenauigkeit des Modells geben kann. Wir lassen deshalb die Region in diesem Modell nicht als Faktor zu (auch der Übersicht halber).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.3990     0.1812  62.915  < 2e-16
GE2          -1.6123     0.2790  -5.779 7.61e-09
AB2           6.6802     0.2031  32.888  < 2e-16
AB3          -1.2969     0.3049  -4.253 2.12e-05
AB4          11.1562     0.3623  30.794  < 2e-16
AB5          12.6755     0.2971  42.657  < 2e-16
AB6          17.9670     0.2554  70.359  < 2e-16
AB7           2.9414     0.2388  12.318  < 2e-16
GE2:AB2      -1.5843     0.3129  -5.063 4.15e-07
GE2:AB3       2.0128     0.6548   3.074  0.00212
GE2:AB4      -3.3307     0.5163  -6.452 1.13e-10
GE2:AB5      -3.8079     0.4611  -8.258  < 2e-16
GE2:AB6      -4.5009     0.4342 -10.365  < 2e-16
GE2:AB7      -1.8652     0.3593  -5.192 2.10e-07
```

In der unteren Tablle sind die Ergebnisse aufgeführt.

Frauen verdienen ausbildungsübergreifend ca. 20 % weniger. Auffällig ist hierbei der Wert für Abiturienten ohne Universitäts-/Fachhochschulabschluss und ohne Berufsausbildung: die Differenz des BSL bzgl. einer männlichen Person mit der gleichen Ausbildung ist knapp positiv. Die P-Werte für die entsprechenden Parameter (GE2,AB3, und

Ausbildung:	Absolute Differenz (Euro)	Differenz in Prozent
mittlere Reife	-1,6	-14
mittlere Reife + BA	-3,2	-18
Abitur	0,4	4
Abitur + BA	-4,94	-21
Fachhochschulabs.	-5,4	-22
Universitätsabs.	-6,1	-21
Unbekannt	-3,48	-24

Tabelle 1: Die Werte beziehen sich auf eine männliche Person mit der gleichen Ausbildung (BA= Berufsausbildung)

der Wechselwirkungsfaktor) sind deutlich größer im Vergleich zu den anderen P-Werten (aber immer noch kleiner als das α) und es lässt sich dadurch die Vermutung aufstellen, dass das Geschlecht in dieser Ausbildungsklasse kaum Auswirkungen auf den BSL hat. Dies liegt daran, dass es einfach sehr wenige Personen gibt, die nur Abitur haben und sonst keine weitere Ausbildung abgeschlossen haben.
Generell verdienen Personen mit einer besseren Ausbildung mehr.

Die Residuen dieser linearen Regression sind normalverteilt und homoskedastisch.

3.2.3 Modell Nr.3

Der Faktor Art des Arbeitsvertrages wird nun untersucht. Wir fassen die Ergebnisse kurz in einer Tabelle zusammen. Die Parametertabelle mit den P-Werten befindet sich im Anhang. Für dieses Modell gilt $R^2_{adj} = 0,286$. Die Art des Arbeitsvertrages unterteilt die Personen im Datensatz praktisch in 2 Gruppen: einmal in eine Gruppe gering(er) bezahlter Personen und eine Gruppe von Besserverdienern. In letzterer Gruppe ist die Differenz des BSL bzgl. des Geschlechts aber auch deutlich größer. Es liegt die Vermutung nahe, dass je höher das Einkommen ist, desto größer auch die Diskrepanz zwischen den Gehältern von Mann und Frau. Die Vermutung versuchen wir mit dem 4.ten-Modell zu bestätigen.

Art des Arbeitsvertrages:	Absolute Differenz (Euro)	Differenz in Prozent
unbefristet	-4,38	-21,6
befristet	-1	-7,5
Azubis	-0,08	-1,9
Altersteilzeit	-6,5	-22
geringfügig Beschäftigte	0,06	0,7

Tabelle 2: Die Werte beziehen sich auf eine männliche Person mit der gleichen Art des Arbeitsvetrages

3.2.4 Modell Nr.4

Der Faktor IS gibt Einblick in die berufliche Position der Person. So können wir die in 3.2.3 aufgestellte Vermutung überprüfen, dass Personen in hohen Postionen mehr verdienen, die Einkommensdifferenz zwischen den Geschlechtern dabei aber ebenfalls größer wird.
Die Berufsgruppe nach ISCO erlaubt die Kategorie 'keine Angabe'. Wir 'verschieben' diese Kategorie auf den Intercept und werden sie nicht weiter beachten. Von Interesse ist vor allem die Einkommensdifferenz in den anderen Kategorien.

In diesem Modell gibt es Parameter, die den partiellen F-Test nicht bestehen, bzw. die Nullhypothese wird abgelehnt. Wir befassen uns deshalb nur mit den Parametern, die einen signifikanten Einfluss auf die Response haben.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.2543     0.2170  84.109  < 2e-16
GE2          -3.5933     0.3112 -11.548  < 2e-16
IS1          14.6281     0.3893  37.571  < 2e-16
IS2           8.6151     0.2737  31.481  < 2e-16
IS3           5.4600     0.2599  21.011  < 2e-16
IS4          -5.2689     0.2647 -19.902  < 2e-16
IS5          -5.4114     0.3823 -14.153  < 2e-16
IS6          -8.7302     0.8728 -10.002  < 2e-16
IS7          -2.5052     0.2544  -9.847  < 2e-16
IS8          -3.9867     0.2736 -14.574  < 2e-16
IS9          -7.7678     0.3033 -25.610  < 2e-16
GE2:IS1      -5.5085     0.6812  -8.086 6.43e-16
GE2:IS2      -3.2263     0.4117  -7.837 4.79e-15
GE2:IS3      -2.3408     0.3795  -6.168 7.03e-10
GE2:IS4       4.7451     0.3738  12.693  < 2e-16
GE2:IS5       0.8485     0.4864   1.745  0.08105
GE2:IS6       5.2971     3.3959   1.560  0.11881
GE2:IS7      -1.6283     0.5194  -3.135  0.00172
GE2:IS8       1.1466     0.4856   2.361  0.01822
GE2:IS9       2.3402     0.4264   5.488 4.10e-08
```

ISCO-Schlüssel:	Absolute Differenz (Euro)	Differenz in Prozent
Führungskräfte	-9,1	-27,67
Akademische Berufe	-6,81	-25,3
Techniker u.ä.	-5,9	-25
Bürokräfte u.ä.	1,15	8,87
Dienstleistungsberufe	-2,7 (nicht signifikant)	-21,3
Land-/Forstwirtschaft	1,7 (nicht signifikant)	17,89
Handwerksberufe	-5,2	-33,1
Bediener von Maschienen u.ä.	-2,4	-17,15
Hilfsarbeiter	-1,25	-11,9

Tabelle 3: Die Werte beziehen sich auf eine männliche Person mit dem gleichen ISCO-Schlüssel

Bei diesem Modell liegt der R^2_{adj}-Wert bei 0,357, der bisher größte Wert. Neben der am Anfang von 3.2.4 aufgestellten Vermutung, die sich bedingt mit Tabelle 3 bestätigen lässt (dazu gleich mehr), kann eine weitere Interpretation von Modell 4 gegeben werden. In geschlechtsspezifischen Berufen werden Personen, die nicht dem typischen Geschlecht angehören, deutlich schlechter bezahlt. Die erste Vermutung ist in diesem Fall 'abhängig' von der zweiten Vermutung und umgekehrt. Um hier genauere Ergebnisse zu erzielen, müssten weitere Faktoren hinzugenommen werden (z.B. weitere Berufsgruppen, die sich möglichst nicht überschneiden).

4 Diskussion

Wir fassen die Ergebnisse der Analyse kurz zusammen: Das Geschlecht hat einen deutlichen Einfluss auf das Einkommen. Frauen verdienen teils bis zu 20 % weniger als Männer, abhängig von Berufsgruppe, Art des Arbeitsvertrages, Ausbildung usw..
In dieser Ausarbeitung wurden nicht alle untersuchten Modelle aufgeführt. Es wurden alle im Datensatz befindlichen Faktoren in die Modelle eingebunden und nicht nur bzgl. des Geschlechtes analysiert, sondern auch untereinander, z.B. die Wechselwirkung von Ausbildung und Stellung im Beruf. Allerdings ist hier die Unkorreliertheitsannahme der Faktoren nicht mehr gegeben, welches eine Interpretation erschwert.
Wie in der Einleitung bereits erwähnt, ist der Datensatz einigen Vereinfachungen unterworfen. Dadurch eignet sich die Analyse nicht dazu, konkrete Rückschlüsse auf die Grundgesamtheit zu übertragen. Hierfür würde es mehr Faktoren benötigen, wie z.B. Anzahl der Kinder, Bildung und gesellschaftlicher/finanzieller Status der Eltern (zur Ausbildungszeit), politische Regelungen (Mindestlohn u.ä.) usw.. Mit diesen Faktoren könnte man ein sehr genaues Modell aufstellen.
Unsere linearen Modelle könnte man auch weiter verbessern, indem man bestimmte Faktoren gewichtet. Eine weitere Möglichkeit der Verbesserung wäre die Verwendung einer *log*-Transformation bzgl. der Response. Auch könnte man nicht-lineare Modelle verwenden, um die Frage in 3.2.3 präziser zu beantworten.

5 Literatur

1. Autor: Helmut Pruscha
   Titel: Statistisches Methodenbuch, Springer-Verlag, 2006
   ISBN-10: 3-540-26006-4

2. Autor: Claudia Czado, Thorsten Schmidt
   Titel: Mathematische Statistik, Springer-Verlag, 2011
   ISBN: 978-3-642-17260-1

3. Autor: L. Fahrmeier, T. Kneib, S. Lang
   Titel: Regression, Springer-Verlag, 2007
   ISBN: 978-3-540-33932-8

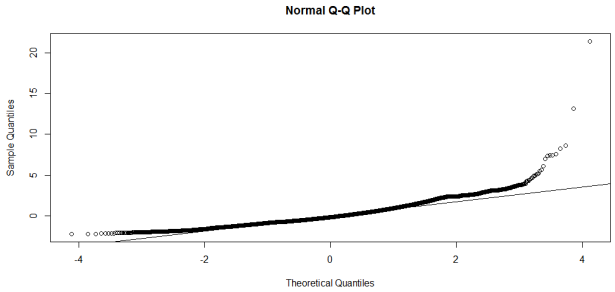
4. Autor:Thorsten Dickhaus
   Titel: Methoden der Statistik, Skript für das Wintersemester 2013/14, Version: 8. Oktober 2013

6 Anhang

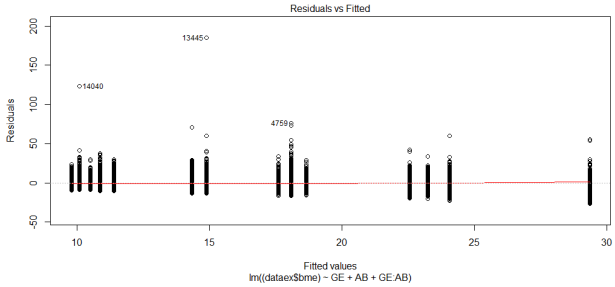
6.1 Residualanalyse

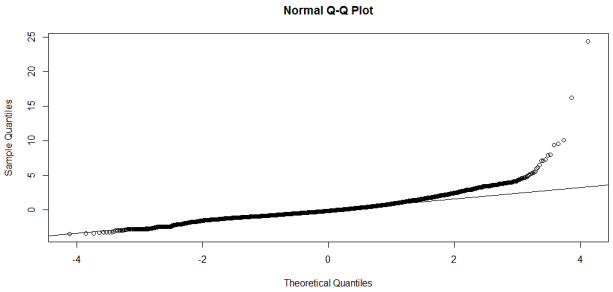
Es sind einige QQ-Plots für die Modelle angegeben. Die Punkte sollen, wenn die Normalverteilungsannahme zutrifft, auf der Geraden liegen. Bei uns ist dies nicht immer gegeben. Am Ende machen sich die Ausreißer bemerkbar. Dennoch sind die Modelle aussagekräftig, da die Datenmenge sehr groß ist (ca. 25000 Personen) und die Dichte der Punkte am Ende der Geraden abnimmt.

6.1.1 Model Nr.1



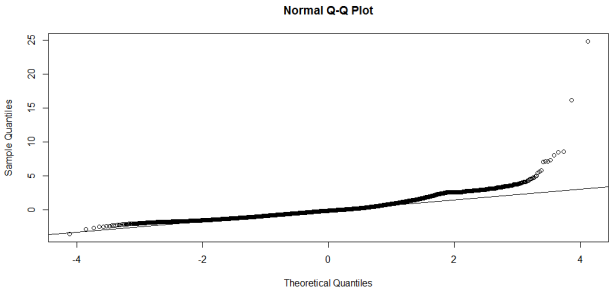
6.1.2 Model Nr.2





6.1.3 Model Nr.3

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.28254    0.07184 282.345  < 2e-16
GE2          -4.37810    0.11386 -38.453  < 2e-16
AA2          -6.64465    0.22545 -29.473  < 2e-16
AA3         -16.03686    0.28942 -55.410  < 2e-16
AA4           8.70942    0.39521  22.038  < 2e-16
AA5         -11.77410    0.22293 -52.814  < 2e-16
GE2:AA2       3.35311    0.34556   9.703  < 2e-16
GE2:AA3       4.29693    0.45172   9.512  < 2e-16
GE2:AA4      -2.13131    0.59455  -3.585 0.000338
GE2:AA5       4.44397    0.32693  13.593  < 2e-16
```



6.1.4 Model Nr.4