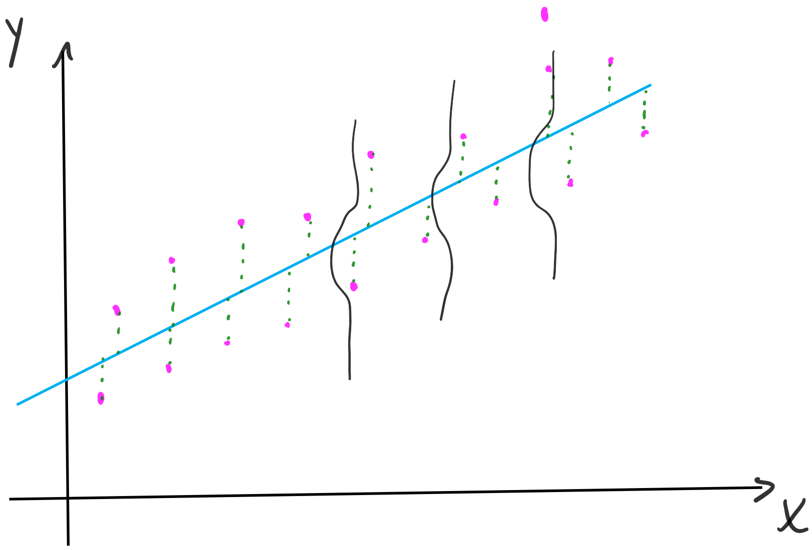
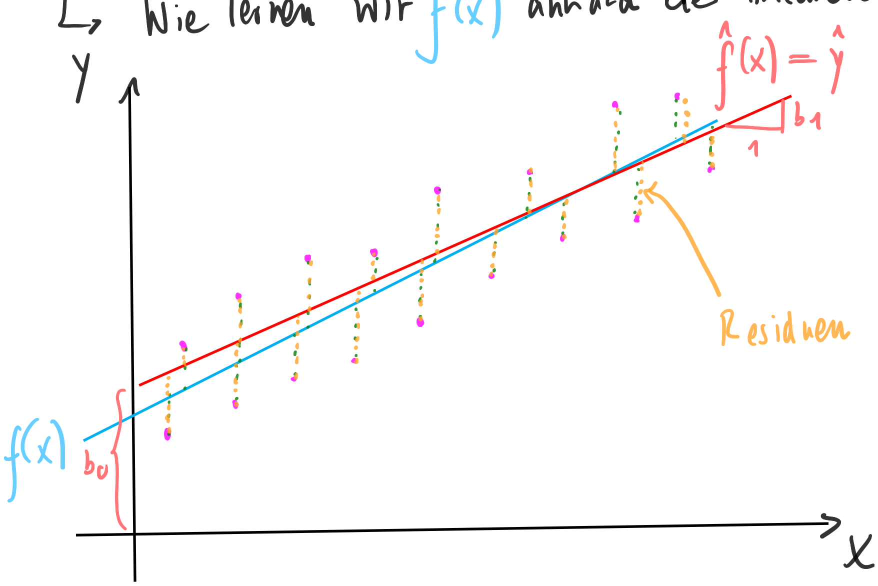
Kurzzusammenfassung (Regression)

· <u>Modell der Realität</u>: $y = f(x) + \varepsilon$, wobei ε einen stochastischen Fehler darstellt und $f(x)$ das Signal (die wahre Beziehung)



↳ Wir versuchen $f(x)$ zu lernen, indem wir eine statistische Methode anwenden. Wir beschränken uns v.a auf die lineare Regression, wobei es eine riesige Menge an Methoden gibt (z.B Neuronale Netzwerke).

↳ Wie lernen wir $f(x)$ anhand der linearen Regression?



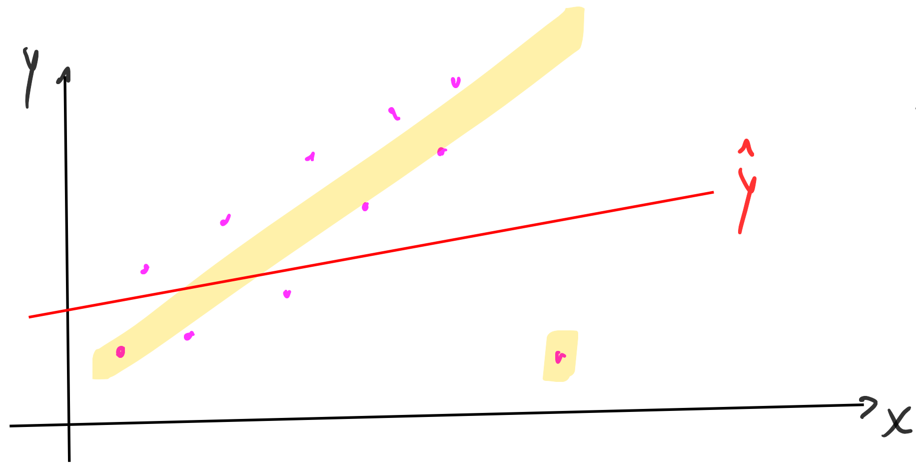
· Wir wählen (x eindimensional) b_0, b_1 so, dass:

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - b_0 - b_1 x_i\right)^2$$

minimiert wird (KQ).

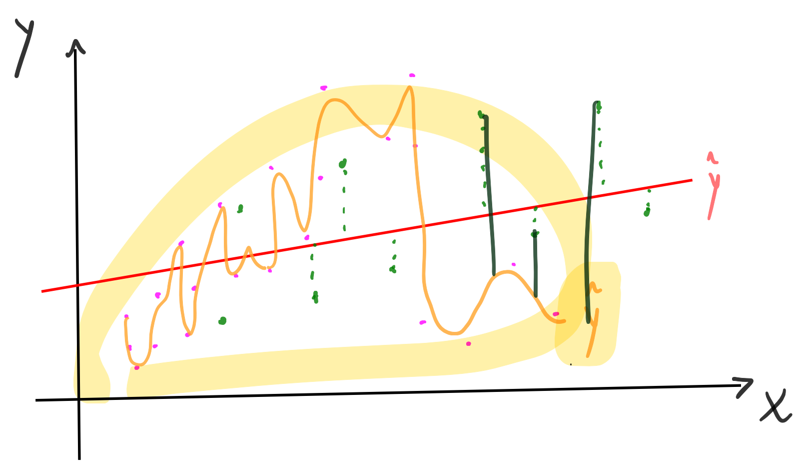
- ## Unterteilung in Training und Test-Daten:

  - Wir wählen b_0, b_1 so, dass sich $b_0 + b_1 x = \hat{f}(x) = \hat{y}$ möglichst gut den Daten anpasst (MSE minimieren).

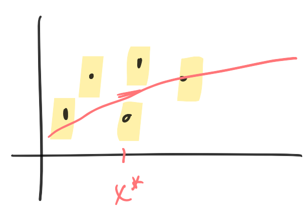


$y = f(x) + \varepsilon$

↳ Das kann zu unerwünschten Folgen führen, vor allem bei "flexibleren" Methoden wie z.B KNN:



KNN (sehr flexibel):



↳ Deshalb unterteilen wir die Daten in Trainingsdaten und Testdaten und nehmen diejenigen Parameter (auf den Trainingsdaten gelernt), die den kleinsten MSE auf den Testdaten erreichen.

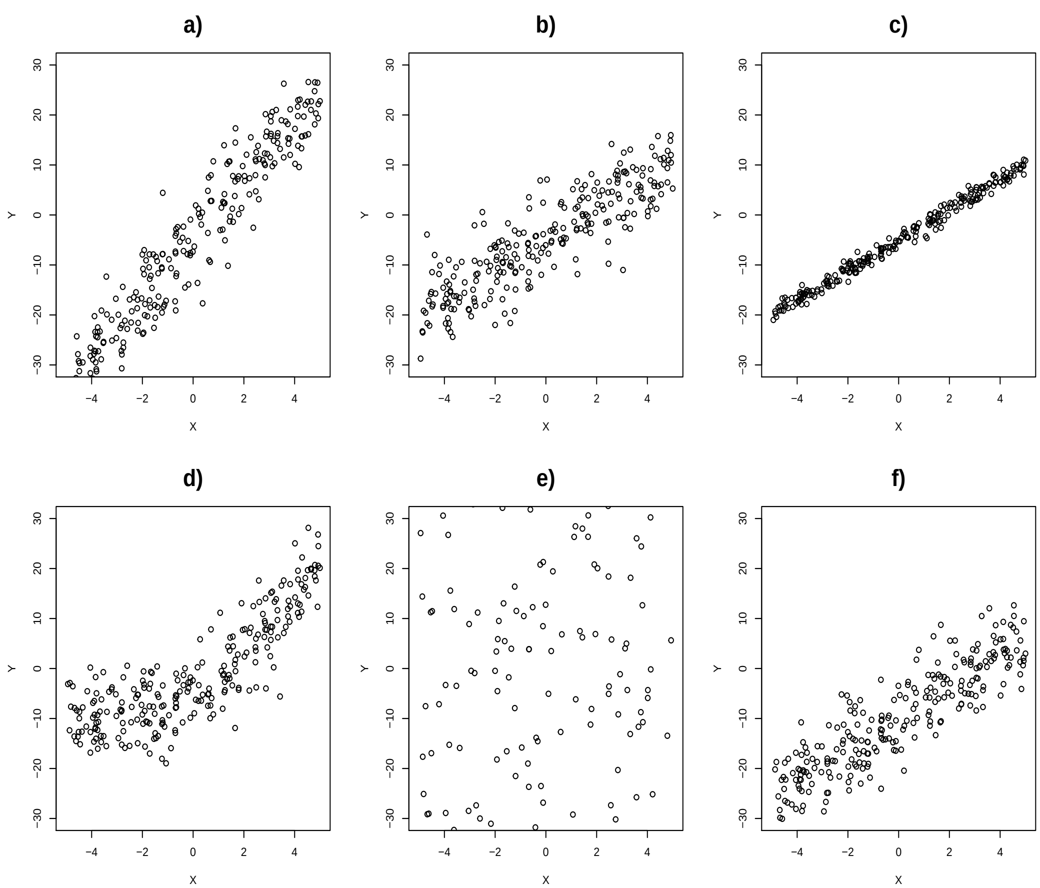
Einführung in die Empirische Wirtschaftsforschung

Übungsaufgaben 3

Statistische Modelle für Y - Kausalität vs. Korrelation - Mean Squared Error

1. Ordnen Sie die folgenden Modelle den Bildern zu. Jedes der Bilder enthält $N = 250$ Realisationen.

   1. $Y = -5 + 3X + \epsilon, \;\; \epsilon \sim Normal(0, 1)$
   2. $Y = -5 + 3X + \epsilon, \;\; \epsilon \sim Normal(0, 25)$
   3. $Y = -10 + 3X + \epsilon, \;\; \epsilon \sim Normal(0, 25)$
   4. $Y = -5 + 6X + \epsilon, \;\; \epsilon \sim Normal(0, 25)$
   5. $Y = -5 + 3X + \epsilon, \;\; \epsilon \sim Normal(0, 2500)$
   6. $Y = -5 + 3X + 0.5X^2 + \epsilon, \;\; \epsilon \sim Normal(0, 25)$



2. Betrachten Sie den Datensatz “nlsy.csv”.

a) Identifizieren Sie in R geeignete Prädiktoren, um die unterschiedlichen Gehälter zu erklären.
Tipp: Die R-Funktion `pairs` verschafft einen guten Überblick.

b) Was sind Ihre Gedanken bezüglich Korrelation vs. Kausalität in den von Ihnen identifizierten Prädiktoren?

c) Gibt es Unterschiede im Zusammenhang zwischen der abhängigen Variable “Earnings” und den Prädiktoren für “Weisse” und “Nicht-Weisse” (Die Variable “white” nimmt nur zwei Werte an: Eine Person ist “weiss”, falls sie den Wert 1 annimmt und “nicht-weiss”, falls sie den Wert 0 annimmt)?

3. Der Spiegel behauptet, dass durch einen höheren Schokoladenkonsum die Wahrscheinlichkeit erhöht wird Nobelpreisträger zu werden. Als Beweis benutzt er folgende Grafik:

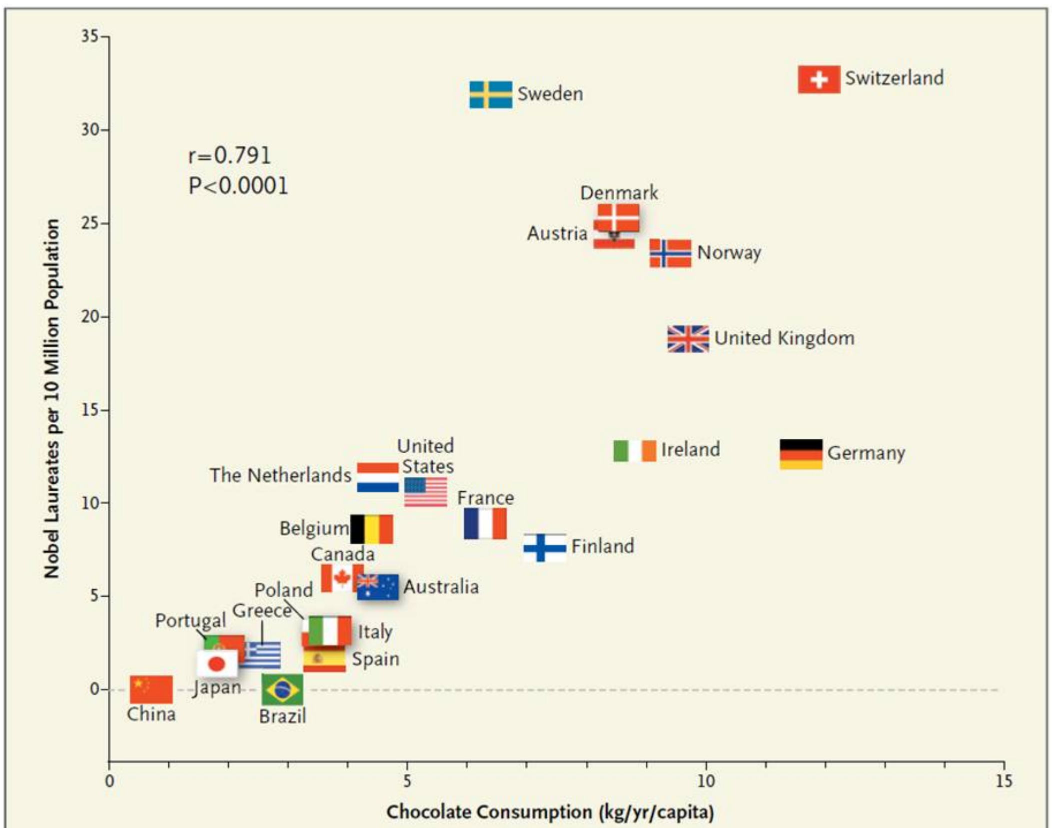


Figure 1. Correlation between Countries’ Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Glauben Sie diesem Artikel?

4. Betrachten Sie den Datensatz “cars”, welcher schon im base R installiert ist. Verschaffen Sie sich einen Überblick über den Datensatz mit dem command ?cars. Wir wollen eine Gerade finden, mit der wir die Distanz möglichst gut anhand des Tempos beschreiben können. Hierfür suchen wir die Gerade, die den Training-MSE minimiert. Nehmen Sie an, dass der Achsenabschnitt a mit -17.5791 gegeben ist. Finden Sie mit einem passenden trial-and-error Vorgehen einen Steigungsparameter b mit kleinstmöglichem Training-MSE.
Später im Kurs werden wir sehen, wie wir den Achsenabschnitts- und Steigungsparameter analytisch bestimmen können mit Hilfe des Kleinsten Quadrate (KQ) Schätzers.

5. Die folgende Aufgabe benutzt den Datensatz “luft.csv” aus der Vorlesung. Für die Beziehung zwischen Lufttemperatur (X) und Ozon (Y) wurden die folgenden Modelle geschätzt (in rot):

a) Linear (Polynom 1. Ordnung)

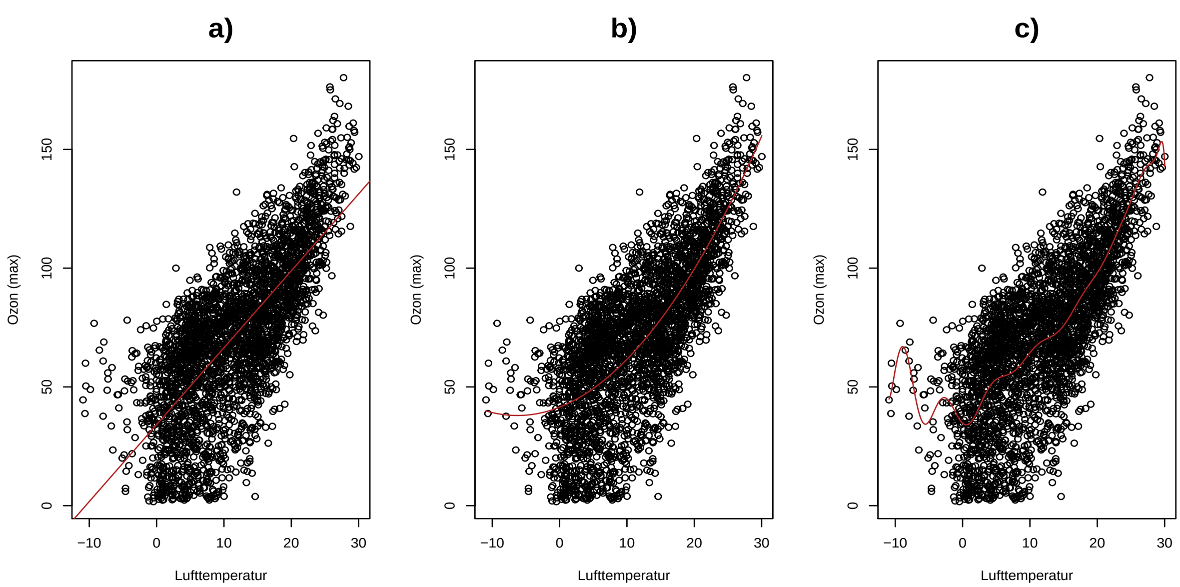
b) Polynom 2. Ordnung

c) Polynom 20. Ordnung

Ordnen Sie die Modellschätzungen den untenstehenden Aussagen zu. Begründen Sie Ihre Wahl.

1. Mittlerer Training-MSE und tiefster Test-MSE.
2. Höchster Training-MSE und mittlerer Test-MSE (underfitting).
3. Tiefster Training-MSE und höchster Test-MSE (overfitting).

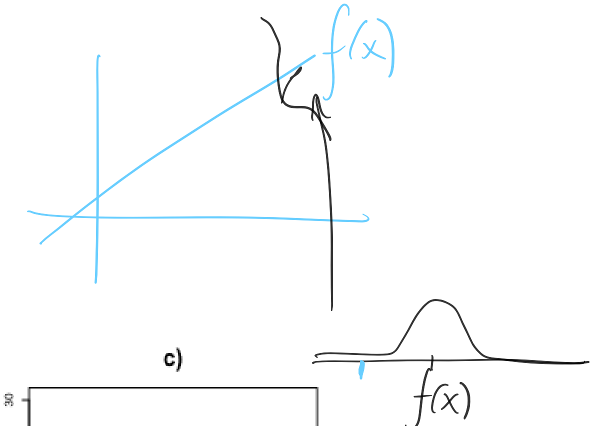
Welches Modell würden Sie für eine Prognose heranziehen?

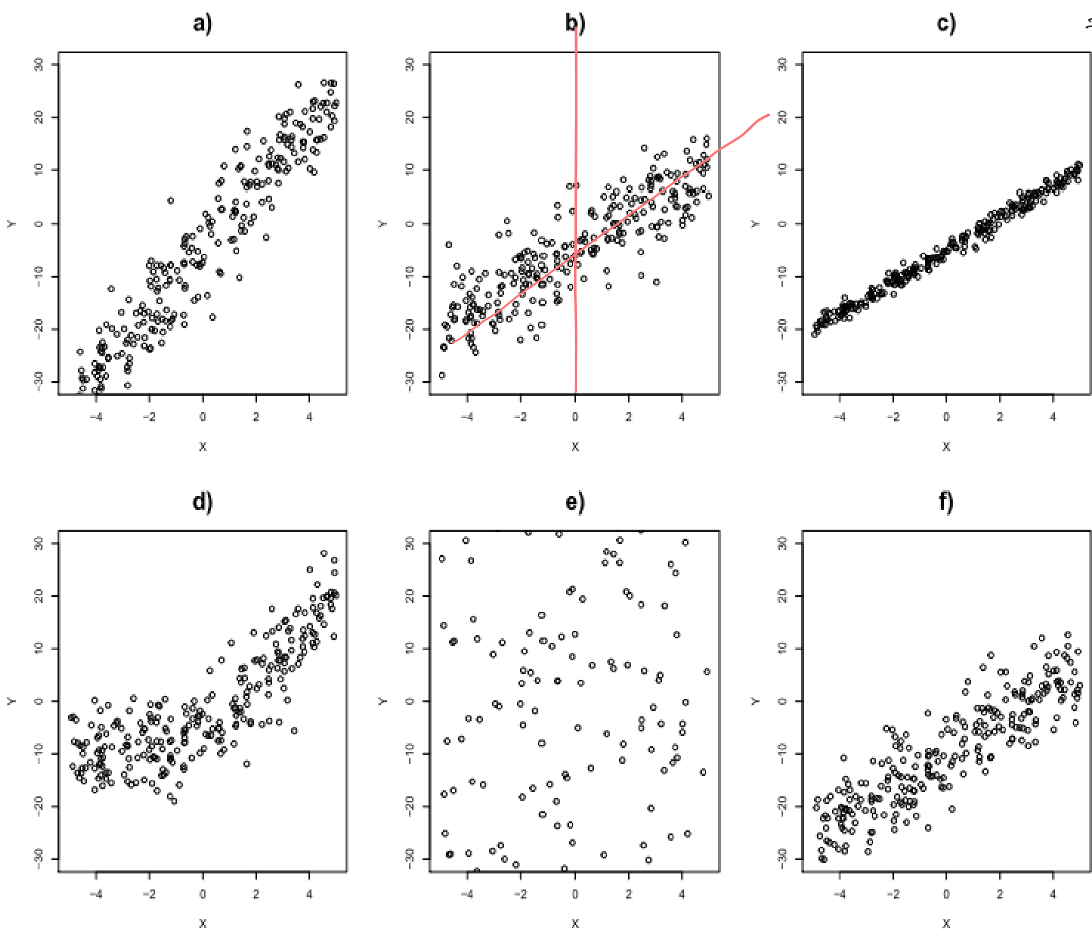


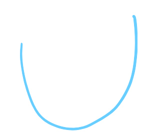
1. Ordnen Sie die folgenden Modelle den Bildern zu. Jedes der Bilder enthält $N = 250$ Realisationen.

1. $Y = -5 + 3X + \epsilon, \quad \epsilon \sim Normal(0, 1)$ c)
2. $Y = -5 + 3X + \epsilon, \quad \epsilon \sim Normal(0, 25)$ b)
3. $Y = -10 + 3X + \epsilon, \quad \epsilon \sim Normal(0, 25)$ f)
4. $Y = -5 + 6X + \epsilon, \quad \epsilon \sim Normal(0, 25)$ a)
5. $Y = -5 + 3X + \epsilon, \quad \epsilon \sim Normal(0, 2500)$ e)
6. $Y = -5 + 3X + 0.5X^2 + \epsilon, \quad \epsilon \sim Normal(0, 25)$ d)

$y = f(x) + \varepsilon, \quad \varepsilon \sim Normal(\mu, \sigma^2)$

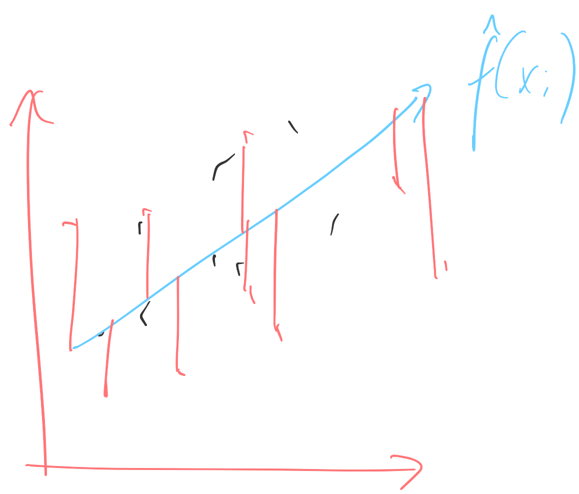






$\sqrt{f(x)}$

$$\mathbb{E}(\varepsilon) = \underbrace{\mathbb{E}[\underbrace{\mathbb{E}[\varepsilon | X]}_{=c}]}_{=0}$$



2. Betrachten Sie den Datensatz “nlsy.csv”.

   a) Identifizieren Sie in R geeignete Prädiktoren, um die unterschiedlichen Gehälter zu erklären.
      Tipp: Die R-Funktion `pairs` verschafft einen guten Überblick.

   b) Was sind Ihre Gedanken bezüglich Korrelation vs. Kausalität in den von Ihnen identifizierten Prädiktoren?

   c) Gibt es Unterschiede im Zusammenhang zwischen der abhängigen Variable “Earnings” und den Prädiktoren für “Weisse” und “Nicht-Weisse” (Die Variable “white” nimmt nur zwei Werte an: Eine Person ist “weiss”, falls sie den Wert 1 annimmt und “nicht-weiss”, falls sie den Wert 0 annimmt)?

3. Der Spiegel behauptet, dass durch einen höheren Schokoladenkonsum die Wahrscheinlichkeit erhöht wird Nobelpreisträger zu werden. Als Beweis benutzt er folgende Grafik:

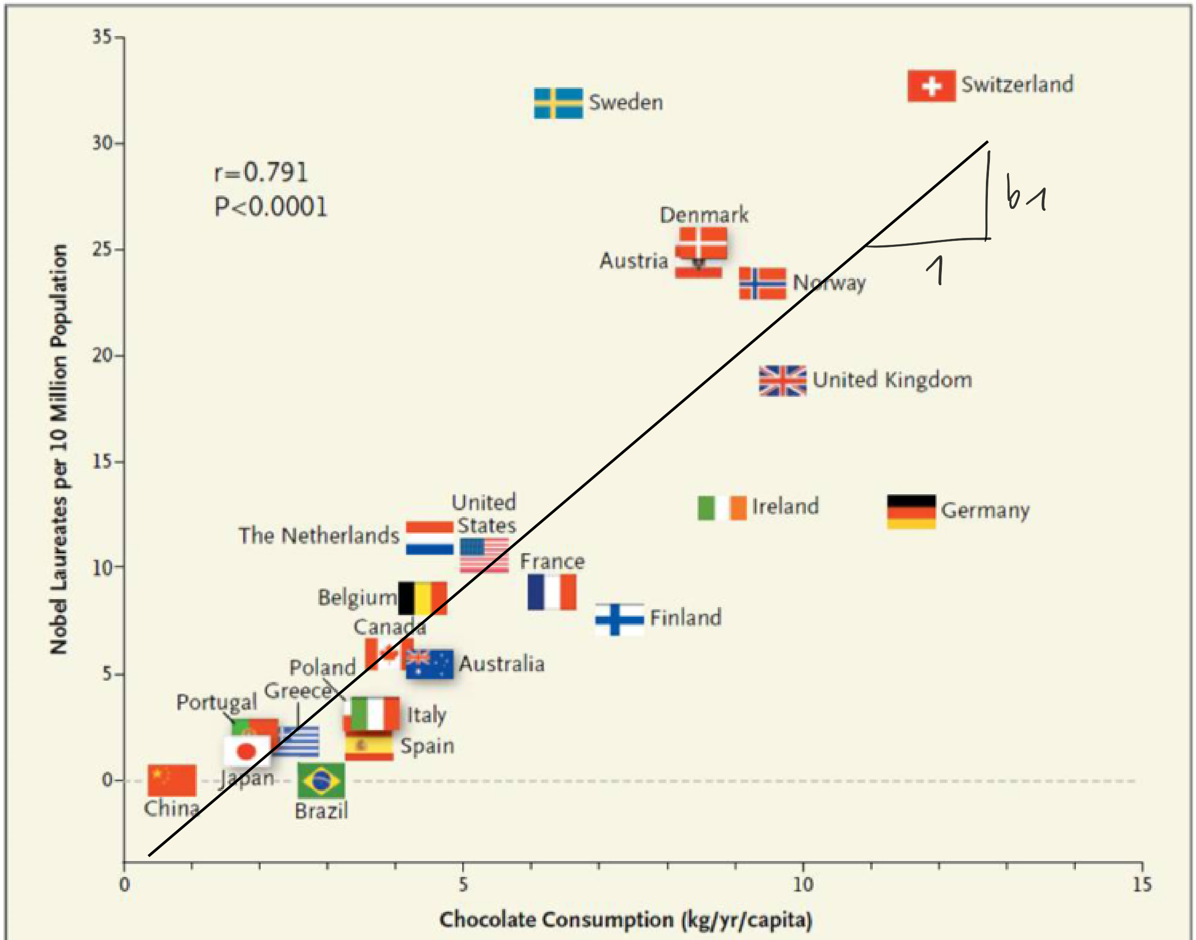
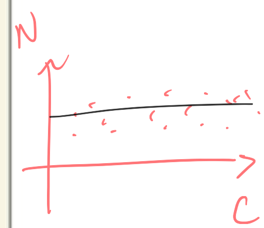


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.



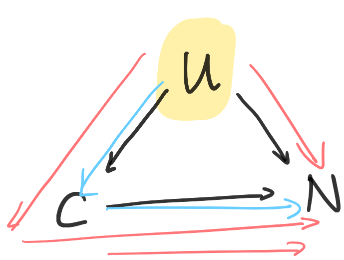
Glauben Sie diesem Artikel?

Wir definieren:

$C :=$ Chocolate

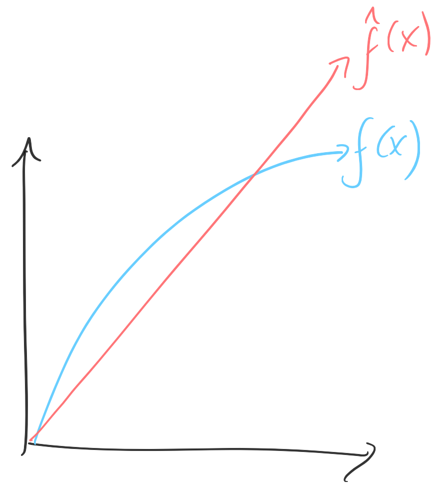
$N :=$ Nobel laureates

Problem:



, wobei wir U nicht observieren. Eine mögliche Variable U ist hier

$$U := BIP$$

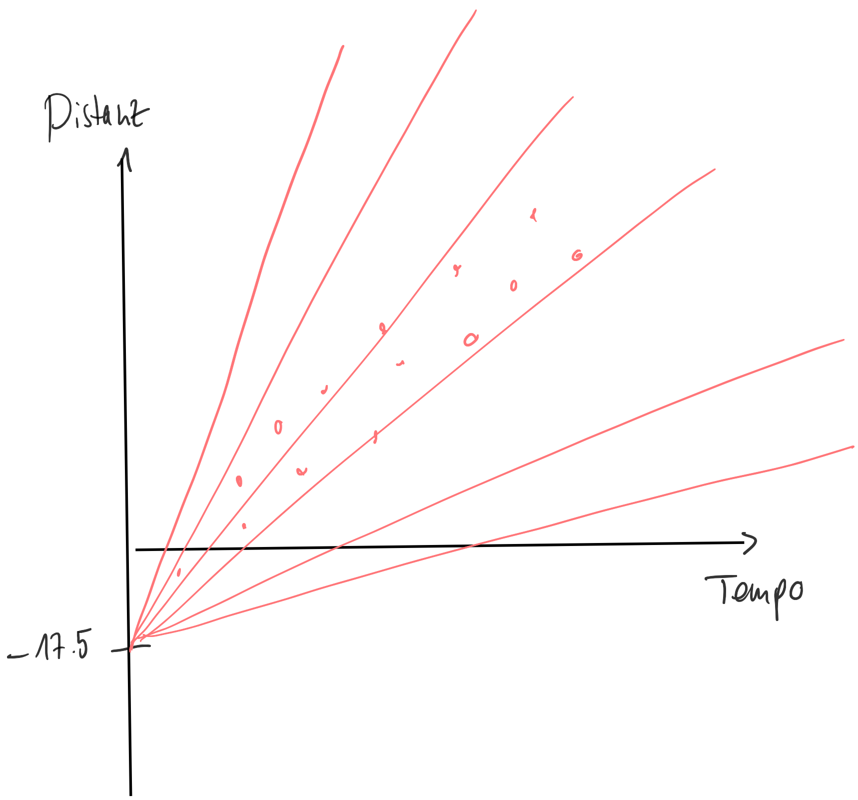


Zudem:

1) Spez. von $f(x)$

2) Messfehler

4. Betrachten Sie den Datensatz "cars", welcher schon im base R installiert ist. Verschaffen Sie sich einen Überblick über den Datensatz mit dem command ?cars. Wir wollen eine Gerade finden, mit der wir die Distanz möglichst gut anhand des Tempos beschreiben können. Hierfür suchen wir die Gerade, die den Training-MSE minimiert. Nehmen Sie an, dass der Achsenabschnitt a mit -17.5791 gegeben ist. Finden Sie mit einem passenden trial-and-error Vorgehen einen Steigungsparameter b mit kleinstmöglichem Training-MSE.
Später im Kurs werden wir sehen, wie wir den Achsenabschnitts- und Steigungsparameter analytisch bestimmen können mit Hilfe des Kleinsten Quadrate (KQ) Schätzers.



5. Die folgende Aufgabe benutzt den Datensatz “luft.csv” aus der Vorlesung. Für die Beziehung zwischen Lufttemperatur (X) und Ozon (Y) wurden die folgenden Modelle geschätzt (in rot):

a) Linear (Polynom 1. Ordnung)

b) Polynom 2. Ordnung

c) Polynom 20. Ordnung

Ordnen Sie die Modellschätzungen den untenstehenden Aussagen zu. Begründen Sie Ihre Wahl.

1. Mittlerer Training-MSE und tiefster Test-MSE.
2. Höchster Training-MSE und mittlerer Test-MSE (underfitting).
3. Tiefster Training-MSE und höchster Test-MSE (overfitting).

Welches Modell würden Sie für eine Prognose heranziehen?