Tutorial: Regression Output von R

Eine Firma erzeugt Autositze. Ihr Chef ist besorgt über die Anzahl und die Kosten von Maschinenausfällen. Das Problem ist, dass die Maschinen schon alt sind und deswegen unzuverlässig arbeiten. Die Kosten, diese Maschinen zu ersetzen, sind sehr hoch und folglich ist der Chef unsicher, ob er sie in Anbetracht der schlechten Wirtschaftslage wirklich austauschen sollte.

Um sich die Entscheidung zu erleichtern, sammelt er Daten über die monatlichen Reparaturkosten (Variable Repairs*) und das Alter (in Monaten, Variable* Age*) der 20 in der Fabrik vorhandenen Maschinen. Eine Regressionsgerade gibt den (linearen) Zusammenhang an.*

```
Call:
lm(formula = Repairs ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-74.547 -22.660  -0.981  22.233  83.354

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  114.852     58.685    1.96  0.06603 .
Age            2.473      0.511    4.84  0.00013 ***
---
Signif. codes:  0 ,,***, 0.001 ,,**, 0.01 ,,*, 0.05 ,,., 0.1 ,, , 1
```

```
Residual standard error: 43.3 on 18 degrees of freedom
Multiple R-Squared: 0.566,  Adjusted R-squared: 0.542
F-statistic: 23.5 on 1 and 18 DF,  p-value: 0.000130
```

Welche der folgenden Aussagen sind richtig?

1. Das Bestimmtheitsmaß beträgt $R^2 = 0.566$. Die Wurzel davon wäre der Korrelationskoeffizient.

2. Der p-Wert ist kleiner als jedes vernünftige Signifikanzniveau. Der Koeffizient von *Age* ist daher signifikant von 0 verschieden.

3. Für eine 100 Monate alte Maschine sind Reparaturkosten von $100 * 2.473 = 247.3$ zu erwarten.

1. Das Bestimmtheitsmaß beträgt $R^2 = 0.566$. Die Wurzel davon wäre der Korrelationskoeffizient.

   Richtig. Bei der einfachen Regressionsrechnung gilt diese Beziehung (Achtung: R^2 ist immer positiv, durch ein Quadrieren verschwindet ein etwaiges Minus beim Korrelationskoeffizienten. In diesem Beispiel ist der Korrelationskoeffizient allerdings positiv).

2. Der p-Wert ist kleiner als jedes vernünftige Signifikanzniveau. Der Koeffizient von *Age* ist daher signifikant von 0 verschieden.

   Richtig. Der p-Wert ist sehr klein; das Modell daher signifikant und damit der Koeffizient von *Age* signifikant von 0 verschieden.

3. Für eine 100 Monate alte Maschine sind Reparaturkosten von $100 * 2.473 = 247.3$ zu erwarten.

Falsch. Hier wurde beim Einsetzen in die Regressionsgleichung die Konstante (Intercept) vergessen.

Der Output von Statistikprogrammen nach einer Berechnung eines Regressionsmodells ist auf den ersten Blick unübersichtlich, kann aber in mehrere Teile zergliedert und dadurch für eine Interpretation einfacher zugänglich gemacht werden.

Hier soll der Output von R, wie er als Zusammenfassung einer Regression ausgegeben wird, erläutert werden.

Erster Teil:

```
Call: lm(formula = Repairs ~ Age)
```

Hier steht, wie die Regression aufgerufen wurde (lm(formula = Repairs Age)). In diesem Fall bedeutet es, dass die Responsevariable Repairs in Abhängigkeit zur erklärenden Variable Age gestellt wurde.

Zweiter Teil:

```
Residuals:
    Min      1Q  Median      3Q     Max
-74.547 -22.660  -0.981  22.233  83.354
```

Hier sind die Extrema (Minimum, Maximum) und die Quartile der Residuen angegeben. Das Residuum einer Beobachtung ist die Differenz zwischen dem Wert der Responsevariablen und dem vorhergesagten Wert für die Beobachtung ($e_i = y_i - \widehat{y_i}$).

Die Residuen sollten bei einem guten Modell normalverteilt mit Mittelwert 0 sein. Das läßt sich schwer anhand von fünf Zahlenwerten überprüfen (besser mit einem Histogramm der Residuen). Ein Teilaspekt davon, nämlich die Symmetrie ist leicht zu überprüfen: der

Median sollte etwa 0 sein, das erste (oder untere) und das dritte (oder obere) Quartil sollten symmetrisch um 0 liegen und Ausreißer nach oben und/oder unten sollten auch nicht auftreten.

Für dieses Beispiel bedeutet das: der Interquartilsabstand beträgt

$$Q_3 - Q_1 = 22.233 - (-22.660) = 44.893$$

Der Median ist (gemessen am Interquartilsabstand) nahe bei 0. Erstes und drittes Quartil liegen in etwa symmetrisch um 0 und die Extrema der Residuen sind weniger als das Zweifache des Interquartilsabstandes von 0 entfernt. Es gibt also keinen Grund daran zu zweifeln, dass die Residuen symmetrisch um 0 verteilt sind.

Dritter Teil:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  114.852     58.685    1.96  0.06603 .
        Age    2.473      0.511    4.84  0.00013 ***
---
Signif. codes:  0 ,,***, 0.001 ,,**, 0.01 ,,*, 0.05 ,,., 0.1 ,, , 1
```

Dieser Output-Teil enthält die Werte für de Regressionsgleichung und Angaben, mit denen überprüft werden kann, ob die erklärende Variable signifikant etwas zur Erklärung der Responsevariable beiträgt. In der Spalte, die mit Estimate überschrieben ist, stehen

die Werte für die Regressionsgleichung (Intercept steht für die Konstante), für dieses Beipiel würde also gelten:

$$Repairs = 114.852 + 2.473 \cdot Age$$

Über diese Gleichung können also die Reparaturkosten für eine drei Jahre (=36 Monate) alte Maschine geschätzt werden ($114.852 + 2.473 \cdot 36 = 203.88$).

Die Spalte, die mit Pr($>$|t|) überschrieben ist, enthält den p-Wert für den Test (t-Test), dass der Anstieg der Regressionsgerade signifikant von 0 abweicht, das ist gleich bedeutend damit, dass die erklärende Variable signifikant etwas zur Erklärung der Responsevariable beiträgt. Hier ist der p-Wert von Age mit 0.00013 eindeutig kleiner als übliche Signifikanzniveaus, daher ist der Anstieg signifikant von 0 verschieden. Dies kommt auch durch die folgenden 3 Sterne zum Ausdruck, sie zeigen an, dass bei einem Signifikanzniveau von einem Promille das Ergebnis signifikant wäre.

Vierter Teil:

```
Residual standard error: 43.3 on 18 degrees of freedom

Multiple R-Squared: 0.566,  Adjusted R-squared: 0.542

F-statistic: 23.5 on 1 and 18 DF,  p-value: 0.000130
```

Dieser Teil ist vor allem bei Regressionsmodellen mit mehr als einer erklärenden Variablen von Bedeutung. Die erste Zeile gibt Auskunft über die Varianz der Residuen.

In der zweiten Zeile ist das Bestimmtheitsmaß R^2 als Multiple R-Squared abzulesen, das im Fall der Regression mit nur einer erklärenden Variablen dem quadrierten Korrelationskoeffizienten entspricht (hier $R^2 = 0.566$).

Die dritte Zeile zeigt das Ergebnis des sog. F-Tests, der im Fall von nur einer erklärenden Variablen auch durch den t-Test des dritten Teils erfolgen kann (die p-Werte beider Tests sind mit p=0.00013 ident).