Kurzzusammenfassung Lineare Einfachregression

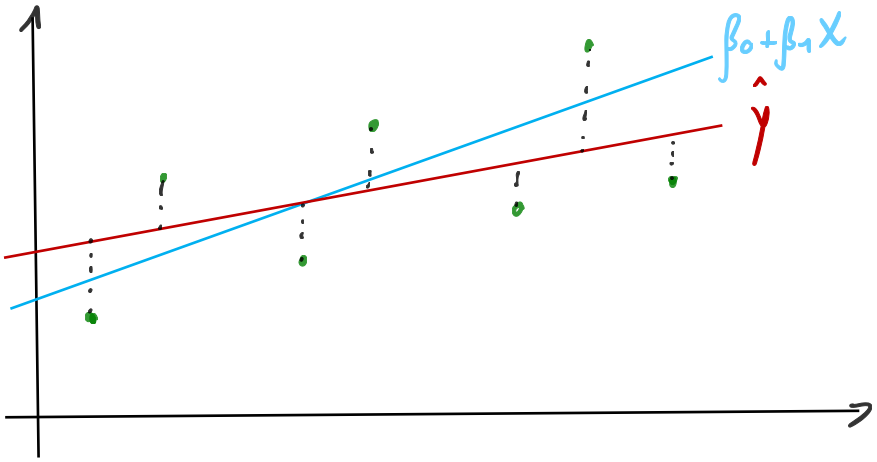
- Annahme:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

- Modell:

$$\hat{y} = b_0 + b_1 X \quad \text{und} \quad y = b_0 + b_1 X + e,$$

wobei $e = y - b_0 - b_1 X :=$ Residuen



$$y = b_0 + b_1 X + e$$

(1) Wir finden b_0, b_1, indem wir

$$SQR = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = N \cdot MSE$$

minimieren (in Abh. von b_0', b_1').

(2) b_0, b_1 sind gegeben durch:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\widehat{Cov}(y,x)}{\widehat{Var}(x)}$$

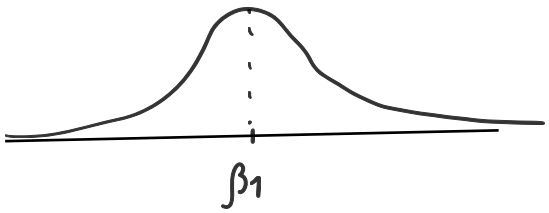
(3) b_0, b_1 sind Zufallsvariablen, da sie von Y abhängen, das aus der Beziehung $Y = \beta_0 + \beta_1 X + \varepsilon$ entsteht und somit auch eine Zufallsvariable ist (wegen ε).

↳ Wir werden also die statistische Maschinerie auf b_0, b_1 an, um ihre Verteilung herzuleiten (samt Erwartungswert, Varianz). So können wir die Unsicherheit der Schätzer b_0, b_1 quantifizieren.

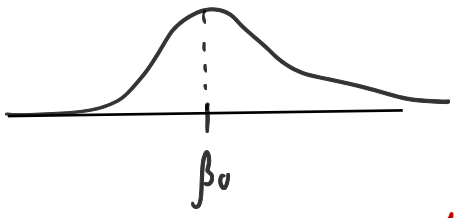
↳ ZGS ($N-2 > 30$):

$E[b_1]$ $\quad$ $Var(b_1)$

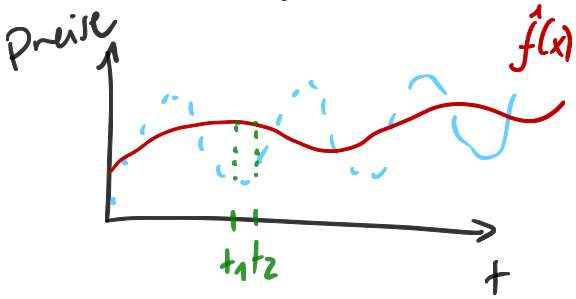
$$b_1 \sim Normal\left(\beta_1, \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)$$



$$b_0 \sim Normal\left(\beta_0, \hat{\sigma}^2\left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]\right)$$



$E(b_0)$ $\quad$ $Var(b_0)$



, wobei $\hat{\sigma}^2 := s^2 = \frac{1}{N-2}\sum_{i=1}^{N} e_i^2$

↳ Anmerkung: Die Varianzen halten so nur, falls

(i) Die Fehler unabhängig sind
(ii) Die Fehler eine konstante Varianz haben
(iii) Die X fix sind (nicht zufällig)

(4) Somit können wir Hypothesentests für β_0, β_1 durchführen und auch Konfidenzintervalle für β_0, β_1 berechnen (Nächste Übung).

Einführung in die Empirische Wirtschaftsforschung

Übungsaufgaben 4

Lineare Einfachregression - Erwartungswert und Varianz der KQ-Schätzer

1. Betrachten Sie den Datensatz "werbung.csv" aus der Vorlesung. Regressieren Sie mit einer linearen Einfachregression in R "sales" (Y) auf "TV" (X).

   a) Interpretieren Sie Ihren R Output (Koeffizienten und Standardfehler).

   b) Plotten Sie die Regressionsgerade im Streuungsdiagramm der Daten.

   c) Ist die lineare Einfachregression ein geeignetes Modell?

2. Betrachten Sie das Modell

$$x_i = \beta_0 + u_i, \quad i = 1, \ldots, N$$

wobei die u_i eine Zufallsstichprobe von einer Verteilung mit Erwartungswert 0 und (unbekannter) Varianz σ^2 darstellen.

   a) Was ist die Interpretation von β_0?

   b) Finden Sie den Kleinste-Quadrate-Schätzer für β_0. Schon mal gesehen?

3. Eine Forscherin möchte herausfinden, wie der Ernteertrag einer Pflanze, Y auf die Menge des verabreichten Düngers, X reagiert. Sie hat dazu $N = 10$ Parzellen (identischen) Bodens zur Verfügung. Auf der i-ten Parzelle wird sie die Menge x_i an Dünger verabreichen, wobei jeweils $x_i \in [0, 100]$ sein muss (in einer angemessenen Einheit).

   a) Die Forscherin geht davon aus, dass es eine lineare Beziehung gibt der Art

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, N$$

      Wie sollte die Forscherin $x_1, \ldots, x_{10}$ wählen, um die Varianz des Kleinste-Quadrate Schätzers b_1 so klein wie möglich zu machen?

   b) Was ist die Gefahr dieser Lösung im Falle, dass die Beziehung eventuell nichtlinear ist?

4. Zeigen Sie, dass die geschätzte lineare Regressionsgerade immer durch den Punkt $(\bar{X}, \bar{Y})$ geht.

5. Betrachten Sie das vereinfachte Regressions-Modell ("Gerade durch den Ursprung")

$$y_i = \beta_1 x_i + u_i, \quad i = 1, \ldots, N.$$

a) Zeigen Sie, dass der KQ-Schätzer b_1 erwartungstreu ist. Welche Annahmen haben Sie für den Beweis benötigt?

b) Zeigen Sie, dass die Varianz von b_1 kleiner oder gleich gross ist im Vergleich zum allgemeinen Modell ("Gerade mit Achsen-Abschitt"), und in der Regel kleiner.

1. Betrachten Sie den Datensatz “werbung.csv” aus der Vorlesung. Regressieren Sie mit einer linearen Einfachregression in R “sales” (Y) auf “TV” (X).

a) Interpretieren Sie Ihren R Output (Koeffizienten und Standardfehler).

b) Plotten Sie die Regressionsgerade im Streuungsdiagramm der Daten.

c) Ist die lineare Einfachregression ein geeignetes Modell?

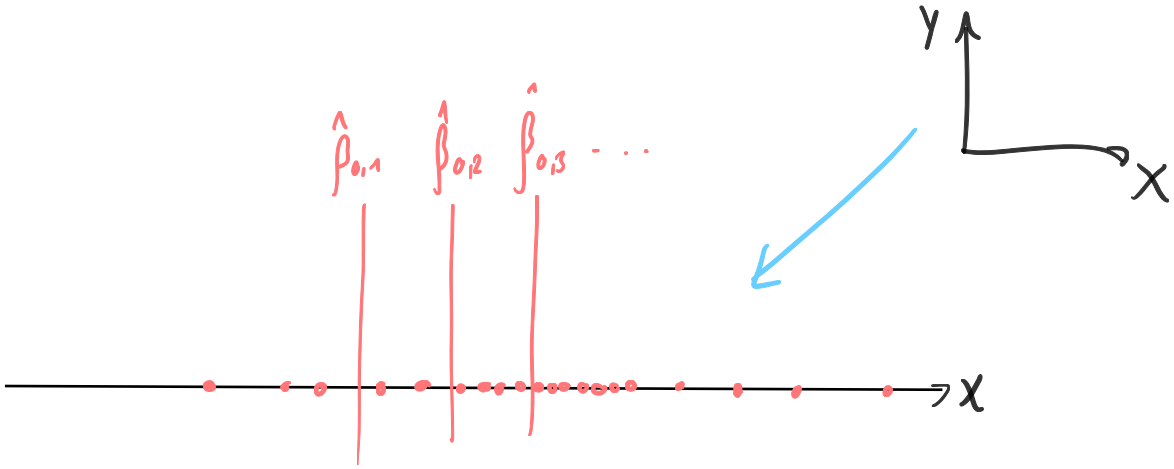
2. Betrachten Sie das Modell

$$x_i = \beta_0 + u_i, \quad i = 1, \ldots, N \qquad , u_i \sim F(0, \sigma^2)$$

wobei die u_i eine Zufallsstichprobe von einer Verteilung mit Erwartungswert 0 und (unbekannter) Varianz σ^2 darstellen.

a) Was ist die Interpretation von β_0?

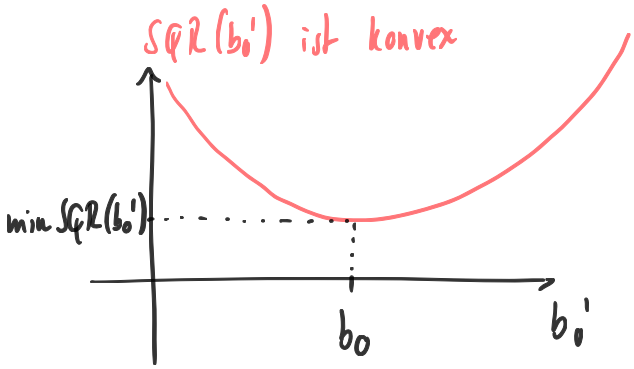
b) Finden Sie den Kleinste-Quadrate-Schätzer für β_0. Schon mal gesehen?



a) $E(x_i) = E(\beta_0 + u_i) = E(\beta_0) + \underbrace{E(u_i)}_{=0} = \beta_0$

b) $b_0 = \underset{b_0'}{\operatorname{argmin}} \sum_{i=1}^{N} (x_i - b_0')^2$

$= \underset{b_0'}{\operatorname{argmin}} \; SQR(b_0')$ $\quad g(f(b_0'))$



$$\frac{dSQR(b_0')}{db_0'} = 2 \sum_{i=1}^{N} (x_i - b_0)(-1) \overset{!}{=} 0$$

$$\sum_{i=1}^{N} x_i - \underbrace{\sum_{i=1}^{N} b_0}_{N \cdot b_0} \overset{!}{=} 0 \quad \Rightarrow \quad b_0 = \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{x}$$

3. Eine Forscherin möchte herausfinden, wie der Ernteertrag einer Pflanze, Y auf die Menge des verabreichten Düngers, X reagiert. Sie hat dazu $N = 10$ Parzellen (identischen) Bodens zur Verfügung. Auf der i-ten Parzelle wird sie die Menge x_i an Dünger verabreichen, wobei jeweils $x_i \in [0, 100]$ sein muss (in einer angemessenen Einheit).

a) Die Forscherin geht davon aus, dass es eine lineare Beziehung gibt der Art

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, N$$

Wie sollte die Forscherin $x_1, \ldots, x_{10}$ wählen, um die Varianz des Kleinste-Quadrate Schätzers b_1 so klein wie möglich zu machen?

b) Was ist die Gefahr dieser Lösung im Falle, dass die Beziehung eventuell nichtlinear ist?

Idee: $y = f(x) + \varepsilon$

a) Repetition:

$$Var(b_0) = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right]$$

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$\rightarrow$ Sie muss $\sum_{i=1}^{N} (x_i - \bar{x})^2$ maximieren. Dazu wählt sie

$x_1 = x_2 = \ldots = x_5 = 0$

$x_6 = x_7 = \ldots = x_{100} = 100$

$\Big\} \bar{x} = 50 \rightarrow \sum_{i=1}^{N} (x_i - \bar{x})^2 = (0-50)^2 + \ldots + (0-50)^2$

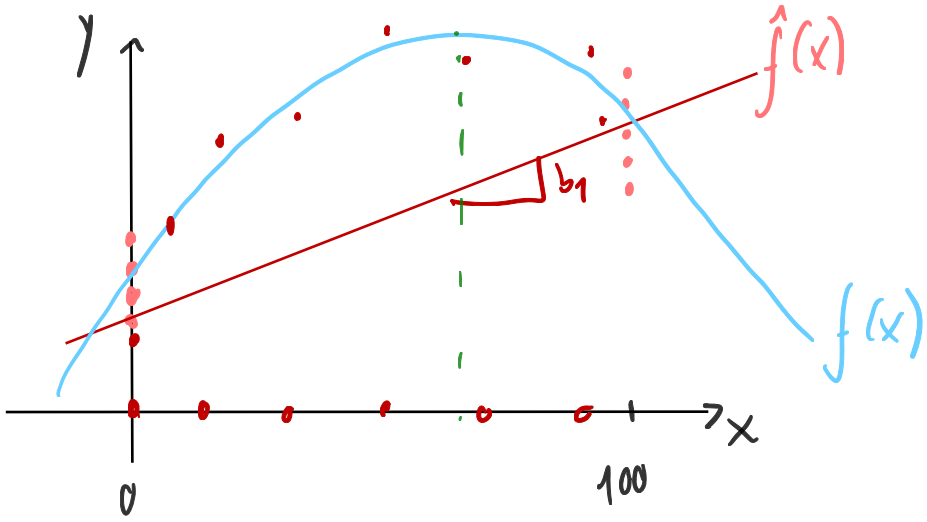
$i = 1$

$+ (100-50)^2 + \ldots$

b) Man erfährt nichts über Linearität:

$(100-50)^2$

$= 10 \cdot 50^2$



↳ Bessere Wahl von X:

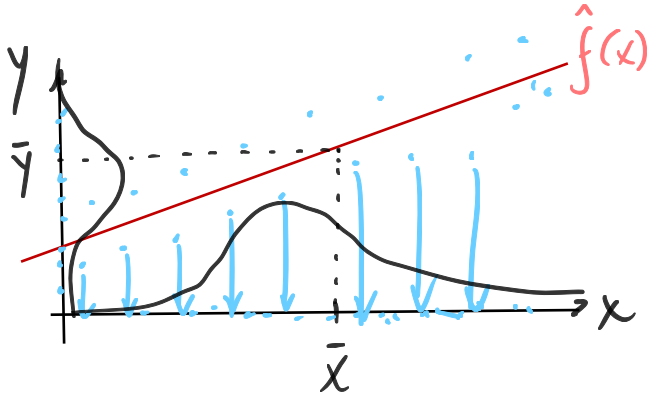
$$x_i \in \{0, 11.1, \ldots, 88.9, 100\}$$

4. Zeigen Sie, dass die geschätzte lineare Regressionsgerade immer durch den Punkt $(\bar{X}, \bar{Y})$ geht.

Repetition:

$$b_0 = \bar{y} - b_1 \bar{x} \iff \bar{y} = b_0 + b_1 \bar{x}$$

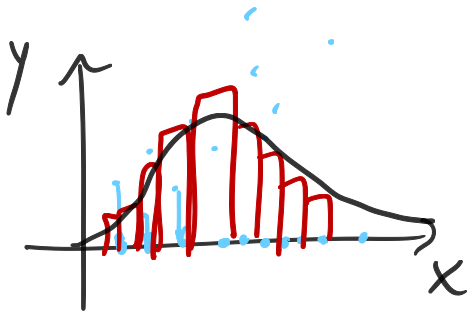
$$b_1 = \frac{\widehat{Cov}(y, x)}{\widehat{Var}(x)}$$



$$D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

Herleitung:

$$b_0 = \underset{b_0'}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2}_{:= SQR \text{ (ist konvex)}}$$



$$\frac{\partial SQR(b_0')}{\partial b_0'} = 2 \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)(-1) \overset{!}{=} 0$$

$$\widehat{Var}(x) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} b_0 - \sum_{i=1}^{N} b_1 x_i \overset{!}{=} 0 \quad \text{(Linearität)}$$

$$\sum_{i=1}^{N} y_i - N b_0 - b_1 \sum_{i=1}^{N} x_i \overset{!}{=} 0$$

$$b_0 = \frac{1}{N} \left(\sum_{i=1}^{N} y_i - b_1 \sum_{i=1}^{N} x_i \right)$$

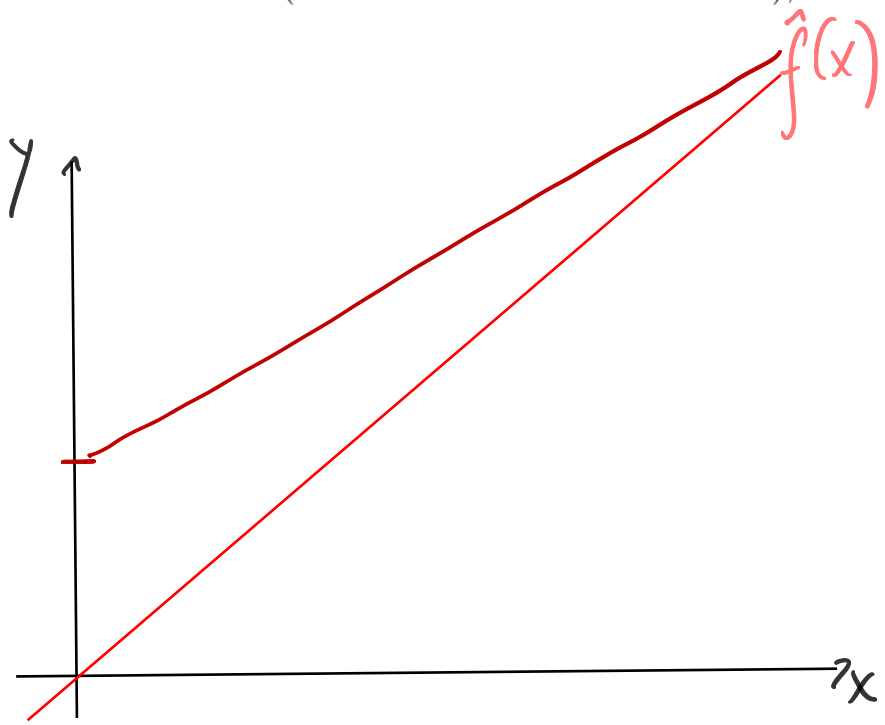
$$= \bar{y} - b_1 \bar{x}$$

5. Betrachten Sie das vereinfachte Regressions-Modell ("Gerade durch den Ursprung")

$$y_i = \beta_1 x_i + u_i, \quad i = 1, \ldots, N.$$

a) Zeigen Sie, dass der KQ-Schätzer b_1 erwartungstreu ist. Welche Annahmen haben Sie für den Beweis benötigt?

b) Zeigen Sie, dass die Varianz von b_1 kleiner oder gleich gross ist im Vergleich zum allgemeinen Modell ("Gerade mit Achsen-Abschitt"), und in der Regel kleiner.



a) $b_1 = \underset{b_1'}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - b_1' x_i)^2$

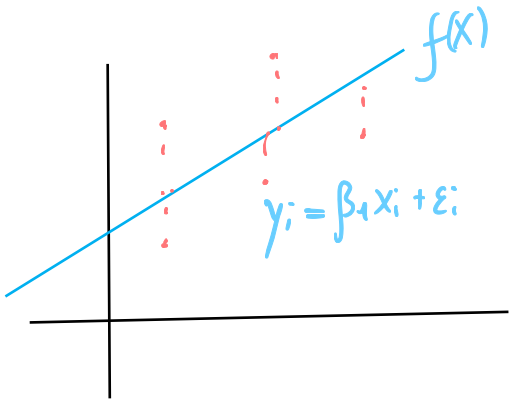
$:= SQR'(b_1')$

$$\frac{d\,SQR'(b_1')}{d\,b_1'} = 2 \sum_{i=1}^{N} (y_i - b_1 x_i)(-x_i) \overset{!}{=} 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i x_i - b_1 \sum_{i=1}^{N} x_i x_i \quad \Rightarrow b_1 = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} x_i^2}$$

Repetition Erwartungstreue:

$$E[\hat{\theta}] = \theta$$



↳ $E[b_1] \overset{!}{=} \beta_1$

$y_i = \beta_1 x_i + \varepsilon_i$

$$E[b_1] = E\left[\frac{\sum y_i x_i}{\sum x_i^2}\right] = E\left[\frac{\sum (\beta_1 x_i + \varepsilon_i) x_i}{\sum x_i^2}\right]$$

$$= E\left[\frac{\sum \beta_1 x_i^2 + \varepsilon_i x_i}{\sum x_i^2}\right]$$, wobei x_i als fix betr. werden

$$= E\left[\beta_1 \frac{\sum x_i^2}{\sum x_i^2}\right] + E\left[\frac{\sum \varepsilon_i x_i}{\sum x_i^2}\right]$$

$$= \beta_1 + \frac{\sum E[\varepsilon_i] x_i}{\sum x_i^2} = \beta_1 + 0 = \beta_1$$

$\sum x_i = N\bar{x}$

$N\left(\frac{1}{N}\sum x_i\right)$

$\sum_{i=1}^{N} x_i^2 - 2x_i\bar{x} + \bar{x}^2$

$\sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2$

$N\bar{x}^2$

b) Hier:

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{N} x_i^2}$$ über $Var(b_1) = E[b_1^2] - E[b_1]^2$

$E[b_1]E[b_1]$

Sonst:

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

mit $\sum_{i=1}^{N} (x_i - \bar{x})^2 \leq \sum_{i=1}^{N} x_i^2$

mit Gleichheit falls $\bar{x} = 0$

$$\sum x_i^2 - 2N\bar{x}^2 + N\bar{x}^2$$

$$\sum x_i^2 - N\bar{x}^2$$

$$\hat{y} = b_0 + b_1 x_{NEU} + b_1 (x - x_{NEU})$$

$$\hat{y} = b_0 + b_1 x_{NEU} - b_1 x_{NEU} + b_1 x$$

$$= b_0 + b_1 x$$

@ enussl auf Github

$$\sum (y_i - b_0 - b_1 x_i) x_i = 0$$

$$\sum y_i x_i - b_0 x_i - b_1 x_i^2 = 0$$

$$\sum y_i x_i - b_0 \sum x_i - b_1 \sum x_i^2 = 0$$