

# Teil 4

## Regression

Rebecca Karwen

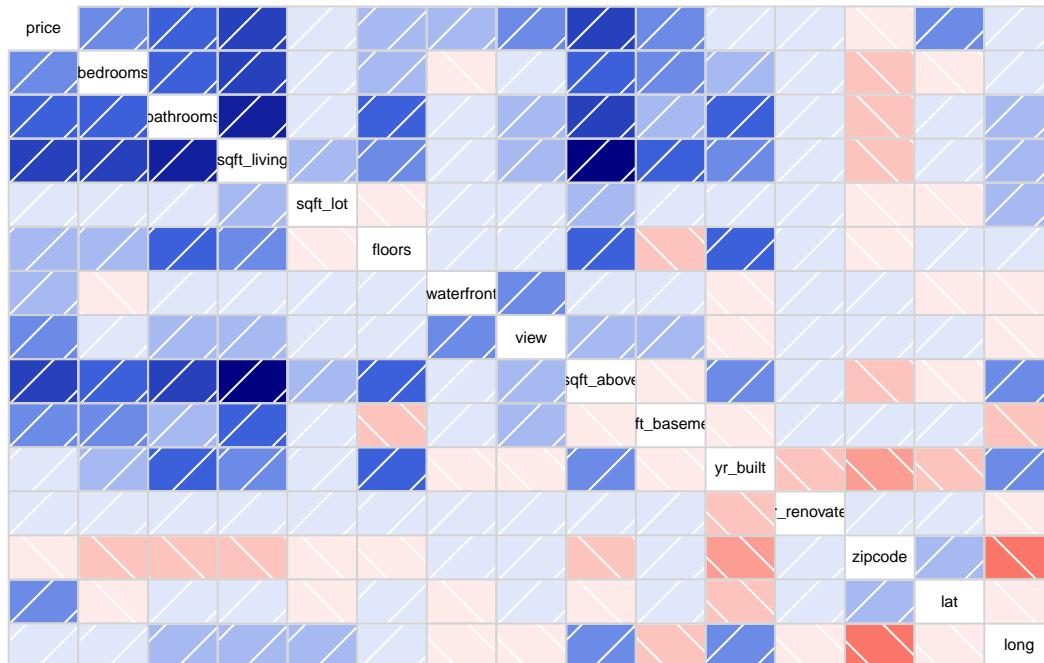
### Contents

Aufgabe 1	1
Aufgabe 2	3
Aufgabe 3	4
Aufgabe 4	5

### Aufgabe 1

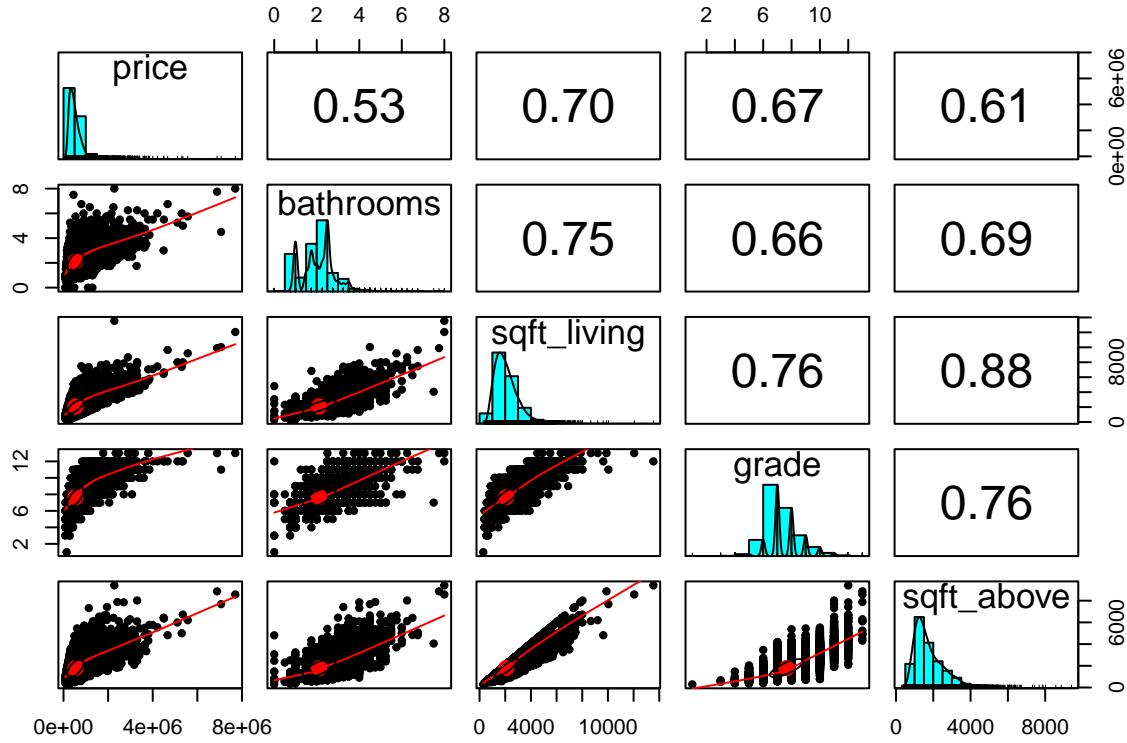
Zunächst wird die Funktion corrgram() verwendet, um einen ersten Eindruck zu bekommen wie stark die Attribute mit dem Verkaufspreis korrelieren.

```
corrgram(haus_verkauf[, c("price", "bedrooms", "bathrooms", "sqft_living", "sqft_lot", "floors",
                           "waterfront", "view", "condition", "grade", "sqft_above", "sqft_basement",
                           "yr_builtin", "yr_renovated", "zipcode", "lat", "long")])
```



Es scheint, dass der Verkaufspreis am stärksten mit der Größe der Wohnung und der Fläche eines Obergeschosses korreliert. Um dies noch einmal zu testen, wird pairs.panel() Funktion verwendet. Hinzu nehmen wir noch das Attribut grade, da es als Faktor nicht in der vorherigen Grafik erscheint. Hierdurch wird ersichtlich, dass mit der Wohnfläche die größte Korrelation besteht.

```
pairs.panels(haus_verkauf[, c("price", "bathrooms", "sqft_living", "grade", "sqft_above")])
```



Nun berechnen wir die Korrelation zwischen price und sqft\_living.

```
cor.test(haus_verkauf$price, haus_verkauf$sqft_living)
```

```
## 
## Pearson's product-moment correlation
##
## data: haus_verkauf$price and haus_verkauf$sqft_living
## t = 144.92, df = 21611, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6952099 0.7087336
## sample estimates:
##        cor
## 0.7020351
```

Der Korrelationskoeffizient beträgt 0.70 und ist damit moderat positiv. Der P-Wert ist kleiner als 0.05, dadurch ist die Korrelation statistisch signifikant. Das heißt der Preis eines Hauses korreliert mit der Größe der Wohnfläche positiv, steigt also mit zunehmender Wohnfläche an.

## Aufgabe 2

Mit den Werten, mit denen der Preis am stärksten korreliert, erstellen wir ein multiples Regressionsmodell. Diese sind die Wohnfläche und grade, also die Bewertung des Hauses.

```
relation <- lm(price ~ sqft_living + grade , data = haus_verkauf)
print(relation)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + grade, data = haus_verkauf)
##
## Coefficients:
## (Intercept)  sqft_living      grade3      grade4      grade5      grade6
##         96381.0          157.3        15425.9       14101.6      -2541.0       18097.9
##      grade7      grade8      grade9      grade10      grade11      grade12
##      40455.1     102796.0     225954.3      421623.1      709026.9      1234123.0
##      grade13
##     2436095.7
```

Für jeden Sqft Wohnfläche mehr, würde der Preis um 157 Dollar steigen. Für fast jede höhere Bewertungsstufe, würde der Preis steigen, nachdem nach der Wohnfläche kontrolliert wurde. Die einzige Bewertung, bei welcher der Preis sinken würde, ist grade 5 und zwar um 2541 Dollar. Am stärksten beeinflusst die Bewertung 13 den Preis - hier wäre der Preis um 2436095 Dollar höher.

## Aufgabe 3

Wir nutzen unser erstelltes Regressionsmodell um folgendes Haus zuschätzen.

- 1) Anzahl der Schlafzimmer: 4
- 2) Anzahl der Badezimmer: 3
- 3) Wohnfläche in ft<sup>2</sup>: 1900
- 4) Grundstücksfläche in ft<sup>2</sup>: 6000
- 5) Anzahl der Stockwerke: 2
- 6) Lage an Wasser: 1
- 7) Anzahl der Besichtigungen: 3
- 8) condition: 4
- 9) Bewertung des Hauses: 9
- 10) Obergeschossfläche in ft<sup>2</sup>: 1900
- 11) Kellerfläche in ft<sup>2</sup>: 0
- 12) Baujahr: 2009
- 13) Renovierungsjahr: /
- 14) Postleitzahl: 98038
- 15) Breitengrad: 47.380
- 16) Längengrad: -122.020

Zunächst erstellen wir einen Data Frame mit den Werten des Hauses, um dann die predict() Funktion zu verwenden, um durch unser Regressionmodell den Preis zu schätzen.

```
house <- data.frame(bedrooms = c(4), bathrooms = c(3), sqft_living = c(1900), sqft_lot = c(6000), floors = c(2), waterfront = c(1), condition = c(4), grade = c(9), yr_renovated = c(0), yr_built = c(2009), zipcode = c(98038), lat = c(47.380), long = c(-122.020))

predict(relation, newdata = house)

##           1
## 621218.2
```

Der geschätzte Preis beträgt 621218.20 Dollar.

## Aufgabe 4

Da eine starke Korrelation nicht immer etwas darüber aussagt ob sich diese Variablen wirklich stark beeinflussen, versuchen wir nun das Regressionsmodell zu verbessern.

Zunächst verwenden wir die `summary()` Funktion, um mehr nützliche Informationen über unser Regressionsmodell zu erhalten und danach berechnen wir die durchschnittlichen residuals.

```
summary(relation)

##
## Call:
## lm(formula = price ~ sqft_living + grade, data = haus_verkauf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1550528 -129352 - 27914  92324 4677732 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 9.638e+04 2.375e+05 0.406 0.6849    
## sqft_living 1.573e+02 2.781e+00 56.567 < 2e-16 *** 
## grade3     1.543e+04 2.743e+05 0.056 0.9551    
## grade4     1.410e+04 2.416e+05 0.058 0.9535    
## grade5    -2.541e+03 2.380e+05 -0.011 0.9915    
## grade6     1.810e+04 2.376e+05 0.076 0.9393    
## grade7     4.046e+04 2.376e+05 0.170 0.8648    
## grade8     1.028e+05 2.376e+05 0.433 0.6653    
## grade9     2.260e+05 2.377e+05 0.951 0.3418    
## grade10    4.216e+05 2.378e+05 1.773 0.0762 .  
## grade11    7.090e+05 2.381e+05 2.978 0.0029 ** 
## grade12    1.234e+06 2.393e+05 5.158 2.52e-07 *** 
## grade13    2.436e+06 2.473e+05 9.851 < 2e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 237500 on 21600 degrees of freedom
## Multiple R-squared:  0.5817, Adjusted R-squared:  0.5814 
## F-statistic:  2503 on 12 and 21600 DF,  p-value: < 2.2e-16

mean(residuals(relation))

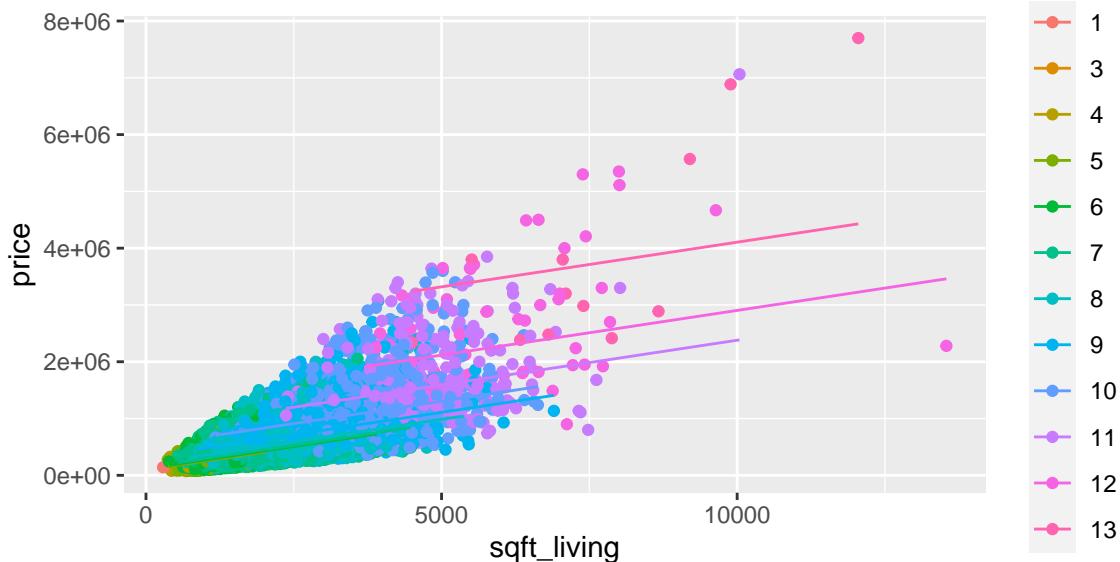
## [1] 2.977578e-11
```

Der P-Wert ist kleiner als 0.005 und damit statistisch signifikant. Der Wert von  $R^2$  mit 59 % ist im Mittelfeld. Der Wert bedeutet, dass 58,2 % des Verkaufspreises durch die von uns gewählten Werte bestimmt wird. Der Residual Standard Error ist mir 234800 Dollar sehr hoch.

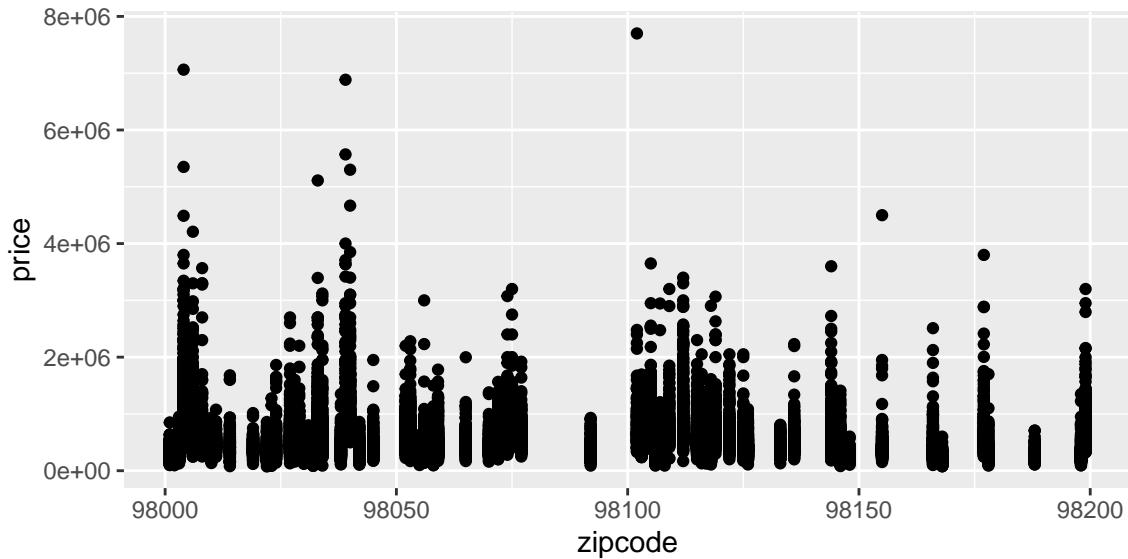
Ein Visualisierung der Werte kann auch helfen einen besseren Überblick zu erhalten. Dafür wird ein Diagramm für die Wohnfläche mit Preis und gerade erstellt und noch ein weiteres für die Verteilung des Zipcodes.

```
augment_haus <- augment(relation)
```

```
ggplot(augment_haus, aes(x=sqft_living, y=price, color = grade)) +
  geom_point() +
  geom_line(aes(y = .fitted))
```



```
ggplot(haus_verkauf, aes(x=zipcode, y=price)) +
  geom_point()
```



Es scheint das bis auf einige Ausläufer in unseren Daten die Postleitzahl, nicht zu sehr mit dem Preis zusammen hängt und es es bei den meisten eine Verteilung von Häusern in verschiedenen Preisklassen gibt.

Eine Sache, bei der oft angenommen wird ist, dass sie den Preis beeinflusst, ist die Lage am Wasser. Dafür schauen wir uns den durchschnittlichen Verkaufspreises eines Hauses am Wasser und nicht am Wasser an - und stellen dabei fest, dass der durchschnittliche Wert mit Lage am Wasser höher ist. Da dies anscheinend eine Auswirkung hat, werden wir die Variable mit in unser Regressionmodell aufnehmen.

Das gleiche machen wir auch mit dem Zipcode. Hier werden dabei aber nicht genug Anhaltspunkte gefunden, um die Variable mit in unser Modell zu nehmen.

```
wasser <- haus_verkauf %>%
  group_by(waterfront) %>%
  summarize(mean_price = mean(price))
wasser

## # A tibble: 2 x 2
##   waterfront mean_price
##       <dbl>      <dbl>
## 1          0     531564.
## 2          1    1661876.

zipcode <- haus_verkauf %>%
  group_by(zipcode) %>%
  summarize(mean_price = mean(price)) %>%
  arrange(desc(mean_price))
zipcode

## # A tibble: 70 x 2
##   zipcode mean_price
##       <dbl>      <dbl>
## 1  98039    2160607.
## 2  98004    1355927.
## 3  98040    1194230.
## 4  98112    1095499.
## 5  98102     901258.
## 6  98109     879624.
## 7  98105     862825.
## 8  98006     859685.
## 9  98119     849448.
## 10 98005     810165.
## # ... with 60 more rows
```

Nun versuchen wir mehr Variablen mit einzubinden, damit wir das Modell verbessern, indem wir Lage am Wasser und Interaktion zwischen Wohnfläche und Lage am Wasser hinzufügen.

```
relation2 <- lm(price ~ sqft_living + grade + waterfront + sqft_living:waterfront , data=haus_verkauf)

summary(relation2)

## 
## Call:
## lm(formula = price ~ sqft_living + grade + waterfront + sqft_living:waterfront,
##      data = haus_verkauf)
## 
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -1557403 -126561  -26800   91921 3334845 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.004e+05 2.236e+05  0.449  0.65354  
## sqft_living  1.435e+02 2.636e+00 54.461 < 2e-16 ***
## grade3      1.965e+04 2.582e+05  0.076  0.93935  
## grade4      1.920e+04 2.274e+05  0.084  0.93272  
## grade5      6.700e+03 2.241e+05  0.030  0.97615  
## grade6      2.973e+04 2.237e+05  0.133  0.89427  
## grade7      5.856e+04 2.237e+05  0.262  0.79348  
## grade8      1.246e+05 2.237e+05  0.557  0.57766  
## grade9      2.549e+05 2.238e+05  1.139  0.25471  
## grade10     4.427e+05 2.239e+05  1.977  0.04801 *  
## grade11     6.952e+05 2.242e+05  3.102  0.00193 ** 
## grade12     1.097e+06 2.253e+05  4.872  1.12e-06 ***
## grade13     2.535e+06 2.328e+05 10.888 < 2e-16 ***
## waterfront   -2.495e+05 3.931e+04 -6.346  2.25e-10 ***
## sqft_living:waterfront 3.276e+02 1.120e+01 29.246 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 223600 on 21598 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.629 
## F-statistic:  2618 on 14 and 21598 DF,  p-value: < 2.2e-16

predict(relation2, newdata = house)

##      1
## 1000962
```

Das verbessert das Modell um einige Prozentpunkte für Multiple R-squared und verringert den Residual standard error etwas.

Der erwartete Wert des Hauses wäre dann 1 Millionen Dollar.