

# Statistik

CH.12 - Multiple Regression

SS 2021 | | Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

- 1 Multiple Lineare Regression
- 2 Hypothesentests
- 3 Residualdiagnostik
- 4 Multikollinearität
- 5 Nichtlinearität

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon$$

- Mit Hilfe der **einfachen linearen Regression** kann der Zusammenhang einer abhängigen Variablen  $Y$  mit **einer** unabhängigen Variablen  $X$  modelliert werden.
- Die **multiple lineare Regression** erlaubt das Modellieren des Zusammenhangs einer abhängigen Variablen  $Y$  mit **mehreren** unabhängigen Variablen  $X_1, X_2, \dots, X_p$ .

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

- Die zuvor diskutierte einfache lineare Regression kann als **Spezialfall** der multiple linearen Regression aufgefasst werden bei der gilt  $p = 1$ .
- Wir nehmen weiterhin an, dass *innerhalb des Wertebereichs* der Daten, der wahre Zusammenhang zwischen  $Y$  und den Prädiktoren durch ein lineares Modell **approximiert** werden kann.
- Jeder Regressor geht mit einem eigenen Koeffizienten  $\beta_0, \beta_2, \dots, \beta_p$  in die Gleichung ein. Der Fehlerterm  $\epsilon$  enthält zudem keine **systematischen Informationen** zur Erklärung der Streuung von  $Y$  die nicht bereits durch die Regressoren abgebildet wurden.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

$$i = 1, 2, \dots, n$$

- Aus der Modellgleichung folgt die obige Darstellung für jede Beobachtung. Dabei repräsentiert  $y_i$  die  $i$ -te Beobachtung der abhängigen Variablen  $Y$ . Die Werte  $x_{i1}, x_{i2}, \dots, x_{ip}$  sind die Werte der zugehörigen Regressoren, für die  $i$ -te Beobachtung in der Stichprobe (üblicherweise  $i$ -te Zeile im Datensatz).
- Der Wert  $\epsilon_i$  ist der Anpassungsfehler (Fehlerterm) der linearen Approximation für die  $i$ -te Beobachtungseinheit.

# Beispiel: Autodaten

d					
##	l100km	weight	hp	cyl	hub
## Mazda RX4	11.200714	1.1884110	110	6	2.621936
## Mazda RX4 Wag	11.200714	1.3040770	110	6	2.621936
## Datsun 710	10.316447	1.0523334	93	4	1.769807
## Hornet 4 Drive	10.991355	1.4582983	110	6	4.227872
## Hornet Sportabout	12.578342	1.5603565	175	8	5.899356
## Valiant	12.995304	1.5694283	105	6	3.687098
## Duster 360	16.448601	1.6193234	245	8	5.899356
## Merc 240D	9.639959	1.4469585	62	4	2.403988
## Merc 230	10.316447	1.4288148	95	4	2.307304
## Merc 280	12.250781	1.5603565	123	6	2.746478
## Merc 280C	13.214326	1.5603565	123	6	2.746478
## Merc 450SE	14.342378	1.8461194	180	8	4.519562
## Merc 450SL	13.596243	1.6918982	180	8	4.519562
## Merc 450SLC	15.474671	1.7145778	180	8	4.519562
## Cadillac Fleetwood	22.616827	2.3813580	205	8	7.734711
## Lincoln Continental	22.616827	2.4602830	215	8	7.538066
## Chrysler Imperial	16.001020	2.4244492	230	8	7.210324
## Fiat 128	7.259722	0.9979024	66	4	1.289665
## Honda Civic	7.737336	0.7325511	52	4	1.240503
## Toyota Corolla	6.938496	0.8323413	65	4	1.165123
## Toyota Corona	10.940233	1.1181043	97	4	1.968091
## Dodge Challenger	15.175161	1.5966438	150	8	5.211098
## AMC Javelin	15.474671	1.5580885	150	8	4.981678
## Camaro Z28	17.685338	1.7417933	245	8	5.735485
## Pontiac Firebird	12.250781	1.7440612	175	8	6.554840
## Fiat X1-9	8.615934	0.8777005	66	4	1.294581
## Porsche 914-2	9.046731	0.9706869	91	4	1.971368

## Datenbeschreibung

l100km Kraftstoffverbrauch  
in Litern pro 100km bei  
normaler Fahrweise.

weight Fahrzeuggewicht in  
Tonnen.

hp Motorleistung in PS.

cyl Anzahl der Zylinder des  
Fahrzeugmotors.

hub Hubraum des Motors in  
Litern.

# Beispiel: Autodaten

```
dim(d) # Anzahl Beobachtungen und Anzahl Variablen
```

```
## [1] 32 5
```

```
t(sapply(d, summary)) # Deskriptive Statistik für alle Variablen
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
## l100km	6.9384956	10.316447	12.250781	12.755060	15.250039	22.616827
## weight	0.6862847	1.170834	1.508193	1.459319	1.637467	2.460283
## hp	52.0000000	96.500000	123.000000	146.687500	180.000000	335.000000
## cyl	4.0000000	4.000000	6.000000	6.187500	8.000000	8.000000
## hub	1.1651228	1.979971	3.216788	3.780862	5.342195	7.734711

## Beispiel: Autodaten

```
round(var(d),4)           # Varianz-Kovarianz-Matrix
```

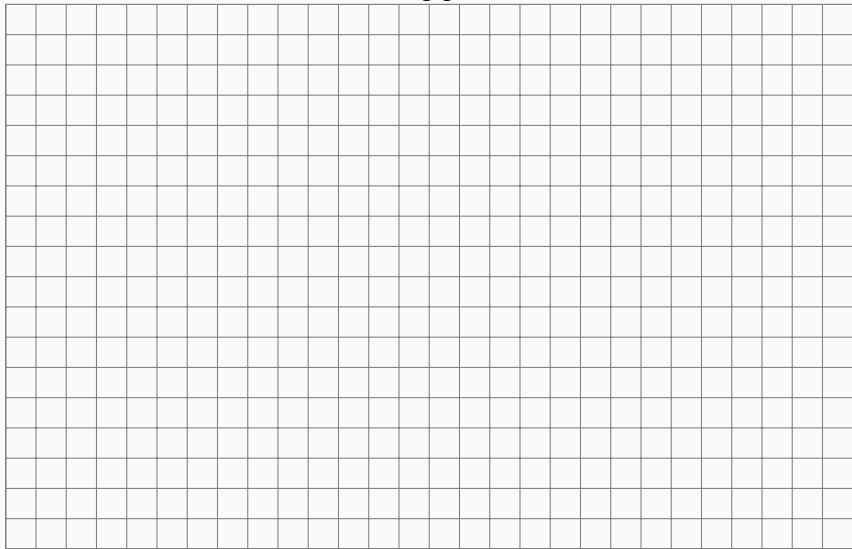
```
##           l100km weight           hp           cyl           hub
## l100km  14.9247  1.5258  202.0862   5.6144   6.9033
## weight   1.5258  0.1970   20.0454   0.6202   0.8004
## hp       202.0862 20.0454 4700.8669 101.9315 110.1403
## cyl       5.6144  0.6202  101.9315   3.1895   3.2719
## hub       6.9033  0.8004  110.1403   3.2719   4.1249
```

```
round(cor(d),4)           # Paarweise Korrelationskoeffizienten
```

```
##           l100km weight           hp           cyl           hub
## l100km  1.0000  0.8899  0.7629  0.8137  0.8798
## weight  0.8899  1.0000  0.6587  0.7825  0.8880
## hp       0.7629  0.6587  1.0000  0.8324  0.7909
## cyl       0.8137  0.7825  0.8324  1.0000  0.9020
## hub       0.8798  0.8880  0.7909  0.9020  1.0000
```



Wovon ist die Größe  $l_{100\text{km}}$  abhängig?



$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ \text{Kraftstoffverbrauch} &= \beta_0 + \beta_1 \text{Gewicht} + \beta_2 \text{Motorleistung} + \epsilon \end{aligned}$$

- Wir nehmen an, dass  $Y$  linear von (mindestens) zwei erklärenden Variablen abhängig ist.
- Diese Annahme muss verifiziert werden (was wir zunächst ignorieren), da sonst nicht sichergestellt ist, dass diese Variablen einen Einfluss haben oder entscheidende Variablen im Modell fehlen.

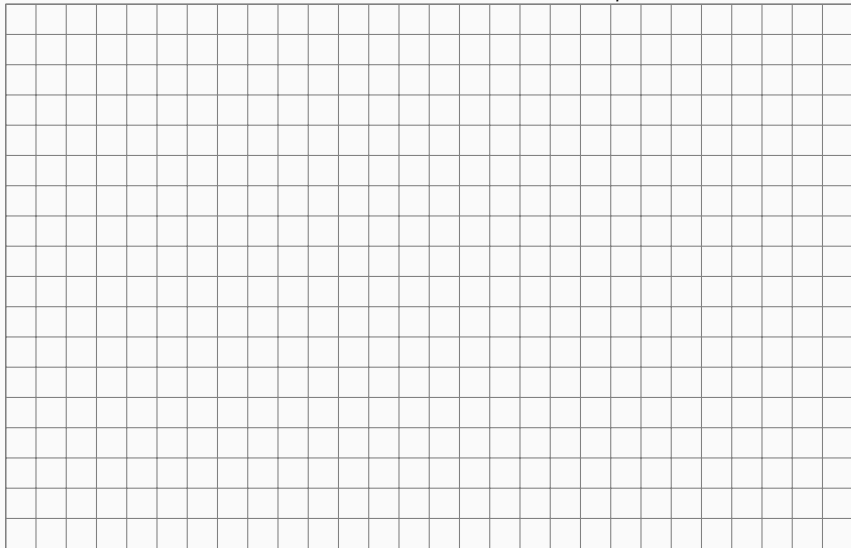
# Multiple Lineare Regression

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ \text{Kraftstoffverbrauch} &= \beta_0 + \beta_1 \text{Gewicht} + \beta_2 \text{Motorleistung} + \epsilon \end{aligned}$$

- Wir nehmen an, dass  $Y$  linear von (mindestens) zwei erklärenden Variablen abhängig ist.
- Diese Annahme muss verifiziert werden (was wir zunächst ignorieren), da sonst nicht sichergestellt ist, dass diese Variablen einen Einfluss haben oder entscheidende Variablen im Modell fehlen.

Wie können die Parameter  $\beta_0, \beta_1, \dots, \beta_p$  bei der multiplen linearen Regression bestimmt werden?

Wie können die Regressionsparameter  $\beta_0, \beta_1, \dots, \beta_p$  bestimmt werden?



- **Lösung:** Minimieren der Fehlerquadratsumme nach dem Prinzip der kleinsten Quadrate (Kleinste-Quadrate-Schätzung).
- Der Anpassungsfehler für jede Beobachtung ergibt sich aus der umgestellten Beobachtungsgleichung:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip}$$

- Die zu minimierende Funktion in Abhängigkeit der Parameter  $\beta_0, \beta_1, \dots, \beta_p$  ergibt sich damit wie folgt:

$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2$$

- $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  bzw.  $b_0, b_1, \dots, b_p$  sind die Werte, die die Funktion  $S(\ )$  minimieren.

## Your Turn

Schreiben Sie die zugehörige Regressionsgleichung auf.

```
mod <- lm(l100km ~ 1 + weight + hp, data=d)
mod

##
## Call:
## lm(formula = l100km ~ 1 + weight + hp, data = d)
##
## Coefficients:
## (Intercept)      weight          hp
##      1.48306      5.95580      0.01759
```

- Multiple lineare Regressionsmodelle können in R ebenfalls mit Hilfe der Funktion `lm()` geschätzt werden.
- R greift für die Bestimmung der Parameterschätzer ebenfalls auf die Methode der kleinsten Quadrate zurück.

- Einfache Regressionsmodelle (nur  $X_1$ ) können als Gerade dargestellt werden. Multiple Regressionsmodelle ( $X_1$  und  $X_2$ ) können mit einer Ebene oder als Hyperebene (mehr als zwei Prädiktoren) dargestellt werden. Diese Darstellung wird sehr schnell unübersichtlich.
- $\beta_0$  ist der Achsenabschnitt und der abgebildete Wert von  $Y$ , wenn  $X_1 = X_2 = \dots = X_p = 0$ .
- Die Steigungskoeffizienten  $\beta_j$  haben mehrere Interpretationen:
  - $\beta_j$  ist die **Veränderung** in  $Y$  wenn sich  $X_j$  um eine Einheit erhöht und alle anderen Prädiktoren konstant gehalten werden (ceteris paribus).
  - $\beta_j$  wird also als **Partialeffekt** bezeichnet, weil er den Effekt von  $X_j$  auf  $Y$  abbildet, nachdem die Zielvariable um die Effekte der anderen Variablen adjustiert wurde.

## Your Turn

Interpretieren Sie die Schätzergebnisse ( $\alpha = 0.05$ ) und das Gütemaß des Regressionsmodells.

```
mod <- lm(l100km ~ 1 + weight + hp, data=d)
summary(mod)

##
## Call:
## lm(formula = l100km ~ 1 + weight + hp, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.9678 -1.1667  0.1802  0.9415  3.3444
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.483062    0.962884   1.540  0.13435
## weight       5.955801    0.840115   7.089 8.45e-08 ***
## hp           0.017592    0.005438   3.235 0.00303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.562 on 29 degrees of freedom
## Multiple R-squared:  0.8471, Adjusted R-squared:  0.8365
## F-statistic: 80.33 on 2 and 29 DF,  p-value: 1.494e-12
```



- In wissenschaftlichen Aufsätzen werden Regressionsmodelle häufig schrittweise aufgebaut und übersichtlich in Tabellen dargestellt.

	Model 1	Model 2	Model 3
(Intercept)	1.45 (1.10)	6.45*** (1.07)	1.48 (0.96)
weight	7.75*** (0.72)		5.96*** (0.84)
hp		0.04*** (0.01)	0.02** (0.01)
R <sup>2</sup>	0.79	0.58	0.85
Adj. R <sup>2</sup>	0.78	0.57	0.84
Num. obs.	32	32	32

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; \* $p < 0.05$

Statistical models

- 1 Multiple Lineare Regression
- 2 Hypothesentests
- 3 Residualdiagnostik
- 4 Multikollinearität
- 5 Nichtlinearität

- Ergänzend zum t-Test für die einzelnen Koeffizienten ( $H_0 : \beta_j = 0$  vs.  $H_1 : \beta_j \neq 0$ ) gibt es auch die Möglichkeit **alle Koeffizienten auf einmal** einem Hypothesentest zu unterziehen.
- Das Szenario ob alle Regressoren zusammen genommen einen Effekt auf die abhängige Variable Y haben kann mit Hilfe des **F-Tests** untersucht werden.
- Die Idee dieses simultanen Testens ist zu prüfen, ob mit hoher Wahrscheinlichkeit davon auszugehen ist, dass nicht alle Parameter  $\beta_1, \beta_2, \dots, \beta_p$  gleich 0 sind, also zu prüfen:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0 \text{ für min. ein } j$$

$$\text{FM:} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

$$\text{RM:} \quad Y = \beta_0 + \epsilon$$

- Die Nullhypothese ist gleichbedeutend mit der Tatsache, dass auch ein **reduziertes Modell** (RM) ohne Regressoren den gleichen Erklärungsgehalt liefert wie das volle Modell (FM) mit allen  $p$  Regressoren.
- Dieser fehlende **Fit** kann mit Hilfe der Fehlerquadratsumme (SSE), für die beiden Modelle messbar gemacht werden.

$$SSE(\text{FM}) = \sum (y_i - \hat{y}_i)^2 \quad SSE(\text{RM}) = \sum (y_i - \hat{y}_i^*)^2$$

$$F = \frac{[SSE(RM) - SSE(FM)]/(p + 1 - k)}{SSE(FM)/(n - p - 1)}$$

- Die Differenz  $SSE(RM) - SSE(FM)$  gibt die Erhöhung der Residualstreuung durch Rückgriff auf das reduzierte Modell an. Wenn diese Differenz groß ist, ist das RM mit  $k$  Parametern **nicht adäquat**.
- Wenn der beobachtete F-Wert größer ist als der kritische Wert, ist der F-Test signifikant zum Level  $\alpha$ .
- Das bedeutet, dass das reduzierte Modell (RM) nicht zufriedenstellend ist und die Nullhypothese (und die entsprechenden Werte für die  $\beta$ 's) verworfen werden kann.
- Verwerfe  $H_0$  wenn gilt:

$$F \geq F_{(p+1-k, n-p-1; 1-\alpha)} \quad \text{oder} \quad p(F) \leq \alpha$$

```
mod <- lm(l100km ~ 1 + weight + hp, data=d)
summary(mod)

##
## Call:
## lm(formula = l100km ~ 1 + weight + hp, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.9678 -1.1667  0.1802  0.9415  3.3444
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.483062   0.962884   1.540  0.13435
## weight       5.955801   0.840115   7.089 8.45e-08 ***
## hp           0.017592   0.005438   3.235 0.00303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.562 on 29 degrees of freedom
## Multiple R-squared:  0.8471, Adjusted R-squared:  0.8365
## F-statistic: 80.33 on 2 and 29 DF,  p-value: 1.494e-12
```

- 1 Multiple Lineare Regression
- 2 Hypothesentests
- 3 Residualdiagnostik
- 4 Multikollinearität
- 5 Nichtlinearität

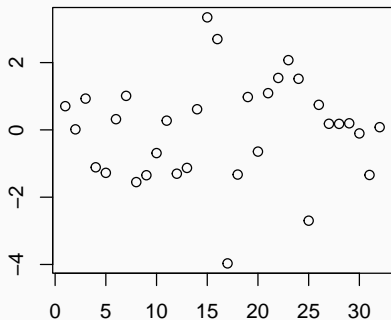
- Modellierungsprobelme, wie inkorrekt spezifizierte Modelle, fehlende und vergessene Variablen **äußern sich häufig in einer Verletzung der Annahmen der Residuen.**
- Um zu überprüfen ob das ausgewählte Regressionsmodelle den theoretischen Anfroderungen genügt, müssen daher die Residuen inspiziert werden. Dieses Prozess nennt man Resdidualdiagnostik.
- Residuen sollten (annähernd) Normalverteilt sein, keine Zusammenhangsstrutkur aufweisen (i.i.d.) und frei von Ausreißern sein. Diese Eigenschaften werden häufig in **grafischen Darstellungen** der Residuen sichtbar.
- **R-Funktion:** `residuals()` erlaubt das Extrahieren von Residuen aus der Rückgabe der `lm()` Funktion.



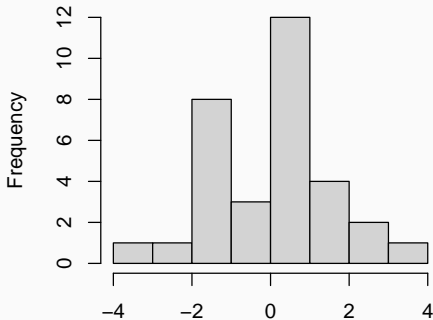
- Darstellung Residuen der Regression Kraftstoffverbrauch  $Y$  erklärt durch Fahrzeuggewicht ( $X_1$ ) und Motorleistung ( $X_2$ ).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

**Indexplot der Residuen**



**Histogramm der Residuen**



**1** Multiple Lineare Regression

**2** Hypothesentests

**3** Residualdiagnostik

**4** Multikollinearität

**5** Nichtlinearität

- Die Interpretation der Koeffizienten eines multiplen Regressionsmodells setzt voraus, dass die Prädiktoren keinen ausgeprägten Zusammenhang untereinander haben, da die (ceteris paribus) Interpretation der Koeffizienten dann nicht mehr greift.
- Wenn eine starke Abhängigkeitsstruktur zwischen den Prädiktoren vorhanden ist, dann bezeichnet man dieses Problem als **Multikollinearität**. Multikollinearität ist ein Problem in den Daten und kein Problem der Modellierung.
- Multikollinearität führt zu unplausiblen Werten der Koeffizientenschätzer und wird durch spezielle Maßzahlen, wie Varianzinflationsfaktoren (VIF), messbar.

```
cor(d)
```

```
##          l100km    weight         hp         cyl         hub
## l100km 1.00000000 0.8898927 0.7629477 0.8137493 0.8798217
## weight 0.8898927 1.0000000 0.6587479 0.7824958 0.8879799
## hp      0.7629477 0.6587479 1.0000000 0.8324475 0.7909486
## cyl     0.8137493 0.7824958 0.8324475 1.0000000 0.9020329
## hub     0.8798217 0.8879799 0.7909486 0.9020329 1.0000000
```

- Um Multikollinearitätsprobleme zu diagnostizieren, müssen die Zusammenhangsstrukturen zwischen den Prädiktoren untersucht werden. Das beinhaltet die Analyse des  $R^2$ , dass aus der Regression jedes Prädiktors auf alle verbleibenden Prädiktoren resultiert.

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad \text{with } j = 1, \dots, p$$

- $R_j^2$  bezeichnet das Bestimmtheitsmaß bei der Erklärung von  $X_j$  durch alle verbleibenden  $p - 1$  Prädiktoren. Wenn  $X_j$  gut durch die anderen Variablen erklärt werden kann, wird das  $R_j^2$  nah bei 1 sein und in einem großen Wert des  $\text{VIF}_j$  resultieren.

Ein Wert von  $\text{VIF} > 10$  wird oft als Grenzwert gesehen, ab dem man von Multikollinearität in problematischem Ausmaß ausgeht.

- Varianzinflationsfaktoren können mit der **R-Funktion** `vif()` aus dem Zusatzpaket `car` berechnet werden.

```
mod1 <- lm(l100km ~ 1 + weight + hp, data=d)
car::vif(mod1)
```

```
##    weight      hp
## 1.766625 1.766625
```

```
mod2 <- lm(l100km ~ 1 + weight + hp + hub + cyl, data=d)
car::vif(mod2)
```

```
##    weight      hp      hub      cyl
## 4.848016 3.405983 10.373286 6.737707
```

**1 Multiple Lineare Regression**

**2 Hypothesentests**

**3 Residualdiagnostik**

**4 Multikollinearität**

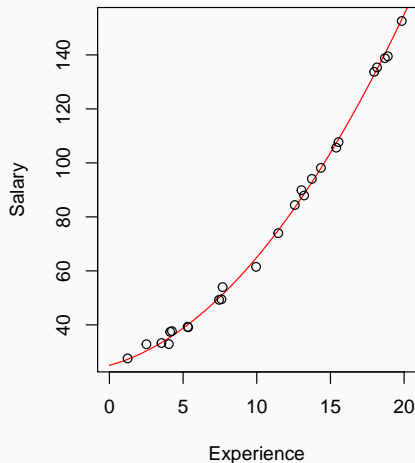
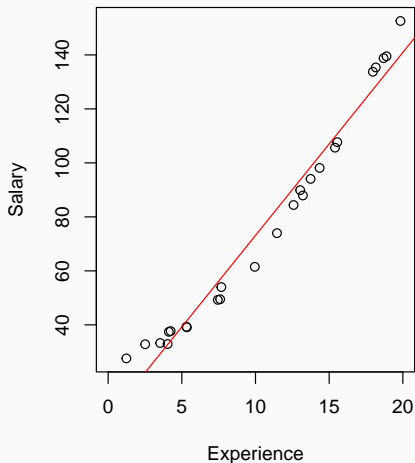
**5 Nichtlinearität**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

- Die lineare Regression ist linear im Bezug auf die Tatsache, dass die Parameter  $\beta_0, \beta_1, \dots, \beta_p$  **linear** in das Modell eingehen.
- Mit der linearen Regression können dennoch **nicht-lineare** Zusammenhänge modelliert werden indem **nicht-lineare Transformationen** als zusätzliche unabhängige Variablen in das Modell integriert werden.



Welches Modell passt besser zu den gezeigten Daten?



```
mod1
```

```
##  
## Call:  
## lm(formula = y ~ 1 + x)  
##  
## Coefficients:  
## (Intercept)          x  
##      5.334      6.779
```

```
mod2
```

```
##  
## Call:  
## lm(formula = y ~ 1 + x + x_sq)  
##  
## Coefficients:  
## (Intercept)          x      x_sq  
##      25.5809      1.4377      0.2498
```

- Wie ist das Bestimmtheitsmaß  $R^2$  bei der multiplen Regression zu interpretieren?
- Was ist damit gemeint, dass die diskutierten Regressionsmodelle lineare Modelle sind?
- Was ist Multikollinearität?