STATISTIK FÜR DIE SOZIALWISSENSCHAFTEN

Multiple lineare Regressionen

Asymmetrische Zusammenhänge zwischen metrischen Variablen

Meine Notizen:

Themen der Woche

- Erweiterung der Regression auf drei Variablen
- Regressionskoeffizienten der trivariaten Regression
- Der Determinationskoeffizient in der trivariaten Regression
- Signifikanztests in der trivariaten Regression
- Erweiterung der Regression auf mehr als drei Variablen

Meine Notizen:

Die Grenze der Bivariatheit überkommen: Trivariate lineare Regressionen

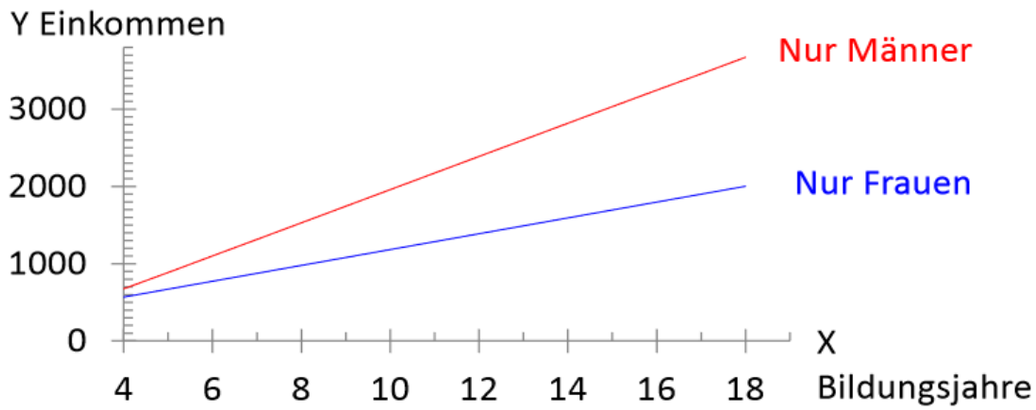
- Wollen wir den Einfluss einer dritten Variablen prüfen, ist es für Dichotome Drittvariablen leicht, einzelne Regressionen für beide Gruppen der Drittvariable zu rechnen. Wir nennen sie konditionale Regressionsmodelle. Es entstehen dabei aber mehrere Probleme:
  - Wir haben nicht mehr ein Modell zur Beschreibung von Zusammenhängen, sondern zwei. Was ist dann also die allgemeine Aussage?
  - Die beiden Regressionen sind nur schwer zu vergleichen.

  - Wie sollen Drittvariablen berücksichtigt werden, die nicht Dichotom, vielleicht sogar metrisch sind? Pro Ausprägung der Drittvariable müsste ein eigenes konditionales Modell gerechnet werden.

Meine Notizen:

Die Grenze der Bivariatheit überkommen: Konditionale Regression am Beispiel

Bsp: Ist Geschlecht eine relevante Drittvariable für den Einfluss von Bildung auf Einkommen?



Bivariate Regr.: nur Männer Y = Einkommen	b
Konstante	−184.31
Bildungsjahre	214.36

Daten: Allbus 2012, n=916

Bivariate Regr.: nur Frauen Y = Einkommen	b
Konstante	158.32
Bildungsjahre	102.47

Daten: Allbus 2012, n=781

- Es bleibt offen:
  - Wie groß ist der Einfluss von Geschlecht auf den Zusammenhang Bildung-Einkommen?
  - Ist der Einfluss des Geschlechts relevant?
  - Wie gut erklärt das konditionale Modell das Einkommen einer Person?

Meine Notizen:

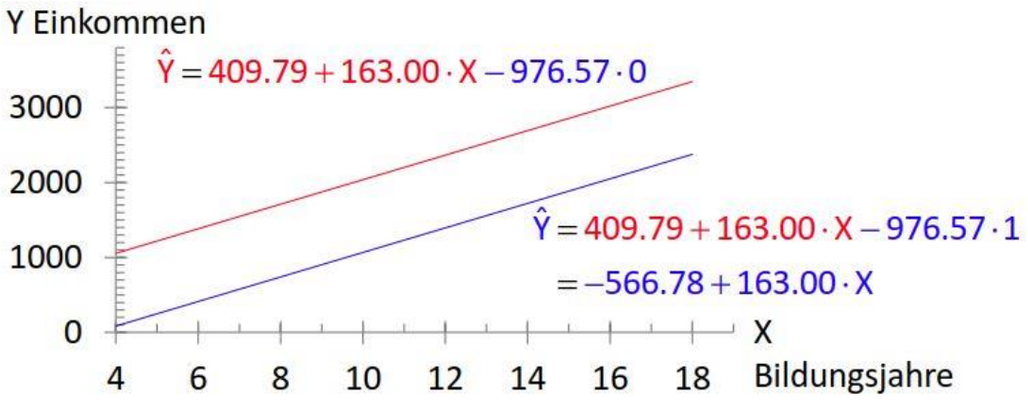
Trivariate lineare Regressionen

- Zur Behebung dieser Probleme lassen sich die beiden konditionalen Regressions-modelle zu einem gemeinsamen trivariaten Regressionsmodell zusammenführen.
- In der trivariaten Regression wird eine abhängige Variable (AV) durch zwei unabhängige Variablen (UVs) erklärt:

  Bedingte Mittelwerte / Vorhersagewerte der abhängigen Variable Y

  = lineare Funktion von 2 (erklärenden) Variablen X und W.

$$Y = \underbrace{b_0 + b_1 \cdot X + b_2 \cdot W}_{=\hat{Y}} + E$$



Trivariate Regression Y = Einkommen	b
Konstante	409.79
Bildungsjahre	163.00
Geschlecht	–976.57

Daten: Allbus 2012, n=1701

Meine Notizen:

Vorhersagen durch Trivariate lin. Regressionen

- Durch Einsetzen in die Regressionsgleichung können wie im bivariaten Fall Vorhersagen für bestimmte Personen oder Gruppen getroffen werden:

z. B.: *männliche (W=0) Person mit Realschulabschluss (X= 10):*

$\hat{Y}$ = 409.79 +163.00·**10** –976.57·**0** = 2039.79 € *= prognostiziertes Einkommen bei männlichen Realschulabsolventen*

$\hat{Y}$ = 409.79 +163.00· **9** –976.57·**0** = 1876.79 € *= prognostiziertes Einkommen bei männlichen Hauptschulabsolventen*

Differenz: **163.00** € = b_1

→ Interpretation Regressionsgewicht: **b_1 gibt Veränderung an, wenn X um +1 Einheit ansteigt**!

$\hat{Y}$ = 409.79 +163.00·10 –976.57·**1** = 1063.22 € *= prognostiziertes Einkommen bei **weiblichen** Realschulabsolventinnen*

$\hat{Y}$ = 409.79 +163.00·10 –976.57·**0** = 2039.79 €*= prognostiziertes Einkommen bei **männlichen** Realschulabsolventen*

Differenz: **–976.57** € = b_2

→ Interpretation Regressionsgewicht: **b_2 gibt Veränderung an, wenn W um +1 Einheit ansteigt**!

Meine Notizen:

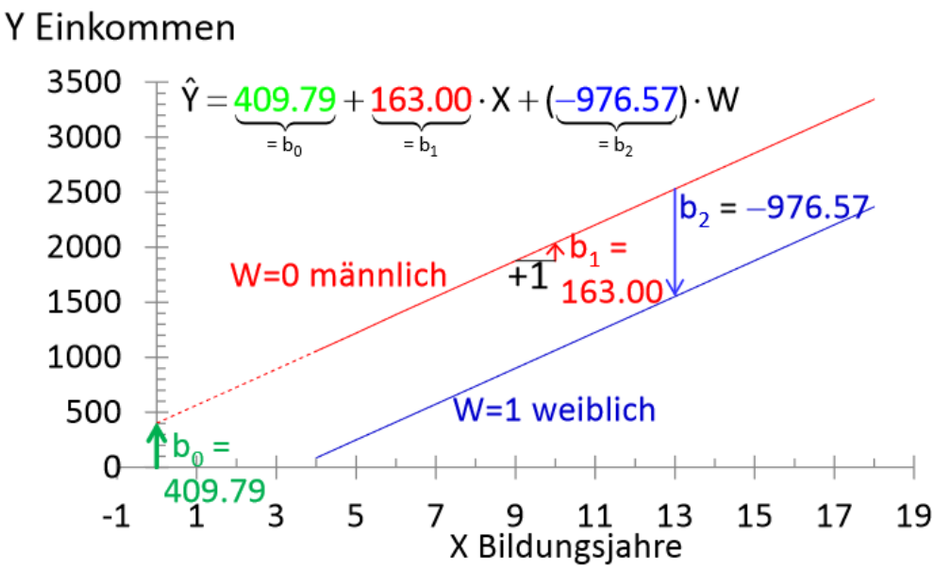
Bedeutung der Regressionskoeffizienten

- Ein trivariates Regressionsmodell liefert wie das bivariate eine Vorhersage von Y, in diesem Fall auf Basis der Vorinformation in zwei anderen Variablen:

b_0 = Regressionskonstante/Intercept: Vorhersagewert, wenn X=0 und W=0;

b_1 = Regressionsgewicht von X: Vorhersagewertveränderung, wenn X um +1 Einheit ansteigt bei Kontrolle von W;

b_2 = Regressionsgewicht von W: Vorhersagewertveränderung, wenn X um +1 Einheit ansteigt bei Kontrolle von X.



Meine Notizen:

Lineradditivität in Trivariater lin. Regressionen

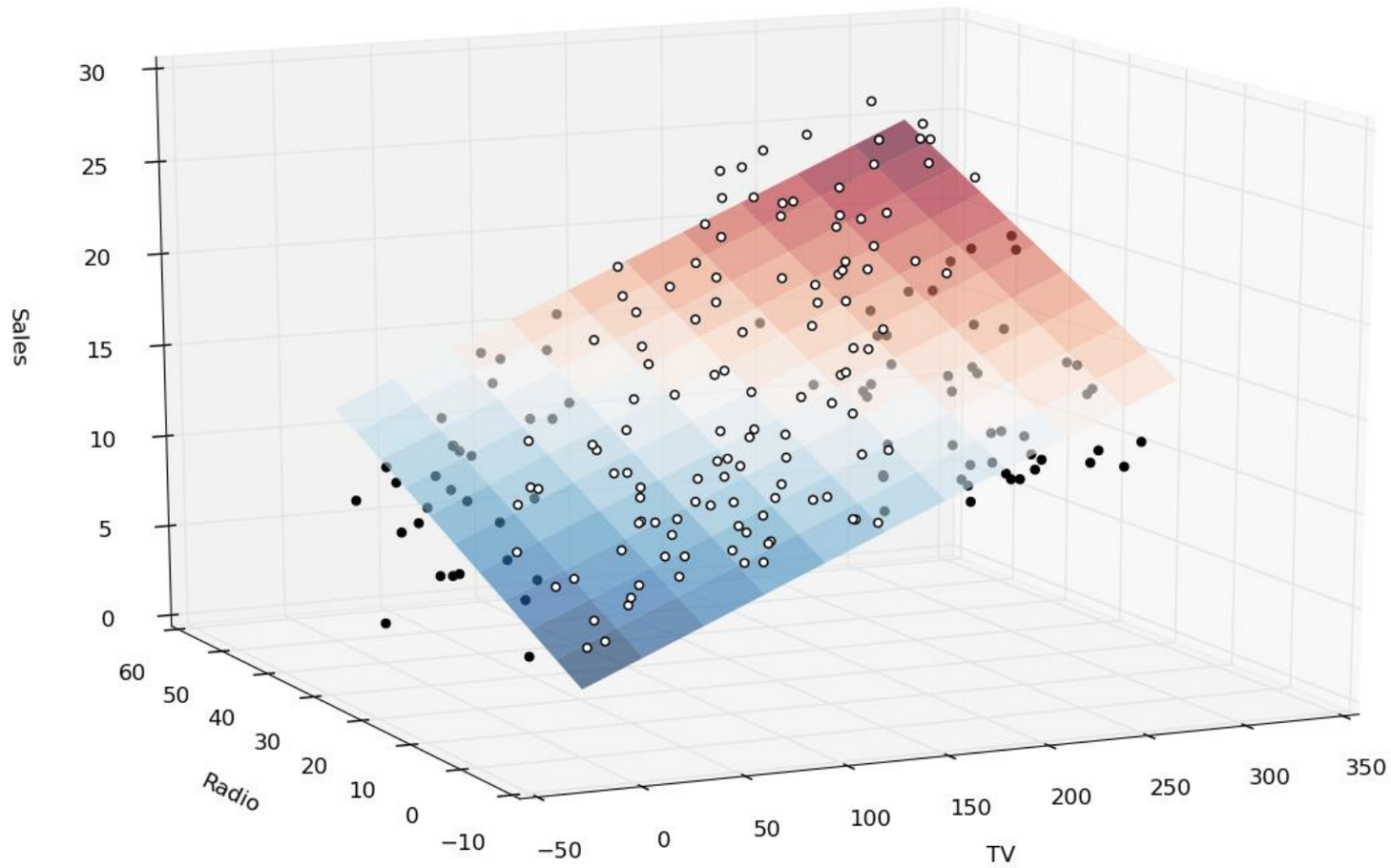
- Jede unabhängige Variable ist jeweils Drittvariable für die andere unabhängige Variable.
- Bivariate Regressionsmodelle einer unabhängigen Variable (gegeben ein Wert der jeweils anderen unabhängigen Variable) unterscheiden sich nur bei der Regressions-konstanten. Im Beispiel bedeutet das: Mehr Bildung wirkt in beiden Geschlechtern gleich. Sollte Bildung unterschiedliche Effekte haben, kann das hier nicht gemessen werden (Linearadditivität).
- die bedingten Regressionsgewichte für eine gegebene Drittvariable sind i.A. ≠ bivariate Regressionsgewichte

Meine Notizen:

Trivariate lin. Regressionen graphisch darstellen

- Ist die zu prüfende Drittvariable metrisch, funktioniert alles wie gesehen.

  Aufwendiger ist nur die graphische Darstellung:



Meine Notizen:

Trivariate Regressionsgewichte bestimmen

- Die Bestimmung der Regressionsgewichte folgt der Grundidee des bivariaten Falles.

  Jedoch müssen nun alle Kovariationen berücksichtigt werden:

$$b_1 = \frac{SS_W * SP_{YX} - SP_{YW} * SP_{XW}}{SS_W * SS_X - (SP_{XW})^2} = \frac{s_W^2 * s_{YX} - s_{YW} * s_{XW}}{s_W^2 * s_X^2 - (s_{XW})^2} = \frac{\hat{\sigma}_W^2 * \hat{\sigma}_{YX} - \hat{\sigma}_{YW} * \hat{\sigma}_{XW}}{\hat{\sigma}_W^2 * \hat{\sigma}_X^2 - (\hat{\sigma}_{XW})^2}$$

$$b_2 = \frac{SS_X * SP_{YW} - SP_{YX} * SP_{XW}}{SS_W * SS_X - (SP_{XW})^2} = \frac{s_X^2 * s_{YW} - s_{YX} * s_{XW}}{s_W^2 * s_X^2 - (s_{XW})^2} = \frac{\hat{\sigma}_X^2 * \hat{\sigma}_{YW} - \hat{\sigma}_{YX} * \hat{\sigma}_{XW}}{\hat{\sigma}_W^2 * \hat{\sigma}_X^2 - (\hat{\sigma}_{XW})^2}$$

$$b_0 = \bar{y} - b_1 * \bar{x} - b_2 * \bar{w}$$

Meine Notizen:

Der Determinationskoeffizient R^2 in der trivariaten Regression

- Natürlich trifft auch die trivariate Schätzfunktion nicht jeden Punkt y genau.
- Die genauen Ausprägungen von y lassen sich darstellen als

$$y_i = b_0 + b_1 * x + b_2 * w + e_i = \hat{y} + e_i$$

- Wie im bivariaten Fall können wir also die Variation von Y zerlegen in einen Teil, der durch X und W erklärt werden kann und einen verbleibenden Fehler:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Variation von } y} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Variation der Residuen}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Variation der Regresswerte}}$$

- Wir können die Frage stellen, wie viel Prozent der Variation von Y durch X und W erklärt werden kann und die Güte des Modells mit diesem Wert beschreiben.

Meine Notizen:

Determinationskoeffizient R^2

- Der Anteil der Variation von Y, der durch Variation von X erklärt werden kann, ist ein Qualitätskriterium für Regressionen. Er heißt Determinationskoeffizient R^2 oder auch Bestimmtheitsmaß.
- Da: Gesamtvariation von Y = erklärte Variation + nicht erklärte Variation gilt:

$$R^2 = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Meine Notizen:

R^2 in der trivariaten Regression: Die Berechnung

- Wie gesehen gilt:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{Variation von } y} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{Variation der Residuen}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{Variation der Regresswerte}}$$

- Das heißt kurz gesagt: $SS_Y = SS_E + SS_{\hat{Y}}$

- SS_Y ist bekannt. Außerdem gilt:

$$SS_{\hat{Y}} = b_1^2 * SS_X + b_2^2 * SS_W + 2 * b_1 * b_2 * SP_{XW} = b_1 * SP_{XY} + b_2 * SP_{WY}$$

$$s_{\hat{Y}}^2 = b_1^2 * s_X^2 + b_2^2 * s_W^2 + 2 * b_1 * b_2 * s_{XW} = b_1 * s_{XY} + b_2 * s_{WY}$$

- Daraus lässt sich dann zusammenfassen:

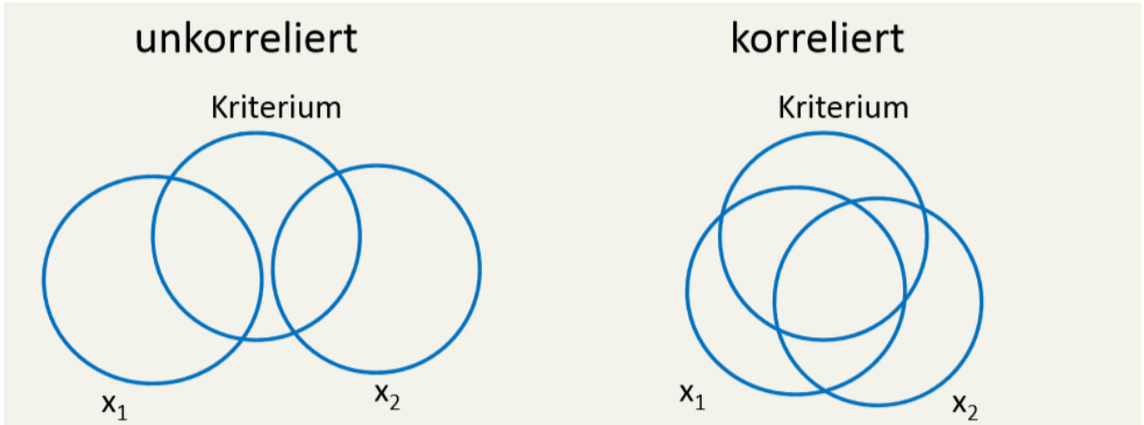
$$R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{b_1 * SP_{XY} + b_2 * SP_{WY}}{SS_Y} = \frac{b_1 * s_{XY} + b_2 * s_{WY}}{s_Y^2}$$

$$= 1 - \frac{SS_E}{SS_Y} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}$$

Meine Notizen:

Besonderheiten des Determinationskoeffizient R^2 in der trivariaten Regression

- Der Determinationskoeffizient liefert also auch im trivariaten Fall eine Prozentangabe, wie viel Variation von Y das Modell erklärt.
- Da durch die dritte Variable eine Mehrinformation gegeben ist gegenüber dem bivariaten Modell, kann die Erklärung von Y nur besser werden als beide R^2 der bivariaten Regressionen.
- Gleichzeitig kann das R^2 des trivariaten Modells nicht besser sein, als die Summe der beiden bivariaten R^2.



Meine Notizen:

Regressionsgewichte trivariater Reg. schätzen

- Im Schätzen von Populationswerten werden Punkt- und Intervallschätzung unterschieden. Zuverlässiger ist das Kofidenzintervall:
- Konfidenzintervall = Parameter der Stichprobe +/- Standardfehler * Quantilwert
- Wichtig zu kennen ist also der Standardfehler eines Regressionsgewichtes. Er berechnet sich mit folgenden Formeln:

$$\hat{\sigma}(b_1) = \sqrt{\frac{SS_W}{SS_X * SS_W - (SP_{XW})^2} * \frac{SS_E}{n-3}}$$

$$\hat{\sigma}(b_2) = \sqrt{\frac{SS_X}{SS_X * SS_W - (SP_{XW})^2} * \frac{SS_E}{n-3}}$$

Meine Notizen:

Regressionsgewichte trivariater Reg. testen

- Der Test eines Regressionsgewichts erfolgt wie jeder Test über die Teststatistik.
- Wie schon beim Test bivariater Regressionsgewichte kann neben dem Test gegen null auch jeder andere feste Wert leicht getestet werden, z.B. ob das Regressionsgewicht β_1 größer ist als der feste Wert β=0,2.
- Die Teststatistik eines Regressionsgewichts berechnet sich als

$$T = \frac{b_i - \beta}{\hat{\sigma}(b_i)}$$

- Liefert der Test gegen null ein Verwerfen der Nullhypothese, so spricht man von einem signifikanten Regressionsgewicht bzw. einem signifikanten Einfluss.

Meine Notizen:

Die Grenze der Trivariatheit überkommen: Multiple lineare Regressionen

- Wollen wir den Einfluss von mehr als zwei unabhängigen Variablen auf eine abhängige Variable püfen, stehen wir zunächst vor dem gleichen Phänomen, wie in der Entwicklung der trivariaten Regression:
  - Wir haben nicht mehr ein Modell zur Beschreibung von Zusammenhängen. Was ist dann also die allgemeine Aussage? Wie lassen sich die Modelle vergleichen?
  - Wie sollen weitere Variablen berücksichtigt werden, die nicht Dichotom, vielleicht sogar metrisch sind? Pro Ausprägung der Variablen müsste ein eigenes konditionales Modell gerechnet werden.

Meine Notizen:

Multiple lineare Regressionen

- Wie bereits von bivariat zu trivariat lassen sich Regressionsmodelle einfach erweitern:

$$Y = \underbrace{b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_K \cdot X_K}_{=\hat{Y}} + E = \underbrace{b_0 + \sum_{k=1}^{K} b_k \cdot X_k}_{=Y} + E$$

- Wir erweitern in diesem Sinne unser Beispiel:

  Unabhängige Variablen (Prädiktoren):

  X_1 vorgesehene Bildungsjahre

  X_2 Geschlecht: 0 = männlich, 1 = weiblich

  X_3 Alter in Jahren

  X_4 durchschnittliche Wochenarbeitszeit

  X_5 Region: 0 = Westen; 1 = Osten;

Abhängig: Persönliches Netto-Einkommen (Y)

Prädiktor	Regressionskoeffizient
Konstante	$b_0 = -2034.98$
Bildungsjahre (X_1)	$b_1 = 149.11$
Geschlecht (weiblich; X_2)	$b_2 = -604.68$
Alter (in Jahren; X_3)	$b_3 = 23.51$
Wochenarbeitszeit (in Std., X_4)	$b_4 = 38.79$
Region (Ost; X_5)	$b_5 = -377.89$

Daten: Allbus 2012, nur Berufstätige, n=1697

Meine Notizen:

Interpretation der partiellen Regressionsgewichte

- Bei linear-additiver Kontrolle von Geschlecht, Alter in Jahren, Wochenarbeitszeit und Region steigt das durchschnittliche Nettoeinkommen pro Bildungsjahr um 149.11 € an.

⇒ Personen mit Hauptschulabschluss (9 Schuljahre) haben ein um 745.55 € höheres Einkommen als Personen ohne Schulabschluss (4 Schuljahre): 745.55 = 149.11 · (9-4)

⇒ Personen mit Abitur (13 Bildungsjahre) haben im Durchschnitt ein um 447.33 € =149.11 · (13-10) höheres Einkommen als Personen mit Realschulabschluss

⇒ Personen mit MA (18 Bildungsjahre) haben im Durchschnitt ein um 298.22 € =149.11 · (18-16) höheres Einkommen als Personen mit BA.

Abhängig: Persönliches Netto-Einkommen (Y)	
Prädiktor	Regressionskoeffizient
Konstante	$b_0 = -2034.98$
Bildungsjahre (X_1)	$b_1 = 149.11$
Geschlecht (weiblich; X_2)	$b_2 = -604.68$
Alter (in Jahren; X_3)	$b_3 = 23.51$
Wochenarbeitszeit (in Std., X_4)	$b_4 = 38.79$
Region (Ost; X_5)	$b_5 = -377.89$

Daten: Allbus 2012, nur Berufstätige, n=1697

Meine Notizen:

Interpretation der partiellen Regressionsgewichte

- Bei linear-additiver Kontrolle von Bildungsjahren, Alter in Jahren, Wochenarbeitszeit und Region sinkt das durchschnittliche Nettoeinkommen bei einem Anstieg des Geschlechts um +1 Einheit um -604,68 €.

Abhängig: Persönliches Netto-Einkommen (Y)	
Prädiktor	Regressionskoeffizient
Konstante	b_0 = –2034.98
Bildungsjahre (X_1)	b_1 = 149.11
Geschlecht (weiblich; X_2)	b_2 = –604.68
Alter (in Jahren; X_3)	b_3 = 23.51
Wochenarbeitszeit (in Std., X_4)	b_4 = 38.79
Region (Ost; X_5)	b_5 = –377.89

Daten: Allbus 2012, nur Berufstätige, n=1697

⇒ Da bei Geschlecht „männlich“ mit dem Wert 0 und „weiblich“ mit dem Wert 1 kodiert ist, bedeutet dies, dass Frauen auch bei Kontrolle der berücksichtigten Drittvariablen im Durchschnitt 604.68 € weniger verdienen als Männer.

- Bei linear-additiver Kontrolle von Bildungsjahren, Geschlecht, Wochenarbeitszeit und Region steigt das durchschnittliche Nettoeinkommen pro Lebensjahr um 23.51 € an: Wenn eine Person 10 Jahre älter ist, verdient sie 235,10 € mehr.

Meine Notizen:

Interpretation der partiellen Regressionsgewichte

- Eine Regression gibt bedingte Mittelwerte von Y für gegebene unabhängige Variablen X_1 bis X_K : Wenn sich zwei Gruppen nur dadurch unterscheiden, dass in einer Gruppe die Ausprägung bei X_i um +1 Einheit höher ist, dann ist der Mittelwert von Y in dieser Gruppe um b_i Einheiten höher.

Abhängig: Persönliches Netto-Einkommen (Y)	
Prädiktor	Regressionskoeffizient
Konstante	b_0 = –2034.98
Bildungsjahre (X_1)	b_1 = 149.11
Geschlecht (weiblich; X_2)	b_2 = –604.68
Alter (in Jahren; X_3)	b_3 = 23.51
Wochenarbeitszeit (in Std., X_4)	b_4 = 38.79
Region (Ost; X_5)	b_5 = –377.89

Daten: Allbus 2012, nur Berufstätige, n=1697

- Eine Regressionsfunktion beschreibt den (gemeinsamen) kausalen Effekt der erklärenden Variablen auf die kausal beeinflusste, abhängige Variable:
  Wenn X_i um +1 Einheit ansteigt, verändert sich Y (im Durchschnitt) um +b_i . Die kausale Interpretation setzt voraus, dass der kausale Zusammenhang korrekt erfasst ist und es keine Konfundierung durch unberücksichtigte Drittvariablen gibt!

Meine Notizen:

Vorhersagen in multiplen linearen Regressionen

- Einsetzen der Ausprägungen für alle unabhängigen Variablen: z.B.: Vorhersage für 25jährige (X_3=25) Frau (X_2=1) mit Abitur (X_1=13), die 40 Stunden Woche arbeitet (X_4=40) die in einem der alten Bundesländer lebt (X_5=0):

Abhängig: Persönliches Netto-Einkommen (Y)	
Prädiktor	Regressionskoeffizient
Konstante	b_0 = –2034.98
Bildungsjahre (X_1)	b_1 = 149.11
Geschlecht (weiblich; X_2)	b_2 = –604.68
Alter (in Jahren; X_3)	b_3 = 23.51
Wochenarbeitszeit (in Std., X_4)	b_4 = 38.79
Region (Ost; X_5)	b_5 = –377.89

Daten: Allbus 2012, nur Berufstätige, n=1697

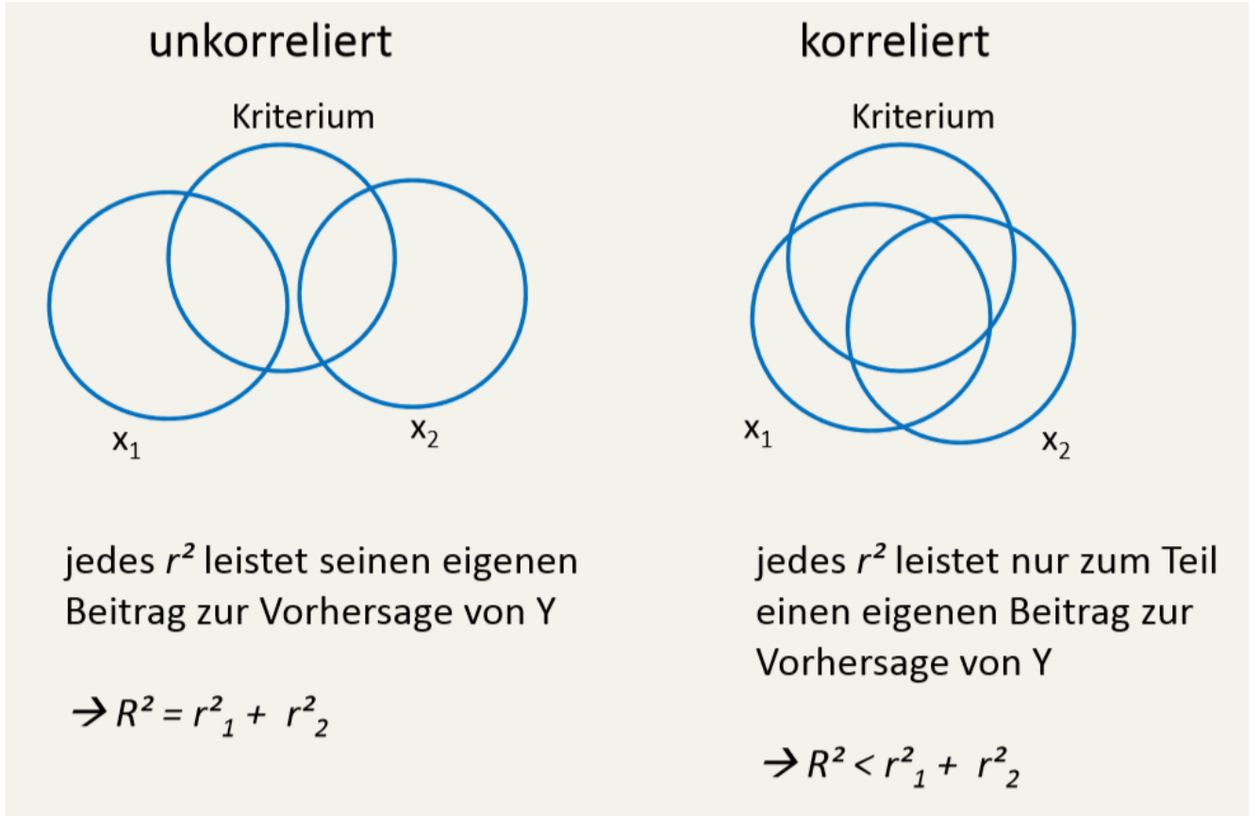
$$\hat{Y} = -2034.98 + 149.11 \cdot \underbrace{13}_{x_1} - 604.68 \cdot \underbrace{1}_{x_2} + 23.51 \cdot \underbrace{25}_{x_3} + 38.79 \cdot \underbrace{40}_{x_4} - 377.89 \cdot \underbrace{0}_{x_5} = 1438.12€$$

- Die Prognose liefert zwei Aussagen:
  - Sie ist das prognostizierte Mittel aller Personen, auf die die Angaben für die X zutreffen.
  - Sie ist die beste Prognose für ein Individuum, auf das die Angaben für die X zutreffen.

Meine Notizen:

Besonderheiten des Determinationskoeffizient R^2 in der multiplen Regression

- Der Determinationskoeffizient liefert also auch im multiplen Fall eine Prozentangabe, wie viel Variation von Y das Modell erklärt.
- Da durch die multiplen Variablen eine Mehrinformation gegeben ist gegenüber dem bivariaten Modell, kann die Erklärung von Y nur besser werden als alle R^2 der bivariaten Regressionen.
- Gleichzeitig kann das R^2 des multiplen Modells nicht besser sein, als die Summe der bivariaten R^2.



Meine Notizen:

Weitergehendes zur multiplen lin. Regressionen

- Eine graphische Vorstellung zu Regressionen mit mehr als zwei unabhängigen Variablen ist schwierig, da jede Variable eine Dimension im Graphen bedeuten würde. Graphen mit mehr als drei Dimensionen sind zwar möglich, aber schwer und wenig anschaulich.
- Die händische Berechnung von Regressionsgewichten, Determinationskoeffizienten, Standardfehlern und Teststatistiken ist sehr aufwendig. Wir werden daher in diesem Kurs darauf verzichten. Freuen Sie sich lieber darauf, dass Sie in Statistik IV lernen werden, wie Software Ihnen das abnimmt. Die Interpretationen hingegen funktionieren leicht, da sie sich wie im trivariaten Fall verhalten.

Meine Notizen:

Was Sie am Ende der Woche können sollten

- **Kern**: Sie bestimmen und interpretieren lineare Regressionsmodelle mit mehr als einer UV und beleuchten die Drittvariablenkontrolle darin.
- Sie bestimmen eine trivariate Regressionsgleichung und interpretieren diese.
- Sie berechnen und interpretieren den Determinationskoeffizienten im trivariaten Fall.
- Sie bestimmen und interpretieren Standardfehler und Teststatistik der trivariaten Regression.
- Sie verstehen die Erweiterung der Regression auf multiple Variablen.
- Sie interpretieren multiple Regressionsmodelle.
- Sie erklären das Prinzip der Linearadditivität.
- Sie diskutieren Drittvariablenkontrolle in multiplen Modellen.

Meine Notizen: