

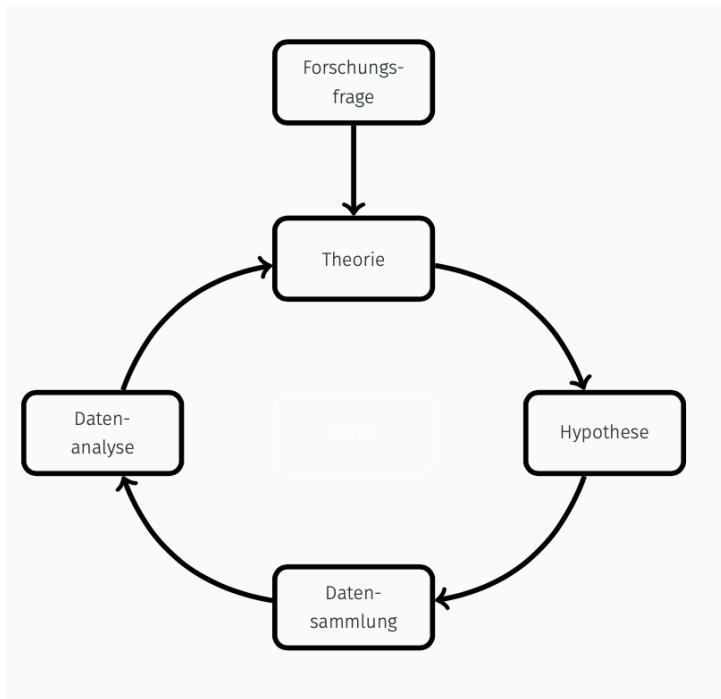
Zusammenfassung Aufbaukurs Methoden und Statistik

Inhaltsverzeichnis

VL 1.....	2
FORSCHUNGSZYKLUS	2
VL 2.....	3
KAUSALITÄT UND COUNTERFACTUALS	3
RANDOMISIERTE EXPERIMENTE UND BEOBSCHAUUNGSSSTUDIEN	8
VL 3.....	12
MOTIVATION UND MODELLE	12
REPETITION KORRELATION	13
DAS LINEARE REGRESSIONSMODELL.....	13
SCHÄTZUNG UND INTERPRETATION.....	15
GÜTEMASS (MEASURE OF FIT)	18
REGRESSION UND MITTELWERTSVERGLEICH	21
VL 4.....	22
GRENZEN DER EINFACHEN LINEAREN REGRESSION	22
MULTIPLE LINEARE REGRESSION	22
MULTIPLE REGRESSION MIT KATEGORIALEN VARIABELN.....	27
MULTIPLE REGRESSION MIT INTERAKTIONEN	28
MULTIPLES ODER ADJUSTED R^2	30
VL 5.....	32
INTERAKTION EINER KONTINUIERLICHEN UND EINER DUMMY VARIABLE.....	32
MULTIPLE LINEARE REGRESSION – EIN BLICK IN DIE MECHANIK.....	32
INTERAKTIONEN VON KONTINUIERLICHEN VARIABELN	34
SPEZIALFALL – INTERAKTION MIT SICH SELBER.....	36
VL 6.....	39
REGRESSION UND RANDOMISIERTE EXPERIMENTE	39
REGRESSION UND BEOBSCHAUUNGSDATEN	42
VL 8.....	48
DESIGN-BASED CAUSAL INFERENCE.....	48
INHALTLCHE MOTIVATION	48
PANEL DATEN	49
FIXED EFFECTS MODELL MIT PANEL DATEN	50
VL 9.....	59

VL 1

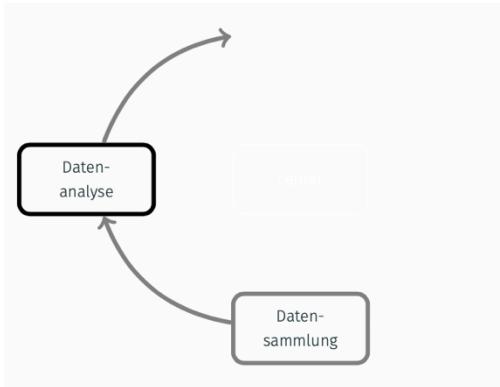
Forschungszyklus



Der Forschungszyklus ist ein big picture was läuft eigentlich? Viele Dinge die wir in der Politikwissenschaft machen kann man so strukturieren.

1. **Forschungsfrage:** Es beginnt irgendwo mit einem Phänomen das zu einer Frage führt.
Z.B. Hat Kontakt mit Minderheiten Folgen wie man mit ihnen umgeht?
2. **Theorie:** Die postuliert das ein Zusammenhang existieren sollte.
z.B. Sinnvoller Kontakt zwischen den Gruppenmitgliedern kann Vorurteile gegenüber Mitgliedern der *out-group* verringern.
Was machen wir damit, wie überprüfen und verfeinern wir diese?
3. **Hypothese:** Dafür leiten wir Hypothesen ab. Diese kann man nicht beweisen. Eine Hypothese ist eine theorie-geleitete Erwartung(en) über einen empirischen Zusammenhang, den wir beobachten können sollten.
Z.B. Personen die öfter Kontakt zu Einwanderern haben, ändern ihre Einstellung zu diesen Personen, z.B. haben sie weniger Vorurteile, unterstützen weniger wahrscheinlich einwanderungsfeindliche Parteien oder diskriminieren weniger wahrscheinlich Einwander*innen
Jetzt haben wir eine konkrete Vorstellung was wir empirisch erwarten würden gegeben einer bestimmten Theorie.
4. **Datensammlung:** Im nächsten Schritt schaffen wir eine Datengrundlage um diese Hypothesen zu überprüfen, das könnte z.B. eine Umfrage sein, wo man versucht diese Ausprägungen zu messen.
5. **Datenanalyse:** Ganzen Prozess machen und dann die gesammelten Daten der Hypothese gegenüberstellen. Die Ergebnisse werden teilweise so sein wie erwartet

teilweise aber auch nicht. Das erklärt auch den Pfeil von der Analyse zurück zur Theorie. Die Analyse kann nämlich auch zu Verfeinerungen oder Verwerfungen von Theorien führen.



In diesem Kurs konzentrieren wir uns auf die letzten beiden Phasen aus quantitativer Sicht, wobei der Schwerpunkt auf der Analysephase liegt.

Quantitative Methoden machen den Unterschied zwischen Politikwissenschaftler*innen (oder allgemein Sozialwissenschaftler*innen) und jemandem der die interessiert die Welt beobachtet!!!

VL 2

Kausalität und Counterfactuals

Ursache-Wirkungs-Beziehungen sind das Herzstück der meisten sozialwissenschaftlichen Theorien

Wir können wichtige Forschungsthemen oft in Form einer einfachen Frage formulieren: Verursacht X kausal Y?

- Führt **wirtschaftliche Entwicklung** zu **Demokratie**?
- Beeinflusst die **Ethnie** einer **Politiker:in** ihre **Chancen**, gewählt zu werden?
- Erhöht **Einwanderung** die Unterstützung für **rechtsgerichtete Parteien**?
- Führt die **Krankenversicherung** zu einer besseren **Gesundheit**?

Der Nachweis von kausalen Wirkungen anhand quantitativer Daten ist schwierig!

Heute konzentrieren wir uns darauf, zu definieren, was kausale Wirkungen sind und wie wir sie mit quantitativen Daten aus Experimenten messen können.

Die wichtigste Frage in den Sozialwissenschaften ist also, ob eine Veränderung von X (unabhängige Variable) zu einer Veränderung von Y führt. Also eine **kausale Frage**.

Der Messschritt ist auch wichtig aber das Kernstück ist die Auswertung von diesen Ursache-Wirkungs Beziehungen. Das schwierige ist es die Kausalität zu beweisen. Also eine kausale Wirkung zwischen einem «outcome» (Y) und einem «treatment» (X) nachzuweisen.

Beispiel aus VL:

- **Y (Abhängige Variable oder "outcome")**: *Gesundheit*
 - Wie ist die selbst eingeschätzte Gesundheit einer Person?
- **X (Unabhängige Variable oder "treatment")**: *Krankenversicherung*
 - Hat die Person eine Krankenversicherung oder nicht?

Der Schlüssel um Kausalität zu verstehen ist über das Counterfactual (kontrafaktischer Zustand) nachzudenken.

- Treatment: key condition, Wo die Veränderung ihren Ursprung hat, z.B. Krankenversicherung ja oder nein (siehe auch Buch S. 56)
- Outcome: Das Ergebnis, Was von der Veränderung betroffen ist, z.B. Gesundheit
- Counterfactual: «Was wäre gewesen, wenn das Treatment nicht stattgefunden hätte?
- Kausale Inferenz: Um die Kausalität zu beweisen vergleichen wir also das Ergebnis wie es gewesen wäre ohne das treatment (also wenn eine key condition anders gewesen wäre) mit dem was wir effektiv beobachtet haben.

-> Das grosse Problem dabei ist das wir counterfactual Ergebnisse nicht beobachten oder messen können! Kausales Schliessen erfordert also das Schätzen von counterfactual Ergebnissen, um sie dann mit den realisierten Ergebnissen zu vergleichen. (Grundproblem kausaler Inferenz: Wir haben nicht beide Daten für eine Person. Ich kann keine individuellen kausalen Effekte feststellen.)

Der kausale Effekt ist also der Unterschied zwischen den beiden möglichen Ergebnissen wenn das treatment stattgefunden hätte oder nicht.

Potential outcomes (potentielle Ergebnisse)

Die potenziellen Ergebnisse von Y sind die Werte von Y , die für verschiedene Werte von X realisiert würden. z.B.

- $Y_i(1)$ = der Wert Y_i , den Y_i gehabt hätte, wenn X_i gleich 1 gewesen wäre
- $Y_i(0)$ = der Wert Y_i , den Y_i gehabt hätte, wenn X_i gleich 0 gewesen wäre

Kausaler Effekt

Für jede einzelne Person kann, wenn wir beide potenziellen Ergebnisse beobachten könnten, der Behandlungseffekt von X auf Y für diese Person wie folgt berechnet werden:

$$Y_i(1) - Y_i(0)$$

Das wäre also der **treatment-Effekt** für Person i

- $Y_i(1)$ = Die Gesundheit, die Person i hätte, wenn die Person eine Krankenversicherung hätte
- $Y_i(0)$ = Die Gesundheit, die Person i hätte, wenn die Person keine Krankenversicherung hätte

Was sind die *treatment* Effekte?

- Wenn $Y_i(1) > Y_i(0)$, dann verbessert die Versicherung die Gesundheit
- Wenn $Y_i(1) < Y_i(0)$, dann verschlechtert die Versicherung die Gesundheit
- Wenn $Y_i(1) = Y_i(0)$, dann hat die Versicherung keinen Einfluss auf die Gesundheit

- $X_i = 1$, wenn die Person versichert ist, und $X_i = 0$, wenn sie nicht versichert ist.
- $Y_i(1)$ ist die Gesundheit der Person, wenn sie versichert wäre.
- $Y_i(0)$ ist die Gesundheit der Person, wenn sie nicht versichert wäre.
- Der *treatment* Effekt für eine Person ist $Y_i(1) - Y_i(0)$.

Individuum	X_i	$Y_i(1)$	$Y_i(0)$	Treatment Effekt
1	1	5	3	2
2	1	5	4	1
3	0	3	3	0
4	0	4	3	1

$$\text{Average treatment effect (ATE)} = \frac{2+1+0+1}{4} = \frac{4}{4} = 1$$

Das sind hypothetische Zahlen

Aber wir können *beide* potenziellen Ergebnisse für keine einzelne Person beobachten!

Individuum	X_i	$Y_i(1)$	$Y_i(0)$	Treatment effect
1	1	5	?	?
2	1	5	?	?
3	0	?	3	?
4	0	?	3	?

$$\text{Average treatment effect (ATE)} = \frac{?+?+?+?}{4} = \frac{?}{4} = ?$$

So sähe es in der echten Welt aus

Das illustriert das Fundamentalproblem kausaler Inferenz. Wir beobachten immer nur ein potenzielles Ergebnis für eine bestimmte Person, und unser beobachtetes Ergebnis hängt vom Status unserer erklärenden Variable ab.

Konsequenzen

- Wir können keine kausalen Effekte für Einzelpersonen berechnen. (*Vielleicht könnten wir das in Filmen: Sliding Doors*)
- Wir müssen *counterfactuals schätzen* für den Vergleich mit realisierten Ergebnissen.

Es ist also nicht möglich das ich für jedes Individuum den kausalen Effekt bestimmen kann, es reicht aber auch schon, wenn ich an einen durchschnittlich kausalen Effekt komme, viele Fragen können wir mit dem bereits beantworten. Wir finden nicht beide outcomes für ein Individuum und müssen also einen Weg finden um das indirekt machen zu können. Das machen wir mit dem ATE.

- Average Treatment Effect (ATE): Differenz der Mittelwerte von denen die ein Treatment erhalten haben und von denen die kein Treatment erhalten haben, also der Mittelwert der Unterschiede

Frage: Ist die Differenz der Mittelwerte gleich dem ATE?

Individuum	X_i	$Y_i(1)$	$Y_i(0)$	Treatment Effekt
1	1	5	3	2
2	1	5	4	1
3	0	3	3	0
4	0	4	3	1

$$\text{Differenz der Mittelwerte} = \frac{5+5}{2} - \frac{3+3}{2} = 5 - 3 = 2$$

$$\text{Average treatment effect (ATE)} = \frac{2+1+0+1}{4} = \frac{4}{4} = 1$$

Nein! Die Differenz der Mittelwerte ist grösser als der ATE. Weshalb?

Der Mittelwertsschätzer ersetzt nicht den wahren ATE. Ob ein Individuum ein treatment erhält oder nicht (in diesem Fall ob ein Individuum eine KK hat oder nicht) kann auch abhängig sein von anderen Faktoren und Einflüssen. Das nennt man **Confounding**, es gibt also andere Unterschiede zwischen der treatment Gruppe (die das treatment erhält) und der Kontrollgruppe (die das treatment nicht erhält) abgesehen vom treatment Effekt. Man hat nie zwei genau gleiche Gruppen von Individuen die sich nur dadurch unterscheiden ob sie eine Krankenkasse haben oder nicht. Das führt dazu das die Mittelwertsdifferenz eine verzerrte Schätzung des ATE liefern kann. (siehe Folie oben) Das nennt man dann **Confounding Bias = Verzerrung**. Diese Verzerrung kann positiv oder negativ sein. Bei Beobachtungsanalysen sollte man sich deshalb IMMER die Frage nach confounding stellen.

Wir wissen nun das man solche Beobachtungsanalysen mit dem Mittelwertsunterschied alleine nicht kausal interpretieren kann. Wir können aber wenn wir noch mehr Informationen in den Daten haben die Gefahr für confounding bias beurteilen. Also ist das Resultat bereits getrieben von confounding. Wir können uns andere Variablen anschauen z.B. das Einkommen. Wenn der Anteil mit und ohne Versicherung gleich hoch ist unabhängig vom Einkommen dann weiss ich das ich diesen Confounder ausschliessen kann. Wir können also, wenn wir andere Variablen haben die beiden Gruppen (treatment/control) auf ihre Ausgeglichenheit (balance) überprüfen. Im Beispiel aus der VL wurde das mit dem Alter gemacht.

Table 7: NHIS data

id	insured	health	age	female	years_educ	non_white	income
1	FALSE	4	29	TRUE	14	FALSE	19282.93
2	FALSE	4	35	FALSE	11	FALSE	19282.93
3	TRUE	3	32	FALSE	12	FALSE	167844.53
4	TRUE	3	34	TRUE	16	FALSE	167844.53
5	TRUE	4	45	FALSE	12	FALSE	85985.78
6	TRUE	4	44	TRUE	12	FALSE	85985.78

Daten aus der Vorlesung mit denen gearbeitet wurde

Eine Folge von *confounding bias*: *Treatment* und Kontrollgruppe unterscheiden sich in Bezug auf andere Merkmale (abgesehen vom *treatment Status*). Wir können das überprüfen, indem wir die Ausgeglichenheit (*balance*) zwischen den beiden Gruppen (*treatment/control*) auf verschiedenen *pre-treatment* Variablen vergleichen.

1. Berechnen Sie $\bar{X}_{T=1}$ (Durchschnitt *pre-treatment* Variable in der *treatment* Gruppe)
2. Berechnen Sie $\bar{X}_{T=0}$ (Durchschnitt *pre-treatment* Variable in der Kontrollgruppe)
3. Große Unterschiede zwischen $\bar{X}_{T=1}$ und $\bar{X}_{T=0}$ weisen auf ein Problem hin.

Zum Beispiel *age*:

```
age.block <- nhis %>%
  group_by(insured) %>%
  summarize(Mean = mean(age, na.rm=TRUE))
```

→ Die Versicherten sind im Durchschnitt 2.4 Jahre älter als die Unversicherten.

Wir sehen, die Leute die eine KK haben sind im Schnitt 2.4 Jahre älter. Wir können das auch noch für andere Variablen tun und sehen folgende Unterschiede.

insured	age	female	years_educ	income
Uninsured	40.9	49.0	11.3	42892.6
Insured	43.3	50.2	14.1	101315.4
Difference	2.4	1.2	2.9	58422.9

Unsere Versicherten sind ...

- ...älter (43) als die Nichtversicherten (41)
- ...haben mehr Bildung (14.3 years) als die Nichtversicherten (12 years)
- ...bedeutend wohlhabender (\$101315) als die Nichtversicherten (\$42893)

Auswirkungen:

1. Verschiedene Variablen sind *nicht ausgeglichen* in Bezug auf den Versichertenstatus
2. Es ist wahrscheinlich, dass wir hier ein *Confounding* Problem haben

Das ist sicher eine Situation, welche wir mit Skepsis anschauen sollten, weil wir eine Reihe von relevanten Faktoren haben die nicht gleich verteilt sind in diesen beiden Gruppen. Es ist nicht zulässig davon auszugehen das ein Mittelwertsvergleich ein guter Schätzer ist für den durchschnittlichen treatment-Effekt. Das Mass der Gesundheit nur daran zu messen ob die Person eine KK hat ist keine gute Idee weil es andere Faktoren gibt die überhaupt beeinflussen ob jemand eine KK hat und die beiden Gruppen nie genau gleich sind. Es ist nicht sinnvoll einfach diese Mittelwerte zu vergleichen und zu meinen wir haben einen kausalen Effekt gemessen. Wir haben etwas gemessen einfach die Differenzen von zwei Mittelwerten aber diese sind nicht so relevant, weil sie nicht kausal interpretiert werden können, weil wir das Problem des *Confounding*s haben.

Das Beispiel mit der Krankenkasse das wir eben gesehen haben ist eine Beobachtungsstudie. Die Zuteilung in das treatment erfolgt hier nicht systematisch durch die Forscher*in erfolgt oder durch eine andere Instanz sondern von der Natur aus. Also durch eigene Entscheide oder irgendwelche systemische Elemente usw. Das Gegenstück dazu ist das randomisierte Experiment. Hier werfen wir z.B. eine Münze wenn Kopf kommt gibt es eine Krankenversicherung und wenn Zahl kommt dann nicht. Das bedeutet aber auch das man hier als Forscher*in ganz direkt in die reale Welt eingreift. Das ist mit ein Grund wieso wir das nicht oft tun. Beobachtungsstudien sind der häufigste Ansatz in den Sozialwissenschaften, gleichzeitig ist unser Ideal was wir gerne hätten ist es treatments zu randomisieren, weil das der Goldstandard ist um kausale Inferenz zu betreiben also kausale Schlüsse zu erarbeiten mit Daten.

Randomisierte Experimente

- Beobachtungen werden *zufällig* unterschiedlichen X-Werten zugewiesen
- Forschende greifen direkt in die reale Welt ein
- Goldstandard für kausale Inferenz

Beobachtungsstudien

- Beobachtungen werden 'von der Natur' unterschiedlichen X-Werten zugewiesen
- Forschende beobachten die Welt
- Sehr häufiger Ansatz in den Sozialwissenschaften

Randomisierte Experimente und Beobachtungsstudien

Wir haben bereits gesehen, dass der Mittelwertsvergleich im Allgemeinen keine unverzerrte Schätzung des ATE erlaubt, da es Probleme mit *confounding* geben kann.

Frage: Wie kann man dieses Problem vermeiden? **Antwort:** Mit einem randomisierten Experiment!

Intuition

Randomisiertes *treatment* beseitigt nicht Unterschiede zwischen Beobachtungen, aber sorgt dafür, dass es keine systematischen Unterschiede nach Gruppen (*treatment, control*) geben sollte.

Konsequenz

Mann kann keine individuellen kausalen Effekte berechnen, aber man kann den durchschnittlichen kausalen Effekt schätzen.

In einem randomisierten Experiment können wir also z.B. dafür sorgen, dass es gleich viele Arme Leute in der treatment und in der Kontrollgruppe gibt und wenn die Einkommensverteilung gleich ist in beiden Gruppen scheidet Einkommen aus als möglicher Confounder aus. Wir haben immer noch nicht den individuellen treatment Effekt (weil wir immer noch nur ein outcome anschauen können) aber wir können ihn für ganze Gruppen

anschauen mit der Mittelwertsdifferenz, welche dann in Erwartung dem wahren ATE entspricht.

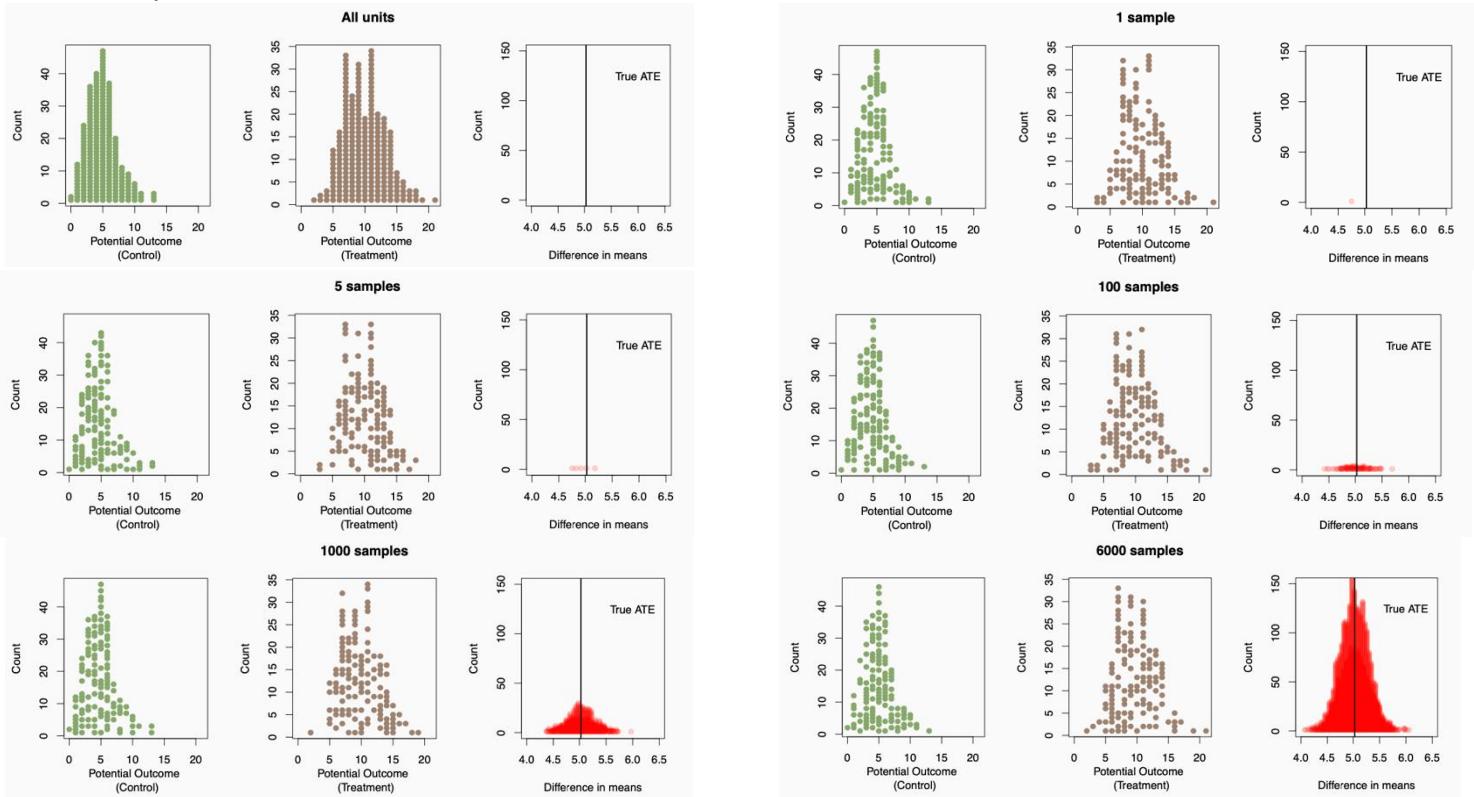
Wenn das *treatment*, X , zufällig zugewiesen wird...

- ... sind die beiden Gruppen (*treatment*, *control*) im Durchschnitt ähnlich in Bezug auf beobachtete und auch unbeobachtete Eigenschaften
- ... der einzige *systematische* Unterschied liegt darin, wer das *treatment* erhalten hat und wer nicht
- ... der durchschnittliche *outcome* in der Kontrollgruppe wird ähnlich sein zu dem, was die *treatment* Gruppe gehabt hätte, wäre sie dem *treatment* nicht ausgesetzt gewesen

→ Wir können die Mittelwertsdifferenz hier verwenden, um den ATE zu schätzen!

Das nennt man Randomisierung und Selektionsverzerrung (selection bias)

Achtung: Wir haben gesagt in Erwartung!!! Also in Erwartung sollten sich diese Werte nicht unterscheiden. Eine Randomisierung ist im Erwartungswert ein perfektes Instrument. Wenn ich z.B. eine Münze 100 Mal je 20 Mal werfe dann werden wir sehen dass es immer irgendwo um 10 Mal Kopf oder Zahl gibt. Wenn ich das 100 Mal mache wird es aber auch Momente geben wo ich 15 Mal ein Kopf und 5 Mal eine Zahl haben. Es kann also sein das eine einzelne Randomisierung nicht zu einem perfekten Ausgleich führt von z.B. Einkommen. Wir müssen uns also bewusst sein, dass es die Gefahr gibt, dass die Randomisierung nicht dazu führt das confounding komplett unmöglich ist, aber im Erwartungswert ist es so. In der Realität machen wir ja oft nicht 100 Mal dasselbe Experiment. Beispiel aus der VL das dies illustriert:



Bei all units oben links sehen wir in diesen theoretischen Daten die künstlich erzeugt wurden, die outcomes mit und ohne treatment sowie den wahren average treatment-Effekt. Dieser liegt ganz knapp über 5.0. Was nun geschieht in einem Experiment sieht man bei 1 sample oben rechts. Man hat eine Stichprobe aus diesen randomisierten Daten der Population und kommt auf einen ATE von etwa 4.8. Wir wissen in diesem Fall das ist nicht ganz richtig aber ich bin schon nahe dran. Man sieht wenn man dies ganz oft wiederholt kann man den Mittelwert dieser ATE's nehmen und dann kommt man auf das Richtige Ergebnis. Also der Mittelwert dieser empirischen Verteilung entspricht dem wahren ATE. Aber in der Realität machen wir halt immer nur ein Experiment und das wird halt ab und an ein bisschen zu tief oder ein bisschen zu hoch sein.

- Die Randomisierung führt dazu, dass der Mittelwertsvergleich ein unverzerrter Schätzer für den ATE ist
- Im Durchschnitt kann man davon ausgehen, dass ein perfektes randomisiertes Experiment einem das richtige Resultat liefert
- Das ist aber keine Garantie, dass die Antwort von einem einzigen Experiment exakt korrekt ist!
- Randomisierte Experiment gelten als der Goldstandard in den Sozialwissenschaften

Das heisst also der Schätzer bei einem randomisierten Experiment nicht verzerrt ist, er ist also auf jeden Fall nicht systematisch zu hoch oder zu tief. Es gibt einfach einen Fehler um diesen herum (Varianz). Das Ganze ist eine Normalverteilung. (Siehe 6000 samples unten rechts) In der Realität kann man das natürlich nicht 6000-mal machen um das zu überprüfen. Man muss nur schon dafür kämpfen die Ressourcen für eine einmalige Durchführung zu bekommen. Trotzdem wäre das wünschenswert also der Goldstandard, weil wir so extrem viele Probleme ausschliessen. Wenn man weitere Angaben zu den Individuen hat kann man ja trotzdem auf confounders überprüfen, also schauen ob sich control und treatment Gruppe im Bezug auf Alter, Einkommen etc. unterscheidet. Das einzige Problem ist, wenn ich das so mit Daten mache kann ich das nur im Bezug auf beobachtete Größen welche in meinen Daten vorhanden sind machen und nicht auf unbeobachtete Größen, welche man vielleicht auch übersehen könnte. Darum ist das randomisierte Experiment der Goldstandard.

Beispiel: Das Experiment mit der Krankenversicherung wurde in den USA tatsächlich randomisiert durchgeführt.

Verbessert eine Krankenversicherung den Gesundheitszustand?

Das RAND Health Insurance Experiment (RAND) war ein Experiment, das zwischen 1974 und 1982 in den USA durchgeführt wurde. Bei diesem Experiment wiesen die Forscher *nach dem Zufallsprinzip* Personen eine Krankenversicherung zu.

- Y (anhängige Variable): Gesundheitszustand
 - "Would you say your health in general is excellent (5), very good (4), good (3), fair (2), or poor (1)?"
- X (unabhängige Variable): Versichert
 - Was the participant *randomly allocated* to receive health insurance? TRUE = Insured, FALSE = Not insured

Wir werden wieder Informationen aus einigen der anderen Fragen der Umfrage verwenden (Geschlecht, Einkommen, Ethnie usw.).

Nun können wir die Mittelwerstdifferenz anhand von experimentellen Daten berechnen:

Mean health level for insured & non-insured individuals

```
rand %>%
  group_by(insured) %>%
  summarize(Mean = mean(health, na.rm=TRUE))

## # A tibble: 2 × 2
##   insured    Mean
##   <lgl>     <dbl>
## 1 FALSE     3.42 
## 2 TRUE      3.41
```

Der Unterschied zwischen den Personen mit und ohne Krankenversicherung ist fast komplett verschwunden in den experimentellen Daten.

	insured	age	female	years_educ	income
Uninsured	331	56.0	121	32597.2	
Insured	33.2	53.4	12.0	31220.9	
Difference	0.1	-2.6	-0.1	-1376.3	

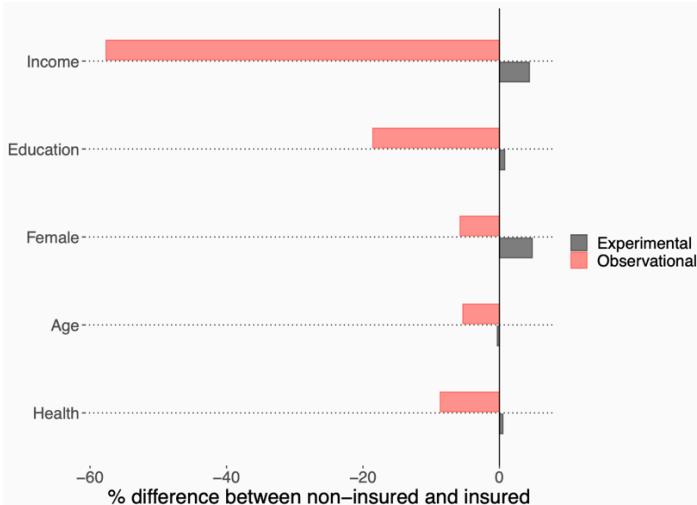
Versicherte und nicht-versicherte Individuen

- haben in etwa dieselbe Wahrscheinlichkeit weiblich zu sein
- haben in etwa dasselbe Alter
- haben in etwa gleich viele Jahre Bildung
- haben in etwa dasselbe Einkommensniveau

Implikationen:

1. In den experimentellen Daten sehen wir keine Hinweise auf Selektionsverzerrungen (*imbalance*)
2. Der geschätzte Effekt in den experimentellen Daten ist viel kleiner

Wir sehen die Personen in der treatment und control Gruppe sind fast gleich alt, Unterschied beim Geschlecht, fast kein Unterschied bei Bildung und nur ein sehr kleiner Unterschied beim Einkommen. Wir haben also **nicht so klare Unterschiede** wie im Fall vorhin wo wir nicht wissen, ob der Unterschied ein Resultat vom Unterschied dieser Gruppen ist und wir sehen **das der Effekt viel kleiner ist**. Wir würden also hier eher davon ausgehen das es keinen kausalen Effekt gibt zwischen der Gesundheit einer Person und davon ob sie eine KK hat oder nicht.



Wir sehen hier den Unterschied zwischen der Beobachtungsstudie und dem Experiment. Unterschiede auf den Variablen in den zwei verschiedenen Studien. Diese Grafik ist die Quintessenz davon wieso Experimente eigentlich besser geeignet sind für uns, weil auf ganz vielen anderen Variablen sich diese beiden Gruppen nicht unterscheiden.

Weshalb verwenden wir nicht immer Experimente, wenn diese doch der Goldstandard in der sozialwissenschaftlichen Forschung sind?

1. Praktische Bedenken

- Oft ist es nicht möglich das *treatment* zu randomisieren
- z.B. wie wollen wir das Wahlsystem von Ländern randomisieren?

2. Ethische Bedenken

- Experimente mit echten Menschen können ethische Regeln verletzen
- z.B. [Manipulieren welche Meldungen Personen auf Facebook zu sehen kriegen](#)

3. Ressourcen

- Experimente können teilw. sehr teuer und aufwendig sein
- Dieser Punkt trifft insbesondere für BA, MA und Doktoratsarbeiten zu

VL 3

Motivation und Modelle

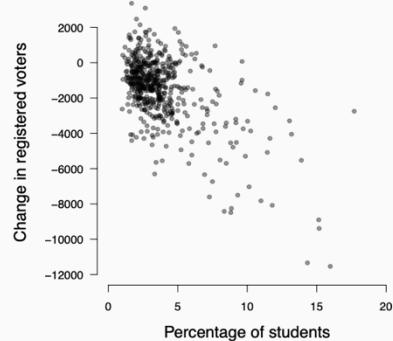
Diese Woche schauen wir uns das Herzstück der quantitativen Sozialwissenschaften an: Das lineare Regressionsmodell. Zunächst zur Beschreibung und Vorhersage. Im ersten Schritt schauen wir uns die lineare Regression nicht dafür an um unmittelbar kausale Inferenz zu generieren, sondern im ersten Schritt ist das jetzt einfach mal ein Instrument zur Beschreibung von Daten und von Zusammenhängen die wir in diesen Daten finden

Student:innen und das Wahlregister

Vor 2015 konnte im Vereinigten Königreich der Haushaltsvorstand alle Mitglieder des Haushalts zur Wahl anmelden. Ab 2015 mussten sich alle separat für die Wahl registrieren lassen.

Es gab Bedenken, dass dies dazu führen würde, dass viele Studenten und junge Menschen aus dem Wählerverzeichnis "herausfallen". Wir erheben Daten zur Wählerregistrierung in 573 britischen Wahlbezirken, um diese Befürchtung zu überprüfen.

- **Analyseeinheit:** 573 parlamentarische Wahlkreise (alle Wahlkreise in England und Wales).
- **Abhängige Variable (Y):** Veränderung der Anzahl registrierten Wähler:innen in einem Wahlkreis (von 2010 zu 2015).
- **Unabhängige Variable (X):** Anteil Student:innen der Population eines Wahlkreises.



- Was sieht man in dieser Grafik?
- Gibt es einen **positiven** oder einen **negativen** Zusammenhang zwischen X und Y?
- Die **lineare Regression** hilft uns solche Zusammenhänge präzise zu beschreiben.

Inhaltliches Beispiel.

Y-Achse: in den meisten Distrikten hat die Anzahl der Wähler*innen abgenommen, auf der X-Achse der Anteil Student*innen an der Bevölkerung in einem Wahlkreis, die meisten würden sagen ja Punktewolke sieht so aus als würde es etwas runtergehen, würde die Aussage also stützen das jüngere Leute nicht so schnell sich wieder registrieren würden, die lineare Regression kann uns helfen zu erklären was der empirische Zusammenhang in dieser Stichprobe in diesen Daten ist. (Zusammenhang ist eine Schätzung) Lineare Regression ist also ein Modell

Was ist ein Modell? Wenn man fragt was ist eine Eisenbahn kann man das anhand verschiedener Modelle machen



Die Modelle unterscheiden sich in ihrer Komplexität, wenn die Frage ist was ist eine Eisenbahn, hätte wahrscheinlich ein viel einfacheres Modell bereits gereicht.

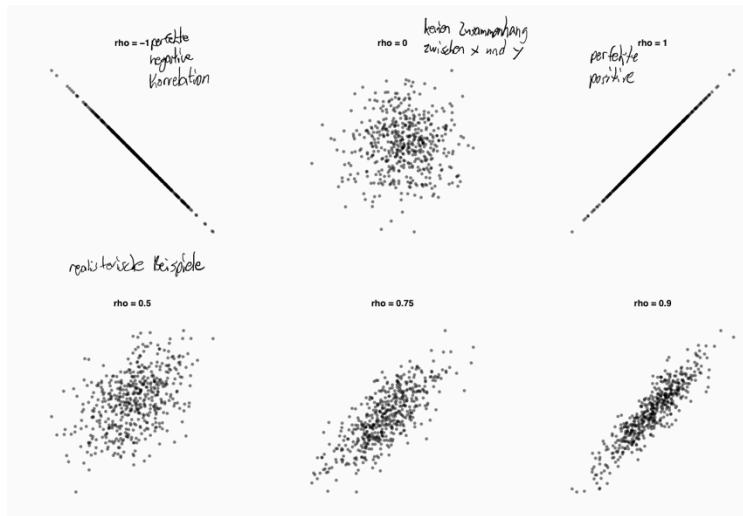
Ein komplizierteres Modell muss nicht unbedingt besser sein, ein Modell soll einfach eine Vereinfachung darstellen. Der Sinn von einem Modell ist also eine Reduktion der Komplexität auf die wesentlichen Elemente/Phänomene die wir beschreiben möchten. Ein wissenschaftliches Modell ist immer eine vereinfachte Abstraktion der Realität.

- Ein Modell ist eine vereinfachte Abstraktion der Realität
- Oft brauchen wir Modelle um Schlüsselemente eines komplizierteren Prozesses zu beschreiben
- Ein Modell soll nicht *alle* Eigenschaften der Realität abbilden, sondern nur diejenigen, die wir brauchen, um ein Phänomen zu beschreiben
- "All models are wrong, but some are useful" – [George Box](#)
- Wir werden statistische Modelle verwenden, die alle 'falsch' sind, aber einige sind hilfreich

Im Kern ist das die Modelle «falsch» sind wenn der Anspruch ist das sie eine gute Abbildung sind und hilfreich, weil sie uns helfen Komplexität zu reduzieren. Wir machen genau das. Alle statistischen Modelle die wir verwenden werden sind eine Reduktion der Realität.

Repetition Korrelation

Wir haben Korrelation letztes Semester kennengelernt.



Das ist ein Mass das wir bereits kennen um in einem Datensatz zu erklären wie das Verhältnis zwischen zwei Variablen ist.

Das lineare Regressionsmodell

Wir möchten das Verhältnis zwischen zwei Variablen beschreiben. In der einfachsten Form gehen wir davon aus das ein linearer Zusammenhang besteht.

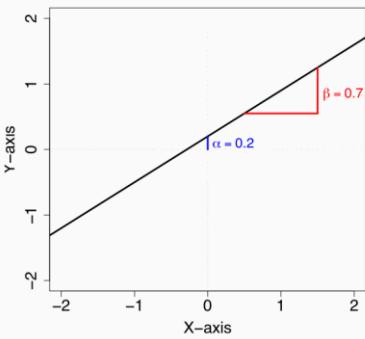
- Die Beziehung zwischen zwei Variablen lässt sich am einfachsten durch eine Gerade beschreiben
- Das **lineare Regressionsmodell** ist eine **Approximation** der Beziehung zwischen unserer **unabhängigen Variable X** und unserer **abhängigen Variable Y**.
Anteil Student:innen
Dann können wir ein Modell schätzen
- In unserem Fall ist die lineare Regression ein Modell, welches die Beziehung zwischen:
 - dem Anteil Student:innen und
 - der Veränderung in der Zahl der registrierten Wähler:innen

beschreibt.

Wir können eine Gerade wie folgt repräsentieren $Y = \alpha + \beta X$

$\alpha = 0.2$ and $\beta = 0.7$

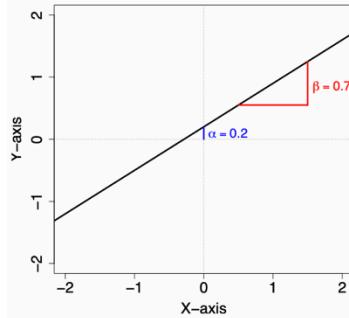
β Steigungskoeffizient



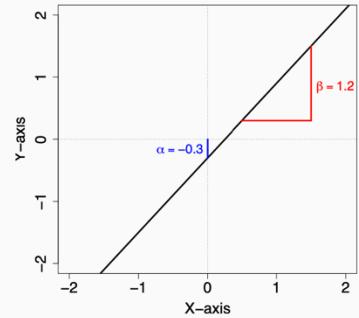
- α ist die Konstante: Der Wert von Y wenn $X = 0$
- β ist die Steigung: Die Veränderung von Y wenn X um eine Einheit ansteigt
- Hier: Ein Anstieg um eine Einheit in X führt zu einem Anstieg von 0.7 Einheiten in Y

Unterschiedliche Werte für α und β definieren unterschiedlichen Geraden

$\alpha = 0.2$ und $\beta = 0.7$



$\alpha = -0.3$ und $\beta = 1.2$



Gerade wird definiert über zwei Elemente. Alpha ist dort wo die Gerade die Y-Achse schneidet. Also gegeben das $X = 0$ ist, den Wert denn wir dann auf der Y-Achse beobachten. Beta ist die Steigung (Steigungskoeffizient). Das ist wie viel Veränderung habe ich auf Y gegeben der Veränderung auf X . Eine steilere Gerade heisst, dass wir einen grösseren Beta Wert haben. Wenn Beta 0 ist dann verläuft die Gerade gerade.

Unser Ziel ist es die Gerade zu schätzen, die am besten zu den Daten passt.

Eine einfache Möglichkeit, den Zusammenhang zwischen zwei Variablen zu beschreiben, ist die Annahme, dass sie linear verbunden sind.

Wir können das mit dem linearen Regressionmodell machen:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

↓
ist konstant für alle Personen im Datensatz
↓
individuelle Person also variiert natürlich über Personen
in Datensatz

- Beobachtungen $i = 1, \dots, n$
- Y ist die abhängige Variable
- X ist die unabhängige Variable
- α ist die Konstante oder der Achsenabschnitt
- β ist die Steigung
- ε_i ist der Fehlerterm oder das Residuum

Unterschied zwischen erwartetem und beobachteten Wert von Y

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

α und β sind die Koeffizienten oder Parameter der Regressionsgeraden.

- α zeigt den durchschnittlichen Wert von Y an, wenn $X = 0$ ist
- β gibt den durchschnittlichen Wechsel von Y an, wenn X sich um eine Einheit ändert
Beta ist der Fokus das ist der wichtigste Alpha beschäftigt uns wenig
- → beschreiben also den durchschnittlichen Zusammenhang zwischen X und Y

ε_i ist der Fehlerterm

Was ist im Fehlerterm? Dort stecken all die Variablen drin die ich nicht bereit explizit modelliert habe und die mit y korrelieren

- ε_i ermöglicht es einer Beobachtung nicht direkt auf der Geraden zu zu liegen
- → repräsentiert alle Faktoren (abgesehen von X), die mit Y korrelieren aber nicht im Modell enthalten sind

Was ist eine Veränderung um eine Einheit und Unterscheidung von Prozenten und Prozentpunkten!

Da β die Veränderung von Y beschreibt, wenn X sich um eine Einheit verändert, müssen wir wissen in welcher Einheit X gemessen wird

Beispiele für eine Veränderung um eine Einheit...

- ...Alter, in Jahren gemessen, ist ein Jahr
- ...Grösse, in cm gemessen, ist ein cm
- ...GDP per capita, in USD gemessen, ist ein USD

Frage: Was ist ein Anstieg um eine Einheit bei "Anteil Student:innen"?

Antwort: Ein Anstieg des Anteils Student:innen um einen Prozentpunkt

Ein häufiger Fehler in der Interpretation liegt darin Prozente mit Prozentpunkten zu verwechseln. Was ist der Unterschied?

Ein Anstieg des Anteils Student:innen von 40% auf 44% repräsentiert:

- Einen Anstieg von 4 Prozentpunkte
- Einen Anstieg von 10 Prozent

Wenn wir Variablen in Prozenten gemessen ins Modell einschliessen, werden wir (fast) immer über Veränderungen in Prozentpunkten sprechen.

Notation mit Alpha und Beta Dach Verstehen:

- α & β repräsentieren die durchschnittliche Beziehung zwischen X und Y

- Es handelt sich dabei um sog. Populationsparameter – also Werte, die existieren

Es gibt den wahren Wert für Beta und Alpha und wenn wir die Daten der gesamten Population hätten könnten wir diese Werte einfach messen, wir arbeiten aber nur mit einem Teil dieser Realität, das sind dann die Sachen die wir schätzen

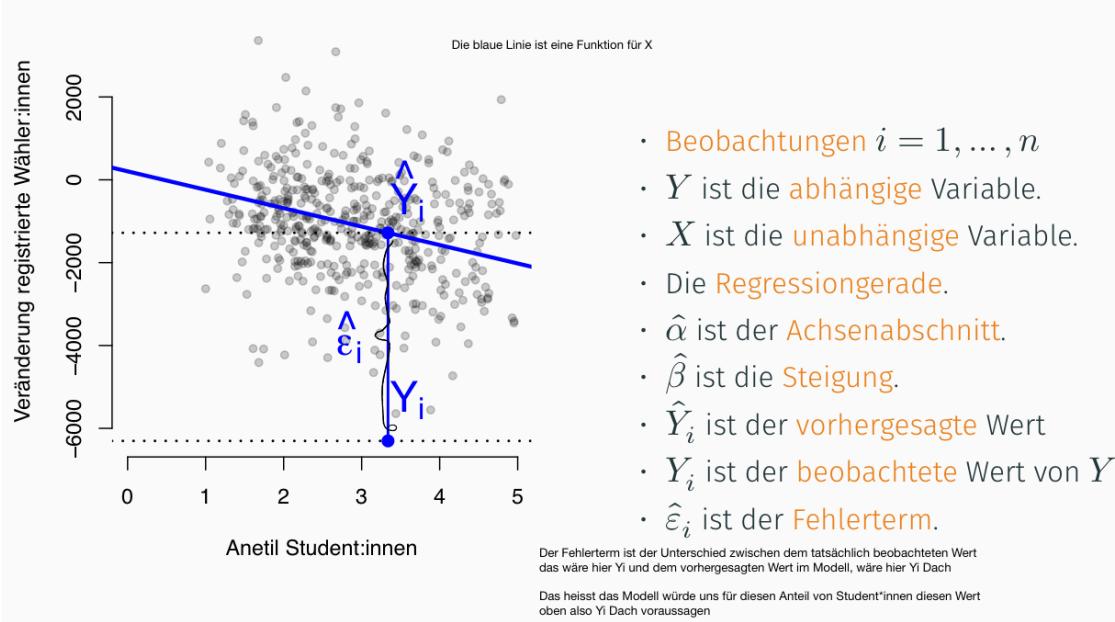
- Wir möchten gerne den genauen numerischen Wert kennen für α und β
- Da wir diese Werte nicht kennen, müssen wir sie schätzen
- Wir schätzen die Werte für die Parameter basierend auf den Daten
- Um geschätzte Parameter als solche zu bezeichnen, verwenden wir eine leicht angepasste Notation

- α wird zu $\hat{\alpha}$, was als "Alpha Dach" gelesen wird

- β wird zu $\hat{\beta}$, was als "Beta Dach" gelesen wird

Die Dachs bezeichnen also das es sich um eine modellbasierte Vorhersage handelt

So sieht das Regressionsmodell dann mit der Gerade aus.

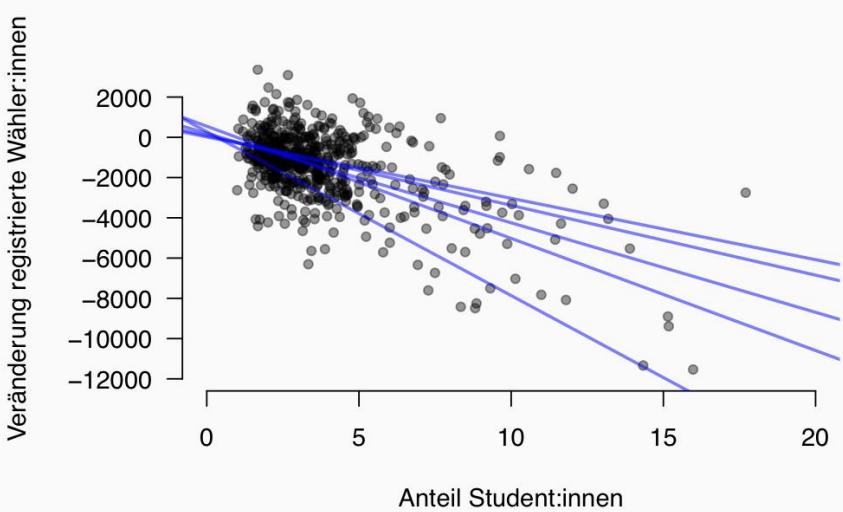


Fehlerterm noch einmal erklärt: Wenn ich die Vorhersagen habe für Y_i , dann kann ich die Residuen (Fehlerterm) generieren. Dann kann ich konkret schauen was war der Wert bei einer bestimmten Beobachtung auf Y . Was ist das was das Modell vorhersagt und das ist eine Abweichung. Der Unterschied zwischen einem tatsächlichen Wert einer Beobachtung auf Y und was das Modell für diese Person vorhersagt.

Schätzung und Interpretation

Das Hauptziel der Regression ist es die ‘beste’ Gerade zu finden für die Daten.

Welche Gerade passt am besten zu diesen Daten?



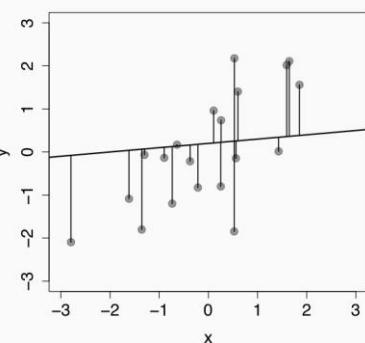
Das vorhin war ja nur ein Ausschnitt aus den Daten (ging nur bis 5% unten). Jetzt nehmen wir wieder die ganzen Daten und versuchen die beste Gerade zu finden. Wir sind uns einig das die unterste Gerade hier z.B. nicht gut den Zusammenhang beschreibt. Wie finden wir also die Beste oder in anderen Worten woher kommt das Alpha und das Beta Dach?

(Ab 41:40 wird das alles erklärt) Wir können für jeden möglichen Wert von Alpha und Beta eine Gerade haben. Das heisst wir können für jeden möglichen Wert von Alpha und Beta auch die Residuen berechnen. Weil wir eine Vorhersage machen können zu Yi Dach, wir können den Unterschied zum tatsächlichen beobachteten Wert nehmen. Was wir nun machen ist wir summieren die quadrierten Abweichungen und wir suchen diejenigen Werte von Alpha und Beta die diese Summe minimieren. Also die Linie die am Ende des Tages die geringsten quadrierten Abweichungen zu den tatsächlichen Daten ermöglicht. Das ist unsere Regressionsgerade. Das macht weil die Datensätze sehr gross sind R für uns. Wir nennen die Summe der quadrierten Residuen **SSR** (sum of square residuals) Unser Ziel ist es das zu minimieren. Wieso quadrieren wir? Es kommt ja nicht darauf an ob die Veränderung nach unten oder nach oben ist sondern einfach wie gross der ist. Wir nehmen nicht den absoluten Wert, weil es keine eindeutige Lösung gibt. Absolute Distanz muss nicht zwingend eine Lösung geben.

- Der häufigste Ansatz, um die Parameter der linearen Regression zu schätzen ist die **ordinary least squares (OLS)** Methode.
- Der OLS Schätzer bestimmt die Gerade so, dass sie am besten zu den Daten passt
- Formal kann man das wie folgt beschreiben: Von allen möglichen Werten α und β , werden die Werte $\hat{\alpha}$ und $\hat{\beta}$ resultieren, die die **Summe der quadrierten Residuen** minimiert.
Die Linie welche die kleinste Abweichung zu den tatsächlichen Werten hat ist die Regressionsgerade

$$SSR = \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}X_i)]^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

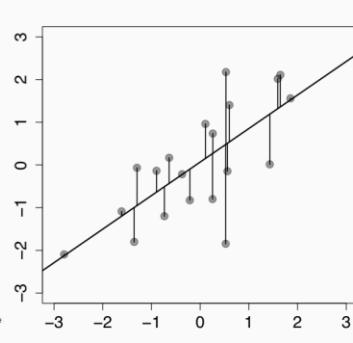
- OLS führt also zu einer Geraden, die den Unterschied zwischen beobachteten Werten (Y_i) und vorhergesagten Werten (\hat{Y}_i) für alle Beobachtungen minimiert



- Man nimmt einen Datensatz
- Zeichnet eine Gerade durch die Punkte
- Für diese Gerade kann man die Summe der quadrierten Distanzen von Y_i und \hat{Y}_i berechnen:

$$\sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}X_i)]^2 = 30.54$$

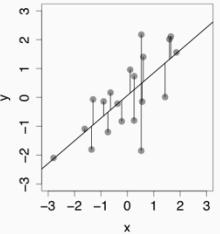
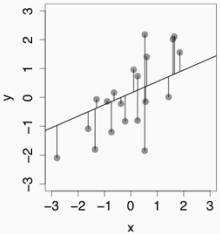
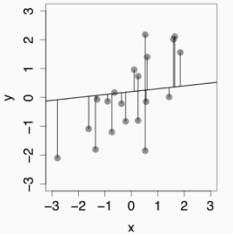
Welche Alphas und Betas müssen rein damit die Zahl am Schluss möglichst klein ist



- Man nimmt einen Datensatz
- Zeichnet eine Gerade durch die Punkte
- Für diese Gerade kann man die Summe der quadrierten Distanzen von Y_i und \hat{Y}_i berechnen:

$$\sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}X_i)]^2 = 16.95$$

Man sieht hier an den Beispielen nur schon visuell. Die gerade links ist keine gute Beschreibung der Daten. Rechts hingegen sind wir am «sweet spot», weil diese Gerade den Besten fit zu den Daten herstellt. Man versucht also Alphas und Betas zu finden um die **Zahl** möglichst klein zu machen.



In einem einfachen Modell mit nur einer unabhängigen Variable, können wir den OLS Schätzer einfach zeigen:

$$\hat{\beta} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2} = \frac{\text{Kovarianz}}{\text{Varianz}}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

wobei \bar{X} und \bar{Y} die Stichprobenmittelwerte von X und Y sind.

→ OLS führt zu einer Regressiongeraden, welche die Summe der quadrierten Distanzen zwischen Gerade und jedem Punkt minimiert

So sähe das Ganze noch formal aus, einfach das man sieht was dahinter steht aber wir müssen das nicht von Hand können. Das macht R für uns und sieht dann so aus:

Das geht ganz einfach:

```
simple_ols_model <- lm(voters_change ~ students,
                         data = constituencies)
```

Zuerst innen der **Brücke** \sim
also abhängige Variable, also Y

```
## Call:
## lm(formula = voters_change ~ students, data = constituencies)
##
## Coefficients:
## (Intercept)      students
##           205.1        -445.0
##             ^          ^  
             |          |
##             Gest. Konstante  Gest. Steigung
```

wobei $(\text{Intercept}) = \alpha$ und $\text{students} = \beta$

Visualisierung und Interpretation mit den Daten aus dem R output, also wenn man Alpha Dach und Beta Dach herausgefunden hat

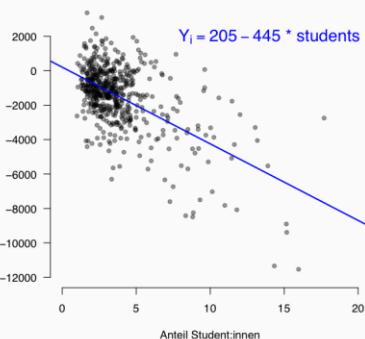
Der geschätzte Zusammenhang zwischen Anteil Student:innen und Veränderung

in der Anzahl registrierter Wähler:innen ist
Mit den Daten aus unserem R output:

$$\widehat{Voters}_i = 205 - 445 \times Students_i$$

$\widehat{\alpha}$ $\widehat{\beta}$

- *Voters* ist die Veränderung in der Zahl der registrierten Wähler:innen
- *Students* ist der Anteil Student:innen



Was ist die Interpretation von $\widehat{\beta} = -445$?

- **Generisch:** Ein Anstieg in X um eine Einheit geht im Durchschnitt einher mit einem Anstieg von Y um $\widehat{\beta}$.
- **Spezifisch:** Einen Anstieg von einem Prozentpunkt im Anteil der Student:innen in einem Wahlkreis geht im Durchschnitt einher mit einer Abnahme von -445 registrierten Wähler:innen in einem Wahlkreis.

Man kann nun basiert auf diesem Modell Vorhersagen treffen um die Resultate besser verständlich zu machen:

Wir können die vorhergesagten Werte für *irgendeinen X* Wert berechnen

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

- Was ist die erwartete Veränderung in der Zahl der registrierten Wähler:innen, wenn ein Wahlkreis zu 10% aus Student:innen bestehen würde?

$$\hat{Y}_i = 205 - 445 * 10 = -4245$$

- Was ist die erwartete Veränderung in der Zahl der registrierten Wähler:innen, wenn ein Wahlkreis zu 20% aus Student:innen bestehen würde?

$$\hat{Y}_i = 205 - 445 * 20 = -8695$$

Man kann das einfach in R umsetzen:

```
predict(simple_ols_model, newdata = data.frame(students = 10))
```

Vorhersage für 10%

```
##          1
## -4244.566
```

Vorhersage für 20%

```
predict(simple_ols_model, newdata = data.frame(students = 20))
```

```
##          1
## -8694.281
```

- `predict` ist der Befehl für Vorhersagen (basierend auf einem Modell)
- Das Argument `newdata` erlaubt es uns spezifische Werte vorzugeben für welche wir gerne einen Vorhersage berechnen möchten

In der Realität machen wir das natürlich mit R (rechtes Bild)

Es gibt einen Zusammenhang zwischen der Regression und der Korrelation, Beta ist eigentlich nichts anderes als eine reskalierte Variante von der Korrelation von x und y.

Was ist der Zusammenhang zwischen dem Korrelationskoeffizienten, ρ und dem Regressionsparameter β ?

$$\hat{\beta} = \text{Korrelation von } X \text{ und } Y \times \frac{\text{Standardabweichung von } Y}{\text{Standardabweichung von } X}$$

Wichtigster Takeaway: Wenn Korrelation positiv (oder negativ) ist, sollte Beta das gleiche sein

Implikationen: *Geschätzte Variante von der Korrelation von x und y*

- Wenn die Korrelation positiv (negativ) ist, dann ist β ebenfalls positiv (negativ)
- Falls X um eine Standardabweichung steigt, dann steigt Y um ρ Standardabweichung

Bei der Regression haben wir zusätzlich eine Realweltliche Grösse (Beta) das mir erlaubt Zu sagen im Schnitt pro zus. Prozentpunkt Student*in geht es um 445 reg. Wähler*innen runter. Diese Aussage schafft man nicht mit der Korrelation allein.

1. Die Regression ist das bessere Instrument, da man auch eine Aussage über die Magnitude machen kann

- Korrelation gibt Aufschluss darüber, ob X und Y positiv/negativ miteinander verbunden sind und sagt etwas über die "Stärke" dieses Zusammenhangs aus
- β gibt an um wie viele Einheiten sich Y ändert, wenn sich X um eine Einheit ändert

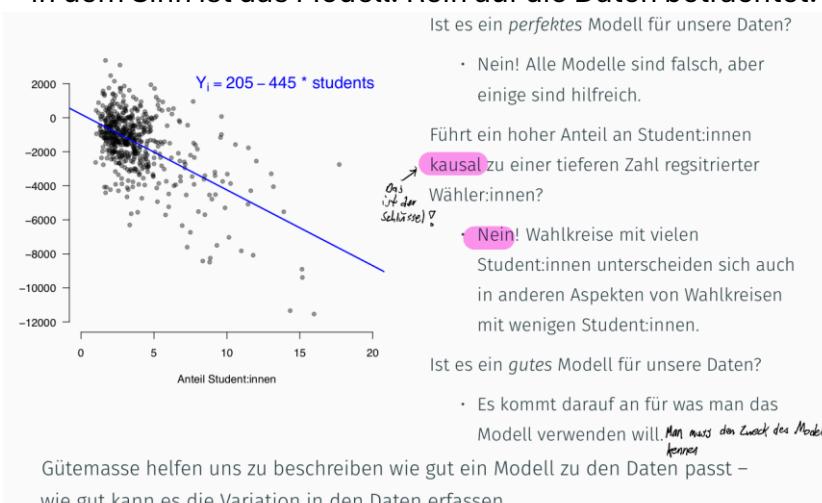
2. Regression ist flexibler

- Man ist nicht beschränkt auf Zusammenhänge von zwei Variablen – man kann mehrere unabhängige Variablen ins Modell einschliessen
- Man kann auch nicht-linearen Zusammenhänge untersuchen

Deshalb verwenden wir so viel mehr Zeit auf die Regression als auf die Korrelation.

Gütemass (Measure of Fit)

Wie gut ist ein bestimmtes Modell? Also der Fit von der Geraden mit den Punkten. Je näher die Punkte an diese Gerade herankommen, desto höher ist dieser Fit und desto «besser» in dem Sinn ist das Modell. Rein auf die Daten betrachtet.



Wie messen wir die Frage vom Ende dieser Folie? -> Mit R Quadrat

Die Frage ist wieviel der Varianz mein Modell auffängt. Wenn mein Modell einen riesengrossen Anteil der Varianz erfasst und erklären kann, dann weiss ich es gibt nicht mehr so viel andere Faktoren, die reinspielen. Das Mass hier ist immer zwischen 0 und 1. Wenn es perfekt ist wird das Resultat 1 sein, wenn es überhaupt keinen Zusammenhang gibt 0. Wenn wir den summary Befehl in R verwenden, wird das R Quadrat immer auftauchen. Was steht dahinter von wo kommt diese Zahl?

Wir können Y_i auch beschreiben als $Y_i = \text{Dach} + \epsilon_i$. Das ergibt sich aus dem, was wir vorher gemacht haben. Was wir dem Gegenüberstellen (also unserem Modell wo wir irgendeine Variable drin haben) ist das naivste oder dümmste, was man machen kann. Oder statistisch gesehen, aus einer datenwissenschaftlichen Perspektive, ein Modell das keine andere Information als Y hat, also es gibt kein X . Das ist das einfachste Modell. Wenn ich hierhinkomme und ich sehe, 150 Personen und es gibt nichts dass ich sie unterscheiden könnte und ich weiss von ihnen allen das Alter und dann stellen sie mir eine Person hin und fragen Herr Leemann was schätzen sie wie alt ist diese Person? Das beste, was man machen kann, ist der Durchschnitt von allen Personen nehmen und sagen das ist meine Schätzung, also in einer Welt wo die Personen nicht unterschieden werden können, das heisst auf statistische Modellsprache übersetzt es gibt kein X , also ich sehe weder gross noch klein, noch das Gesicht etc. -> Das ist genau das was das naive Modell macht, es nimmt einfach den Mittelwert von Y und macht für jede Person die genau gleiche Vorhersage, das erlaubt uns hier eine Varianz zu beschreiben die existiert, das ist die Bewegung die in Y drin ist, diese vergleiche jetzt mit der Bewegung die ich im Modell drin habe, ich kann ja mit OLS mein Alpha und mein Beta schätzen und darauf basierend die Unterschiede anschauen zwischen meinem Modell (das eben ein X verwendet wird) und dem naiven Modell. Also R Quadrat vergleicht einfach diese beiden Grössen, was ich sehen will ist das OLS Modell viel mehr erklären kann als ein naives Modell das einfach den Mittelwert reintut. Der Wert von R Qudrat ist immer zwischen 0 und 1 zwingend. Wenn der Wert 0 ist, dann sind die Fehler in meinem Modell genau so gross wie im naiven Modell, das heisst die X die ich ins Modell geschmissen habe, die haben genau nichts genutzt um bessere Vorhersagen zu generieren und wenn die X die ich verwende eine gute Wahl war, dann wird das SSR viel kleiner sein als das TSS, der Wert wird also steigen und näher an 1 sein

Das R^2 ist eine Statistik, die angibt wie viel besser die Modellvorhersagen eines Modells sind im Vergleich zu einem naiven Modell (also ein Modell ohne X).

Definition:

Das R^2 ist wie folgt definiert

$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

wobei

- TSS (Total sum of squares) ist das hier $\sum_{i=1}^n (Y_i - \bar{Y})^2$
- SSR (Sum squared residuals) ist das hier $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

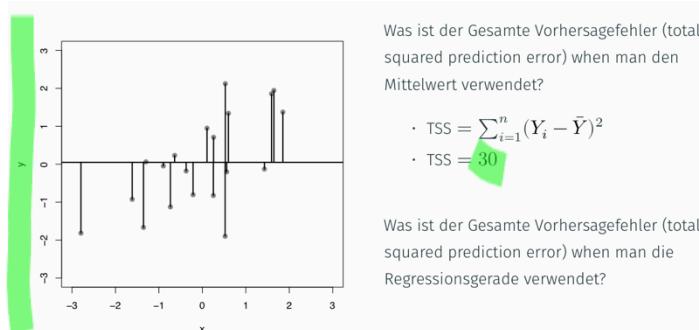
Das R^2 ist eine Statistik, die angibt wie viel besser die Modellvorhersagen eines Modells sind im Vergleich zu einem naiven Modell (also ein Modell ohne X).

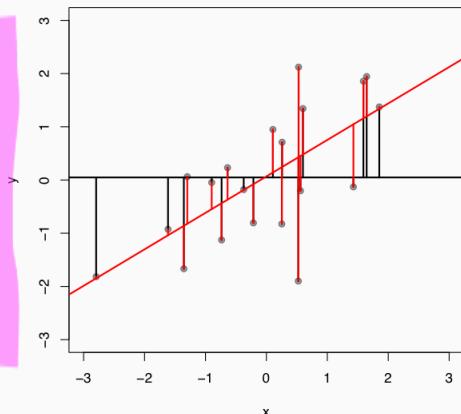
Intuition:

- R^2 ist zwischen 0 and 1
- Wenn die Residuen (Vorhersagefehler) unseres Modelles gross sind, dann ist R^2 näher bei 0
- Wenn die Residuen (Vorhersagefehler) unseres Modelles klein sind, dann ist R^2 näher bei 1

Hier noch formal und die Erklärung aus den Vorlesungsfolien

Hier das Ganze noch visualisiert:





Was ist der Gesamte Vorhersagefehler (total squared prediction error) wenn man den Mittelwert verwendet?

- $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$
- $TSS = 30$

Was ist der Gesamte Vorhersagefehler (total squared prediction error) wenn man die Regressionsgerade verwendet?

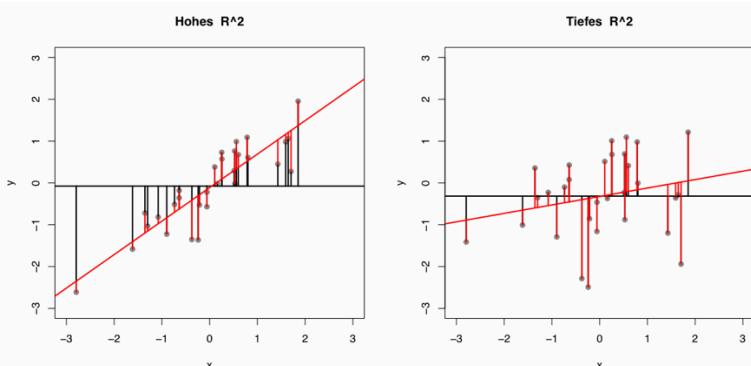
- $SSR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
- $SSR = 17$

$$R^2 = \frac{30 - 17}{30} = 0.44$$

Im oberen Bild (eine Seite vorher unten) sehen wir das naive Modell, also die Linie die raus kommt wenn wir unabhängig von X jeder Beobachtung den Mittelwert auf Y als vorhergesagten Wert geben. Wenn wir die Abweichungen berechnen zwischen naiven Modell und tatsächlich beobachteten Daten das wäre in diesem Beispiel 30.

Im unteren Bild (auf dieser Seite) sehen wir nun unsere geschätzte Gerade wenn ich X zur Hand habe (rote Linie) und nun sehen wir die Fehler werden kleiner, die roten Linien sind kürzer als die schwarzen Linien, weil wir gesamthaft näher an den Daten sind und wenn wir das nun berechnen also wieder den Unterschied zwischen diesem Modell und den tatsächlich beobachteten Daten, sehen wir es gibt 17. Der Abstand ist also gesamthaft kleiner als beim naiven Modell wo er 30 war. Mit diesen Zahlen können wir nun R Quadrat berechnen, gemäss der Formel und kommen auf 0.44. Das ist der Anteil der Varianz in Y die durch X erklärt werden kann.

Hier noch 2 Beispiele für ein hohes und ein tiefes R Quadrat



Was sagt uns das R^2 ?

- Hohe Werte → unabhängige Variable kann Y gut vorhersagen
- Tiefe Werte → unabhängige Variable kann Y schlecht vorhersagen

Was sagt uns das R^2 nicht?

- Hohe Werte für R^2 implizieren keine Kausalität
- Tiefe Werte für R^2 implizieren nicht, dass das Modell keinen Wert hat

Auf der linken Seite (im linken Bild) sehen wir das die Regressionsgerade viel näher an den Punkten zu liegen kommt, im linken Fall ist das X sehr gut geeignet um das Y vorherzusagen im rechten Fall ist dies weniger stark ausgeprägt. Deshalb werde ich links ein hohes R Quadrat finden und rechts ein tieferes R Quadrat. Hier wird also genau mit diesen Epsilons (Fehlerterm) gearbeitet (Was ist der Fehlerterm? Siehe Seite 14), ich generiere sie zweimal, einmal für mein Modell und einmal für ein naives Modell. Das hilft mir dann zu bestimmen, wie gut mein eigentliches Modell mit dem X ist.

```
## We can find out more detail about our estimated model using "summary"
summary(simple_ols_model)

...
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 205.15     119.46   1.717   0.0865 .
## students    -444.97     26.99  -16.489  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1525 on 571 degrees of freedom
## Multiple R-squared:  0.3226, Adjusted R-squared:  0.3214
## F-statistic: 271.9 on 1 and 571 DF,  p-value: < 2.2e-16
...
summary(simple_ols_model)$r.squared

## [1] 0.3225678
```

Ein Modell, welches den Anteil der Student:innen verwendet, kann bereit 32% der Variation in der Veränderung der registrierten Wähler:inne erklären

Nochmals auf das Beispiel bezogen mit den reg. Wähler*innen und Anzahl Student*innen, hier mit R output

Regression und Mittelwertsvergleich

Wir haben begonnen mit dem Hinweis, dass das lineare Modell uns erlaubt den Zusammenhang zwischen zwei kontinuierlichen Variablen zu untersuchen.

Das lineare Modell ist aber viel flexibler

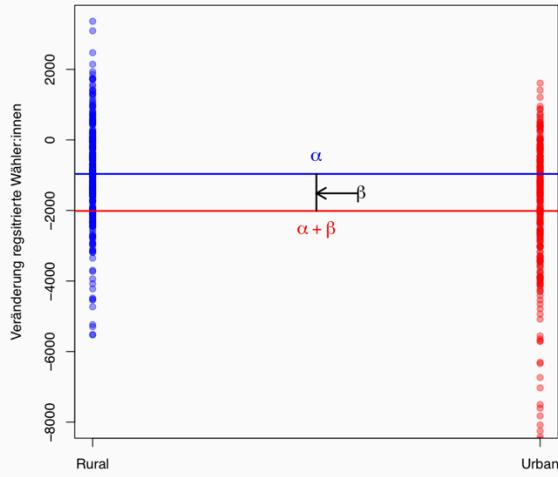
- Y sollte immer (annähernd) kontinuierlich sein
- X kann **jedes** Skalenniveau annehmen

Wenn X eine **binäre** Variable (auch: Dummy Variable) ist, dann ist $\hat{\beta}$ äquivalent zum Mittelwertsunterschied

Binäre Variable ("Dummy" Variable)

Dummy Variablen sind binäre Indikatoren, die = 1 sind, falls eine Beobachtung eine spezifische Eigenschaft aufweist und sonst = 0.

Beispiel: X_{male} , X_{labour} , X_{urban}



->Lineare Regression funktioniert auch mit binärer X Variable!

Hier ist ein lineares Regressionsmodell mit einer binären X Variable:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Wie kann man den Parameter α in diesem Modell interpretieren?

- α ist der Durchschnitt von Y wenn $X = 0$
- α ist der Durchschnitt der Veränderung der Registrierten für *ländliche* Wahlkreise

Wie kann man den Parameter β in diesem Modell interpretieren?

- β ist die durchschnittliche Veränderung in Y wenn X sich um eine Einheit erhöht
- Was ist eine Veränderung von einer Einheit in *urban*? Man geht von ländlich zu urban!
- β is also die durchschnittliche Veränderung in der Registrierung zwischen ländlichen und urbanen Wahlkreisen.

→ β ist das gleiche wie die Mittelwertsdifferenz!

```
urban_change
## [1] -2013.686
rural_change
## [1] -964.8212
urban_change - rural_change
## [1] -1048.865
urban_ols <- lm(voters_change ~ urban,
                 data = constituencies)
urban_ols
...
## Coefficients:
## (Intercept)      urban
##           -964.8          -1048.9
...
```

$\hat{\alpha}$ ist dasselbe wie *rural_change*

- Registrierung nahm im Durchschnitt um 965 ab in ländlichen Gegenden

$\hat{\beta}$ ist dasselbe wie *urban_change - rural_change*

- Registrierung nahm um 1049 mehr ab in urbanen Wahlkreisen

VL 4

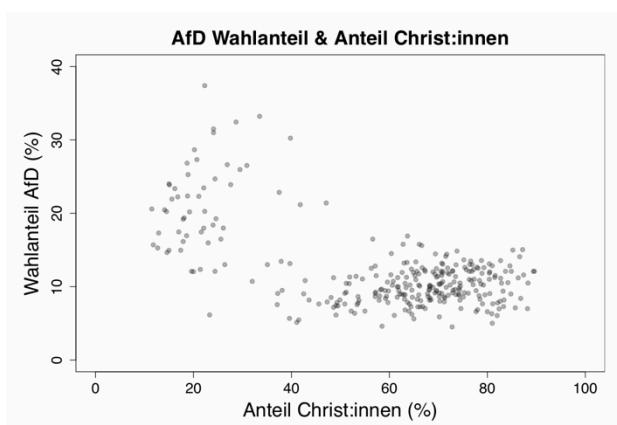
Grenzen der einfachen linearen Regression

Beispiel für diese Vorlesung:

Christentum und AfD Stimmenstärke

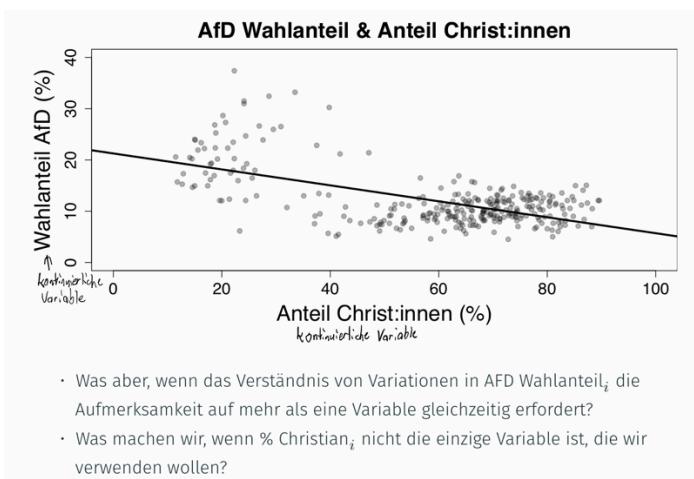
Bei der Bundestagswahl 2017 fiel vielen Beobachter:innen auf, dass die rechtsextreme Partei AfD in Gebieten mit weniger (selbstdeklarierten) christlichen Bürger:innen deutlich mehr Unterstützung erhielt. Wir werden den Zusammenhang zwischen dem "Christentum" einer Region und dem Stimmenanteil der AfD untersuchen, indem wir Daten zu den Wahlergebnissen von 299 Wahlkreisen sammeln.

- Analyseeinheit: 299 Wahlbezirke
- Abhängige Variable (Y_i): Wahlanteil der AfD im Wahlbezirk
- Unabhängige Variable (X_i): Anteil der Bevölkerung in einem Wahlbezirk, die sich als Christ:innen bezeichnen.

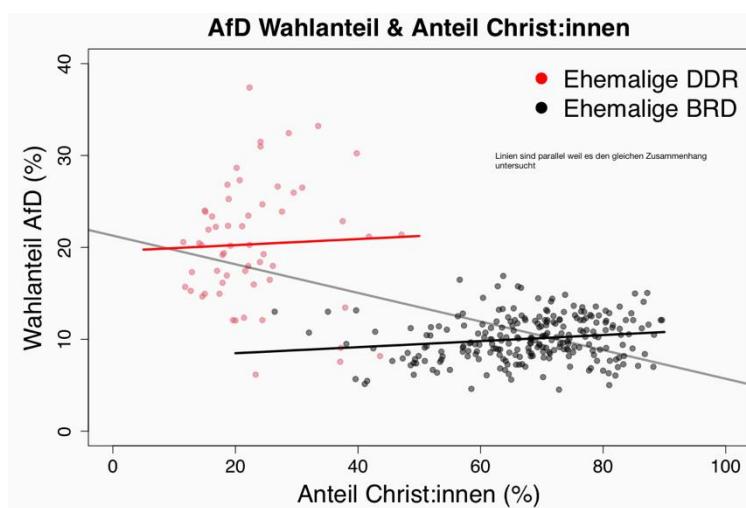


Man sieht es gibt irgendwie einen negativen Zusammenhang.

Wie wir diesen Zusammenhang darstellen können, haben wir mit der linearen Regression bereits gesehen. Wir haben zwei kontinuierliche Variablen und können mit der Regressionsgerade diese Daten analysieren. Das Problem ist aber das ganz viele Faktoren einen Einfluss haben können auf den Wahlanteil einer Partei. Das auf eine einzige Variable zu reduzieren ist oft sehr gefährlich, weil wir nicht in der Lage sind das Ganze Spektrum der möglichen erklärenden Faktoren zu erfassen und wir dann nicht die relevante Variation analysieren sondern die Gesamtvariation.



Multiple Lineare Regression



Die graue Linie ist die Regressionsgerade die wir bereits kennen, also wenn ich keinen Unterschied mache vom Wahlanteil der AfD ausser der Anteil Christ*innen, wir sehen das ist vielleicht nicht das Einzige das relevant ist, wir wollen den Ost-West Faktor ins Modell einschliessen(rote und schwarze Linie)

Wenn ich jetzt ein Modell mache zwischen Anteil Christi*innen und Wahlanteil AFD, und davon ausgehe, dass dieser konstant ist, aber einfach die grundsätzlichen Niveaus sich unterscheiden im Osten und Westen, dann haben wir ein Resultat mit der roten und schwarzen Linie wie wir es oben sehen. Die beiden Linien sind parallel, weil sie den gleichen Zusammenhang zeigen. Also wenn man sagt was ist die erwartete Veränderung wenn der Anteil Christi*innen um 10% steigt, das ist dasselbe für ein Fall im Westen und im Osten, aber wir geben dem Modell eine Flexibilität das es unterschiedliche Niveaus haben kann. Wir sehen sofort das dies relevant ist, wir hatten bis jetzt einen negativen Zusammenhang, geben wir dem Modell die Flexibilität, jetzt ist der Zusammenhang positiv. Ost-West korreliert also nicht nur mit dem outcome sondern korreliert auch mit der anderen erklärenden Variable, ist das was dazu führt, das sich der Zusammenhang im Vergleich zum einfachen Modell ändert. Was wir eigentlich gemacht haben, wenn wir zwei Geraden reinzeichnen, ist das wir den Zusammenhang zwischen Anteil Christi*innen und Wahlanteil der AFD innerhalb dieser beiden Gruppen geschätzt haben, also fast als hätte man den Datensatz getrennt und man hätte es einmal auf Osten und einmal auf Westen laufen lassen. Das ist natürlich nicht genau was wir gemacht haben, weil die Geraden ja parallel sind, aber was wir ausnützen ist die Variation die wir innerhalb dieser zwei Gruppen finden um diesen Zusammenhang zu schätzen.

- Die Multiple Regression ermöglicht die Beschreibung von Variation in einer **abhängigen Variablen** unter Verwendung von mehr als einer **unabhängigen Variable** gleichzeitig

- Y : AfD Wahlanteil
- X_1 : Anteil Christ:innen
- X_2 : Ehemaliger Osten *Dummy Variable : 1 = Osten
0 = Westen*

Generell sind wir aber nicht beschränkt darauf das wir nur 2 Variablen reinnehmen können. Dann erhalten wir am Schluss einfach für jede Variable die wir ins Modell einschließen einen anderen Koeffizienten. Diese Koeffizienten sind ja eine Art Übersetzung, wenn etwas in X sich verändert was geschieht nachher auf Y, das wird gesteuert durch das jeweilige Beta, deshalb kriegt jede Variable ihr eigenes Beta, wir haben weiterhin einen Fehlerterm, weil das ist ja einfach der Unterschied zwischen der Modellvorhersage für eine bestimmte Beobachtung und den tatsächlichen Wert den wir messen. Das sieht dann so aus

Das Multiple Regressionsmodell kann man wie folgt schreiben:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

- **Beobachtungen** $i = 1, \dots, n$
- Y ist die **abhängige Variable**
- X_1, \dots, X_k sind k **unabhängige Variablen**
- α ist der **Achsenabschnitt** oder die **Konstante**
- β_1, \dots, β_k sind die **Koeffizienten**
- ε_i ist der **Fehlerterm**

Wie interpretieren wir nun Beta? Letzte Vorlesung haben wir gelernt Beta zeigt uns an wie wird sich Y verändern wenn X um eine Einheit ansteigt. Jetzt haben wir hier verschiedene Variablen drin, man interpretiert eigentlich genau gleich es braucht einfach noch einen Zusatz. Was ist die erwartete Veränderung im Schnitt auf Y wenn X sich um eine Einheit verändert UND alle anderen Variablen konstant gehalten werden. Das einzige was sich also verändert ist ein einziges X im Modell, dann gibt uns dieses Beta die erwartete Veränderung in Y an. Das sieht dann so aus:

- Jeder Koeffizient β beschreibt den Zusammenhang in den Daten zwischen dieser unabhängigen Variable und der abhängigen Variable, während man für die anderen unabhängigen Variablen **kontrolliert**

- Die Koeffizienten zeigen deshalb **partielle Zusammenhänge**

- Nehmen wir ein Modell mit drei erklärenden Variablen:

$$Y_i = (\underbrace{\alpha + \beta_1 X_1 + \beta_2 X_2}_{\text{Bei } X_3 \text{ festgehalten}}) + \beta_3 X_3 = (\text{Others}) + \beta_3 X_3$$

- Hier bezeichnet "(Others)" der Teil des Modells, welcher von X_1 und X_2 abhängt aber nicht X_3

- Wenn X_3 um eine Einheit ansteigt, und X_1 und X_2 konstant bleiben, ist die erwartete (d.h. *durchschnittliche*) Veränderung von Y nach dem Modell genau β_3

Für unser Modell das wir schätzen wollen heisst das nun folgendes.

- **Abhängige Variable (Y):** AfD Wahlanteil (kontinuierlich)
- 1. **Unabhängige Variable (X_1):** Anteil Christ:innen (kontinuierlich)
- 2. **Unabhängige Variable (X_2):** Ost-West (binär, Ost = 1)

$$\text{AfD Share}_i = \alpha + \beta_1 \cdot \text{Christian Share}_i + \beta_2 \cdot \text{East}_i$$

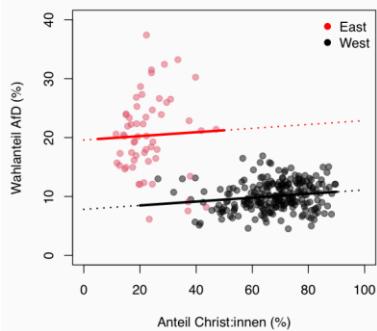
Ein Anstieg von einer Einheit bei **Christian Share** führt in unserem Modell zu einer Veränderung von β_1 in der abhängigen Variable **AfD Share im Durchschnitt**, wenn man die Ost-West Variable **konstant hält**

Östliche Wahlkreise haben einen um β_2 höheren Wert **im Durchschnitt** auf **AfD Share** relativ zu westlichen Wahlkreisen, wenn man den **Christian Share** **konstant hält**.

Interpretation von Beta1: Zeigt die erwartete Veränderung an, wenn der Anteil Christ*innen um eine Einheit steigt und ich ich die andere Variable (also Ost-West) konstant halte. Also ich befinde mich dann auf einer dieser beiden Geraden der roten oder der schwarzen die wir vorhin gesehen haben.

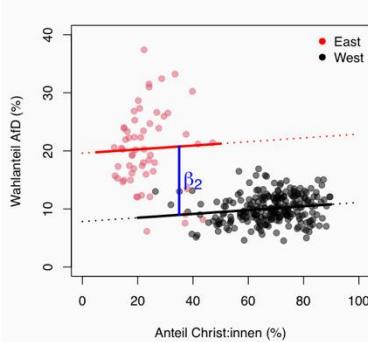
Interpretation von Beta2: Das ist der Unterschied um eine Einheit, wenn ich also von einem westlichen Distrikt in einen östlichen wechsle aber ich halte den Anteil Christ*innen konstant, das ist auch die Differenz die wir gesehen haben zwischen dieser roten und schwarzen Geraden, was wir vielleicht als Niveauunterschied bezeichnen wollen, also was sich generell unterscheidet zwischen dem Osten und dem Westen. Diese beiden Interpretationen sehen grafisch dann so aus:

AfD Wahlanteil & Anteil Christ:innen



- Wenn wir die East Variable 'konstant halten'
- Beschreibt β_1 die Steigung der beiden Linien innerhalb der östlichen und westlichen Wahlkreise

AfD Wahlanteil & Anteil Christ:innen



- Wenn wir nun die Variable Christian 'konstant halten'
- denn beschreibt β_2 die Distanz zwischen den beiden Linien

Wie schätzt man eine multiple lineare Regression in R?

Ursprüngliches Modell mit einer erklärenden Variable

```
linear_model_1 <- lm(AfD ~ christian, data = results)
```

Neues Modell mit zwei erklärenden Variablen

```
linear_model_2 <- lm(AfD ~ christian + east, data = results)
```

- Einfach mit Plus weitere Variablen hinzufügen!
→ Und hier noch die Interpretation der Werte für das Beispiel:

	AfD
christian	0.033 (0.019)
east	11.767* (0.994)
Constant	7.825* (1.300)
Observations	299
R ²	0.564

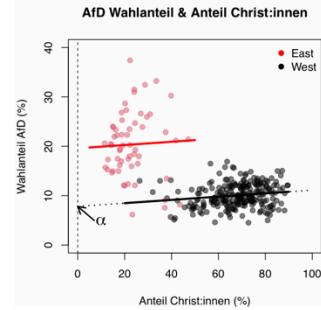
Note: *p<0.05

- Ein Anstieg des Anteils der Christ:innen um einen Prozentpunkt ist im Durchschnitt mit einem Anstieg des Stimmenanteils der AfD um 0.03 Prozentpunkte verbunden, wenn wir die Ost-West-Lage konstant halten
- Wahldistrikte im ehemaligen Osten haben im Durchschnitt einen AfD Wahlanteil, der um 11.8 Prozentpunkte höher ist, wenn wir den Anteil Christ:innen konstant halten

Technisch gesprochen geht die Interpretation von alpha analaog zum linearen Regressionsmodell, die Modellkonstante ist der erwartete Wert in y, wenn alle erklärenden Faktoren gleich 0 sind. Wieso technisch gesprochen? Weil man kann sich zwar vorstellen das der Ost West Dummy 0 ist, das sind dann alle westlichen Distrikte, aber wenn wir die Grafik anschauen dann würden sehen wollen das wir tatsächlich auch Beobachtungen haben wo der Anteil der Christ*innen auch wirklich 0 ist. Das haben wir nicht. Beim tiefsten Punkt haben wir immer noch über 10% Prozentpunkte Christ*innen. Man kann es also technisch so sagen, aber es macht keinen Sinn das substanziell interpretieren zu wollen, weil das ist ausserhalb des Bereiches für den wir Daten haben und mit einer gewissen Zuverlässigkeit eine Aussage machen können.(extrapolieren) Desto mehr Variabeln man hat desto eher ist alpha irrelevant. Mit der Ausnahme von ein paar speziellen Fällen zum Beispiel wenn wir nur kategoriale Variablen haben (siehe später), sonst interpretieren wir alpha deshalb auch nicht und konzentrieren uns auf das Beta. Das macht auch Sinn, weil

wir machen das ganze ja nicht als eine Messübung sondern weil uns Zusammenhänge zwischen x und y interessieren und die werden in den Betas erfasst. Also was alpha ist können wir berechnen, aber es ist bedeutungslos da wir ja nicht wissen, ob dieser Prozess sich so darstellen würde hätten wir tatsächlich Bezirke wo es ein Anteil von genau 0% bekennenden Christi*innen hat. Die Konstante (constant) die im Modell rauskommt wenn wir es schätzen, ist also der Ort wo die Westgerade (weil diese ist ja $X = 0$) die y-Achse schneidet. Würde man diesen Punkt für die Ostgerade haben wolle, müsste man diese Konstante noch plus Beta2 rechnen, weil das X dann ja nicht 0 wäre bei einem Ostbezirk.

- Letzte Woche haben wir gesehen, dass – technisch gesprochen – man $\hat{\alpha}$ als den Wert von Y interpretieren kann, wenn $X = 0$
- Nun haben wir aber mehrere X: $\hat{\alpha}$ repräsentiert nun den durchschnittlichen Wert von Y wenn alle X Variablen gleich 0 sind (technisch gesprochen)
- Je mehr Variablen man in einem Modell hat, desto eher verliert $\hat{\alpha}$ eine relevante substantielle Bedeutung



- Hier ist X_2 die Ost/West Variable und ist gleich 1 für Wahlbezirke im Osten und gleich 0 für Wahlbezirke im Westen
- $\hat{\alpha}$ ist also der Punkt in dem die schwarze Regressionsgerade die Y-Achse schneidet

Man kann auch noch weitere Variablen ins Modell einschliessen. Das wichtigste Thema für die AfD ist das Thema Migration. Wir könnten uns also fragen ob der Anteil Migrant*innen in einem Bezirk auch mit dem Stimmenanteil der AfD zusammenhängt.

- Man könnte sich fragen, ob der Anteil der Migrant:innen in einem Bezirk auch mit dem Stimmenanteil der AfD zusammenhängt
- Wenn man diese Variable ins Modell aufnimmt, dann haben wir folgendes Modell:

$$\text{AfD Share}_i = \alpha + \beta_1 \cdot \text{Christian}_i + \beta_2 \cdot \text{East}_i + \beta_3 \cdot \text{Migrant}_i + \varepsilon_i$$

- ... und die Interpretation bleibt immer dieselbe:

- β_k ist die durchschnittliche Veränderung von Y, wenn sich X_k um eine Einheit erhöhte und wir alle anderen Variablen konstant halten = *ceteris paribus*

Ceteris paribus = Wir halten alles andere konstant und fragen uns was ist die erwartete Veränderung in dieser Grösse, wenn diese eine Grösse sich verändert.

```
linear_model_3 <- lm(AfD ~ christian + east + migrantfraction, data = results)
library(texreg)
screenreg(list(linear_model_1, linear_model_2, linear_model_3), digits = 2)

##
## =====
##          Model 1    Model 2    Model 3
## -----
## (Intercept) 21.29 ***   7.82 ***  11.78 ***
##             (0.76)    (1.30)    (1.90)
## christian -0.16 ***   0.03      0.00
##             (0.01)    (0.02)    (0.02)
## eastTRUE        11.77 ***   9.14 ***
##                 (0.99)    (1.35)
## migrantfraction           -0.09 **
##                             (0.03)
## -----
## R^2            0.36      0.56      0.58
## Adj. R^2       0.35      0.56      0.57
## Num. obs.     299       299       299
## =====
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Hinweis: mit screenreg (in texreg library) kann man die Resultate sehr gut in R darstellen, mit «list» sogar noch eine Tabelle wo wir alle Modelle hinzufügen können.

Achtung: Wir beschreiben hier nur Korrelationen, also Zusammenhänge zwischen Variablen (also wie beschreiben einen Datenzusammenhang), diese müssen überhaupt nicht kausal sein. Wie wir dann abschätzen können ob diese kausal sind lernen wir in einer späteren Vorlesung in diesem Semester. Das ist dann die harte empirische Währung.

Frage aus der Vorlesung: Wie würde so ein Plot aussehen? -> Mit 3 Variablen wäre das ein vierdimensionaler Raum (diesen können wir nicht zeigen), mit zwei Variablen ein dreidimensionaler Raum in den wir eine Ebene zeichnen.

Die Vorhersage der Werte funktioniert genau gleich wie beim einfachen Modell. Wir können also basierend auf unseren X im Modell eine Vorhersage machen für eine Beobachtung.

Wie beim einfachen Regressionsmodell kann mit dem Modell auch vorhergesagte Werte generieren:

- Die vorhergesagten Werte (\hat{Y}) sind:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$$

- Interpretation:** Die vorhergesagten Werte sind Ihre beste Schätzung für Y für spezifische Werte von X_1, X_2, \dots, X_k

Die Ergebnisse in diesem rechten Bild sind
Also die modellbasierten Vorhersagen für
diese spezifisch beschriebenen Wahlbezirke.

$$\begin{aligned}\widehat{\text{AfD}}_i &= \hat{\alpha} + \hat{\beta}_1 \text{Christian}_i + \hat{\beta}_2 \text{East}_i + \hat{\beta}_3 \text{Migrant}_i \\ \widehat{\text{AfD}}_i &= 11.78 + 0.004 \cdot \text{Christian}_i + 9.14 \cdot \text{East}_i - 0.09 \cdot \text{Migrant}_i\end{aligned}$$

Frage: Was ist der vorhergesagte AfD Wahlanteil in einem Wahlbezirk im Osten, mit 40% Christ:innen und 5% Migrant:innen?

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot 40 + 9.14 \cdot 1 - 0.09 \cdot 5 = 20.63$$

Frage: Was ist der vorhergesagte AfD Wahlanteil in einem Wahlbezirk im Westen, mit 20% Christ:innen und 15% Migrant:innen?

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot 20 + 9.14 \cdot 0 - 0.09 \cdot 15 = 10.47$$

Multiple Regression mit kategorialen Variablen

Kategoriale Variablen sind eine Möglichkeit für uns qualitative Informationen ins Modell zu geben. Weshalb? Weil Ost und West steht ja nicht einer hierarchischen Beziehung das ist nicht eine ordinale Variable, nicht eine metrisch skalierte Variable, das ist eine reine Kategorisierung andere Beispiele sind Farben, Haarfarben, Familiennamen. Diese kann man nicht in eine Ordnung stellen, sondern nur sagen ich habe so viele Leute die heißen Müller, so viele Leute die heißen Milic, mehr kann man nicht machen. Diese Informationen kann man ins Modell einfügen, haben wir bereits gemacht mit Ost und West aber manchmal ist es vielleicht nicht nur binär sondern es sind 6,7 oder was weiß ich wie viele Kategorien, wenn es zum Beispiel Farben sind, also stellt sich die Frage, wie kann ich das in ein solches Modell reinbringen.

- Hier ein Beispiel von Dummy Variablen mit vier Regionen: Hamburg, Bayern, Berlin, Brandenburg

- The reference category is Brandenburg:

	X_{Hamburg}	X_{Bayern}	X_{Berlin}
Bayern	0	1	0
Bayern	0	1	0
Berlin	0	0	1
Hamburg	1	0	0
Brandenburg	0	0	0
Bayern	0	1	0
:	:	:	:

- R wird automatisch eine `factor variable` zu einer Gruppe von Dummy Variablen umcodieren und eine Kategorie als Referenzkategorie setzen

Das gleiche haben wir vorhin eigentlich für Ost und West gemacht einfach nur mit 2 Variablen. West war dabei die Referenzkategorie.

Wir können nun also für jede Dummy Variable, also für jedes Bundesland in diesem Beispiel den Koeffizientenschätzer ausgeben lassen.

Wir können die Variable `region` (die das Bundesland angibt) dafür verwenden:

```
## 
## BB BE BW BY HB HE HH MV NI NW RP SH SL SN ST TH
## 10 12 38 46 2 22 6 6 30 64 15 11 4 16 9 8
```

Man sieht nun direkt wie viele Wahlbezirke in jedem Bundesland existieren.

Wir können nun auch gleich ein Modell damit schätzen:

```
linear_model_4 <- lm(AfD ~ christian + region ,
                      data = results)
```

```
##             Estimate Std. Error   t value   Pr(>|t|)    
## (Intercept) 19.38519398 0.98617499 19.6569516 8.906315e-55
## christian    0.01948636 0.01956459  0.9960014 3.201033e-01
## regionBE    -7.88621573 1.23122444 -6.4051813 6.268269e-10
## regionBW    -9.19481347 1.39715459 -6.5810996 2.269137e-10
## regionBY    -9.99382116 1.46489653 -6.8222028 5.466454e-11
## regionHB   -10.76355393 2.29222915 -4.6956710 4.154088e-06
## regionHE   -9.78162918 1.38528012 -7.0615207 1.286179e-11
## regionHH   -11.89533599 1.52194233 -7.8158914 1.089623e-13
## regionMV   -1.60037325 1.47391164 -1.0858000 2.784948e-01
## regionNI   -11.90378488 1.36936299 -8.6929360 2.951986e-16
## regionNW   -11.25031787 1.34613899 -8.3575960 2.953719e-15
## regionRP   -10.56147953 1.58461659 -6.6650963 1.388249e-10
## regionSH   -13.04309714 1.45043829 -8.9926453 3.612896e-17
## regionSL   -11.64505844 2.06676178 -5.6289992 4.374645e-08
## regionSN    6.40523554 1.15243488  5.5580021 6.321846e-08
## regionST   -1.26241031 1.31281745 -0.9616038 3.370724e-01
## regionTH   2.49520321 1.336963016  1.8218080 6.954336e-02
```

Nun können wir ganz einfach Informationen für die verschiedenen Bundesländern aus dem Modell ablesen. Wir sehen das der AFD Wahlanteil in Berlin (BE) im Durchschnitt 7.9 Prozentpunkte tiefer liegt als in Brandenburg, wenn die Anzahl der Christ*innen gleich ist. In R wird die Variable oft schon so sein (also as factor) und sonst hat man die Möglichkeit den Befehl `as.factor` zu verwenden. Siehe Tutorat.

- In unserem Datensatz ist `region` eine kategoriale Variable, da sie als `factor variable` gespeichert ist

• Die erste Kategorie, Brandenburg (BB), wird dann zur Referenzkategorie

Multiple Regression mit Interaktionen

Extrem wichtiges Thema weil das da ist wo unser sozialwissenschaftliches Interesse an Grenzen stossen kann und womit viele Leute die das nicht vernünftig gelernt haben, später hadern. Ganz viele wissenschaftliche Erklärungen sind eben nicht nur Zusammenhänge mit *ceteris paribus* sondern es ist kontextualisiert, das ist gruppenspezifisch. Zum Beispiel dieser Effekt ist stärker bei Leuten mit höherer Bildung oder der Effekt dieser Intervention funktioniert primär bei Leuten mit knappen finanziellen Mitteln, diese Aussagen sind konditional. Wir brauchen also modellierungstechnisch eine Möglichkeit realistische Zusammenhänge zu modellieren. Das bedeutet eben das nicht alles andere konstant bleiben sondern das gewisse Sachen abhängen von etwas anderem die Art und Weise wie wir das in so ein Modell reinkriegen ist mit Interaktionen.

Das bedeutet das ich zum Beispiel die Daten anschauen und das Gefühl habe eine gute Beschreibung dieser Daten ist eben nicht das es einen generellen Zusammenhang zwischen x und y gibt (Also Wahlanteil AFD und Anzahl Christ*innen) sondern das dieser anders ist in den alten Bundesländern als in den neuen. Das heisst ich will jetzt weg von diesen parallelen Linien. Ich will sozusagen zwei verschiedene Regressionslinien haben, eine die gut zu Ostdeutschland passt und eine die gut zu Westdeutschland passt.

Interaktionen

Eine **Interaktion** besteht zwischen zwei erklärenden Variablen, wenn die Beziehung zwischen (einer) von ihnen und der abhängigen Variable vom Wert der anderen abhängt

- Wir können diese Intuition in das lineare Regressionsmodell einbauen, indem wir das **Produkt** von zwei erklärenden Variablen in unser Modell aufnehmen

- Das einfache Modell bisher hat ‘konstante Zusammenhänge’ (z.B. die Beziehung zwischen X_1 und Y hängt nicht von anderen X ’en ab)

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- Wir können die Annahme einer konstanten Assoziation lockern, indem wir das Produkt der erklärenden Variablen zu einem Modell hinzufügen:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \cdot X_{2i} + \varepsilon_i$$

- In unserem Beispiel wäre das nun das neue Modell:

$$\text{AfD} = \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} + \varepsilon$$

Wir müssen also dem Modell die Möglichkeit geben, Linien zu schätzen die nicht unbedingt parallel sind. Wir machen das in dem wir einen Interaktionsterm einschliessen.

```
m1 <- lm(AfD ~ migrantfraction, data = results)
m2 <- lm(AfD ~ migrantfraction + east, data = results) ohne Interaktion
m3 <- lm(AfD ~ migrantfraction * east, data = results) mit Interaktion
screenreg(list(m1,m2,m3))

##
## =====
## Model 1      Model 2      Model 3
## -----
## (Intercept)    18.36 ***   12.13 ***   9.69 ***
##                 (0.60)       (0.66)       (0.66)
## migrantfraction -0.34 ***  -0.10 ***   0.02
##                   (0.03)       (0.03)       (0.03)
## eastTRUE        8.91 ***  14.44 ***  -0.54 ***
##                   (0.67)       (0.88)       (0.06)
## migrantfraction:eastTRUE
##                   (0.06)
## -----
## R^2            0.32       0.58       0.66
## Adj. R^2       0.32       0.57       0.66
## Num. obs.      299        299        299
## =====
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Wir können uns noch anschauen was R im Hintergrund macht, das sieht dann so aus:

Wie ist der Zusammenhang zwischen `migrantfraction` und `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- Wie ist der Zusammenhang im Westen (i.e., `east=0`)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0 \\ \text{AfD} &= \frac{9.69}{\text{Intercept}} + \frac{0.02}{\text{Slope}} * \text{migrant} \end{aligned}$$

- Wie ist der Zusammenhang im Osten (i.e., `east=1`)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 1 + \beta_3 * \text{migrant} * 1 \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 + -0.54 * \text{migrant} \\ \text{AfD} &= 9.69 + 14.44 + (0.02 - 0.54) * \text{migrant} \\ \text{AfD} &= \frac{24.13}{\text{Intercept}} - \frac{(0.52)}{\text{Slope}} * \text{migrant} \end{aligned}$$

Wenn wir das in R eingeben müssen wir nur anstatt ein Plus ein Mal machen. Wir sehen dies in diesem Beispiel unter m3. Das Modell schätzt dann automatisch migrantische Population, Ost-West UND das Produkt der beiden.

East	Intercept	Slope
0 = west	α	β_1
	9.69	0.02
1 = east	$\alpha + \beta_2$	$\beta_1 + \beta_3$
	9.69 + 14.44 = 24.13	0.02 + -0.54 = -0.52

β_1

- Im **Westen** gilt, dass wenn es in einem Bezirk einen Prozentpunkt mehr Migrantinnen hat, der Wahlanteil der AfD im Durchschnitt um 0.02 Prozentpunkte steigt.

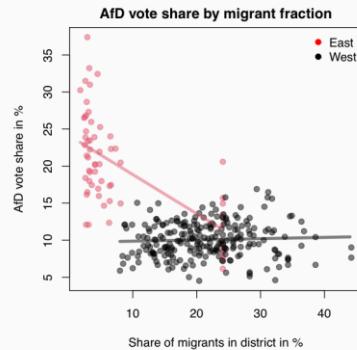
$\beta_1 + \beta_3$

- Im **Osten** gilt, dass wenn es in einem Bezirk einen Prozentpunkt mehr Migrantinnen hat, der Wahlanteil der AfD im Durchschnitt um 0.52 Prozentpunkte sinkt.

Die Interpretation des Ganzen sieht dann so aus:

	Model 1
(Intercept)	9.69 (0.66)
migrantfraction	0.02 (0.03)
east	14.44 (0.88)
migrantfraction:east	-0.54 (0.06)
Adj. R ²	0.66
Num. obs.	299

- β_1 (Koeffizient von migrantfraction) ist der durchschnittliche Zusammenhang zwischen Anteil Migrant:innen und Wahlanteil, falls die Dummy Variable Ost gleich 0 ist, d.h. in westlichen Bezirken
- β_2 (Koeffizient von east) ist der durchschnittliche Unterschied zwischen Osten und Westen, falls der Anteil der Migrant:innen 0% ist
- β_3 (Koeffizient der Interaktion) ist der durchschnittliche Unterschied im Zusammenhang zwischen migrantfraction und AfD zwischen Ost und West



- β_2 (Koeffizient von east) ist nicht der durchschnittliche Unterschied zwischen Ost und West
- β_1 (Koeffizient von migrantfraction) ist nicht der unkonditionale Zusammenhang zwischen Anteil Migrant:innen und AfD Wahlanteil
- Es gibt keinen unkonditionalen (oder: generellen) Zusammenhang mehr!



Multiples oder adjusted R²

Man kann das genau so berechnen mit einer multiplen linearen Regression. Das Problem ist allerdings das diese Zahl fast immer steigt, je mehr X man ins Modell einschliesst.

R^2 ist eine hilfreiche Statistik:

- Einfache Lineare Regression
 - R^2 = Anteil der Varianz von Y , die im Modell von X erklärt werden kann
- Multiple Lineare Regression
 - R^2 = Anteil der Varianz von Y , die im Modell von den Variablen X_1, \dots, X_k erklärt werden kann

Wir hätten aber gerne ein Gütemass welches Grössere Modelle nicht bevorzugt. Also ein Mass wo R² nicht steigt wenn wir mehr X ins Modell einschliessen. Die Lösung dafür ist das angepasste R² also das adjusted R². Wir ergänzen deshalb das R² mit dem Unterschied der hier rechts rot ist. Desto grösser die Anzahl der Erklärenden Variablen sein wird desto Kleiner wird das R². Deshalb ist das Ein Strafmass, k ist hier unter dem Bruch, wenn das kleiner wird wird der Rote Ausdruck grösser aber es ist Minus vorher das heisst wenn k steigt Sinkt das angepasste R². Zudem ist Es natürlich nicht stumpfsinnig also Konstant wie es steigt oder fällt Sondern dieses Mass ist natürlich

Auch beeinflusst von der Fallzahl, der Strafterm hier hat n drin, das heisst 4 X Variablen zu verwenden hat nicht Den gleichen Einfluss wenn ich nur

20 Beobachtungen habe als wenn ich 2000 Beobachtungen habe. Unglücklicherweise gibt es ein kleines Issue man kann dieses angepasste R² nicht mehr so schön interpretieren wie das normale, es hat keine realweltlichen Entsprechung, das wir irgendwie sagen können das ist der Anteil der erklärten Varianz, weil wir jetzt diesen Strafterm drin haben.

$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

- R^2 wird fast immer steigen, wenn man eine zusätzliche Variable X einfügt
- R^2 wird nie fallen, wenn man eine zusätzliche Variable X einfügt

Implikation:

- Man sollte nicht ein Modell nur wegen des höheren R^2 wählen
- Wir brauchen ein Mass, welches einen Strafterm für die Anzahl erklärender Variablen einführt

$$\text{adj. } R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{n - 1}{n - k - 1} \frac{SSR}{TSS}$$

wobei

- TSS (Total sum of squares) gleich $\sum_{i=1}^n (Y_i - \bar{Y})^2$ ist
- SSR (Sum squared residuals) gleich $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ ist
- k ist die Zahl der erklärenden Variablen im Modell
- n ist die Zahl der Beobachtungen

$$\text{adj. } R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{n - 1}{n - k - 1} \frac{SSR}{TSS}$$

Intuition

- adj. R^2 steigt nicht immer, wenn neue X ins Modell eingefügt werden
- adj. R^2 wird immer kleiner sein als das R^2
- Das adj. R^2 hat keine so kalre Interpretation wie das normale R^2

$$adj.R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS}$$

- Sie müssen diese Formeln nicht auswendig können, aber Sie sollten den Unterschied kennen (R berechnet diese Werte für Sie).

```
summary(m3)$r.squared
```

```
## [1] 0.6594207
```

```
summary(m3)$adj.r.squared
```

```
## [1] 0.6559571
```

} Model 3, Folie 40
letzte Woche

Wir sehen nun
ganz marginale
Unterschiede zwischen
 R^2 und $adj.R^2$

VL 5

Interaktion einer kontinuierlichen und einer Dummy Variable

Haben wir eigentlich schon letzte Vorlesung behandelt einfache kurze Repetition. Wir geben dem Modell noch mehr Flexibilität und sagen das der Zusammenhang zwischen Wahlanteil AFD und Anteil Christi*innen oder sonst einer erklärenden Variable, das der variieren kann vom Osten zum Westen. Also das es nicht zwingend dieselbe Gerade sein muss die das beschreibt. Das heisst wir versuchen einen konditionalen Zusammenhang herauszufinden. Also der Zusammenhang zwischen X1 und Y nicht einfach generell konstant ist, sondern von etwas anderem abhängt, brauchen wir die Möglichkeit das in einem Modell abzubilden und schätzen zu können und das machen wir mit Interaktionen. Wenn wir also glauben das der Zusammenhang zwischen X1 und Y nicht immer gleich ist sondern vom Wert von X2 abhängt, fügen wir wir das Produkt dieser beiden Variablen (die sogenannte Interaktion) ins Modell ein.

- Das einfache Modell bisher hat 'konstante Zusammenhänge' (z.B. die Beziehung zwischen X_1 und Y hängt nicht von anderen X 'en ab)

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- Wir können die Annahme einer konstanten Assoziation lockern, indem wir das Produkt der erklärenden Variablen zu einem Modell hinzufügen:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \cdot X_{2i} + \varepsilon_i$$

- In unserem Beispiel wäre das nun das neue Modell:

$$\text{AfD} = \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} + \varepsilon$$

```
m1 <- lm(AfD ~ migrantfraction, data = results)
m2 <- lm(AfD ~ migrantfraction + east, data = results)
m3 <- lm(AfD ~ migrantfraction * east, data = results) 3. Modell mit Interaktion
screenreg(list(m1,m2,m3))

##
## ===== Model 1 Model 2 Model 3 =====
## -----
## (Intercept) 18.36 *** (0.60) (0.66) (0.66)
## migrantfraction -0.34 *** (0.03) (0.03) (0.03)
## eastTRUE 8.91 *** (0.67) (0.88) (0.88)
## migrantfraction:eastTRUE -0.54 *** (0.06)
## -----
## R^2 0.32 0.58 0.66
## Adj. R^2 0.32 0.57 0.66
## Num. obs. 299 299 299
## =====
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

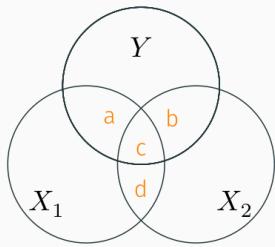
Wir sehen im rechten Bild wie sich die verschiedenen Modelle auf das Ergebnis auswirken. Wenn ich nicht mal zwischen Ost und West unterscheide finde ich einen klar negativen Zusammenhang zwischen Anteil Migrant*innen und Wahlanteil AFD. Wenn ich nun nur schon den Ost-West Dummy reinlege, wird die fallende Gerade schon viel flacher. Und im dritten Modell mit der Interaktion wo wir erlauben, dass die Art und Weise wie AFD Wahlanteil und Anteil Migrant*innen zusammenhängen, das soll sich nun unterscheiden dürfen zwischen Westen und Osten, wir erreichen das dadurch das wir Migrant*innen multiplizieren mit Ost-West und wir sehen den Koeffizientenschätzer für Beta3 ist -0.54. Das gibt uns den Unterschied an von der Steigung von Migrant*innen zu AFD im Westen zum Anstieg von Migrant*innen zu AFD im Osten. Es sagt uns also wie stark sich diese Geraden verändern.

Multiple lineare Regression – Ein Blick in die Mechanik

Wie kann man sich vorstellen, wie diese Koeffizienten geschätzt werden und was geschieht wenn unterschiedliche X Variablen die ich ins Modell gebe, ähnlich sind zueinander oder nicht und insbesondere der Spezialfall wenn ich eigentlich zweimal genau die gleiche Variable ins Modell gebe, respektive eine Variable eine perfekte lineare Transformation der anderen ist.

Wie werden also die Koeffizienten geschätzt? Wir haben bereits gesehen, dass wir in Y Varianz erklären und je mehr Varianz von Y durch X erklärt werden kann desto grösser ist das R2. Wir haben auch gesehen, wie man das berechnen kann. (mit diesen Abständen

welche die geschätzten Punkte von der Regressionsgerade haben.) Eine andere Art das zu machen ist das wir uns Varianz einer Variable vorstellen als Kreis (als Menge) und in diesem Fall haben wir eine Kovarianz (Schnittmenge von X und Y). Je grösser die Kovarianz ist, desto grösser ist R², weil ein grösserer Teil der Varianz wird erklärt durch dieses X. Wenn wir nun mehr als eine Variable ins Modell geben gibt es eine Kovarianz zwischen X₁ und Y, eine zwischen X₂ und Y und X₁ und X₂ die korrelieren auch. Was ist das Element das nun genutzt wird um die Koeffizienten zu schätzen? Es ist eben nicht so das alle Information in a und c relevant ist wenn ich den Koeffizienten für X₁ schätzen will oder das b und c relevant ist wenn ich das Beta für X₂ schätzen will.



- $a+c$ Kovarianz von X_1 und Y
- $b+c$ Kovarianz von X_2 und Y
- $d+c$ Kovarianz von X_1 und X_2

- Wir schätzen die Koeffizienten mit OLS (minimieren die Summe der quadrierten Residuen)
- Welche Kovariation ist relevant?
 - Für β_1 ist nur a relevant
 - Für β_2 ist nur b relevant

Bei c sind wir uns gar nicht so einig. Gehört das nun zu X₂ oder X₁ dazu, ist nicht so klar. Was am Ende relevant ist um eine Zahl bestimmen zu können für Beta1 und Beta2, stellt sich heraus ist dieses a und dieses b das wir hier haben. Das sind die beiden einzigen Elemente die einen Einfluss haben auf Beta1 und Beta2.

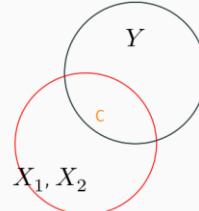
→ Der Anteil c wird nicht genutzt für die Schätzung des Koeffizienten.

Jetzt sehen wir vielleicht auf eine andere Art, wieso es ein Problem ist wenn X₁ und X₂ perfekt miteinander korrelieren, weil dann fehlt uns dieser unabhängige Teil, denn wir verwenden können, der eben nur zwischen Y und X₁ besteht, um Beta1 bestimmten zu können. Das ist was wir als perfekte Multikollinearität bezeichnen, wenn zwei erklärende Faktoren die ins lineare Regressionsmodell gegeben werden eine Korrelation aufweisen von plus oder minus 1, wichtig ist das es perfekt ist. (nicht 0.9 oder so) Wenn ich das nun grafisch darstelle überlappen sich das X₁ und X₂ perfekt. Das rote in diesem Bild sind zwei Kreise die perfekt übereinanderliegen. Es gibt

Also keinen Teil zwischen X₁ und Y der nur von den beiden abgedeckt wird sondern alle Varianz in X₁ existiert genau so in X₂. Die Konsequenz davon ist, dass wir das nicht schätzen können. Es bezieht sich auch nicht nur auf Dummy Variablen wenn wir beiden reintun, es kann auch mit Masseneinheiten zu tun haben, nehmen wir den Anteil Migrant*innen geben wir es ein als Anteil also eine Zahl zwischen 0 und 1 und dann geben sie die Variable nochmals ein aber in Prozentpunkten.

Die grösser c im Vergleich zu a ist, desto weniger Information steht zur Verfügung um β_1 zu schätzen

- Im Extremfall sind korrelieren X_1 und X_2 perfekt → perfekte Multikollinearität



- Konsequenz: Man kann β_1 und β_2 nicht schätzen
- Beispiel 1: Anteil Migrant:innen in %-Punkten und Anteil Migrant:innen
- Beispiel 2 (Dummy Variable Falle): East und West

Das ist die genau gleiche Information einfach in verschiedenen Einheiten, das ist natürlich eine lineare Transformation, das eine ist gleich 100* dem anderen. Das heisst wir werden wieder die Situation haben wo man nicht in der Lage ist zwei Koeffizienten zu schätzen sondern man hat maximal einen Koeffizienten denn man identifizieren kann. Wenn das

trotzdem passiert gibt uns R für das zweite X welches perfekt korreliert einfach NA zurück. Ist ein sehr häufiger Fehler auch wenn man z.B. die Bachelorarbeit macht.

- Perfekte Multikollinearität ist *fast immer* eine Konsequenz eines Fehlers der Analyst:in
- Glücklicherweise löst es R für und gleich

Interaktionen von kontinuierlichen Variablen

Nochmals: Wir haben teilweise konditionale Zusammenhänge und wir wollen, dann einen Zusammenhang zwischen Y und X1 modellieren das sich je nach Werten von X2 ändern kann. Wir hatten bis jetzt Beispiele wo X1 kontinuierlich ist und X2 eine Dummy Variable. Es gibt aber auch Fälle wo X2 auch eine kontinuierliche Variable ist. Wie würden wir das angehen? Eigentlich genau gleich wie bisher was sich ändert ist das Vorgehen wie wir die Resultate interpretieren. Jetzt können wir nicht mehr so einfach sagen was ist die Regression gerade wenn die Dummy Variable 0 oder 1 ist, weil wir ja jetzt eine andere konditionierende Variable haben die verschiedene Werte annehmen kann. Wir nehmen wieder das Gleiche Beispiel als X2 nehmen wir nun aber die Arbeitslosenrate in einem Wahlkreis.

- Wir können folgendes Modell mit zwei kontinuierlichen Variablen verwenden:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \cdot X_{2i} + \varepsilon_i$$

- In unserem Beispiel könnte das bspw. das folgende sein:

$$AfD_i = \alpha + \beta_1 * \text{migrant}_i + \beta_2 * \text{unemployment}_i + \dots \beta_3 * \text{migrant}_i * \text{unemployment}_i + \varepsilon_i$$

```
m1 <- lm(AFD ~ migrantfraction, data = results)
m2 <- lm(AFD ~ unemploymentrate, data = results)
m3 <- lm(AFD ~ migrantfraction + unemploymentrate, data = results)
m4 <- lm(AFD ~ migrantfraction * unemploymentrate, data = results)
screenreg(list(m1,m2,m3,m4))

##
## =====
##               Model 1   Model 2   Model 3   Model 4
##
## (Intercept) 18.36 ***  8.48 ***  14.96 ***  8.68 ***
##              (0.60)    (0.74)    (0.81)    (1.71)
## migrantfraction -0.34 ***  -0.33 ***  -0.01
##                      (0.03)    (0.03)    (0.08)
## unemploymentrate 0.58 ***  0.52 ***  1.38 ***
##                      (0.11)    (0.09)    (0.22)
## migrantfraction:unemploymentrate -0.04 ***
##                                     (0.01)
##
## R^2          0.32      0.09      0.39      0.42
## Adj. R^2     0.32      0.08      0.38      0.42
## Num. obs.    299       299       299       299
##
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Dieser Zusammenhang wäre konsistent wenn man ein Argument hat das es einen Zusammenhang geben sollte, zwischen Anteil migrantischer Population und AFD Wahlanteil, dieser Zusammenhang sich aber je nachdem ändert ob es mehr oder weniger Arbeitslosigkeit in einem Distrikt gibt. Wenn dies das substanziale Argument ist welches gemacht wird, dann wäre dies (das linke Bild) das adäquate Modell dafür. Wenn wir das in R durchlaufen lassen sieht das so aus wie im rechten Bild oben. Das Modell 4 ist das mit der Interaktion. Wie interpretieren wir nun diese Interaktion? Was kann ich nun sagen wie der Zusammenhang zwischen migrantischer Population und AFD-Wahlanteil variiert je nach Höhe der Arbeitslosigkeit.

Wir haben das schon einmal gemacht. Wir rechnen bei unserem X1 noch etwas dazu und dann schauen wir uns die Unterschiede an und wie sich unser Y verändert. Der erste Schritt ist also bei diesem X1 noch etwas dazuzurechnen und wenn ich das vorne mache muss ich dies hinten auch machen (überall wo X1 ist). Und bei Y wird sich eine Veränderung ergeben, das fangen wir ebenfalls mit diesem Delta Y auf. Jetzt muss ich im nächsten Schritt nichts anderes machen als mir die Differenz dieser beiden Gleichungen anschauen. Dann habe ich: Die Differenz von Y ist nichts anderes als Beta1 * DeltaX1 plus Beta3 * DeltaX1 * X2. Weshalb ist das genau so? Einfach mathematisch ganz simpel, wenn wir die beiden voneinander abziehen was übrig bleibt!

- Nun ist die Interpretation der Zusammenhänge etwas schwieriger, da der Zusammenhang (bspw. von X_1 und Y) vom Wert von X_2 abhängt
- Wir können aber wieder mit unserem Δ -Ansatz arbeiten:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2 \quad (1)$$

$$Y + \Delta Y = \alpha + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 + \beta_3 (X_1 + \Delta X_1) \cdot X_2 \quad (2)$$

- Wir nehmen nun wieder die Differenz von Gleichung (1) und (2):

$$\Delta Y = \beta_1 \Delta X_1 + \beta_3 \cdot \Delta X_1 \cdot X_2 \quad (3)$$

- Nehmen wir mal $\Delta X = 1$, dann gilt:

$$\Delta Y = \beta_1 + \beta_3 \cdot X_2 \quad (4)$$

- → Der Zusammenhang zwischen X_1 und Y hängt vom Wert für X_2 ab!

Jetzt sagen wir einfach Mal das dieses DeltaX1 (das Ding was ich dazu addiert habe) 1 ist. Also wie ein Anstieg um eine Einheit in X1. Dann bleibt das Delta Y = Beta1 + Beta3 * X2. Das heisst ich kann jetzt sagen die Veränderung vom Wahlanteil AFD wird Beta1 + Beta3 * X2 sein. Wieso ist das wichtig? Weil nun kann man zum eine Arbeitslosenrate angeben und ich kann das gleich einfügen (weil das ist ja X2). Beta1 und Beta3 haben wir weil wir es geschätzt haben, wir haben nun also Delta Y und schon können wir sagen auf dem Niveau von Arbeitslosigkeit wenn der Anteil migrantischer Population rauftgeht, erwarten wir im Durchschnitt diese Veränderung im AFD Wahlanteil.

→ Der Zusammenhang zwischen X1 und Y hängt vom Wert für X2 ab.

Um diese Veränderungen mit R festzustellen benutzen wir die predict Funktion. Das sieht dann so aus:

- Was geschieht, wenn sich der Anteil Migrant:innen um einen Prozentpunkt erhöht in einem Distrikt mit 2% Arbeitslosigkeit?

```
predict(m4, newdata = data.frame(migrantfraction = 5,unemploymentrate=2))
```

```
##      1
## 10.95201
predict(m4, newdata = data.frame(migrantfraction = 6,unemploymentrate=2))

##      1
## 10.85637
```

- Was geschieht, wenn sich der Anteil Migrant:innen um einen Prozentpunkt erhöht in einem Distrikt mit 10% Arbeitslosigkeit?

```
predict(m4, newdata = data.frame(migrantfraction = 5,unemploymentrate=10))
```

```
##      1
## 20.22534
predict(m4, newdata = data.frame(migrantfraction = 6,unemploymentrate=10))

##      1
## 19.78084
```

→ Der Unterschied liegt also in der Höhe der Arbeitslosenquote welche wie gesagt X2 entspricht!

Wir können uns das auch einfach in R visualisieren lassen.

Hinweis: In der zweiten Zeile dieses Befehls kommt es darauf an was zuerst kommt (ob migrantfraction oder unemploymentrate sonst sind sie vertauscht)

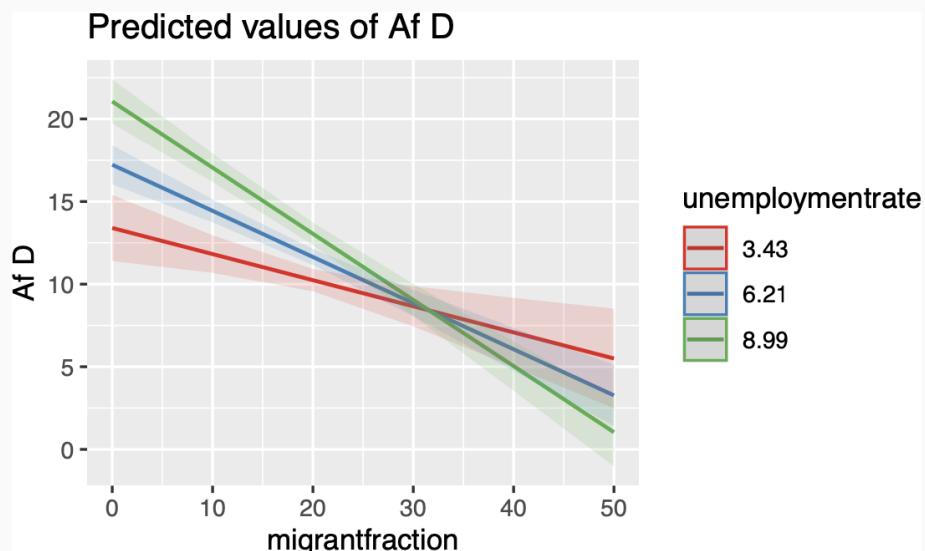
Wir sehen beim ersten Beispiel das der Wahlanteil für die AFD um knapp 0.1 runtergeht.

Wir sehen im zweiten Beispiel (in diesem hypothetischen Wahl-district) ist die Arbeitslosenquote höher als im ersten Beispiel. Wir sehen der Unterschied ist jetzt etwa bei 0.44. Wir haben aber den Anteil Migrant*innen ebenfalls nur um einen Prozentpunkt verändert! Nur die Arbeitslosenquote ist höher.

```

library(sjPlot)
m4 <- lm(AfD ~ migrantfraction*unemploymentrate, data = results)
plot_model(m4, type = "pred", terms = c("migrantfraction","unemploymentrate"))

```



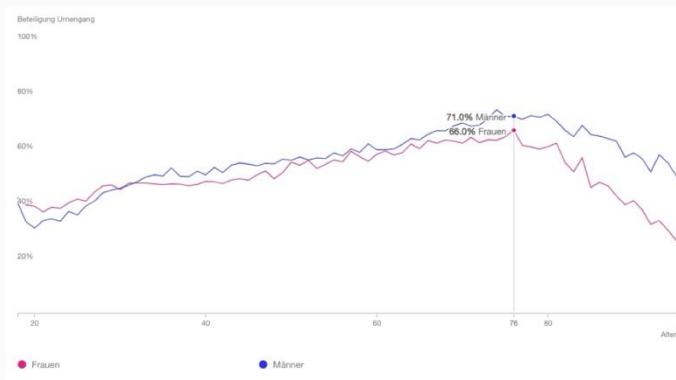
Weiterer Hinweis: Skaleniveaus kennen! Nominal, Ordinal, Metrisch sollte man kennen!
Wir haben jetzt also ein Beispiel gesehen wie wir zwei kontinuierliche Variablen miteinander interagieren können. Wir haben nun das gesamte Spektrum durchgespielt. Es fehlt noch der Spezialfall, wo wir eine Variable mit sich selber interagieren.

Wir brauchen für das ganz wenig Mittelschulmathematik. Wir müssen ableiten können. Wenn wir eine Interaktion modelliert haben und versucht haben den Zusammenhang zu beschreiben, haben wir bis jetzt mit dem Delta Ansatz gearbeitet. Wir können das aber eigentlich noch direkter machen. Wir können einfach sagen wir nehmen die Ableitung von X1 wenn wir wissen wollen was ist der Zusammenhang zwischen X1 und Y, dann nehme ich meine Regressionsfunktion und davon die Ableitung nach X1 und was dort rauskommt ist genau der marginale Zusammenhang zwischen X1 und Y. Das ist die Steigung dieser Funktion für X1 und die ist hier in diesem Modell so bezeichnet als Beta1 + Beta3 * X2. Man nennt diese Größe Beta1 + Beta3 * X2 den marginalen Effekt, das ist die erwartete Veränderung in Y wenn sich X1 um eine Einheit verändert oder anders formuliert die Steigung an der Stelle der Funktion.

Spezialfall – Interaktion mit sich selber

Wieso würden wir das überhaupt wollen? Weil bis jetzt haben wir nur Modelle gesehen die sind ein bisschen steif. Die soziale Wirklichkeit ist nicht immer so. In diesem Beispiel sehen wir Teilnahmen an der Wahl nach Alter und Geschlecht und man hat für jeden Jahrgang die Teilnahmequote

- Bis jetzt haben wir einfache lineare Zusammenhänge modelliert
- Mit Interaktionen konnten wir dem Modell mehr Flexibilität geben
- Aber was machen wir, wenn wir einen solchen Zusammenhang sehen?



Wenn ich diesen Zusammenhang der hier abgebildet ist modellieren will, wenn ich hier eine Gerade reinlege ist das nicht befriedigend. Ich brauche also die Möglichkeit einen kurvlinearen Zusammenhang modellieren zu können. Genau das erreiche mit der Interaktion einer Variable mit sich selber. Das sieht dann so aus:

- Glücklicherweise kann man das mit einer speziellen Interaktion angehen
- Man kann X auch mit sich selber interagieren, um einen kurv-linearen Zusammenhang zu modellieren:

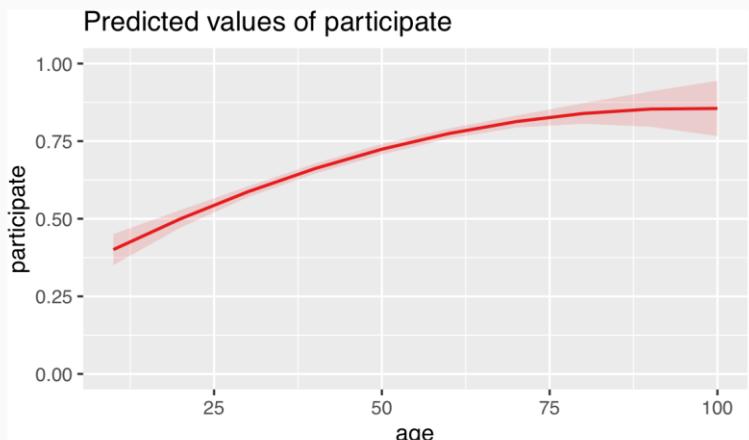
$$Y = \alpha + \beta_1 \cdot X + \beta_2 \cdot X^2 \quad (7)$$

Wenn wir das in R schätzen und visualisieren lassen sieht das so aus:

```
m5 <- lm(participate ~ age, data = selects)
m6 <- lm(participate ~ age + I(age*age), data = selects)
screenreg(list(m5,m6), digits=5) Alter im Quadrat

## -----
##          Model 1      Model 2
## -----
## (Intercept) 0.41309 ***   0.28958 ***
##             (0.01747)    (0.04073)
## age         0.00581 ***   0.01172 ***
##             (0.00034)    (0.00179)
## age * age   -0.00006 ***  (0.00002)
##             (0.00002)
## -----
## R^2          0.05373     0.05576
## Adj. R^2    0.05355     0.05540
## Num. obs.   5250        5250
## -----
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
plot_model(m6, type = "pred", terms = c("age")) +
  scale_y_continuous(limits = c(0, 1),
                     breaks = c(0,.25, .5, .75,1))
```



Wir sehen beim Model 2 wir haben zuerst einen positiven Koeffizientenschätzer für Beta1 und einen negativen Koeffizientenschätzer für Beta2, wir werden sehen das dies heisst, dass wir etwas abflachendes betrachten. Man kann den Zusammenhang nun beschreiben (also was man sieht) oder auch wieder mit predict berechnen.

- "Positiver Zusammenhang von Alter und Teilnahme, wobei der Zusammenhang mit steigendem Alter abnimmt. Ab ca. 82 Jahren findet sich kein Zusammenhang mehr zwischen durchschnittlicher Beteiligungsrate und Alter in diesen Daten."

- Man kann das auch berechnen – ein Jahr höheres Alter für 20-jährige

```
x1 <- predict(m6, newdata = data.frame(age=20))
x2 <- predict(m6, newdata = data.frame(age=21))
x2-x1
```

```
##           1
## 0.009235475
```

- Ein Jahr höheres Alter für 60-jährige

```
x1 <- predict(m6, newdata = data.frame(age=60))
x2 <- predict(m6, newdata = data.frame(age=61))
x2-x1
```

```
##           1
## 0.00438525
```

Wir sehen ein zusätzliches Jahr im Alter von 20 Jahren hat einen viel stärkeren Effekt auf die Teilnahmequote wie ein zusätzliches Jahr mit 60 Jahren. Der Zusammenhang ist nun nur noch etwa halb so gross. Das ist die sinkende Steigung dieser Funktion.

Wir können es natürlich aber auch wieder «von Hand» machen.

- Wir können nun wieder den Zusammenhang zwischen Alter und Teilnahme berechnen
- Dafür nehmen wir die erste Ableitung der Regressionsfunktion nach X_{Alter}

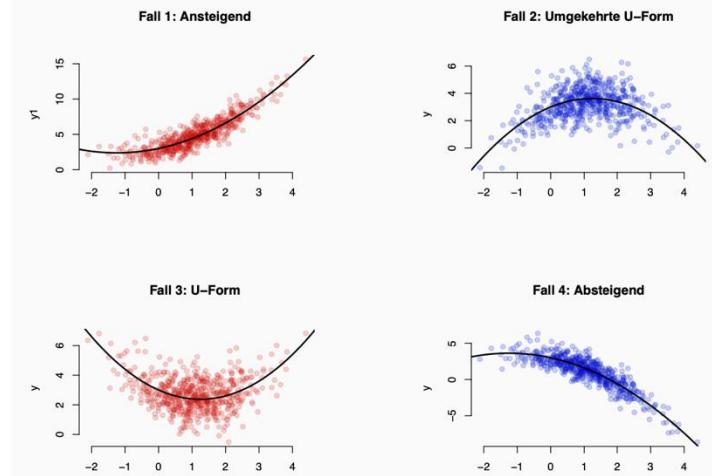
$$\begin{aligned} Y &= \alpha + \beta_1 \text{Alter} + \beta_2 \text{Alter}^2 \\ \frac{\partial Y}{\partial \text{Alter}} &= \beta_1 + 2 \cdot \beta_2 \cdot \text{Alter} \\ \frac{\partial Y}{\partial \text{Alter}} &= 0.01172 - 2 \cdot 0.00006 \cdot \text{Alter} \end{aligned}$$

- Wir sehen also, dass der Zusammenhang mit Alter schwächer wird

Wir haben den Ausdruck Y ist eine Funktion von Alter und Alter Quadrat. Wir wollen wissen wie verändert sich Y als eine Funktion von Alter. Also nehme ich die Ableitung dieser Funktion zu Alter. Wenn ich das durchspiele erhalte ich $\beta_1 + 2 * \beta_2 * Alter$. Damit können wir nun jedes Alter das wir wollen einsetzen plus die Koeffizientenschätzer und wir können sozusagen genau den marginalen Effekt von Alter auf Teilnahme berechnen in einem kurvlinearen Modell.

Es ist nun natürlich nicht immer so wie in unserem Beispiel, dass wir etwas haben, was zuerst ansteigend ist und mit der Zeit fällt. Es gibt zwei Koeffizienten, die bestimmen was für eine Form für einen Zusammenhang besteht, das ist β_1 und β_2 . Diese können positiv oder negativ sein, das heisst es gibt vier Fälle. So sehen diese vier Fälle gelistet und visualisiert aus:

- Interaktionen mit sich selber können unterschiedliche Formen annehmen, bspw...
 - Fall 1: Wenn beide Koeffizienten positiv sind ($\beta_X > 0 \& \beta_{X^2} > 0$)
 - Fall 2: Wenn β_X positiv ist und β_{X^2} negativ ist
 - Fall 3: Wenn β_X negativ ist und β_{X^2} positiv ist
 - Fall 4: Wenn beide Koeffizienten negativ sind ($\beta_X < 0 \& \beta_{X^2} < 0$)



Damit haben wir jetzt viel mehr Flexibilität, um ein Modell zu schätzen das sich den Daten anpassen kann. Wir sehen bereits jetzt das die lineare Regression nicht linear heisst weil wir Linien gezeichnet haben sondern weil sie linear in den Parametern ist. Wir sind überhaupt nicht beschränkt nur Geraden zu machen.

Noch ein paar wichtige Bemerkungen zu Interaktionen:

- Wenn man $X_1 \cdot X_2$ ins Modell einschliesst, muss man auch immer X_1 und X_2 einschliessen!
- Wenn eine Interaktion von X_1 und X_2 im Modell ist, kann man β_1 und β_2 nicht direkt interpretieren.
- In Woche 9 kommen wir zu diesem Thema zurück und fragen uns, wann man eine Interaktion einschliessen soll.

Zum zweiten Punkt: Immer wenn es eine Interaktion hat gibt es vor der Interpretation einen Zwischenschritt! Teilweise ist es ganz einfach wie im Ost-West Beispiel wo wir einfach die Fallunterscheidung machen. Teilwei-

se aber komplexer wenn wir zwei kontinuierliche haben. Ich kann die Koeffizienten nicht direkt aus dem R Code interpretieren bei Interaktionen, das ist FALSCH! Was richtig sein

könnte ist wenn ich sage unter der Annahme das 0% Migrant*innen hat, aber jetzt habe ich bereits mehr gemacht als eine Zahl abgelesen und es konditional formuliert.

VL 6

Wir verbinden in diesem Teil, die Regression mit der Kausalität und fragen uns, wann wir die geschätzten Parameter kausal schätzen können. Beispiele für diesen Teil kennen wir schon aus früheren Vorlesungen. Das Outcome ist die Gesundheit und das treatment ob jemand eine Krankenversicherung hat. Wir haben dabei verschiedene Datenquellen nämlich einmal Beobachtungsdaten und zweitens Experimentelle Daten.

Regression und randomisierte Experimente

Wieso sollten wir eine Regression machen bei randomisierten Experimenten? Wenn wir von Experimenten sprechen meinen wir damit, dass das treatment zufällig zugewiesen wurde. Es bezieht sich also nicht darauf das wir in einem Labor stehen oder einer künstlichen Umgebung sondern der Begriff Experiment bezieht sich darauf das wir die Kontrolle darüber haben wer das treatment erhält und das wir das randomisiert machen.

- **Repetition (1):** Die Differenz der Mittelwerte liefert eine unverzerrte Schätzung des kausalen Effekts, wenn unser 'treatment' den Beobachtungen zufällig zugewiesen wird (Woche 2).
- **Repetition (2):** Der Koeffizient für eine binäre Variable in einer einfachen linearen Regression ist gleich der Schätzung der Mittelwertdifferenz (Woche 4).
- **Implikation:** Wenn das 'treatment' randomisiert wird, liefert der lineare Regressionskoeffizient eine unverzerrte Schätzung des kausalen Effekts!

Zu Rep (2): Der Koeffizient den wir schätzen auf dieser binären Variable gibt uns den Mittelwert an das heisst, wenn ich den Mittelwert bestimmen will für diejenigen die eine 1 haben auf dieser binären Variable, rechne ich alpha plus dieses Beta. Das heisst dieses Beta ist dann genau dieser Mittelwertsunterschied. Wir sehen also, wenn ich die Regression auf einen Datensatz verwende, der aus einem Experiment stammt, dann kann ich auch so gleich den treatment Effekt bestimmen. Das heisst in einem randomisierten Experiment kann ich das Regressionsmodell verwenden, die erklärende Variable die ich ins Modell gebe, ist ob eine Beobachtung das treatment erhalten hat oder nicht und der Koeffizientenschätzer ist dann ein unverzerrter Schätzer für den Effekt dieser Intervention. Hier ist der Beweis dafür in R:

Wir können die Differenz der Mittelwerte in diesen experimentellen Daten berechnen:

```
## Mean health level for insured and uninsured individuals
mean_health_insured <- mean(rand$health[rand$insured == TRUE])
mean_health_uninsured <- mean(rand$health[rand$insured == FALSE])
mean_health_insured - mean_health_uninsured

## [1] -0.01895885

## Regression of health on insurance status
lm(health ~ insured, rand)

...
## (Intercept) insuredTRUE
##      3.40702      -0.01896
...
```

Es gibt verschiedene Vorteile der Regression bei der Analyse von experimentellen Daten:

1. Heterogene treatment Effekte

- Sind die Effekte des ‘treatments’ je nach Gruppe unterschiedlich?
- Sie wissen bereits wie man das analysieren kann: Interaktionen!

2. Nicht-binäre treatments

- Ist das ‘treatment’ kontinuierlich? Kategorial?
- Sie wissen bereits wie man das analysieren kann: Kontinuierliche Variable im Modell oder dann mit ‘factor’ Dummies generieren!

3. Erhöhung der “Präzision” unserer Schätzungen

Woche 9 oder 10 nachvollziehbar sein!!

- Die Kontrolle anderer Faktoren, die das Ergebnis bestimmen, kann die Schätzungen der ‘treatment’ Effekte präzisieren
- Das werden wir in den kommenden Wochen behandeln!

Heterogene treatment Effekte:

Wir können zum Beispiel mit unterscheiden ob die Effekte des treatments unterschiedlich sind für verschiedene Gruppen. Wir können also unterscheiden ob die marginal erwartete Veränderung in Y, wenn X1 sich ein bisschen verändert, das kann jetzt von X2 abhängen, falls das Modell in den Daten wirklich ein solches Muster findet. Wieso diese Formulierung? Es kann natürlich auch sein das ich ein Interaktionsmodell schätze aber der geschätzte Parameter Beta3 plus minus Null ist. Ich habe also gedacht es gibt eine Interaktion aber die ist so in den Daten einfach nicht vorhanden. Aber ich gebe dem Modell die Möglichkeit von den Daten zu lernen das ein solcher Interaktionseffekt besteht. Wenn wir das illustrieren wollen (letzte 2 Vorlesungen haben wir das gemacht) können wir beispielsweise für unterschiedliche Werte von X2 zeigen was geschieht, wenn sich X1 um eine Einheit erhöht und dann sieht man zum Beispiel an dieser Effekt beim selbsteingeschätzten Gesundheitsstatus geht runter wenn das Einkommen raufgeht.

Zusammenhang wenn wir für unser Beispiel eine Interaktion in R schätzen:

```
heterogeneous_effects_model <- lm(health ~ insured * income, rand)
summary(heterogeneous_effects_model)

...
## Coefficients:
##               geschätzter
##               Koeffizient
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.195337  0.072245 44.229 < 2e-16 ***
## insuredTRUE 0.098969  0.079757  1.241 0.214754
## income      0.006551  0.001963  3.338 0.000855 ***
## insuredTRUE:income -0.003536  0.002181 -1.621 0.105126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.767 on 2698 degrees of freedom
## Multiple R-squared:  0.007878,   Adjusted R-squared:  0.006775
## F-statistic: 7.141 on 3 and 2698 DF,  p-value: 8.932e-05
...
```

Zusammenhang: Wenn das Einkommen um eine Einheit steigt, nimmt für versicherte Personen der selbsteingeschätzte Gesundheitsstatus um 0.0035 Skalenpunkte ab.

Wenn ich illustrieren will wie unterscheiden sich diese Effekte kann ich das mit predict machen:

```
# Treatment effect for low-income individuals
predict(heterogeneous_effects_model,
        newdata = data.frame(insured = c(F,T),
                             income = 10)) # income in thousands
##      1      2 + Wert (Gesundheit) steigt wenn Personen
## 3.26085 3.32446 versichert sind

# Treatment effect for high-income individuals
predict(heterogeneous_effects_model,
        newdata = data.frame(insured = c(F,T),
                             income = 30)) # income in thousands
##      1      2 - Wert (Gesundheit) sinkt wenn Personen
## 3.391874 3.384768 versichert sind
```

Implikation: Es gibt Hinweise darauf, dass das 'treatment' den Gesundheitszustand von Personen mit niedrigem Einkommen *leicht erhöht*, nicht aber den von Personen mit hohem Einkommen.

Nicht binäre treatments:

Sehr oft ist es allerdings so das die treatments nicht wie bis anhin angenommen binär sind. Sie könnten zum Beispiel auch kontinuierlich oder kategorisch sein. In unserem Beispiel aus dem Gesundheitswesen war die Behandlung im Experiment tatsächlich kategorisch.

```
table(rand$insured, rand$plantype)
```

```
##          Catastrophic Coinsurance Deductible Free
## FALSE       491           0         0   0
## TRUE        0        727     593 891
```

- "Catastrophic" – Einzelpersonen zahlen für alle Gesundheitskosten
- "Coinsurance" – Einzelpersonen zahlen 25-50% der Kosten
- "Deductible" – Kostenobergrenze von \$150
- "Free" – Einzelpersonen zahlen nichts

Frage: Sind die kausalen Effekte für alle drei Behandlungsbedingungen gleich?

```
categorical_treatment_model <- lm(health ~ plantype, rand)
summary(categorical_treatment_model)

...
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.407016  0.034684 98.231  <2e-16 ***
## plantypeCoinsurance  0.044112  0.044893  0.983  0.3259
## plantypeDeductible -0.009908  0.046893 -0.211  0.8327
## plantypeFree        -0.076444  0.043196 -1.770  0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7685 on 2698 degrees of freedom
## Multiple R-squared:  0.003769,   Adjusted R-squared:  0.002661
## F-statistic: 3.403 on 3 and 2698 DF,  p-value: 0.01701
```

Das gute an diesem R Output rechts ist, hier haben wir es gleich so erhalten das die Personen die keine Versicherung haben, das ist die Referenzkategorie (baseline) wir erhalten dann gleich, drei Koeffizientenschätzer für diese drei möglichen treatment Gruppen die wir haben. Wir können also gleich schauen wie unterscheiden die sich von den Leuten denen man keine Krankenversicherung zugeteilt hat. Wir sehen zum Beispiel beim untersten Koeffizientenschätzer (free heisst man kann die ganze Zeit zum Arzt und es wird alles bezahlt) das wir in dieser Gruppe einen leicht tieferen Durchschnitt der Selbsteinschätzung der Gesundheit haben als bei denen die keine Krankenversicherung im Expe-

riment erhalten haben. Wir können uns dann mit predict auch die Mittelwerte der verschiedenen Gruppen anzeigen lassen. (Siehe Vorlesungsfolien) Wir sehen es gibt keine grossen Unterschiede.

Zusammenfassung Randomisierung und Regression:

Die Regressionsanalyse des RAND-Gesundheitsexperiments hat ergeben, dass:

1. ...der durchschnittliche kausale Effekt der Krankenversicherung auf die Gesundheitsergebnisse sehr gering ist
2. ...der kausale Effekt ist für Personen mit geringem Einkommen etwas positiver, wenn auch immer noch gering
3. ...die geschätzten kausalen Auswirkungen der verschiedenen Versicherungstypen unterscheiden sich kaum

Nebenbemerkung: Trotz dieser bescheidenen Effekte zeigte das RAND-Experiment viel grössere Effekte in Bezug auf die Inanspruchnahme von Gesundheitsdiensten. Der selbstberichtete Gesundheitszustand ist vielleicht nicht das wichtigste Ergebnis der Gesundheitsversorgung!

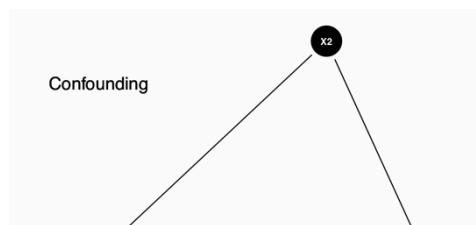
Bemerkung unten: Dieses Beispiel zeigt, es ist sehr schwierig und teuer hier gute Forschung zu machen. Man hat nur den selbsteingeschätzten Gesundheitszustand gemessen. Die waren nicht alle vorher und nachher beim Doktor, das wäre noch viel teurer geworden und die Studie war so schon sehr teuer, weil es ein Experiment war.

Frage aus dem Publikum: Wann weiss man ob man ein Regressionsmodell schätzen muss oder nicht: Faktisch eigentlich immer auch wenn man Daten aus einem Experiment hat (was man nicht oft hat) weil man eigentlich immer herausfinden will/muss ob die Effekte je nach Gruppe unterschiedlich sind.

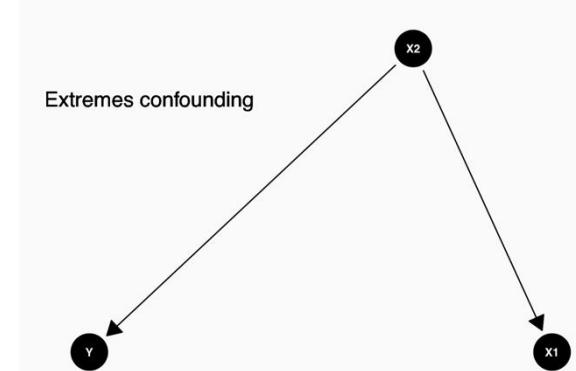
Regression und Beobachtungsdaten

Die meisten Daten in der Politikwissenschaft sind Beobachtungsdaten. Ein Grund wieso die Regression hier für uns interessant ist, weil sie uns helfen kann einen Rahmen für die Kontrolle von confounding Faktoren, welche in Beobachtungsdaten auftreten, zu bieten. Was ist confounding noch einmal? Ist die Situation wo es zwischen der treatment und control Gruppe nicht nur im Bezug auf das treatment sondern auch auf andere potenzielle erklärende Variablen einen Unterschied gibt zwischen diesen beiden Gruppen. Das heisst wenn das passiert ist unser Schätzer den wir erhalten verzerrt geschätzt. Wir nennen das omitted variable bias. Das ist der bias, also die Verzerrung, die daher stammt das wir eine relevante Variable, omitted, also nicht ins Modell eingeführt haben, also das wir eine relevante Variable weggelassen haben. Das eine omitted variable bias entsteht müssen zwei Bedingungen erfüllt sein. Die Verzerrung tritt also nur dann ein, wenn diese Variable die wir nicht im Modell drinhaben, wenn die auch tatsächlich mit diesem anderen X1 korreliert. Also was auch immer fehlend ist bei Nobelpreisen und Schokolade, das muss korrelieren mit per Capita Schokoladenkonsum (Beispiel aus Vorlesung das der nationale Schokoladenkonsum angeblich mit der Anzahl der Nobelpreise verknüpft ist), dass es zu einer Verzerrung führen kann. Und die zweite Bedingung ist diese ausgelassene Variable muss auch mit dem outcome korrelieren. Nur wenn wir beide Bedingungen erfüllt haben ist eine so ausgelassene Variable eine sogenannte confounding Variable. Weshalb ist das so? Überlegen wir uns mal, wenn es einen Zusammenhang gibt zwischen X2 und Y und das X2 korreliert mit X1 und ich gebe nur das X1 ins Modell, dann wird das Modell alle Variation die eigentlich dem X2 zugeschrieben werden soll, wird es auf einen Teil von X1 zuschreiben, da die beiden ja miteinander korrelieren und dann führt das dazu das so ein Koeffizientenschätzer verzerrt ist, also das wir einen bias drin haben.

Das ist die Situation vor der wir uns fürchten.



Das es eine Variable gibt im Hintergrund, welche Wir nicht im Modell haben, die X_1 und Y beeinflusst, wenn wir dann nur Y und X_1 integrieren, finden wir irgendwas und es kann sein das der tatsächliche Zusammenhang zwischen X_1 und Y 0 ist und nur getrieben wird durch dieses X_2 was wir dann nicht sehen würden. Das wäre dann dieses extreme Beispiel.



Wir können aber mehr sagen als nur das es zu einer Verzerrung kommt. Je nachdem wie diese Korrelationen sind können wir auch gleich sagen ob wir einen Koeffizienten zu gross oder zu klein schätzen. Es kann sein das ich einen Zusammenhang finde aber er existiert gar nicht aber umgekehrt kann es auch sein das ich keinen Zusammenhang finde obwohl er existiert. Das sind bad news, das ist ein Problem.

Wenn OVB vorhanden ist und wir nichts dagegen unternehmen, dann wird unsere Schätzung von β_1 verzerrt (falsch) sein

	$cor(X_1, X_2) > 0$	$cor(X_1, X_2) < 0$
$cor(X_2, Y) > 0$	$\hat{\beta}_1$ zu gross	$\hat{\beta}_1$ zu klein
$cor(X_2, Y) < 0$	$\hat{\beta}_1$ zu klein	$\hat{\beta}_1$ zu gross

- $cor(X_1, X_2)$ ist die Korrelation zwischen X_1 und X_2
- $cor(X_2, Y)$ ist die Korrelation zwischen X_2 und Y

Implikation: Je nachdem, in welchem Verhältnis X_2 zu X_1 und Y steht, kann $\hat{\beta}_1$ entweder zu gross oder zu klein sein!

Wenn wir uns nun unseren Beispiel Beobachtungsdatensatz anschauen sehen wir das beide Bedingungen erfüllt sind. Wenn wir also versuchen den Zusammenhang anzuschauen zwischen selbsteingeschätztem Gesundheitszustand und ob jemand eine Versicherung hat oder nicht, ist das was wir aus der Regression rauskriegen, das reflektiert eher Einkommensunterschiede im Hintergrund als tatsächlich kausale Effekte einer Versicherung. Das sehen wir hier unten im rechten Bild.

OVB ist ein Problem, wenn ...

1. Die unabhängige Variable mit anderen Faktoren korreliert ist und
2. Die abhängige Variable auch mit diesen Faktoren korreliert ist

Also können wir nun untersuchen, ob es im NHIS-Datensatz zu einem 'Confounding' kommen könnte:

- Ist die **unabhängige** Variable korreliert mit anderen Faktoren? *Ja ganz klar*

insured	health	age	female	years_educ	income
Insured	3.9	43.3	50.2	14.1	101.3

OVB ist ein Problem, wenn ...

1. Die unabhängige Variable mit anderen Faktoren korreliert ist und
2. Die abhängige Variable auch mit diesen Faktoren korreliert ist

Also können wir nun untersuchen, ob es im NHIS-Datensatz zu einem 'Confounding' kommen könnte:

- Ist die **abhängige** Variable mit diesen Faktoren korreliert?

	age	female	years_educ	income
$corr(\text{health}, X)$	-0.162	-0.001	0.255	0.269

Im rechten Bild oben sehen wir die Korrelation von Gesundheit (abhängigen Variable, outcome) mit den Faktoren. Wir sehen bei income ebenfalls eine starke Korrelation. Das ist der Punkt wo ich mir Sorgen machen würde. Ich habe bei income also einseitig eine deutliche Korrelation von Einkommen und Gesundheitszustand und eine Seite vorher (im rechten Bild) haben wir gesehen das Einkommen sich auch sehr stark unterscheidet in den beiden Gruppen (äquivalent zu der Aussage das es auch dort sehr stark korreliert), das heißt hier liegt ein confounding Problem vor wenn wir Einkommen nicht modellieren. Wir haben zwei Lösungsansätze für das, also zwei Strategien wie wir diese potenziellen Störfaktoren kontrollieren können.

1. Subklassifikationen bilden, also zum Beispiel den Zusammenhang zwischen Versicherungsstatus und Gesundheitszustand nur innerhalb derselben Einkommensgruppe anschauen. Innerhalb dieser Gruppen ist der Unterschied im Einkommen untereinander ja logischerweise viel ähnlicher. Also ich mache zum Beispiel Gruppen der Einkommen 0-20'000, 20'000-40'000 usw.
2. Wir verwenden eine multiple Regression, diese erlaubt uns ja einfach verschiedene Faktoren in ein Modell zu integrieren.

Wenn wir die Subklassifikation in unserem Beispiel machen sehen wir, also wenn wir das Einkommen miteinbeziehen in unsere Analyse, dann wird dieser angeblich geschätzte Effekt viel kleiner.

```
insured_low_mean - uninsured_low_mean
## [1] 0.1010293
insured_mid_mean - uninsured_mid_mean
## [1] 0.04954519
insured_high_mean - uninsured_high_mean
## [1] 0.1542666

Konsequenz: Sobald wir für das Einkommen kontrollieren, sind die Auswirkungen der Versicherung auf die Gesundheit viel geringer als bei dem naiven Mittelwertvergleich.

mean(nhis$health[nhis$insured == T]) -
  mean(nhis$health[nhis$insured == F])
## [1] 0.3262003
```

Ein Problem könnte sein, dass die Varianz dadurch nicht richtig erfasst wird. Es könnte ja sein, dass in einer Einkommenskategorie immer noch eine Varianz gibt, die nicht vollständig erfasst wird. Es könnte immer noch sein, dass in einer Einkommensgruppe die Leute mit einer Versicherung eher ein höheres Einkommen haben. Das andere Problem ist, wenn ich nun auch noch Bildung einschließen wollte. Nur schon wenn ich auch drei Bildungsniveaus hätte, müsste ich insgesamt schon 9 Subklassifikationen bilden. Das wird schnell unübersichtlich. Deshalb verwenden wir oft die Alternative und machen das über das Regressionsmodell.

Wenn wir das mit der Regression machen: So das wir sagen wir kontrollieren für dieses X2 und wir machen das indem wir das X2 ins Modell integrieren. Wir erinnern uns in Woche 4 haben wir den Zusammenhang zwischen X1 und Y beschreiben unter der Annahme als würden wir X2 konstant halten. Das ist das was wir machen wollen, wir wollen wissen was ist der Effekt dieser Krankenversicherung ohne das sich gleichzeitig das Einkommen ändert, sondern wenn ich das Einkommen wie stillhalten könnte für eine Person und nur die Versicherung ändern. Das wollen wir haben und das können wir holen über die Regression, indem wir das aufsetzen als multiples Modell wo wir einerseits den Versicherungsstatus als erklärende Variable drin haben und auf der anderen Seite die Einkommensvariable. Der Vorteil ist natürlich wenn ich nun auch noch Bildung oder andere Variablen kontrollieren will, kann ich das machen, ist der Vorteil gegenüber der Subklassifikation wo ich schnell an die Grenzen komme. Wir schätzen nun also die Regressionen,

```
# Naive model
nhis_model <- lm(health ~ insured, data = nhis)

# Model controlling for income only
nhis_model_with_income <- lm(health ~ insured + income,
                               data = nhis)

# Model controlling for many variables
nhis_model_with_covariates <- lm(health ~ insured + income +
                                   age + female +
                                   years_educ,
                                   data = nhis)
```

	insured	(1)	(2)	(3)	health
age					-0.02
female					-0.05
years_educ					0.05
Constant	3.62	3.43	3.80		
Observations	19,996	19,996	19,996		
R ²	0.02	0.07	0.13		

- In diesem Beispiel nimmt der $\hat{\beta}$ -Koeffizient für "versichert" ab, wenn für andere Variablen kontrolliert wird
- $\hat{\beta}_{\text{versichert}}$ in Modell 3 ist viel näher an der experimentellen Schätzung

Wir sehen diesen Wert den wir erhalten wenn wir einfach nur den Mittelwertsunterschied bestimmen mit der Regression(0.33), sobald wir das Einkommen reinnehmen geht dieser Wert um mehr als den Faktor 5 zurück (0.06), dies weil was in diesen 0.33 steckt ist ganz viel ist der Zusammenhang zwischen Einkommen auf den Gesundheitszustand und nicht wirklich Krankenversicherung auf Gesundheitszustand. Wenn ich dann noch Alter, Geschlecht und Bildung ins Modell gebe, dann sehen wir das dieser Koeffizient noch einmal runtergeht auf 0.02. Diese Zahl ist jetzt viel näher dran beim Schätzer denn wir in den experimentellen Daten haben. Wovon wir ausgehen das dies ein guter Schätzer ist für den wahren durchschnittlichen kausalen Effekt einer Versicherung ist. Wir sehen also den Wert davon, das wir versuchen für confounders zu kontrollieren. Würde ich das nicht machen, hätte ich viel zu grosse Zahlen bekommen. Die Frage im Modell 3 (im rechten Bild oben) ist dieser Schätzer nun der kausale Effekt?

	health		
	(1)	(2)	(3)
insured	0.33	0.06	0.02
income		0.004	0.004
age			-0.02
female			-0.05
years_educ			0.05
Constant	3.62	3.43	3.80
Observations	19,996	19,996	19,996
R ²	0.02	0.07	0.13

Frage: Stellt $\hat{\beta}_{\text{insured}}$ den kausalen Effekt der Versicherung auf die selbst angegebene Gesundheit dar?

45

Antwort: Nur wenn wir davon auszugehen, dass wir für alle Störfak-

Es gibt einfach gewisse Grenzen zur Erhebung von Variablen die aber durchaus relevant sein könnten. Zum Beispiel die Risikofreudigkeit von Personen, oder das Personen sich verschieden gut schauen und um ihre eigene Gesundheit kümmern. Das ist unmöglich zu messen, also können wir es auch nicht ins Modell einschliessen. Es ist also schwierig abschliessend anzukommen. Und ein Experiment durchzuführen ist ethischen, finanziellen, politischenn Gründen oft nicht möglich. Aber das ist die Welt in der wir leben und wir müssen versuchen damit klarzukommen.

Hier noch die Illustration wieso:

Beachten Sie, dass dies bei experimentellen Daten nicht der Fall ist!

	health		
	(1)	(2)	(3)
insured	-0.02	-0.01	-0.01
income		0.004	0.003
age			-0.01
female			-0.03
years_educ			0.05
Constant	3.41	3.29	3.19
Observations	2,702	2,702	2,702
R ²	0.0001	0.01	0.09

- Der Koeffizient für "versichert" ist in den Modellen 1, 2 und 3 nahezu identisch. **Warum?**
- OVB liegt vor, wenn die ausgelassenen Variablen mit **beiden**, unabhängigen und abhängigen Variablen, korreliert sind.
- Der Versicherungsstatus wird zufällig zugewiesen, kann also nicht mit anderen Faktoren korreliert sein.

Das ist genau der Grund, wieso wir randomisieren.

Confounders, kontrollieren und kausale Inferenzen

Um zu behaupten, dass Schätzungen, die auf multiplen linearen Regressionen unter Verwendung von Beobachtungsdaten beruhen, **kausale** Effekte darstellen, müssen Sie argumentieren können, dass Sie für **alle** möglichen Störvariablen kontrolliert haben.

Dies ist schwierig, weil Sie möglicherweise nicht:

- alle Störfaktoren *kennen* können
- alle Störfaktoren *messen* können
- alle Störfaktoren *beobachten* können

Wenn die Annahme, dass man für **alle** möglichen Störvariablen kontrollieren kann, plausibel ist, kann man die multiple lineare Regression verwenden, um kausale Effekte aus Beobachtungsdaten zu schätzen.

- Wir können solche Fälle als Fälle bezeichnen, in denen eine "**selection on observables**" möglich ist - das heisst, in diesen Fällen unterscheiden sich die treatment und control Gruppe nur durch eine Reihe von *beobachtbaren* Merkmalen.

Selection on observables heisst wir müssen alles beobachten können für den outcome. Das ist nicht immer gegeben, oder nicht immer einfach glaubhaft zu argumentieren das dies der Fall ist.

VL 8

Design-based causal inference

Das ist eine von verschiedenen Varianten wo wir durch das Modell bereits, einen bedeutenden Schritt hin zur Kausalität machen können. Das heisst wir erreichen eine kausale Interpretierbarkeit der geschätzten Koeffizienten, nicht nur darüber, das wir uns jeden möglichen confounder messen und ins Modell integrieren, sondern wir können auch Modelle so erstellen, das ganze Klassen von möglichen confounders keinen Einfluss haben auf das Ergebnis.

- In Woche 02 haben wir das Fundamentalproblem der kausalen Inferenz gesehen. Der entscheidende Punkt war, dass randomisierte Experimente uns erlauben den *average treatment effect* zu schätzen.
- Aber Sozialwissenschaftler:innen, die mit Beobachtungsdaten arbeiten (müssen), können sich nicht auf die Randomisierung verlassen und es besteht ein potentielles Problem von *confounding*.
- Vor Ostern (VL 06) haben wir gesehen, dass *nur falls* wir für *alle* potentiellen confounders kontrollieren, wir die multiple lineare Regression verwenden können, um das Problem mit Querschnittsdaten anzugehen.
- Heute sehen wir eine weitere Modellkategorie, die sehr hilfreich sein kann für die Identifizierung von kausalen Effekten, wenn man sogenannte Panel-Daten hat.

Diese weitere Kategorie erlaubt es uns nicht unbedingt jeden einzelnen Faktor zu messen und zu integrieren, weil es gibt Faktoren die sind extrem schwierig zu messen, zu operationalisieren. Wir sehen also eine Modellkategorie die es uns erlaubt eine ganze Kategorie von möglichen confounders zu vergessen, weil wir wissen selbst wenn die existieren, wenn ich dieses Modell verwende, können sie nicht zu einer Verzerrung des Koeffizienten führen. Das erreichen wir am Ende mit diesen fixed effects Modellen und dafür brauchen wir Panel Daten.

Inhaltliche Motivation

Beispiel das uns in dieser Veranstaltung begleitet:

Gibt es einen Resource Curse (Ressourcenfluch)

Der Ressourcenfluch, auch bekannt als das Paradox des Überflusses oder das Armutsparadox, beschreibt das Phänomen, dass Länder mit einer Fülle an natürlichen Ressourcen (wie fossilen Brennstoffen und bestimmten Mineralien) ein geringeres Wirtschaftswachstum, weniger Demokratie oder schlechtere Entwicklungsresultate haben als Länder mit weniger natürlichen Ressourcen.

- **Y (Abhängige Variable oder "outcome"):** Qualität der Institutionen
 - Indexwerte von *Quality of Government Institute at the University of Gothenburg*, die die Qualität der politischen Institutionen messen (z.B. Korruption)
- **X (Unabhängige Variable oder "treatment"):** Natürliche Ressourcen
 - Einnahmen aus Ölförderung in Prozenten des BIP

Beobachtungsdaten

Wir werden dieser Frage auf Basis eines Datensatzes mit Variablen von 58 Ländern zwischen 1996 und 2010 nachgehen:

- **Y (Outcome): institutions**
 - Institutional Quality index
- **X (Treatment): oil**
 - Öleinnahmen (% des BIPs)
- **country**
- **year**
- **aid** - Entwicklungshilfe (% des BIPs)
- **gdp.capita** - BIP pro Kopf in USD
- **polity2** - Polity IV Indexwert

Panel Daten

Bis jetzt hatten wir immer Querschnittsdaten. Dort haben wir zu einem fixen Zeitpunkt für verschiedene Einheiten, Werte beobachtet auf den unterschiedlichen Variablen. Die Art und Weise wie wir dann eine ganz einfache lineare Regression geschrieben haben sah dann so aus:

- Bei Querschnittsdaten können wir die einzelnen Beobachtungen dadurch unterscheiden, dass jede Beobachtung einen anderen Wert auf dem Index i hat.
- Beispiel der einfachen linearen Regression:¹

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

wobei

- X_i das X der i -ten Beobachtung ist
- Y_i das Y der i -ten Beobachtung ist
- ε_i der Fehlerterm der i -ten Beobachtung ist

¹Beachten Sie, dass man auch oft β_0 statt α für die Konstante verwendet.

Wir haben also nur das Subskript i , das uns anzeigt, dass es unterschiedliche Länder sind. Hinweis: Hier mal gewechselt anstatt alpha haben wir Beta 0, sehen wir oft auch so in Forschungspaper, einfach das wir das auch schon gesehen haben. Wenn wir das ganze für Paneldaten anpassen wollen müssen wir noch irgendwie die Zeit reinbekommen. Weil wir haben ja jetzt pro Land mehrere Beobachtungen (hier jedes Jahr eine), wir wollen also zeigen, dass es nicht nur Unterschiede gibt über i , sondern über etwas anderes und dafür verwenden wir das Subskript t (für time) und ändern dann die Schreibweise:

- Panel Daten sind wiederholte Beobachtungen derselben Einheiten, bspw. 58 Länder über 15 Jahre
- In Panel Daten haben wir aber *mehrere* Beobachtungen für jede Einheit i .
- Wir brauchen also ein weiteres Element in der Notation, um unterschiedliche Zeitpunkte der Messung zu beschreiben. Wir machen das dadurch, dass wir zwei Indizes verwenden:
 - Wir benutzen weiterhin i um die Einheiten zu bezeichnen
 - Und wir ergänzen es mit einem Subskript t um die unterschiedlichen Zeitpunkte der Messung zu bezeichnen

$$Y_{i,t} = \beta_0 + \beta_1 X_{i,t} + \varepsilon_{i,t}$$

wobei

- $X_{i,t}$ bezeichnet die Beobachtung der Variable X für Einheit i zum Zeitpunkt t
- $Y_{i,t}$ bezeichnet die Beobachtung der Variable Y für Einheit i zum Zeitpunkt t
- $\varepsilon_{i,t}$ bezeichnet den Fehlerterm für Einheit i zum Zeitpunkt t

Beta 0 und Beta 1 haben kein Subskript, wir schätzen diese beiden genau einmal über eine Zeit über alle Länder. Es sind also zunächst mal nur die Daten die variieren über Raum und Zeit.

Fixed effects Modell mit Panel Daten

Wie können wir solche Daten mit einer Regression analysieren?

- Wenn wir auf einem "pooled dataset" eine Regression schätzen *ohne die Zeitdimension weiter zu berücksichtigen*, dann nennt man das resultierende Regressionsmodell ein "pooled model"
- Wir "poolen" alle Länder-Jahr Beobachtungen zusammen

$$\text{institutions}_{i,t} = \beta_0 + \beta_1 \text{oil}_{i,t} + \varepsilon_{i,t}$$

Wir sehen beim R output den Zusammenhang den wir schon bei der Motivation gesehen haben. Wenn der Anteil an natürlichen Ressourcen (Öl) steigt, scheint die Qualität der Institutionen zu sinken. Wenn ich versuche die Qualität der Institutionen nur mit diesem einen Faktor zu erklären dann kann ich etwas mehr als 11 Prozentpunkte der Varianz in dieser outcome Variable mit diesem Modell erklären.

Das Problem ist wenn ich das kausal interpretieren will, dann muss ich eine sehr starke Annahme machen, nämlich das es keine confounding Variablen gibt, es gibt also beispielsweise nichts das dahinter steht und einen Einfluss auf die institutionelle Qualität hat und einen Einfluss hat wie gross der Anteil ist der aus natürlichen Ressourcen kommt.

Die Beziehung zwischen Öl und institutioneller Qualität kann durch viele Faktoren "gestört" werden

- d.h. die Beobachtungen mit hohen Ölernnahmen könnten aus Ländern stammen, die sich (in Bezug auf die Qualität der Institutionen) von denen mit niedrigen Ölernnahmen unterscheiden.
- Wir können Kontrollvariablen hinzufügen, um zu versuchen potenzielle Störfaktoren zu berücksichtigen, genau wie zuvor: z. B. das Pro-Kopf-Einkommen, ob ein Land eine Demokratie ist, Auslandshilfe usw.

Wir sehen mit Kontrollvariablen ist dieser Koeffizientenschätzer von oil schon mal runtergekommen auf 0.015. Wir sehen also nur noch einen halb so grossen Effekt. Aber auch hier gilt das ist nur ein guter Schätzer für den kausalen Effekt falls es keine anderen möglichen confounder gibt, die einen Einfluss haben können auf die Qualität der Institutionen.

Frage: Vergleichen wir Beobachtungen, die im Durchschnitt in Bezug auf alle möglichen Confounders identisch sind?

Antwort: Wahrscheinlich nicht! Es könnte viele andere mögliche Confounders für Länder mit mehr/weniger Öl geben

- Unterschiede in der bisherigen institutionellen Geschichte
- Ökonomische Unterschiede
- Kulturelle Unterschiede → wie will man das messen?
- Etc.

→ Wir wissen bereits, wie wir confounding Variablen kontrollieren können, die wir messen können....Aber was ist mit denen, die wir nicht (leicht) messen können?

```
pooled_model <- lm(institutions ~ oil,
                     data = resource_curse)

## =====
## Model 1
## -----
## (Intercept) 0.106 **
##             (0.034)
## oil         -0.028 ***
##             (0.003)
## -----
## R^2          0.114
## Adj. R^2    0.113
## Num. obs.   682
## =====
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
pooled_model_2 <- lm(institutions ~ oil + log(gdp.capita) +
                         + polity2 + aid,
                     data = resource_curse)

## =====
## Model 1     Model 2
## -----
## (Intercept) 0.106 ** -3.356 ***
##             (0.034)  (0.099)
## oil         -0.028 *** -0.015 **
##             (0.003)  (0.002)
## log(gdp.capita) 0.403 ***
##                   (0.013)
## polity2      0.031 ***
##                   (0.003)
## aid          0.012 ***
##                   (0.003)
## -----
## R^2          0.114    0.744
## Adj. R^2    0.113    0.742
## Num. obs.   682     672
## =====
```

Mit Kontrollvariablen!



Es gibt die Möglichkeit das wir Faktoren, die wir vielleicht nicht einmal messen können, trotzdem kontrollieren oder blockieren können, damit sie unseren Schätzer nicht beeinflussen. Das können wir beispielsweise machen wenn wir Paneldaten vorliegen haben, die uns erlauben, auszunützen das ich ein und dasselbe Land also ein und dieselbe Einheit, mehrmals in meinem Datensatz drin habe. Das eröffnet uns eine neue Möglichkeit. Es erlaubt uns ein fixed effects Modell zu schätzen. Um genau zu sein ein unit fixed effects Modell, weil wir für die Einheiten ein fixed effect drin haben. Technisch gesehen was machen wir? Ganz einfach eigentlich wir geben einfach für jedes Land außer für die Referenzkategorie eine Dummy Variable ins Modell. Es stellt sich heraus das uns dies sehr hilft bei der Isolation des kausalen Effekts. Wir erinnern uns eine Dummy Variable das einzige was die macht wenn ich sie ins Modell gebe ist das sich der Y-Achsenabschnitt verschieben kann. (Beispiele mit Ost und Westdeutschland z.B.) Für jede Gruppe die einen eigenen Dummy hat muss ich dann jeweils den spezifischen Achsenabschnitt berechnen. Der Koeffizient der Dummy Variable ist ja sozusagen wie der zusätzliche Shift denn wir mit der Konstante addieren können. Wir können so also Unterschiede die generell über die Länder existieren schonmal abfangen. Was dieser Ansatz aber nicht kann ist confounders abzufangen die sich über die Zeit ändern. Das sieht also Formel dann so aus:

Wir könnten das so formulieren:

$$Y_{i,t} = \beta_0 + \beta_1 X_{1i,t} + \gamma_2 C_{2i} + \gamma_3 C_{3i} + \dots + \gamma_n C_{ni} + \varepsilon_{i,t}$$

wobei

- $Y_{i,t}$ ist institutionelle Qualität für Land i im Jahr t
- β_0 ist die Modellkonstante
- $X_{1i,t}$ sind die Öleinnahmen für Land i im Jahr t
- β_1 ist der Koeffizientenschätzer für die Öleinnahmen
- γ_n ist der Koeffizientenschätzer für die binäre Variable, die gleich 1 ist, wenn das Land $i = n$.
- C_n ist die binäre Variable, die gleich 1 ist, wenn das Land $i = n$.
- $\varepsilon_{i,t}$ ist der Fehlerterm für Land i im Jahr t

Linkes Bild: Der Anfang ist gleich. Das Gamma $2 * C2i$ zeigt an das dies die Dummy Variable ist für das zweite Land in meinem Datensatz, das heißt dieses $C2i$ wird immer 0 sein außer wenn das Land die Nummer 2 wäre. Ich habe also für alle Länder eine Dummy Variable, dieses Gamma ist dann nichts anderes als der Shift beim Intercept denn ich für dieses Land verwenden muss. Man beachte all diese griechischen Buchstaben die wir sehen haben keine Subskripte i oder t. Das sind fixe Zahlen die wir schätzen die bleiben konstant in diesem Datensatz je nachdem welches Land ich anschau habe ich einen andern Achsenabschnitt.

Rechtes Bild: Oft schreiben wir das Ganze etwas vereinfacht, wir ersetzen die Konstante einfach durch diesen Ausdruck mit dem Subskript i um anzuzeigen, jede Einheit in diesem Fall jedes Land, einen anderen Achsenabschnitt hat. Der Achsenabschnitt ist nichts anderes als im linken Bild Beta0 plus das entsprechende Gamma für jedes Land, also bei Land 2 ist es Gamma 2 usw. das ich jeweils zu diesem Beta 0 hinzuaddiere.

Anstatt jedoch alle Dummy-Variablen auszuschreiben, definieren wir das Modell der unit fixed effects wie folgt:

$$Y_{i,t} = \alpha_i + \beta_1 X_{1i,t} + \varepsilon_{i,t}$$

wobei

- $Y_{i,t}$ ist institutionelle Qualität für Land i im Jahr t
- $\alpha_i = \beta_0 + \gamma_i$ ist die Konstante für Land i
 - für Land 1, ist die Gleichung $\beta_0 + \beta_1 X_{1i,t}$
 - für Land i , ist die Gleichung $\beta_0 + \gamma_i + \beta_1 X_{1i,t}$ – also $\alpha_i = \beta_0 + \gamma_i$
- $X_{1i,t}$ sind die Öleinnahmen für Land i im Jahr t
- β_1 ist der Koeffizientenschätzer für die Öleinnahmen
- $\varepsilon_{i,t}$ ist der Fehlerterm für Land i im Jahr t

Diese unit fixed effects, in unserem Fall hier sind das länderspezifische Dummy Variablen die wir ins Modell tun, sind uns nicht fremd wir haben diese schon kennengelernt. (z.B. Ost-West oder 16 Bundesländer) Wir haben hier 58 Länder also werden wir 57 Dummy Variablen ins Modell geben. Wir müssen das zum Glück nicht von Hand machen sondern können das spezifizieren lassen weil wir eine Variable Country haben im Datensatz, wir

Es ist möglich, eine Dummy Variable für jedes Land zu erstellen, dann kann man `summary(country_fe_model)` schreiben und R gibt uns die Schätzungen für jede einzelne Variable.

• Fügen Sie einen Faktor mit den Ländernamen hinzu, der als Variable `country` benannt ist.
 Faktorvariable `country` erstellen
`country_fe_model = lm(institutions ~ oil + as.factor(country), data = resource curse)`

deshalb eigentlich genau gleich mit `lm`, der einzige Zusatz ist ich will das R mir ein Modell schätzt mit fixed effects für die Länder, wie erreiche ich das? Ich schreibe noch plus und `as.factor(country)` dazu. Was dann geschieht ist das für jedes Niveau dieses Faktors, je-

country_fe_model = lm(institutions ~ oil +
`as.factor(country), # the unit FE`
`data = resource curse)`

	##	Min	1Q	Median	3Q	Max
	##	-0.58818	-0.06665	-0.00369	0.06641	0.48773
	##					
	## Coefficients:					
	##					
	## (Intercept)					
	## oil					
	## as.factor(country)Argentina					
	## as.factor(country)Armenia					
	## as.factor(country)Australia					
	## as.factor(country)Azerbaijan					
	## as.factor(country)Belarus					
	## as.factor(country)Belgium					
	## as.factor(country)Benin					

Wir erhalten also einen gleichen output wie sonst auch. Einfach noch mit den verschiedenen Gammas drin und wir sehen auch gleich für welches Land ist das der Gamma. Ich kann also für jedes Land den Achsenabschnitt berechnen. Wenn ich zum Beispiel den Achsenabschnitt für Argentinien ermitteln will, rechne ich einfach die Konstante plus den Koeffizienten des Dummies für Argentinien, das ist dann der Achsenabschnitt für Argentinien.

Den Beta Koeffizienten für Öl interpretieren wir wie wir es bis anhin gelernt haben. Das ist die erwartete Veränderung in Y würde man dieses X um eine Einheit erhöhen. Konkret in diesem Beispiel wie verändert sich die institutionelle Qualität in Erwartung, wenn die Ressourceneinnahmen um einen Prozentpunkt ansteigen. Wir sehen wenn wir das mit dem fixed Modell machen um einen kausalen Zusammenhang zu interpretieren scheint es keinen Zusammenhang zu geben! (Oil -0.00)

```
# Wenn Sie die länderspezifischen fixed effects
# nicht in Ihrem Output sehen wollen
screenreg(country_fe_model,
          custom.coef.map = list("(Intercept)"="Intercept",
                                 "oil"="Oil"))
```

```
##
## =====
##               Model 1
## -----
## Intercept   -0.44 ***
##              (0.03)
## Oil        -0.00
##              (0.00)
## -----
## R^2         0.98
## Adj. R^2    0.98
```

Der zweite Punkt hier bezieht sich auf die Achsenabschnitte. Wie können wir diese interpretieren? Theoretisch so: Die durchschnittliche institutionelle Qualität beim ersten Land (bei unserem Basisland, der Referenzkategorie), wenn es kein Öl einnimmt. Das ist in der Praxis nicht so sinnvoll, weil dies gar nicht unsere grundsätzliche Frage war. Desto mehr Faktoren wir ins Modell aufnehmen, desto schwieriger oder unklarer wird eigentlich was dieser Achsenabschnitt bedeutet. Das ist wichtig, weil es viel Arbeit im Hintergrund im Modell macht aber der Koeffizient als solches hat nicht unbedingt eine relevante realweltliche Bedeutung.

Was bedeutet es ein Land konstant zu halten? Wir haben ja bis anhin confounders blockiert oder gesagt wir modellieren die, das ist so als würden wir diese anderen X konstant halten und dann schauen wir was geschieht da auf dieser anderen Variable. Jetzt haben wir das gemacht mit Länder fixed effects die uns erlauben unterschiedliche Konstanten zu schätzen. Wir haben also weiterhin parallele Linien wir haben einfach den Startpunkt geändert. Jedes Land erhält dann einen anderen Wert. Die numerische Berechnung des Achsenabschnittes für jedes einzelne Land wie wir es vorhin gesehen haben wäre das. Was wir machen ist wir nützen nicht mehr diese Unterschiede zwischen Ländern aus sondern wir nutzen die Varianz auf Y und X innerhalb eines Landes über die Zeit aus. Deshalb kommt dieser Schätzer auch mit einem anderen Namen das wird sehr oft als ein within estimator bezeichnet, weil der Schätzer identifiziert ist durch die Variation innerhalb der Einheiten und nicht über die Einheiten hinweg. Wie gesagt, es ist sehr schwierig die Größen der fixed effects zu interpretieren und je mehr Variablen ins Modell hinzugefügt werden.

Was haben wir nun mit den fixed effects wegkontrolliert? Oder was sind die Faktoren worin sich diese Länder unterscheiden können die aber nicht einen Einfluss ausüben können auf das geschätzte Beta für Öl.

Vorteile

- Fixed effects für alle zeitlich *konstanten* Variablen eines Landes.
 - Kolonialvergangenheit (grossteils)
 - Wetter / Klima (grossteils)
 - Kultur (grossteils)
 - Wohlstand (nicht wirklich)
- Über längere Zeiträume hinweg ist es weniger wahrscheinlich, dass potenzielle Confounder wirklich konstant bleiben.
 - Wenn ein unbeobachteter Confounder *fast* konstant bleibt, werden durch fixed effects zwar viele, aber nicht alle potenziellen Störeffekte (weg-)kontrolliert.

- Ermöglicht die Kontrolle für alle "zeitinvarianten" (d. h. über die Zeit konstanten) confounder

- Dies ist besonders wichtig, wenn es sich um Variablen handelt, die schwer oder gar nicht zu beobachten sind.

Nachteile

- Im Gegenzug dazu haben wir unsere Schätzungen geändert
 - Wir vergleichen nicht mehr verschiedene Länder.
 - Wir suchen nach Beziehungen innerhalb eines Landes zwischen X und Y im Zeitverlauf.

Der Vorteil ist ohne bestimmte confounders zu messen haben wir eine ganze Kategorie von ihnen wegkontrolliert. Alles, was sich nicht ändert über die Zeit in den Ländern und einen Effekt haben könnte ist einfach weg, das ist kein möglicher Störfaktor mehr denn wir hier haben. Das ist ein riesen Vorteil gegenüber wir müssen uns jeden confounder überlegen, messen und ins Modell tun. Die Konsequenzen dieses Modells sind wir schauen uns nicht mehr unterschiedliche Länder an sondern benutzen Variation inner-

halb der Länder um etwas über diesen Zusammenhang zu lernen. Das ist in diesem Beispiel sinnvoll, es gibt aber auch Beispiele wo uns eben gerade dies interessiert, also die Unterschiede die in den Ländern bestehen, z.B. wenn uns interessiert was sind die Effekte von politischen Institutionen, diese verändern sich sehr selten. Solche Fragen sind schwierig anzuschauen mit fixed effects in Paneldaten.

Was ist nun mit der Annahme das es überhaupt keine confounders mehr gibt? Das einzige was unit fixed effects macht das blockiert das Faktoren die über die Zeit konstant sind einen Einfluss auf den Koeffizientenschätzer haben können. Sind das die einzigen möglichen confounder? Wahrscheinlich nicht. Wir sehen zum Beispiel globale Ölpreisschocks, werden dieses Y verändern, diese betreffen alle Länder aber sind nicht über die Zeit konstant. Ein weiteres Beispiel wären Konflikte. Diese finden in einzelnen Ländern in einzelnen Jahren statt. Können wir also etwas machen gegen Faktoren die nicht unbedingt über die Länder variieren, die eigentlich für jedes Land in einem gegebenen Jahr konstant sind aber sich über die Zeit ändern? Eigentlich gleiche Logik wie mit den Ländern einfach aus der Zeitperspektive. Wir wollen also Faktoren die für die Länder in einem gegebenem Jahr konstant bleiben aber die über die Zeit sich ändern. Sind also konstant innerhalb eines Jahres aber verändern sich über die Jahre sich und diese müssen für alle Länder gleich sein. Ein gutes Beispiel dafür wäre der globale Ölpreis. Das können wir kontrollieren mit time fixed models. Mit diesen kontrollieren wir alle gemeinsamen Schocks, die alle Länder zur gleichen Zeit betreffen. Anstatt für jedes Land einen unterschiedlichen Achsenabschnitt zu modellieren, modellieren wir für jedes Jahr einen unterschiedlichen Achsenabschnitt.

Anstatt alle Dummy-Variablen auszuschreiben, definieren wir das Modell mit festen time-fixed-effects wie folgt:

$$Y_{i,t} = \delta_t + \beta_1 X_{1i,t} + \varepsilon_{i,t}$$

where

- $Y_{i,t}$ ist institutionelle Qualität für Land i im Jahr t
- $X_{1i,t}$ sind die Ölennahmen für Land i im Jahr t
- δ_t ist die Konstante für Zeit t
- β_1 ist der Koeffizientenschätzer für die Ölennahmen
- $\varepsilon_{i,t}$ ist der Fehlerterm für Land i im Jahr t

Delta t ist die jahresspezifische Konstante. Das einzige was wir dem Modell als Freiheit geben, ist das jedes Jahr ein anderer Achsenabschnitt vorliegen kann. Wir schliessen also für jedes Jahr ausser für das erste einen sogenannten Jahresdummy ins Modell ein.

Bei R geben wir nun nicht as.factor(country) an sondern as.factor(year) um die richtigen Dummy Variablen zu generieren.

Die Schätzung von time-fixed-effects in R ist genauso einfach wie die Schätzung von Länder-fixed-effects.

```
time_fe_model <- lm( institutions ~ oil +
  as.factor(year), # the time FE
  data = resource_curse
)
```

Der einzige Unterschied zum Modell mit länderspezifischen fixed effects besteht darin, dass wir jetzt für das Jahr kontrollieren, da wir an zeitlich fixed effects interessiert sind.

```
summary(time_fe_model)

##
## Call:
## lm(formula = institutions ~ oil + as.factor(year), data = resource_curse)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -1.6598 -0.5339 -0.2172  0.5678  1.8682 
## 
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)    
## (Intercept) 0.1252002  0.1048933  1.194    0.233    
## oil        -0.0282864  0.0030299 -9.336 <0.0000000000000002 ***
## as.factor(year)1998 -0.0755021  0.1471882 -0.513    0.608    
## as.factor(year)2000  0.0555346  0.1470753  0.378    0.706
```

Wir erhalten fast den gleichen Koeffizientenschätzer für oil wie beim gepoolten Modell. Das ist ein Anzeichen dafür das wir uns vor zeitlichen confounders nicht fürchten müssen, weil selbst wenn wir für diese blockieren erhalten wir de facto die genau gleiche Zahl wie wenn wir die Dummy Variablen auch ins Modell geben. Aber all das wissen wir nicht

- Wir können die Koeffizientenschätzungen nach der gleichen Logik wie zuvor interpretieren, wobei wir daran denken, dass wir jetzt das Jahr konstant halten..
- Das Modell mit zeitlichen fixed effects ergibt eine Schätzung für die Variable oil, die der des gepoolten Modells sehr ähnlich ist

	Pooled	Pooled	Country FE	Time FE
## Intercept	0.106 ** (0.034)	-3.356 *** (0.099)	-0.443 *** (0.035)	0.125 (0.105)
## Oil	-0.028 *** (0.003)	-0.015 *** (0.002)	-0.003 (0.001)	-0.028 *** (0.003)
## Log(GDP pc)		0.403 *** (0.013)		
## Polity IV		0.031 *** (0.003)		
## Aid		0.012 *** (0.003)		
## R^2	0.114	0.744	0.980	0.116
## Adj. R^2	0.113	0.742	0.979	0.100
## Num. obs.	682	672	682	682

- Die zeitlichen fixed effects bedeuten, dass die Regressionslinien für jedes Jahr die gleiche Steigung haben und parallel zueinander verlaufen.
- Mit anderen Worten, wir nutzen die Unterschiede zwischen den Ländern in Bezug auf oil und institutions innerhalb desselben Jahres, um die Auswirkungen der Ölannahmen (in Prozent des BIP) auf die institutionelle Qualität zu schätzen.
- Die Grösse der fixed effects sind weniger gut interpretierbar, desto mehr Variablen ins Modell hinzugefügt werden (analog zum Achsenabschnitt), aber die Unterschiede zwischen den Einheiten können nützliche Informationen liefern.

Genau gleich wie beim unit fixed modell macht es keinen Sinn die Grösse der fixed effects zu interpretieren. Diese beiden Modelle die wir nun kennengelernt haben ist eigentlich einfach der technische Schritt der uns erlaubt dafür zu sorgen das Faktoren die in einem Jahr konstant sind über alle Länder und sich nur über die Zeit verändern das solche Variablen nicht ein confounding darstellen für das Beta das uns inhaltlich interessiert. Oder umgekehrt mit den Ländern. Auch damit haben wir wahrscheinlich nicht alle confounder beseitigt.

Frage: Vergleichen wir Beobachtungen, die im Durchschnitt in Bezug auf alle Confounder identisch sind?

Antwort: Wahrscheinlich nicht! Es könnte viele andere mögliche Faktoren für Länder mit mehr/weniger Öl geben, die die Institutionen beeinflussen

→ Wenn wir doch nur die Möglichkeit hätten, zeit- und länderspezifische fixed effects gleichzeitig zu verwenden!

Das führt zur Frage gibt es eine Möglichkeit das wir beide Sachen kontrollieren können? Also zeitspezifische und länderspezifische fixed effects verwenden um im gleichen Modell diese zwei Kategorien von Variablen zu blockieren, also Variablen die nur über die Zeit variieren aber nicht über die Länder und Variablen die über die Länder aber nicht über die Zeit variieren. Das gibt es, das ist nichts anders als die Kombination dieser beiden Modelle.

->two way fixed effects, erlaubt uns beide Kategorien zu blockieren.

Wir geben dafür alpha i und delta t ins Modell. Das heisst ich zeige an es gibt ein Element der Konstante das über die Zeit (delta t) und ein Element das über die Länder variiert (alpha i). Danach wieder Beta 1 * X1i,t, man beachte wieder beim Beta kein Subskript, das ist eine Zahl die wir schätzen im Gesamtmodell, das ist die Zahl die uns substantiell interessiert, was ist die erwartete Veränderung in Y, wenn sich X um eine Einheit ändert. Das ist die Frage gibt es einen Kausalzusammenhang zwischen Ressourcenreichtum und institutioneller Qualität.

Wir definieren das two-way fixed effects Modell wie folgt:

$$Y_{i,t} = \alpha_i + \delta_t + \beta_1 X_{1i,t} + \varepsilon_{i,t}$$

wobei

```

twoway_fe_model <- lm( institutions ~ oil +
  as.factor(country) + # the unit FE
  as.factor(year), # the time FE
  data = resource_curse
)

```

Hinweis zum R output, unten an den Ländern kommt natürlich noch der output für die Jahre. Die Zahl welche uns hier wieder interessiert Ist der Koeffizientenschätzer für oil. Das ist die Zahl wo wir wenn möglich ein Modell schätzen Möchten welches uns erlaubt das kausal zu Interpretieren. Wenn wir uns das anschauen

Sind wir wieder irgendwo bei -0.0017, also irgendwie 14 Mal kleiner als wir in unserem allerersten Beispiel mit dem gepoolten Modell hatten. Das ist de facto eine flache Linie also der Zusammenhang ist eigentlich 0. Wir sehen also keine offensichtliche Evidenz dafür das Ressourcenreichtum, einen Effekt hat auf die Qualität der Institutionen. Wie bei den beiden Beispielen vorher gibt es auch hier wieder die Möglichkeit zu beschränken was im R output dargestellt wird. Einfach um es besser darzustellen ist diese Folie.

```

# Wenn Sie die fixed effects
# nicht in Ihrem Output sehen wollen
screenreg(twoway_fe_model,
  custom.coef.map = list("(Intercept)"="Intercept",
    "oil"="Oil"), digits=3)

##
## =====
## Model 1
## -----
## Intercept -0.409 ***
## (0.038)
## Oil -0.002
## (0.002)
## -----
## R^2 0.981
## Adj. R^2 0.979
## Num. obs. 682
## =====
## *** p < 0.001; ** p < 0.01; * p < 0.05

```

Wir können die Koeffizientenschätzer nach der gleichen Logik wie zuvor interpretieren, wobei wir daran denken, dass wir jetzt *Jahr und Land konstant* halten.

```

##
## =====
## Pooled Country FE Time FE Twoway FE
## -----
## Intercept 0.106 ** -0.443 *** 0.125 -0.409 ***
## (0.034) (0.035) (0.105) (0.038)
## Oil -0.028 *** -0.003 -0.028 *** -0.002
## (0.003) (0.001) (0.003) (0.002)
## -----
## R^2 0.114 0.980 0.116 0.981
## Adj. R^2 0.113 0.979 0.100 0.979
## Num. obs. 682 682 682 682
## =====

```

```

summary(twoway_fe_model)

##
## Call:
## lm(formula = institutions ~ oil + as.factor(country) + as.factor(year),
## data = resource_curse)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -0.42456 -0.06482  0.00200  0.06699  0.45720 
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)                 -0.408568  0.037613 -10.862
## oil                            -0.001730  0.001648 -1.049
## as.factor(country)Argentina  0.245153  0.048805  5.023
## as.factor(country)Armenia    0.103465  0.048717  2.124

```

Wir sehen die 4 Modelle die wir in dieser Vorlesung gesehen haben. Wir sehen der Koeffizient für oil ist jeweils plus minus Null wenn country fixed effects drin sind die time fixed effects machen eigentlich keinen grossen Unterschied gegenüber dem gepoolten Modell. Wie können wir nun die Konstante in den verschiedenen Modellen interpretieren? Weil wir sehen der Wert dieser springt herum. Diese Frage wird uns Leemann nicht stellen weil Konstanten wichtig sind wenn wir Werte vorhersagen, substantiell interessieren sie uns höchst selten. Was wir natürlich machen können müssen um eine Vorhersage zu erstellen ist wissen wie man ein solches Modell nimmt und den den richtigen Wert für zum Beispiel Brasilien 1990 generiert.

Was probieren wir hier eigentlich? Wir versuchen zwei grundlegende Kategorien von Störfaktoren auszuschalten. Also alles was innerhalb eines Land konstant bleibt über die Zeit und alles was innerhalb eines Jahres konstant bleibt über die verschiedenen Ländern, jede Variable die in eine dieser beiden Kategorien fällt, wird keinen bias also keine Verzerrung in unserem Beta erzeugen. Der Vorteil ist wir haben natürlich noch mehr Variablen blockiert und um die wir uns keine Sorgen machen müssen. Aber dafür haben wir auch einen grossen Teil der Variation weggenommen. Das bedeutet wenn ich auch andere erklärende Faktoren ins Modell einschliessen möchte, die beispielsweise nur über Länder variieren, z.B. mit proportionalem Wahlrecht ist dieser Zusammenhang viel schwächer, das kann ich jetzt nicht machen wenn wir eine Variable ins Modell geben die innerhalb eines Landes nicht variiert, dann fliegt dieses Modell raus, weil es perfekt multikolinear ist mit den Dummy Variablen. Es schränkt also ein was ich genau machen kann.

Die ultimative Frage zum Schluss nocheinmal, haben wir nun alle confounder erwischt die ein Problem darstellen können? Kann sein kann aber auch nicht sein. Gibt es nicht bestimmte Effekte die über die Zeit variieren und die die Länder unterschiedlich betreffen, wenn wir also so einen Faktor haben müssten wir diesen natürlich noch zusätzlich ins Modell geben um ihn kontrollieren zu können. Aber es ist natürlich schon eine viel angenehmere Ausgangslage als wenn wir diese Modelle nicht kennen würden und versuchen alles mögliche was uns in den Sinn kommt ins Modell zu integrieren.

Zusammenfassend:

- Institutionelle Qualität ist langsam sich verändern (“sticky”)
 - Sehr geringe Variation innerhalb eines Landes bei Y
- Der offensichtliche Effekt der Ölennahmen auf die Qualität der Institutionen verschwand grösstenteils, wenn man für das Land kontrolliert. Dies zeigt uns, dass die Beziehung im gepoolten Modell...
 - ...resultiert zumeist aus der Tendenz, dass Länder mit durchweg höheren Ölennahmen auch eine geringere institutionelle Qualität aufweisen.
 - ...spiegeln meist nicht die Tendenz wider, dass Länder mit wechselnden Ölennahmen ihre institutionelle Qualität stark verändern.
- Es gibt kaum Anhaltspunkte dafür, dass Zeitschocks (z. B. globale Ölpreise) in dieser speziellen Anwendung von Bedeutung sind.

VL 9

Wann kann man einen Regressionskoeffizienten kausal interpretieren?

1. **Randomisierte Experimente:** Der Koeffizient für ein binäres Treatment ist eine Schätzung des ATE
2. **Beobachtungsstudien:** Wir können Koeffizienten nur kausal interpretieren, wenn wir für alle **Confounders** als zusätzliche X-Variablen kontrolliert haben
3. **Vorher-Nachher Design:** Unter der Annahme, dass es keine zeitvariierenden Confounders gibt, können Sie den durchschnittlichen kausalen Effekt schätzen, indem Sie die Differenz der Mittelwerte des Ergebnisses vor und nach der Behandlung ermitteln.
4. **Difference-in-differences Design:** Unter der Annahme paralleler Trends zwischen Kontroll- und Treatment-Gruppe (falls Treatment-Gruppe das Treatment nicht erhalten hätte) kann man den durchschnittlichen kausalen Effekt schätzen, indem man die Differenz der mittleren Veränderung in der Treatment- und der Kontrollgruppe nimmt.
5. **Panel Daten Design:** Unter der Annahme, dass es keine zeit- *und* einheitsvariierenden Confounders gibt, können Sie sich auf die Variation zwischen Treatment und Outcome in einer Einheit konzentrieren (mit unit-fixed-effects und time-fixed-effects)