Multiple Lineare Regression

Peter Büchel

HSLU I

ASTAT: Block 12

Multiple Lineare Regression

Peter Büchel

HSLU I

ASTAT: Block 12

Multiple lineare Regression

- Einfache lineare Regression: Nützliches Vorgehen, um Output aufgrund *einer* einzelnen erklärenden Variablen vorherzusagen

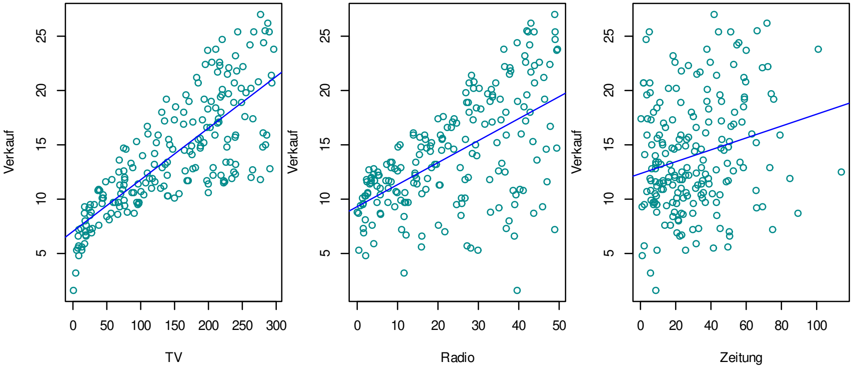
- Praxis: Output hängt oft von mehr als einer erklärenden Variablen ab

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

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

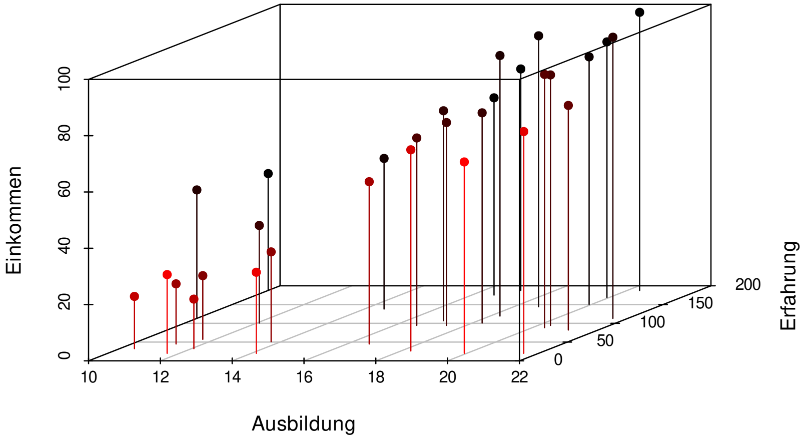
- Multiples lineares Modell verallgemeinert einfaches lineares Modell
- Berechnungen und Interpretationen für multiples Modell ähnlich, wenn auch meist komplizierter als beim linearen Modell
- Graphische Methoden: Entfallen für multiples lineare System praktisch vollends
- Datenpunkte für Beispiel vorher: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

Beispiel: `Einkommen`

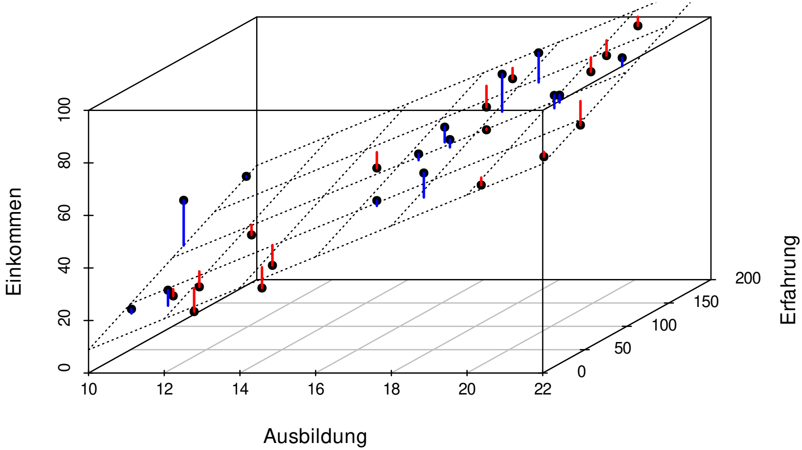
- Graphische Darstellung für zwei erklärende Variablen möglich
- Datensatz `Einkommen`
- Bis jetzt: `Ausbildung` einzige erklärende Variable
- Einkommen auch von `Erfahrung` (Anzahl Berufsmonate) abhängig
- Multiples lineares Modell:

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

- 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:

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

```
coef(lm(Einkommen ~ Ausbildung + Erfahrung))
## (Intercept)  Ausbildung   Erfahrung
## -50.0856388   5.8955560   0.1728555
```

- Multiples lineares Modell:

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

Interpretation der Koeffizienten

- $\widehat{\beta}_0 = -50.086$:
  - 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

- R: Multiples lineares Regressionsmodell für Werbung:

```
coef(lm(Verkauf ~ TV + Radio + Zeitung))
##  (Intercept)           TV        Radio      Zeitung
##  2.938889369  0.045764645  0.188530017 -0.001037493
```

- 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: `coef` durch `summary` ersetzen

```
fit <- lm(Verkauf ~ TV + Radio + Zeitung)

summary(fit)
##
## Call:
## lm(formula = Verkauf ~ TV + Radio + Zeitung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Zeitung     -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

- 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:

```
cor(data.frame(TV, Radio, Zeitung, Verkauf))
##                 TV      Radio    Zeitung   Verkauf
## TV      1.00000000 0.05480866 0.05664787 0.7822244
## Radio   0.05480866 1.00000000 0.35410375 0.5762226
## Zeitung 0.05664787 0.35410375 1.00000000 0.2282990
## Verkauf 0.78222442 0.57622257 0.22829903 1.0000000
```

- 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:

```
summary(lm(Verkauf ~ TV + Radio + Zeitung))
##
## Call:
## lm(formula = Verkauf ~ TV + Radio + Zeitung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Zeitung     -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

- `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

Beispiel

- Warum betrachten wir nicht einfach die einzelnen p-Werte?
- Wenn einer unterhalb des Signifikanzniveaus liegt, dann weiss man, dass mindestens eine Variable Einfluss hat
- Aber: Wegen dem Prinzip des Hypothesentest, ist statistisch signifikanter p-Wert zu 5 % zufällig
- Folgendes Beispiel: Keine Variable ist signifikant
- Alle β_1-Werte in der Nähe von 0
- Aber: Gibt zufällige Abweichungen, wo die zugehörigen p- Werte signifikant sind
- Darum: Wenn sehr viele Variable vorhanden sind, ist praktisch immer eine signifikant, obwohl in Wahrheit keine ist

- Code:

```r
set.seed(4)
v <- 20
d <- 500

df <- matrix(rnorm(v * d), nrow = d)
# head(df)
df <- data.frame(df)

Y <- rnorm(d)
# Y

df$Y <- Y

fit <- lm(Y ~ ., , data = df)
summary(fit)
```

- Output:

```
##
## Call:
## lm(formula = Y ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62976 -0.66857  0.00927  0.64462  2.81840
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.029669   0.047272  -0.628   0.5305
## X1          -0.010970   0.048886  -0.224   0.8225
## X2          -0.036943   0.049150  -0.752   0.4526
## X3          -0.005961   0.047734  -0.125   0.9007
## X4          -0.018073   0.047726  -0.379   0.7051
## X5           0.005827   0.048524   0.120   0.9045
## X6          -0.127798   0.049554  -2.579   0.0102 *
## X7          -0.052386   0.049816  -1.052   0.2935
## X8           0.020574   0.048557   0.424   0.6720
## X9          -0.015178   0.047941  -0.317   0.7517
## X10         -0.015107   0.046988  -0.322   0.7480
## X11          0.005580   0.046517   0.120   0.9046
## X12         -0.004676   0.046583  -0.100   0.9201
## X13         -0.021652   0.049114  -0.441   0.6595
## X14         -0.093800   0.046075  -2.036   0.0423 *
## X15          0.019740   0.047451   0.416   0.6776
## X16          0.042796   0.045267   0.945   0.3449
## X17         -0.074511   0.049061  -1.519   0.1295
## X18          0.041733   0.047568   0.877   0.3808
## X19         -0.078238   0.047492  -1.647   0.1001
## X20         -0.057475   0.048156  -1.194   0.2333
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 479 degrees of freedom
```

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

Beispiel: Vorhersage

- *Vertrauensintervall*, um die Ungewissheit für den *durchschnittlichen* `Verkauf` für eine grosse Zahl von Städten zu quantifizieren

- Nur die erklärenden Variablen `TV` und `Radio` berücksichtigen, da `Zeitung` für `Verkauf` keinen Einfluss hat

- Wenden CHF 100 000 für `TV`-Werbung und CHF 20 000 für `Radio`-Werbung in jeder Stadt auf $\rightarrow$ 95 %-Vertrauensintervall

$$[10\,985,\ 11\,528]$$

```r
predict(lm(Verkauf ~ TV + Radio),
        interval = "confidence",
        data.frame(TV = 100, Radio = 20)

)
##        fit      lwr      upr
## 1 11.25647 10.98525 11.52768
```

- Interpretation lautet wie folgt: 95 % aller Intervalle dieser Form enthalten den wahren Wert $f(X_1, X_2)$

- Was heisst dies?

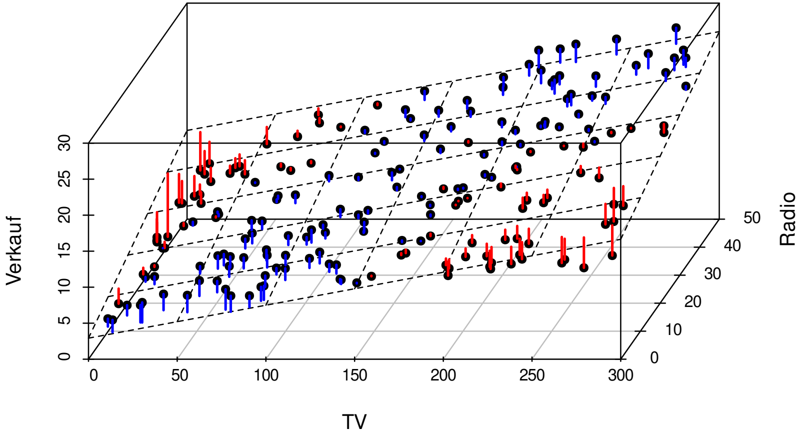
- Sammeln grosse Menge von Datensätzen wie den `Werbung`-Datensatz

- Für jeden Datensatz jeweils das Vertrauensintervall für den wahren durchschnittlichen `Verkauf` berechnen (bei CHF 100 000 für `TV`-Werbung und CHF 20 000 für `Radio`-Werbung)

- In 95 % dieser Intervalle liegt der wahre Wert vom mittleren `Verkauf`

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 <- lm(Verkauf ~ TV + Radio + TV * Radio)

summary(fit)
##
## Call:
## lm(formula = Verkauf ~ TV + Radio + TV * Radio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
## Radio       2.886e-02  8.905e-03   3.241   0.0014 **
## TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678,Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

- 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