Qualitative Variablen
Variablenselektion

Peter Büchel

HSLU I

ASTAT: Block 13

Qualitative erklärende Variablen

- Bisher angenommen: Alle Variablen *quantitativ* in linearem Regressionssystem
- Aber: Oft sind einige erklärenden Variablen *qualitativ*

Beispiel

- Datensatz `Credit` wurde in den USA erhoben
- Enthält für eine grössere Anzahl Individuen:
  - `Balance` (monatliche Kreditkartenrechnung): Zielgrösse, quantitativ
  - `Age` (Alter): erklärend, quantitativ
  - `Cards` (Anzahl Kreditkarten): erklärend, quantitativ
  - `Education` (Anzahl Jahre Ausbildung): erklärend, quantitativ
  - `Income` (Einkommen in Tausenden Dollars): erklärend, quantitativ
  - `Limit` (Kreditkartenlimite): erklärend, quantitativ
  - `Rating` (Kreditwürdigkeit): erklärend, quantitativ

- Datensatz:

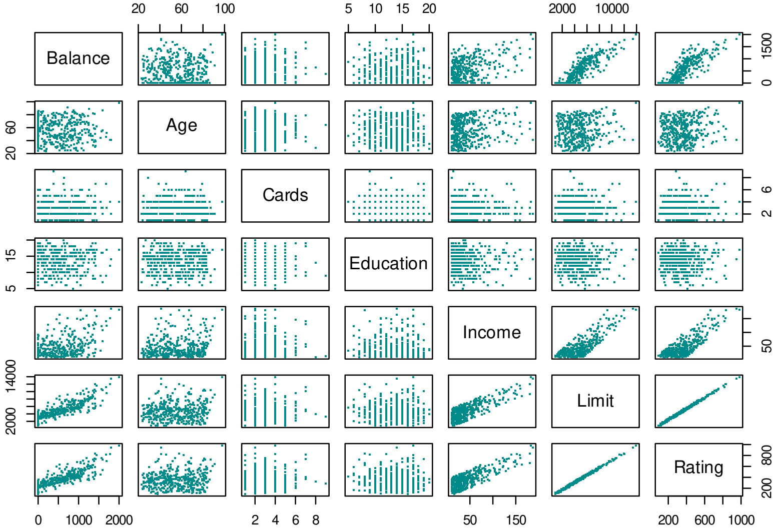
```
Credit <- read.csv("../Data/Credit.csv")[, -1]
head(Credit)

##    Income Limit Rating Cards Age Education Gender Student
## 1  14.891  3606    283     2  34        11   Male      No
## 2 106.025  6645    483     3  82        15 Female     Yes
## 3 104.593  7075    514     4  71        11   Male      No
## 4 148.924  9504    681     3  36        11 Female      No
## 5  55.882  4897    357     2  68        16   Male      No
## 6  80.180  8047    569     4  77        10   Male      No
##   Married Ethnicity Balance
## 1     Yes Caucasian     333
## 2     Yes     Asian     903
## 3      No     Asian     580
## 4      No     Asian     964
## 5     Yes Caucasian     331
## 6      No Caucasian    1151

colnames(Credit)

##  [1] "Income"    "Limit"     "Rating"    "Cards"
##  [5] "Age"       "Education" "Gender"    "Student"
##  [9] "Married"   "Ethnicity" "Balance"
```

- Abbildung:


Balance
Age
Cards
Education
Income
Limit
Rating
20 60 100
5 10 15 20
2000 10000
0 1500
20 60
2 6
5 15
50
2000 14000
200 800
0 1000 2000
2 4 6 8
50 150
200 600 1000

- Code:

```
Credit <- read.csv("../Data/Credit.csv")
pairs(~Balance + Age + Cards + Education + Income + Limit + Rating,
    data = Credit, pch = ".", col = "darkcyan")
```

- Streudiagramme von Paaren von Variablen: Identität gegeben durch entsprechenden Spalten- und Zeilenkennzeichnungen
- Plot direkt rechts des Wortes „Balance": Streudiagramm der Variablen `age` und `balance`
- Streudiagramme:
  - `Age` - `Balance`: Kein Zusammenhang
  - `Education` - `Balance`: Kein Zusammenhang
  - `Income` - `Balance`: Schwacher Zusammenhang
  - `Limit` - `Balance`: Starker Zusammenhang

- Neben quantitativen noch vier erklärende qualitative Variablen:
  - `Gender` (Geschlecht)
  - `Student` (Studentenstatus)
  - `Ethnicity` (Ethnie)
- Qualitativ erklärende Variablen heissen auch *Faktoren*
- Faktoren nehmen *Stufen* oder *Levels* an:
  - `Gender`: male, female
  - `Student`: ja, nein
  - `Ethnicity`: Kaukasier, Afroamerikaner, Asiat

Qualitative erklärende Variable mit nur zwei Levels

- Beispiel `Balance`: Unterschied zwischen Männern und Frauen
- Andere Variablen werden für den Moment ignoriert
- Qualitative erklärende Variable mit zwei *Levels* (mögliche Werte): Hinzunahme dieser Variable in Regressionsmodell sehr einfach
- Führen Indikatorvariable (oder *Dummy-Variable*) ein, die nur zwei mögliche numerische Werte annehmen kann

Beispiel

- Für `Gender`:

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ 0 & \text{falls } i\text{-te Person männlich} \end{cases}$$

- Verwenden diese Variable als erklärende Variable im Regressionsmodell
- Modell:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person männlich} \end{cases}$$

- β_0: durchschn. Kreditkartenrechnungen der Männern
- $\beta_0 + \beta_1$: durchschn. Kreditkartenrechnungen der Frauen
- β_1: durchschn. *Unterschied* der Rechnungen Männern/Frauen

- Tabelle: Koeffizientenschätzungen für unser Modell:

	Koeffizient	Std.fehler	t-Statistik	P-Wert
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[female]`	19.73	46.05	0.429	0.6690

```r
balance <- Credit[, "Balance"]
gender <- Credit[, "Gender"] == "Female"
round(summary(lm(balance ~ gender))$coef, digits = 5)
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 509.80311 33.12808 15.38885 0.00000
## genderTRUE 19.73312 46.05121 0.42850 0.66852
```

- Geschätzte durchschnittliche Rechnungen für Männer: \$ 509.80
- Geschätzter Unterschied zu Frauen: \$ 19.73
- Frauen: $\$\ 509.80 + \$\ 19.73 = \$\ 529.53$
- p-Wert für Indikatorvariable β_1 mit 0.6690 sehr hoch
- Kein statistisch signifikanter Unterschied der `balance` von Frauen und Männern

- Beispiel vorher: Frauen mit 1 und Männer mit 0 kodiert
- Völlig willkürlich
- Kodierung: *Kein* Einfluss auf Grad der Anpassung des Modells an Daten
- Unterschiedliche Kodierung: Unterschiedliche Interpretation der Koeffizienten
- Kodierung Männer mit 1 und Frauen mit 0
- Schätzung für die Parameter β_0 und β_1 \$ 529.53, resp. \$ -19.73
- Entspricht wiederum Rechnungen von:
  - Frauen: \$ 529.53
  - Männer: \$ 529.73 – \$ 19.73 = \$ 509.80
- Dasselbe Resultat wie vorher

Beispiel

- Anstatt der 0/1-Kodierung:

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ -1 & \text{falls } i\text{-te Person männlich} \end{cases}$$

- Regressionsmodell:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person männlich} \end{cases}$$

- β_0: Durchschn. Rechnungen ohne Berücksichtigung des Geschlechts

- β_1: Wert, mit welchem Frauen über dem Durchschnitt liegen und mit welchem Männer unter dem Durchschnitt liegen

- β_0 durch \$ 519.665 geschätzt: Durchschn. Rechnungen von \$ 509.80 für Männer und von \$ 529.53 für Frauen
- Schätzung \$ 9.865 für β_1: Hälfte vom Unterschied \$ 19.73 zwischen Männern und Frauen
- Wichtig: Vorhersagen für d Zielgrösse hängen *nicht* von Kodierung ab
- Einziger Unterschied: Interpretation der Koeffizienten

Qualitative erklärende Variablen mit mehr als zwei Levels

- Qualitative erklärende Variable kann mehr als zwei Levels haben
- *Eine* Indikatorvariable für alle möglichen Werte reicht nicht
- In dieser Situation: Zusätzliche Indikatorvariable hinzufügen

Beispiel

- Variable `Ethnicity`: *Drei* mögliche Levels

- Wählen *zwei* verschiedene Indikatorvariablen

- *Wahl* der 1. Indikatorvariablen:

$$x_{i1} = \begin{cases} 1 & \text{falls } i\text{-te Person asiatisch} \\ 0 & \text{falls } i\text{-te Person nicht asiatisch} \end{cases}$$

- 2. Indikatorvariable:

$$x_{i2} = \begin{cases} 1 & \text{falls } i\text{-te Person kaukasisch} \\ 0 & \text{falls } i\text{-te Person nicht kaukasisch} \end{cases}$$

- Beide Variablen in Regressionsgleichung aufnehmen:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person asiatisch} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person kaukasisch} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person afroamerikanisch} \end{cases}$$

- β_0: Durchschn. Kreditkartenrechnungen von Afroamerikanern
- β_1: Differenz der durchschn. Rechnungen von Afroamerikanern und Asiaten
- β_2: Differenz der durchschn. Rechnungen von Afroamerikanern und Kaukasiern

Bemerkungen

- Es gibt immer eine Indikatorvariable weniger, als es Levels hat
- Level ohne Indikatorvariable (hier Afroamerikaner): *Baseline*
- Folgende Gleichung macht *keinen* Sinn:

$$y_i = \beta_0 + \beta_1 + \beta_2 + \varepsilon_i$$

  - Person müsste asiatisch *und* kaukasisch sein

- Output: Geschätzte `balance` $ 531.00 für Baseline (Afroamerikaner):

```
balance <- Credit[, "Balance"]
ethnicity <- Credit[, "Ethnicity"]
summary(lm(balance ~ ethnicity))
```

```
##
## Call:
## lm(formula = balance ~ ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -531.00 -457.08  -63.25  339.25 1480.50
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          531.00      46.32  11.464   <2e-16 ***
## ethnicityAsian       -18.69      65.02  -0.287    0.774
## ethnicityCaucasian   -12.50      56.68  -0.221    0.826
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
## Multiple R-squared:  0.0002188,Adjusted R-squared:  -0.004818
## F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

- Schätzung für Kategorie Asiaten: \$ -18.69

- Durchschn. Rechnungen um diesen Betrag kleiner als die von Afroamerikanern

- Kaukasier haben um durchschn. \$ 12.50 kleinere Rechnungen als die Afroamerikaner

- p-Werte gross $\quad\rightarrow\quad$ Zufällige Abweichungen

- Kein signifikanter Unterschied bei den Kreditkartenrechnungen zwischen den Ethnien

- Level, für Baseline willkürlich

- Vorhersage der Zielvariable hängt nicht von der Kodierung ab

- p-Werte hängen von der Kodierung ab
- F-Statistik betrachten
- F-Test und testen

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

- p-Wert dieser Statistik hängt *nicht* von der Kodierung ab
- p-Wert 0.96 $\quad \rightarrow \quad$ Relativ hoch
- Vermutung bestätigt: Nullhypothese *nicht* verwerfen
- Es gibt keinen Zusammenhang zwischen `balance` und `ethnicity`

- Indikatorvariablen: Qualitative *und* quantitative erklärende Variablen in Regressionsmodell integrieren
- Regression von `Balance` mit quantitativer erklärenden Variable `Income` und qualitativer erklärenden Variable `student` durchführen
- `Student` mit Indikatorvariablen
- Multiple lineare Regression

Beispiel: Datensatz `Credit`

- Zielgrösse `Balance` durch die erklärenden Variablen `Income` (quantitativ) und `Student` (qualitativ) vorhersagen

- Ohne Interaktionsterm:

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 & \text{falls } i\text{-te Person Student} \\ 0 & \text{falls } i\text{-te Person kein Student} \end{cases} \\
&= \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{falls } i\text{-te Person Student} \\ \beta_0 & \text{falls } i\text{-te Person kein Student} \end{cases}
\end{aligned}
$$

- Output:

```
student <- Credit[, "Student"]
income <- Credit[, "Income"]
summary(lm(balance ~ income + student))

##
## Call:
## lm(formula = balance ~ income + student)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -762.37 -331.38  -45.04  323.60  818.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 211.1430    32.4572   6.505 2.34e-10 ***
## income        5.9843     0.5566  10.751  < 2e-16 ***
## studentYes  382.6705    65.3108   5.859 9.78e-09 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.8 on 397 degrees of freedom
## Multiple R-squared:  0.2775,Adjusted R-squared:  0.2738
## F-statistic: 76.22 on 2 and 397 DF,  p-value: < 2.2e-16
```

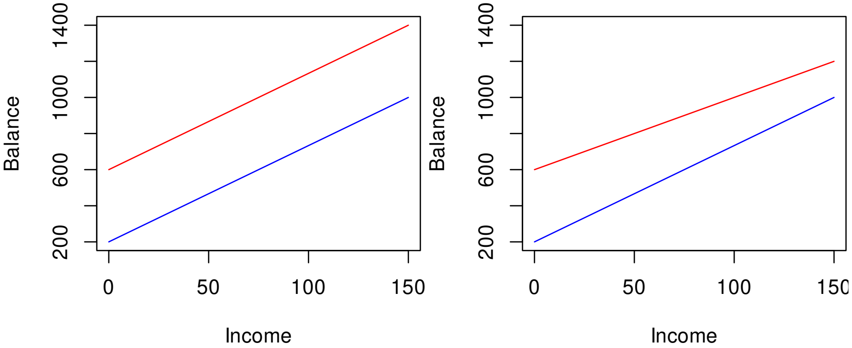
- $\widehat{\beta}_0$:
  Ohne Einkommen und als Nichtstudent zahlt man \$211 monatliche Kreditkartenrechnung

- $\widehat{\beta}_1$:
  Pro \$1000 Einkommen mehr, zahlt man \$6 mehr Kreditkartenrechnung (unabhängig vom Studentenstatus)

- $\widehat{\beta}_2$:
  Studierende zahlen \$383 mehr Kreditkartenrechnung als Nichtstudierende (unabhängig vom Einkommen)

- Modell beschreibt zwei parallele Geraden: eine für Studierende und eine für Nichtstudierende
  - Steigung β_1 ist bei beiden gleich
  - y-Achsenabschnitte sind verschieden ($\beta_0 + \beta_2$ und β_0)

- Abbildung links:



- Durchschn. Zunahme von `Balance` für Vergrösserung von `Income` um eine Einheit hängt nicht davon ab, ob entsprechendes Individuum studiert oder nicht

- Mögliche Einschränkung des Modells: Änderung in `Income` kann eine unterschiedliche Wirkung auf Rechnungen haben kann, ob jemand studiert oder nicht

- Lockerung dieser Einschränkung: Einführung einer Interaktionsvariablen

- `Income` wird mit der Indikatorvariablen für `Student` „multipliziert"

- Modell:

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \texttt{income}_i & \text{falls studierend} \\ 0 & \text{falls nicht studierend} \end{cases}$$
$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \texttt{income}_i & \text{falls studierend} \\ \beta_0 + \beta_1 \cdot \texttt{income}_i & \text{falls nicht studierend} \end{cases}$$

- Zwei unterschiedliche Regressionsgeraden für Studierende und Nichtstudierende (Abbildung oben rechts):
  - Verschiedene Steigungen $\beta_1 + \beta_3$ und β_1
  - Unterschiedliche y-Achsenabschnitte $\beta_0 + \beta_2$ und β_0
- Möglichkeit, Änderung der Zielgrösse (Kreditkartenrechnungen) aufgrund der Änderungen im Einkommen für Studenten und Nichtstudenten getrennt zu betrachten

- Rechte Seite von Abbildung oben: Geschätzter Zusammenhang zwischen `income` und `balance` für Studenierende (rot) und Nichtstudierende (blau)

- Steigung für Studierende ist grösser als für Nichtstudierende

- Deutet an: Zunahme im Einkommen eines Studierenden eine grössere Zunahme der Kreditkartenrechnungen zur Folge hat als für Nichtstudierenden

- Output:

```
summary(lm(balance ~ income * student))
##
## Call:
## lm(formula = balance ~ income * student)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.39 -325.70  -41.13  321.65  814.04
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        200.6232    33.6984   5.953 5.79e-09 ***
## income               6.2182     0.5921  10.502  < 2e-16 ***
## studentYes         476.6758   104.3512   4.568 6.59e-06 ***
## income:studentYes   -1.9992     1.7313  -1.155    0.249
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.6 on 396 degrees of freedom
## Multiple R-squared:  0.2799,Adjusted R-squared:  0.2744
## F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

- p-Wert der Interaktion ist statistisch nicht signifikant
- Somit gibt es keine Interaktion
- Steigungen der beiden Geraden sind nicht signifikant unterschiedlich

Variablenselektion

- Lineares Standardregressionsmodell:

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

- Beschreibung des Zusammenhanges zwischen der Zielvariable Y und den erklärenden Variablen $X_1, X_2, \ldots, X_p$ verwendet

- Schon gesehen: Nicht alle erklärenden Variablen spielen eine Rolle für die Vorhersage der Zielgrösse

Schrittweise Vorwärtsselektion

- *Schrittweise Vorwärtsselektion*: Rechnerisch effiziente Methode, um Variablen zu eliminieren
- Beginnt mit Modell, das gar keine erklärenden Variablen enthält
- Dann wird schrittweise eine Variable um die andere zum Modell hinzugefügt, bis alle Variablen im Modell sind
- In jedem Schritt wird jene Variable ins Modell aufgenommen, die die grösste *zusätzliche* Verbesserung der Anpassung mit sich bringt

Credit

- *Nullmodell* $\mathcal{M}_0$: Enthält keine erklärenden Variablen:

$$\texttt{Balance} = \beta_0 + \varepsilon$$

- Fügen eine erklärende Variable zum Nullmodell hinzu
- R-Befehl `add1`: Jede vorkommende Variable wird getrennt addiert:

```
f.full <- lm(Balance ~ Income + Limit + Rating + Cards + Age +
    Education + Gender + Student + Married + Ethnicity, data = Credit)

f.empty <- lm(Balance ~ NULL, data = Credit)

add1(f.empty, scope = f.full)
```

```
## Single term additions
##
## Model:
## Balance ~ NULL
##           Df Sum of Sq      RSS    AIC
## <none>                 84339912 4905.6
## Income     1  18131167 66208745 4810.7
## Limit      1  62624255 21715657 4364.8
## Rating     1  62904790 21435122 4359.6
## Cards      1    630416 83709496 4904.6
## Age        1       284 84339628 4907.6
## Education  1      5481 84334431 4907.5
## Gender     1     38892 84301020 4907.4
## Student    1   5658372 78681540 4879.8
## Married    1      2715 84337197 4907.5
## Ethnicity  2     18454 84321458 4909.5
```

- Wählen *beste* Variable aus: Kleinster RSS-Wert
- RSS: Summe der Quadrate der Residuen (Abstände von Punkten zu Geraden, Ebene, usw.)
- Je kleiner der RSS-Wert ist, umso besser passen die Daten zum System (Gerade, Ebene, usw.)
- Damit passt diese Variable am besten zu den Daten
- Hier: Variable `Rating`
- Modell $\mathcal{M}_1$:

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Rating} + \varepsilon$$

- Zu diesem Modell fügen wir nun eine weitere Variable hinzu

- update-Befehl und dann wiederum mit dem add1-Befehl aus

```
f.1 <- update(f.empty, . ~ . + Rating)

add1(f.1, scope = f.full)
## Single term additions
##
## Model:
## Balance ~ Rating
##           Df Sum of Sq      RSS    AIC
## <none>                 21435122 4359.6
## Income     1  10902581 10532541 4077.4
## Limit      1      7960 21427162 4361.5
## Cards      1    138580 21296542 4359.0
## Age        1    649110 20786012 4349.3
## Education  1     27243 21407879 4361.1
## Gender     1     16065 21419057 4361.3
## Student    1   5735163 15699959 4237.1
## Married    1    118209 21316913 4359.4
## Ethnicity  2     51100 21384022 4362.7
```

- Wählen wieder diejenige Variable aus, aufgrund welcher das ergänzte Regressionsmodell den kleinsten RSS-Wert hat
- Dies ist in diesem Fall `Income`
- Modell $\mathcal{M}_2$:

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Rating} + \beta_2 \cdot \texttt{Income} + \varepsilon$$

- Verfahren wiederholt sich
- Fügen jene Variable zum Modell $\mathcal{M}_2$ hinzu, aufgrund welcher das neue Regressionsmodell den kleinsten RSS-Wert hat.

```
f.2 <- update(f.1, . ~ . + Income)

add1(f.2, scope = f.full)

## Single term additions
##
## Model:
## Balance ~ Rating + Income
##           Df Sum of Sq      RSS    AIC
## <none>                 10532541 4077.4
## Limit      1     94545 10437996 4075.8
## Cards      1      2094 10530447 4079.3
## Age        1     90286 10442255 4076.0
## Education  1     20819 10511722 4078.6
## Gender     1       948 10531593 4079.4
## Student    1   6305322  4227219 3714.2
## Married    1     95068 10437473 4075.8
## Ethnicity  2     67040 10465501 4078.9
```

- Dies ist hier `Student`
- Modell $\mathcal{M}_3$:

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Rating} + \beta_2 \cdot \texttt{Income} + \beta_3 \cdot \texttt{Student} + \varepsilon$$

- Erhalten 11 Modelle $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{10}$
- Welches ist nun aber das beste unter diesen 11 Modellen?
- Als Entscheidungskriterium: AIC-Wert (letzten Spalte)
- Aufgrund von diesem Wert lassen sich verschiedene Modelle miteinander vergleichen

- Aufgeführtes Verfahren mit `update` und `add1` ziemlich mühsam
- Befehl `regsubsets` aus der library `leaps` führt das gesamte Verfahren automatisch durch

```r
library(leaps)
reg <- regsubsets(Balance ~ ., data = Credit, method = "forward",
    nvmax = 11)

reg.sum <- summary(reg)

reg.sum$which
```

```
##    (Intercept)     X Income Limit Rating Cards   Age
## 1         TRUE FALSE  FALSE FALSE   TRUE FALSE FALSE
## 2         TRUE FALSE   TRUE FALSE   TRUE FALSE FALSE
## 3         TRUE FALSE   TRUE FALSE   TRUE FALSE FALSE
## 4         TRUE FALSE   TRUE  TRUE   TRUE FALSE FALSE
## 5         TRUE FALSE   TRUE  TRUE   TRUE  TRUE FALSE
## 6         TRUE FALSE   TRUE  TRUE   TRUE  TRUE  TRUE
## 7         TRUE FALSE   TRUE  TRUE   TRUE  TRUE  TRUE
## 8         TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
## 9         TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
## 10        TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
## 11        TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
##    Education GenderFemale StudentYes MarriedYes
## 1      FALSE        FALSE      FALSE      FALSE
## 2      FALSE        FALSE      FALSE      FALSE
## 3      FALSE        FALSE       TRUE      FALSE
## 4      FALSE        FALSE       TRUE      FALSE
## 5      FALSE        FALSE       TRUE      FALSE
## 6      FALSE        FALSE       TRUE      FALSE
## 7      FALSE         TRUE       TRUE      FALSE
## 8      FALSE         TRUE       TRUE      FALSE
## 9      FALSE         TRUE       TRUE      FALSE
## 10     FALSE         TRUE       TRUE       TRUE
## 11     FALSE         TRUE       TRUE       TRUE
##    EthnicityAsian EthnicityCaucasian
## 1           FALSE              FALSE
## 2           FALSE              FALSE
## 3           FALSE              FALSE
## 4           FALSE              FALSE
## 5           FALSE              FALSE
## 6           FALSE              FALSE
## 7           FALSE              FALSE
## 8           FALSE              FALSE
## 9            TRUE              FALSE
## 10           TRUE              FALSE
## 11           TRUE               TRUE
```

- Überall, wo `TRUE` steht, kommt entsprechende erklärende Variable vor
- Modell mit drei erklärenden Variablen: `Income`, `Rating` und `Student`
- Formal: Schrittweise Vorwärtsselektion
  - Sei $\mathcal{M}_0$ das Nullmodell, dass keine erklärende Variablen enthält
  - Sei nun $k = 0, \ldots, p-1$:
    - Betrachten alle $p-k$ Modelle, die die Anzahl Variablen in $\mathcal{M}_k$ um eine zusätzliche erklärende Variable erhöhen.
    - Wählen das *beste* aus diesen $p-k$ Modellen aus $\quad\rightarrow\quad \mathcal{M}_{k+1}$
    - „Bestes" Modell: jenes mit dem kleinsten RSS oder grösstem R^2

Schrittweise Rückwärtsselektion

- *Schrittweise Rückwärtsselektion* ist rechnerisch ebenfalls effizient und funktioniert ähnlich wie die schrittweise Vorwärtsselektion
- Beginnen allerdings mit dem vollen Modell, das alle erklärenden Variablen enthält
- Dann wird schrittweise eine Variable um die andere vom Modell entfernt, bis keine erklärende Variable mehr im Modell vorhanden ist
- In jedem Schritt wird jene Variable vom Modell entfernt, die am wenigsten nützlich ist
- Lassen die Variable weg, die den grösstem p-Wert hat

- Auch hier nimmt uns der `regsubsets`-Befehl die ganze Arbeit ab:

```
reg <- regsubsets(Balance ~ ., data = Credit, method = "backward",
    nvmax = 11)
reg.sum <- summary(reg)
reg.sum$which
```

```
##    (Intercept)     X Income Limit Rating Cards   Age
## 1         TRUE FALSE  FALSE  TRUE  FALSE FALSE FALSE
## 2         TRUE FALSE   TRUE  TRUE  FALSE FALSE FALSE
## 3         TRUE FALSE   TRUE  TRUE  FALSE FALSE FALSE
## 4         TRUE FALSE   TRUE  TRUE  FALSE  TRUE FALSE
## 5         TRUE FALSE   TRUE  TRUE   TRUE  TRUE FALSE
## 6         TRUE FALSE   TRUE  TRUE   TRUE  TRUE  TRUE
## 7         TRUE FALSE   TRUE  TRUE   TRUE  TRUE  TRUE
## 8         TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
## 9         TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
## 10        TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
## 11        TRUE  TRUE   TRUE  TRUE   TRUE  TRUE  TRUE
##    Education GenderFemale StudentYes MarriedYes
## 1      FALSE        FALSE      FALSE      FALSE
## 2      FALSE        FALSE      FALSE      FALSE
## 3      FALSE        FALSE       TRUE      FALSE
## 4      FALSE        FALSE       TRUE      FALSE
## 5      FALSE        FALSE       TRUE      FALSE
## 6      FALSE        FALSE       TRUE      FALSE
## 7      FALSE         TRUE       TRUE      FALSE
## 8      FALSE         TRUE       TRUE      FALSE
## 9      FALSE         TRUE       TRUE      FALSE
## 10     FALSE         TRUE       TRUE       TRUE
## 11     FALSE         TRUE       TRUE       TRUE
##    EthnicityAsian EthnicityCaucasian
## 1           FALSE              FALSE
## 2           FALSE              FALSE
## 3           FALSE              FALSE
## 4           FALSE              FALSE
## 5           FALSE              FALSE
## 6           FALSE              FALSE
## 7           FALSE              FALSE
## 8           FALSE              FALSE
## 9            TRUE              FALSE
## 10           TRUE              FALSE
## 11           TRUE               TRUE
```

- Modell mit drei erklärenden Variablen: `Income`, `Limit` und `Student`

- Dieses Modell unterscheidet sich also vom Modell mit drei Variablen, das durch Vorwärtsselektion gewonnen wurde

- Hier kommt `Rating` anstelle von `Limit` vor

- Es gibt noch weitere Selektionsmethoden $\rightarrow$ Nicht hier (Machine Learning)

Wieviele Variablen wählen wir?

- Vorwärts- und Rückwärtsselektion: Nur beschrieben, *wie* Variablen ausgewählt werden, aber nicht *wieviele*
- Problem Vorwärtsselektion: RSS nimmt mit zunehmender Zahl von Variablen ab
- Je mehr Variablen umso besser das System bez. RSS
- Aber: Nicht alle Variablen spielen eine Rolle
- Beispiel: Resultate bei Prüfung abhängig von
  - Zeit fürs Lernen
  - Konzentriertheit
  - ...
  - Haarfarbe

- Nehmen wir die Variable Haarfarbe, dazu so wird RSS grösser
- Aber (nehme ich an) keinen Einfluss auf Prüfungsresultate
- Kann weggelassen werden
- RSS hier kein Mass für die Güte des Modelles
- Sagt nichts aus, wieviele Variablen wir wählen sollen
- Es gibt mehrere Gütekriterien, die abhängig sind von der Anzahl der Variablen
- Heisst: Sie nehmen nicht automatisch mit zunehmender Anzahl Variablen zu oder ab
- Beispiel: Adjusted-R^2, AIC, BIC, Mallow's C_p

- Gleiche Idee wie bei Vorwärtsselektion
- Beispiel: AIC (Akaike information criterion)
- Kleiner AIC-Wert ist besser
- Variablen werden addiert, solange AIC-Wert abnimmt
- Schritte werden hier manuell durchgemacht, damit man sieht, was passiert

- Beginnen wieder mit leerem Modell und addieren jeweils eine Variable

```r
f.full <- lm(Balance ~ Income + Limit + Rating + Cards + Age +
    Education + Gender + Student + Married + Ethnicity, data = Credit)

f.empty <- lm(Balance ~ NULL, data = Credit)

add1(f.empty, scope = f.full)
```

```
## Single term additions
##
## Model:
## Balance ~ NULL
##              Df  Sum of Sq      RSS    AIC
## <none>                     84339912 4905.6
## Income        1   18131167 66208745 4810.7
## Limit         1   62624255 21715657 4364.8
## Rating        1   62904790 21435122 4359.6
## Cards         1     630416 83709496 4904.6
## Age           1        284 84339628 4907.6
## Education     1       5481 84334431 4907.5
## Gender        1      38892 84301020 4907.4
## Student       1    5658372 78681540 4879.8
## Married       1       2715 84337197 4907.5
## Ethnicity     2      18454 84321458 4909.5
```

- Betrachten nun AIC-Wert, anstatt RSS
- Wählen Variable mit kleinstem AIC: `Rating`
- Addieren diese Variable zum leeren System
- Addieren dann jeweils alle anderen Variablen und betrachten AIC:

```
f.1 <- update(f.empty, . ~ . + Rating)
add1(f.1, scope = f.full)
```

```
## Single term additions
##
## Model:
## Balance ~ Rating
##           Df Sum of Sq      RSS    AIC
## <none>                 21435122 4359.6
## Income     1  10902581 10532541 4077.4
## Limit      1      7960 21427162 4361.5
## Cards      1    138580 21296542 4359.0
## Age        1    649110 20786012 4349.3
## Education  1     27243 21407879 4361.1
## Gender     1     16065 21419057 4361.3
## Student    1   5735163 15699959 4237.1
## Married    1    118209 21316913 4359.4
## Ethnicity  2     51100 21384022 4362.7
```

- Income hinzunehmen:

```
f.2 <- update(f.1, . ~ . + Income)

add1(f.2, scope = f.full)
## Single term additions
##
## Model:
## Balance ~ Rating + Income
##           Df Sum of Sq      RSS    AIC
## <none>                 10532541 4077.4
## Limit      1     94545 10437996 4075.8
## Cards      1      2094 10530447 4079.3
## Age        1     90286 10442255 4076.0
## Education  1     20819 10511722 4078.6
## Gender     1       948 10531593 4079.4
## Student    1   6305322  4227219 3714.2
## Married    1     95068 10437473 4075.8
## Ethnicity  2     67040 10465501 4078.9
```

- Student hinzunehmen:

```
f.3 <- update(f.2, . ~ . + Student)

add1(f.3, scope = f.full)
## Single term additions
##
## Model:
## Balance ~ Rating + Income + Student
##           Df Sum of Sq     RSS    AIC
## <none>                 4227219 3714.2
## Limit      1    194718 4032502 3697.4
## Cards      1     10608 4216611 3715.2
## Age        1     44620 4182600 3712.0
## Education  1      1400 4225820 3716.1
## Gender     1     12168 4215051 3715.1
## Married    1     13083 4214137 3715.0
## Ethnicity  2     22322 4204897 3716.1
```

- `Limit` hinzunehmen:

```
f.4 <- update(f.3, . ~ . + Limit)

add1(f.4, scope = f.full)

## Single term additions
##
## Model:
## Balance ~ Rating + Income + Student + Limit
##           Df Sum of Sq     RSS    AIC
## <none>                 4032502 3697.4
## Cards      1    166410 3866091 3682.5
## Age        1     37952 3994549 3695.6
## Education  1      5795 4026707 3698.8
## Gender     1     13345 4019157 3698.0
## Married    1      6660 4025842 3698.7
## Ethnicity  2     17704 4014797 3699.6
```

- Cards hinzunehmen:

```
f.5 <- update(f.4, . ~ . + Cards)

add1(f.5, scope = f.full)

## Single term additions
##
## Model:
## Balance ~ Rating + Income + Student + Limit + Cards
##            Df Sum of Sq     RSS    AIC
## <none>                  3866091 3682.5
## Age         1     44472 3821620 3679.9
## Education   1      5672 3860419 3683.9
## Gender      1     11350 3854741 3683.3
## Married     1      3121 3862970 3684.2
## Ethnicity   2     14756 3851335 3685.0
```

- Age hinzunehmen:

```
f.6 <- update(f.5, . ~ . + Age)

add1(f.6, scope = f.full)
## Single term additions
##
## Model:
## Balance ~ Rating + Income + Student + Limit + Cards + Age
##           Df Sum of Sq     RSS    AIC
## <none>                 3821620 3679.9
## Education  1    5241.7 3816378 3681.3
## Gender     1   10860.9 3810759 3680.7
## Married    1    5450.6 3816169 3681.3
## Ethnicity  2   11517.3 3810102 3682.7
```

- Hier hört der Prozess auf: Nehmen wir noch eine Variable hinzu, so wird der AIC-Wert *grösser*
- Das Modell wird dann schlechter
- Die restlichen Variablen werden weggelassen