Multiple lineare Regression

Peter Büchel

HSLU I

Stat: Block 12

Multiple lineare Regression

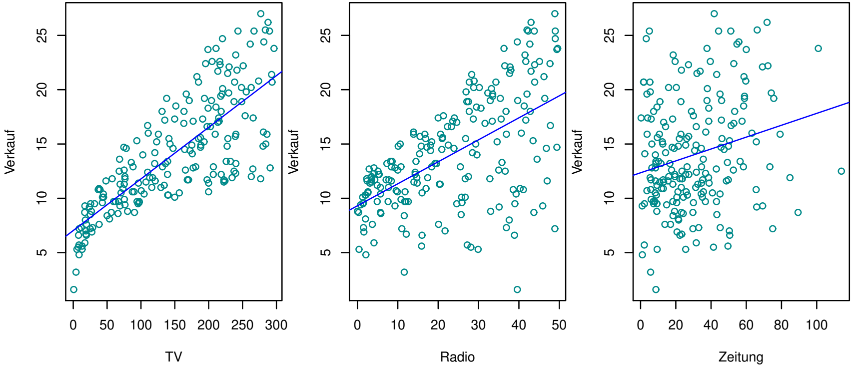
- Einfache lineare Regression: Nützliches Vorgehen, um Output aufgrund *einer* einzelnen erklärenden Variablen vorherzusagen
- Praxis: Output hängt oft von mehr als einer erklärenden Variablen ab

Beispiel

- Datensatz `Werbung`: Zusammenhang zwischen `TV-Werbung` und `Verkauf` untersucht
- Auch Daten für Werbeausgaben für `Radio` und `Zeitung` vorhanden
- Frage: Wirken sich eine oder beide dieser Werbeausgaben auf Verkauf aus?
- Analyse der Verkaufszahlen erweitern: Beiden zusätzlichen Inputs mitberücksichtigen

- Möglichkeit: Für jedes separate Werbebudget eine einfache Regression durchführen

- Abbildung:



- Parameter und weitere wichtige Daten in Tabellen unten aufgeführt

- Einfache Regression von `Verkauf` auf `TV`:

	Koeffizient	Std.fehler	t-Statistik	p-Wert
`Intercept`	7.033	0.458	15.36	< 0.0001
`TV`	0.048	0.003	17.67	< 0.0001

- Einfache Regression von `Verkauf` auf `Radio`:

	Koeffizient	Std.fehler	t-Statistik	p-Wert
`Intercept`	9.312	0.563	16.54	< 0.0001
`Radio`	0.203	0.020	9.92	< 0.0001

- Einfache Regression von `Verkauf` auf `Zeitung`:

	Koeffizient	Std.fehler	t-Statistik	p-Wert
`Intercept`	12.351	0.621	19.88	< 0.0001
`Zeitung`	0.055	0.017	3.30	< 0.0001

- Ansatz separate einfache lineare Regressionen: Nicht zufriedenstellend
- Erstens: Nicht klar, wie man für gegebene Werte der drei erklärenden Variablen eine Vorhersage für den Verkauf machen will:
  - Jeder Input durch *andere Regressionsgleichung* mit Verkauf verknüpft
- Zweitens: Jede der drei Regressionsgleichungen ignoriert die beiden anderen erklärenden Variablen für Bestimmung der Koeffizienten
- Kann zu sehr irreführenden Schätzungen der Wirkung der Werbeausgaben für jedes einzelne Medium auf den Verkauf haben kann, falls die drei erklärenden Variablen miteinander korrelieren

- Besser: Alle erklärenden Variablen direkt mitberücksichtigten

- Jeder erklärenden Variablen wird ein *eigener* Steigungskoeffizient in *einer* Gleichung zugeordnet

- Allgemein: p verschiedene erklärende Variablen

- *Multiples lineares Regressionsmodell*:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- X_j: j-ter Input

- β_j: Zusammenhang zwischen *dieser* erklärenden Variablen und der Zielgrösse Y

- β_j: Durchschnittliche Änderung der Zielgrösse bei Änderung von X_j um eine Einheit, *wenn alle anderen erklärenden Variablen festgehalten werden*

Beispiel

- Multiples lineares Regressionsmodell für den Datensatz `Werbung`:

$$\texttt{Verkauf} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung} + \varepsilon$$

- Also

$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung}$$

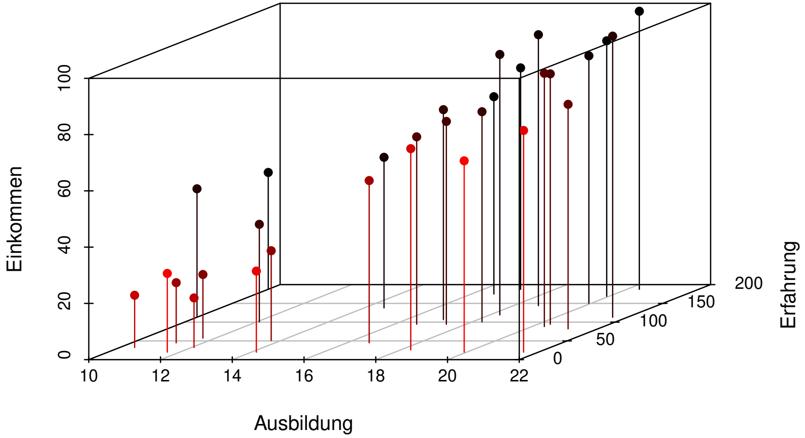
- Multiples lineares Modell verallgemeinert einfaches lineares Modell
- Berechnungen und Interpretationen für multiples Modell ähnlich, wenn auch meist komplizierter als beim linearen Modell
- Graphische Methoden: Entfallen für multiples lineare System praktisch vollends
- Datenpunkte für Beispiel vorher: Nicht darstellbar, da schon für erklärende Variablen drei Achsen gebraucht werden

Beispiel: `Einkommen`

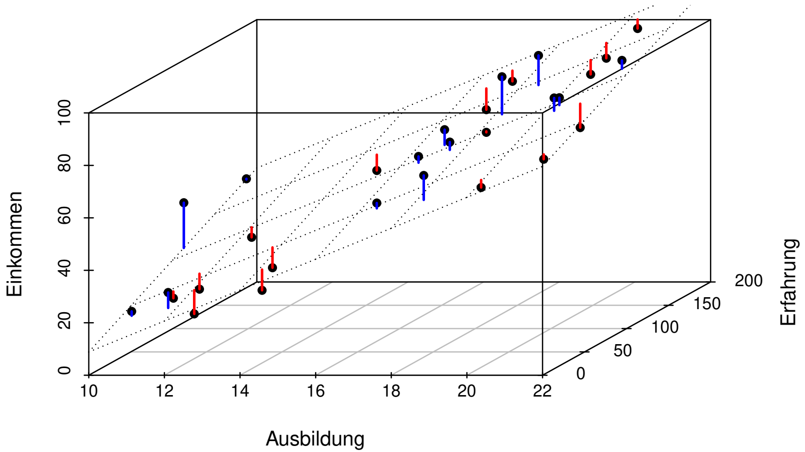
- Graphische Darstellung für zwei erklärende Variablen möglich
- Datensatz `Einkommen`
- Bis jetzt: `Ausbildung` einzige erklärende Variable
- Einkommen auch von `Erfahrung` (Anzahl Berufsmonate) abhängig
- Multiples lineares Modell:

$$\texttt{Einkommen} = \beta_0 + \beta_1 \cdot \texttt{Ausbildung} + \beta_2 \cdot \texttt{Erfahrung} + \varepsilon$$

- Datenpunkte im Raum:



- Analog einfaches lineares Regressionsmodell: Suchen *Ebene*, die am „besten" zu den Datenpunkten passt



- Vorgehen analog zur einfachen linearen Regression
- Bestimmen Ebene so, dass Summe der Quadrate der Abstände der Datenpunkte zur Ebene minimal wird
- Strecken:
  - Blau: Punkte oberhalb der Ebene
  - Rot: Punkte unterhalb der Ebene
- Unterschiede von Punkten zu Ebene: *Residuen*
- Verwenden wieder *Methode der kleinsten Quadrate*

- Schätzung von β_0, β_1 und β_2 mit **R**:

$$\widehat{\beta}_0 = -50.086; \qquad \widehat{\beta}_1 = 5.896; \qquad \widehat{\beta}_2 = 0.173$$

- Code:

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.graphics.regressionplots import abline_plot
import matplotlib.pyplot as plt
import numpy as np

df = pd.read_csv("../Data/Einkommen2.csv").drop("Unnamed: 0", axis=1)

df.head()
```

```
##    Ausbildung   Erfahrung  Einkommen
## 0   21.586207  113.103448  99.917173
## 1   18.275862  119.310345  92.579135
## 2   12.068966  100.689655  34.678727
## 3   17.034483  187.586207  78.702806
## 4   19.931034   20.000000  68.009922
```

- Code:

```
Y = df["Einkommen"]
X = df[["Ausbildung", "Erfahrung"]]
X = sm.add_constant(X)
```

```
## //usr/lib/python3/dist-packages/numpy/core/fromnumeric.py:2495: FutureWarni
##   return ptp(axis=axis, out=out, **kwargs)
```

```
fit = sm.OLS(Y,X).fit()

fit.params
```

```
## const         -50.085639
## Ausbildung      5.895556
## Erfahrung       0.172855
## dtype: float64
```

- Oder:

```
from statsmodels.formula.api import ols
fit = ols("Einkommen ~ Ausbildung + Erfahrung", data=df).fit()

fit.params
```

```
## Intercept    -50.085639
## Ausbildung     5.895556
## Erfahrung      0.172855
## dtype: float64
```

Interpretation der Koeffizienten

- Multiples lineares Modell:

$$\texttt{Einkommen} \approx -50.086 + 5.896 \cdot \texttt{Ausbildung} + 0.173 \cdot \texttt{Erfahrung}$$

- $\widehat{\beta}_0 = -50.086$:
  - Wenn Person keine Ausbildung und keine Erfahrung hat, so „erhält" man CHF $-50\,086$
  - Interpretation macht praktisch natürlich keinen Sinn
- $\widehat{\beta}_1 = 5.896$:
  - Bei konstanter Erfahrung verdient man pro zusätzliches Ausbildungsjahr Ausbildung CHF 5896 mehr
- $\widehat{\beta}_2 = 0.173$:
  - Bei konstanter Ausbildung verdient man pro zusätzlichen Monat Arbeitserfahrung CHF 173 mehr

Allgemein: Schätzung der Regressionskoeffizienten

- Wie einfache linearer Regression: Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_p$ i. A. unbekannt

- Müssen sie aus Daten schätzen:

$$\widehat{\beta}_0, \quad \widehat{\beta}_1, \quad \ldots, \quad \widehat{\beta}_p$$

- Aufgrund der Schätzungen kann man Vorhersagen machen:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \ldots + \widehat{\beta}_p x_p$$

- Parameter wieder mit der Methode der kleinsten Quadrate schätzen

Beispiel

- Python: Multiples lineares Regressionsmodell für Werbung:

```
df = pd.read_csv("../Data/Werbung.csv").drop("Unnamed: 0", axis=1)


Y = df["Verkauf"]
X = df[["TV", "Radio", "Zeitung"]]
X = sm.add_constant(X)

fit = sm.OLS(Y,X).fit()

fit.params
```

```
## const      2.938889
## TV         0.045765
## Radio      0.188530
## Zeitung   -0.001037
## dtype: float64
```

- Es gilt:

$$\texttt{Verkauf} \approx 2.94 + 0.046 \cdot \texttt{TV} + 0.189 \cdot \texttt{Radio} - 0.001 \cdot \texttt{Zeitung}$$

- Koeffizienten interpretieren:
  - Für gegebene Werbeausgaben für Radio und Zeitung werden für zusätzliche CHF 1000 Werbeausgaben für das TV ungefähr 46 Einheiten mehr verkauft
  - Für gegebene Werbeausgaben für TV und Zeitung werden für zusätzliche CHF 1000 Werbeausgaben für das Radio ungefähr 189 Einheiten mehr verkauft
  - Interessant: Bei der Zeitung würde man *weniger* Produkte verkaufen, wenn man *mehr* investiert

- Tabelle: Weitere wichtige Werte:

	Koeffizient	Std.fehler	t-Statistik	P-Wert
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`Radio`	0.189	0.0086	21.89	< 0.0001
`Zeitung`	-0.001	0.0059	-0.18	0.8599

- Code: `params` durch `summary` ersetzen

```
fit.summary()

## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                Verkauf   R-squared:                       0.897
## Model:                            OLS   Adj. R-squared:                  0.896
## Method:                 Least Squares   F-statistic:                     570.3
## Date:                Mon, 11 May 2020   Prob (F-statistic):           1.58e-96
## Time:                        10:25:47   Log-Likelihood:                -386.18
## No. Observations:                 200   AIC:                             780.4
## Df Residuals:                     196   BIC:                             793.6
## Df Model:                           3
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## const          2.9389      0.312      9.422      0.000       2.324       3.554
## TV             0.0458      0.001     32.809      0.000       0.043       0.049
## Radio          0.1885      0.009     21.893      0.000       0.172       0.206
## Zeitung       -0.0010      0.006     -0.177      0.860      -0.013       0.011
## ==============================================================================
## Omnibus:                       60.414   Durbin-Watson:                   2.084
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.241
## Skew:                          -1.327   Prob(JB):                     1.44e-33
## Kurtosis:                       6.332   Cond. No.                         454.
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## """
```

- Koeffizienten der separaten einfachen linearen Regressionen in Slide 5
- Steigungskoeffizienten der multiplen linearen Regression für `TV` und `Radio` sehr ähnlich:
  - `TV`: 0.46 (multiple), 0.48 (einfach)
  - `Radio`: 0.189 (multiple), 0.203 (einfach)
- Geschätzter Regressionskoeffizient $\widehat{\beta}_3$ für `TV` zeigt anderes Verhalten:
  - Einfach: 0.055 (ungleich 0)
  - Multiple: -0.001 (fast gleich 0)
- Entsprechende p-Werte:
  - Einfach: < 0.0001 (hochsignifikant)
  - Multiple: 0.86 (bei weitem nicht mehr signifikant)

- Einfache und multiple Regressionskoeffizienten können sehr verschieden sein

- Einfache Regression: Steigung gibt die Änderung der Zielgrösse `Verkauf` an, wenn man CHF 1000 mehr für die Zeitungswerbung ausgibt, wobei die beiden anderen erklärenden Variablen `TV` und `Radio` *ignoriert* werden

- Multiple lineare Regression: Steigung für `Zeitung` beschreibt die Änderung der Zielgrösse `Verkauf`, wenn man CHF 1000 mehr für Zeitungswerbung ausgibt, wobei die beiden anderen erklärenden Variablen `TV` und `Radio` *festgehalten* werden

- Macht es Sinn, dass die multiple Regression keinen Zusammenhang zwischen `Verkauf` und `Zeitung` andeutet, aber die einfache Regression das Gegenteil impliziert?

- Es macht in der Tat Sinn
- Tabelle mit Korrelationskoeffizienten:

	TV	Radio	Zeitung	Vekauf
TV	1.0000	0.0548	0.0567	0.7822
Radio		1.0000	0.3541	0.5762
Zeitung			1.0000	0.2283
Verkauf				1.0000

- Code:

```
df.corr()
```

```
##                 TV     Radio   Zeitung   Verkauf
## TV        1.000000  0.054809  0.056648  0.782224
## Radio     0.054809  1.000000  0.354104  0.576223
## Zeitung   0.056648  0.354104  1.000000  0.228299
## Verkauf   0.782224  0.576223  0.228299  1.000000
```

- Korrelationskoeffizient `Radio` und `Zeitung`: 0.35
- Was bedeutet dies?
- Zeigt Tendenz bei höheren Werbeausgaben für `Radio` auch mehr in Werbung für `Zeitung` zu investieren
- Annahme: Multiples Regressionsmodell *korrekt*
- Ausgaben für `Zeitung`: Kein direkter Einfluss auf Zielgrösse `Verkauf`
- Werbeausgaben für `Radio`: Höhere Verkäufe
- In Märkten, wo mehr in die Werbung fürs Radio investiert wird, auch Ausgaben für `Zeitung` grösser, da Korrelationskoeffizienten von 0.35

- Einfache lineare Regression: Nur Zusammenhang zwischen `Zeitung` und `Verkauf`, wobei für höhere Werte von `Zeitung` auch höhere Werte für `Verkauf` beobachtet werden

- Aber: Zeitungswerbung beeinflusst Verkäufe *nicht*

- Höhere Werte für `Zeitung` wegen Korrelation auch grössere Werte für `Radio` zur Folge: *Diese* Grösse beeinflusst `Verkauf`

- `Zeitung` schmückt sich hier mit fremden Lorbeeren, nämlich dem Erfolg von `Radio` auf `Verkauf`

- Dieses Resultat steht in Konflikt mit Intuition

- Tritt in realen Situationen aber häufig auf

Absurdes Beispiel

- Einfache Regression: Zusammenhang zwischen Haiattacken und Glaceverkäufen an einem bestimmten Strand
- Je grösser Glaceverkäufe, desto häufiger ereignen sich Haiattacken
- Absurde Idee: Glaceverkäufe an diesem Strand verbieten, damit es keine Haiattacken auf Menschen mehr gibt
- Wo liegt aber der Zusammenhang?
- Real: Bei heissem Wetter kommen mehr Menschen an den Strand $\rightarrow$ mehr Glaceverkäufe $\rightarrow$ mehr Haiattacken
- Confounder: Temperatur
- Multiples Regressionsmodell von Haiattacken mit Glaceverkäufen *und* Temperatur: Glaceverkauf keinen Einfluss mehr auf Haiattacken, Lufttemperatur allerdings schon

Einige wichtige Fragestellungen

- *Ist mindestens eine der erklärenden Variablen* $X_1, \ldots, X_p$ *nützlich, um die Zielgrösse vorherzusagen?*
- *Spielen alle erklärenden Variablen* $X_1, \ldots, X_p$ *für die Vorhersage von* Y *eine Rolle, oder nur eine Teilmenge der erklärenden Variablen?*
- *Wie gut passt das Modell zu den Daten?*
- *Welche Zielgrösse kann man aufgrund konkreter Werte der erklärenden Variablen vorhersagen?*
- *Wie genau ist diese Vorhersage?*

Gibt es einen Zusammenhang zwischen den erklärenden Variablen und der Zielgrösse?

- Hypothesentest:
- Multiple lineare Regression mit p erklärenden Variablen: *Alle* Regressionskoeffizienten ausser β_0 Null sind (keine Variable hat Einfluss):

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0$$

- Nullhypothese

$$H_0 : \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

- Alternativhypothese

$$H_A : \quad \text{mindestens ein } \beta_i \text{ ist ungleich } 0$$

- Berechnung der *F-Statistik* mit p-Wert

Beispiel

- p-Wert für das multiple lineare Modell für den Datensatz `Werbung`:

```
fit.summary()
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                Verkauf   R-squared:                       0.897
## Model:                            OLS   Adj. R-squared:                  0.896
## Method:                 Least Squares   F-statistic:                     570.3
## Date:                Mon, 11 May 2020   Prob (F-statistic):           1.58e-96
## Time:                        10:25:47   Log-Likelihood:                -386.18
## No. Observations:                 200   AIC:                             780.4
## Df Residuals:                     196   BIC:                             793.6
## Df Model:                           3
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## const          2.9389      0.312      9.422      0.000       2.324       3.554
## TV             0.0458      0.001     32.809      0.000       0.043       0.049
## Radio          0.1885      0.009     21.893      0.000       0.172       0.206
## Zeitung       -0.0010      0.006     -0.177      0.860      -0.013       0.011
## ==============================================================================
## Omnibus:                       60.414   Durbin-Watson:                   2.084
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.241
## Skew:                          -1.327   Prob(JB):                     1.44e-33
## Kurtosis:                       6.332   Cond. No.                         454.
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## """
```

- `R`-Ausgabe `p-value` in Zeile für F-Statistik: p-Wert für multiples lineares Modell praktisch null

- Sehr überzeugender Hinweis: Mindestens eine erklärende Variable ist für Zunahme von `Verkauf` bei vergrösserten Werbeausgaben verantwortlich

Bestimmung der wichtigen erklärenden Variablen

- Zuerst entscheiden: Haben erklärende Variablen überhaupt Einfluss auf Zielgrösse
- Entscheid: Mit Hilfe F-Statistik und zugehörigem p-Wert
- Beeinflusst mindestens eine Variable die Zielgrösse: *Welche* erklärende Variablen sind dies?
- Können einzelne p-Werte wie in Tabelle betrachten

- Möglich: Alle erklärenden Variablen beeinflussen Zielgrösse, aber meist sind es nur einige wenige

- Aufgabe: Variablen bestimmen und dann Modell aufstellen, welches nur diese Variablen enthält

- Interessiert an möglichst einfachen Modell, das zu den Daten passt

- Welche Variabeln sind wichtig?

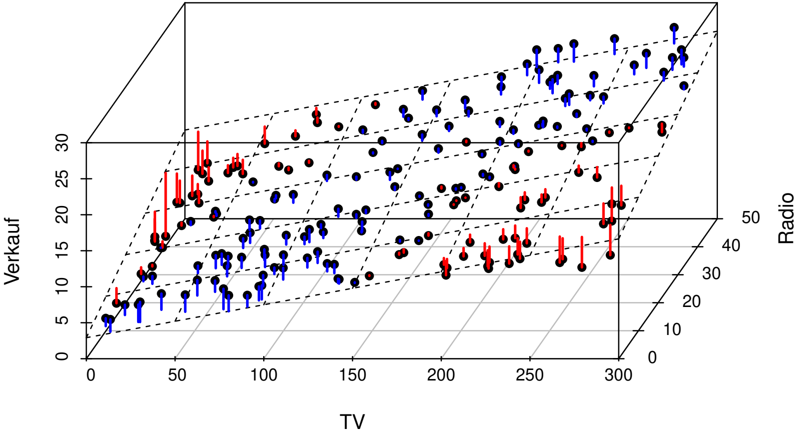
- Prozedere: *Variablenselektion* (nächstes Mal)

Wie gut passt das Modell zu den Daten?

- Bestimmtheitsmass R^2
- Datensatz `Werbung` ist der R^2-Wert 0.8972
- R^2 erhöht sich, je mehr erklärende Variablen berücksichtigt werden

Keine lineare Regression

- Graphischer Überblick: Probleme mit dem Modell aufzeigen, die für die numerischen Werte unsichtbar sind:



- Dreidimensionales Streudiagramm: Nur `TV` und `Radio` berücksichtigt
- Gestrichelt: Regressionsebene
- Beobachtung: Werte der Ebene zu gross, wenn Werbeausgaben ausschliesslich entweder für `TV` oder `Radio` aufgewendet wurden
- Hinten links: Werbung nur für `Radio`
- Vorne rechts: nur für `TV`
- Werte der Ebene sind zu tief, wenn Werbeausgaben gleichmässig auf `TV` und `Radio` verteilt werden
- Nichtlineares Muster: Kann nicht genau durch eine lineare Regression beschrieben werden
- Plot deutet *Interaktion*- oder *Synergieeffekt* an: Grössere Verkäufen, wenn Werbeausgaben aufgeteilt werden

Aufhebung der Annahme bezüglich Additivität

- Interaktionseffekte
- Beispiel Werbung:

```
fit = ols("Verkauf ~ TV + Radio + TV*Radio", data=df).fit()

fit.pvalues
## Intercept    1.541461e-68
## TV           2.363605e-27
## Radio        1.400461e-03
## TV:Radio     2.757681e-51
## dtype: float64
```

- p-Werte zu `TV`, `Radio` und dem Interaktionsterm `TV` $\cdot$ `Radio`: Statistisch signifikant
- Scheint klar: Alle diese Variablen sollten im Modell enthalten sein
- Möglich: p-Wert für den Interaktionterm sehr klein ist, aber die p-Werte der Haupteffekte (hier `TV` und `Radio`) sind es nicht