Erklärung eines R Outputs

Im Folgenden wird ein (multiples lineares) Log-Level-Modell

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathbf{u}|\mathbf{X} \sim IID(0, \sigma^2\mathbf{I}) \quad (\text{ hier: } \mathbf{y} = \log(\mathbf{price}),\ \mathbf{X} = (\mathbf{1}, \mathbf{area}, \mathbf{age}, \mathbf{rooms}))$$

mit Hilfe von `lm()` geschätzt und mittels `summary()` der zu erklärende Output erzeugt:

```
Call:
lm(formula = log(price) ~ 1 + area + age + rooms)

Residuals:
  Min       1Q     Median    3Q     Max
-1.14615 -0.10158 0.01565 0.12887 0.75234

Coefficients:
              Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)   9.890e+00   1.176e-01   84.069    < 2e-16 ***
area          3.100e-04   3.113e-05 ´ 9.958     < 2e-16 ***
age           -3.936e-03  4.519e-04   -8.709    2.25e-15 ***
rooms         1.143e-01   2.049e-02   5.582     8.92e-08 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2259 on 175 degrees of freedom
Multiple R-squared: 0.6275, Adjusted R-squared: 0.6211
F-statistic: 98.26 on 3 and 175 DF, p-value: < 2.2e-16
```

Der R Output ist unterteilt in vier Abschnitte:

`Call` Beziehung von Regressand und Regressoren werden wiederholt; in unserem Fall werden die logarithmierten Preise `log(price)` gegen eine Konstante, die Fläche `area`, das Alter `age` und die Raumanzahl `rooms` regressiert. `1` steht dabei für ein Modell mit Konstante, `-1` wäre eine Option, wenn man dieses Modell ohne Konstante schätzen möchte. Möchte man in (lm()) Regressoren mit arithmetischen Operationen versehen, muss dies innerhalb der Funktion `I()` passieren.

`Residuals` Dieser Abschnitt dient dazu, einen groben Überblick (Symmetrie, Ausdehnung, Quantile) über die Verteilung der Residuen zu erlangen. Es werden Minimum, Maximum, 25%-und 75%-Quantil und der Median angegeben.

`Coefficients` Mit Hilfe der KQ-Methode werden nun die Koeffizienten geschätzt (`Estimate`), deren empirische Standardabweichung (`Std. Error`) wird angegeben, die Teststatistik (`t-value`) zum Test mit $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ (Interpretation: x_i hat keinen Einfluss vs. x_i hat Einfluss) berechnet und der zur Teststatistik gehörende $p-$Wert (`Pr(>|t|)`) notiert (Interpretation siehe unten). Die Sterne (z. B. $\star\star\star$) deuten dabei auf das Signifikanzniveau (mit Legende `Signif. codes`) hin.
Die Zahlen der `Estimate`-Spalte lassen sich für das Log-Level-Modell folgendermaßen interpretieren: Würde z. B. die Anzahl der Räume einer Wohnung um 1 zunehmen, so würde der Preis ceteris paribus im Durchschnitt um approximativ $1.143e-01 \cdot 100\% = 1,143 \cdot 10^{-1} \cdot 100\% = 11,43\%$ zunehmen. Der exakte Anstieg ceteris paribus im Durchschnitt (in diesem Log-Level-Modell) lässt sich über die Formel $(e^{\hat{\beta}_4} - 1) \cdot 100\% = (e^{0,1143} - 1) \cdot 100\% = 12,11\%$ (bei Annahme normalverteilter Fehler) ermitteln.

`RSE, R`^{2}`, F` Die erste Zeile beschreibt die Wurzel der Residuenquadratsumme geteilt durch $n - k$, wobei $n - k$ die Anzahl der Freiheitsgrade (`df`), n die Anzahl der Beobachtungen und k die Anzahl der Regressoren (Konstante wird gezählt; hier also 4) angibt. Man erhält also einen Stichprobenumfang der Größe 179. Die zweite Zeile liefert das unzentrierte und das zentrierte R^2, die einen prozentualen Wert für den Erklärungsgehalt des Modells liefern. Die dritte Zeile beschreibt die Statistik eines $Overall - F-$Tests, der zur Teststatistik $H_0 : \beta_i = 0 \quad \forall i = 2, 3, 4$ vs. $H_1 : \exists i \in \{2, 3, 4\} : \quad \beta_i \neq 0$ (die Konstante wird nicht mitgestestet!) mit 3 Zählergraden (nicht konstante Regressoren) und 175 Nennergraden (=*df*) gehört. (Grundsätzliche Interpretation von p-Werten: die Nullhypothese wird verworfen, wenn der p-Wert kleiner dem vorgegebenen Signifikanzniveau ist.)