Serie 12

Aufgabe 12.1

We investigate further the data set `Boston` from the last exercise sheet.

In order to fit a multiple linear regression model using least squares, we again use the **`ols()`** function. The syntax **`ols("y ~ x1 + x2 + x3", data=...)`** is used to fit a model with three predictors, **`x1`** , **`x2`** , and **`x3`** . The **`summary()`** function now outputs the regression coefficients for all the predictors.

a) Fit a multiple linear regression model with response variable **`medv`** and predictors **`lstat`** and **`age`**.

   Define the model and interpret all values in the **`summary()`** output which we discussed in class (coefficients, its P values, R^2 value, P value of the F-statistics).

b) The `Boston` data set contains 13 variables, and so it would be cumbersome to have to type all of these in order to perform a regression using all of the predictors. Instead, we can use the following short-hand

```
all_columns = "+".join(df.columns.drop("medv"))
all_columns
fit = ols("medv~" + all_columns , data=df).fit()
```

   In the **`summary()`** output interpret the coefficient of **`age`** and the corresponding P value compare this with the output in a) and explain the difference.

c) The R^2 value is bigger than the one calculated in a). Explain.

d) It is easy to include interaction terms in a linear model using the **`lm()`** function. The syntax **`lstat:black`** tells **`R`** to include an interaction term between **`lstat`** and **`black`**.

   The syntax **`lstat * age`** simultaneously includes **`lstat`**, **`age`**, and the interaction term **`lstat × age`** as predictors; it is a shorthand for **`lstat + age + lstat:age`**.

   Again, discuss all the values in the **`summary()`** of **`lstat*age`** as in a).

Aufgabe 12.2

Wir führen noch eine multiple lineare Regression für `Auto` aus der letzten Übung durch.

Bevor wir damit beginnen, entfernen wir alle Variablen, **`"Unnamed: 0", "X1", "name"`**, die qualtitativ oder nicht relevant sind. Dies machen wir mit der **`.drop()`**-Methode.

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.graphics.regressionplots import abline_plot
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
import numpy as np

df = pd.read_csv("../../../Themen/Einfache_Lineare_Regression/Jupyter_Notebooks_de/Auto.csv")

df.columns


## Index(['Unnamed: 0', 'X1', 'mpg', 'cylinders', 'displacement', 'horsepower',
##        'weight', 'acceleration', 'year', 'origin', 'name'],
##       dtype='object')


df = df.drop(["Unnamed: 0", "X1", "name"], axis=1)
df.head()


##      mpg  cylinders  displacement  ...  acceleration  year  origin
## 0  7.650          8         307.0  ...          12.0    70       1
## 1  6.375          8         350.0  ...          11.5    70       1
## 2  7.650          8         318.0  ...          11.0    70       1
## 3  6.800          8         304.0  ...          12.0    70       1
## 4  7.225          8         302.0  ...          10.5    70       1
##
## [5 rows x 8 columns]
```

a) Produzieren Sie mit **`.pairplot()`** Streudiagramme, die alle Variablen des Datensatzes enthält.

```
import seaborn as sns

sns.pairplot(df)
```

Welche Abhängigkeiten stellen Sie fest?

b) Berechnen Sie die Korrelationsmatrix zwischen den Variablen mit **`df.corr()`**.

Interpretieren Sie die Werte für **`horsepower`** und **`displacement`** mit den Streudiagrammen oben.

c) Wir verwenden `lm()` um eine multiple Regression mit der Zielgrösse `mpg` und allen anderen Variablen (ausser `name`) als Prädiktoren durchzuführen. Verwenden Sie wieder Output des `summary()`-Befehls zu interpretieren.

   i) Gibt es einen Zusammenhang zwischen den Prädiktoren und der Zielvariable? Begründen Sie dies mit dem p-Wert zum F-Wert.

   ii) Welche Prädiktoren scheinen statistisch signifikant einen Einfluss auf die Zielvariable zu haben?

   iii) Was deutet der Koeffizient für `year` an?

d) Untersuchen das Modell aus c) noch auf Interaktionseffekte.

Kurzlösungen vereinzelter Aufgaben

Musterlösungen zu Serie 12

Lösung 12.1

a) Model:

$$\texttt{medv} = \beta_0 + \beta_1 \cdot \texttt{lstat} + \beta_2 \cdot \texttt{age}$$

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.graphics.regressionplots import abline_plot
from scipy.stats import uniform, norm
import matplotlib.pyplot as plt
import numpy as np

df = pd.read_csv("../../../Themen/Einfache_Lineare_Regression/Jupyter_Notebooks_de/Boston.csv").drop("U
0", axis=1)

fit = ols("medv~lstat+age", data=df).fit()

fit.summary()

## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   medv   R-squared:                       0.551
## Model:                            OLS   Adj. R-squared:                  0.549
## Method:                 Least Squares   F-statistic:                     309.0
## Date:                Mon, 11 May 2020   Prob (F-statistic):           2.98e-88
## Time:                        10:18:15   Log-Likelihood:                -1637.5
## No. Observations:                 506   AIC:                             3281.
## Df Residuals:                     503   BIC:                             3294.
## Df Model:                           2
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept     33.2228      0.731     45.458      0.000      31.787      34.659
## lstat         -1.0321      0.048    -21.416      0.000      -1.127      -0.937
## age            0.0345      0.012      2.826      0.005       0.011       0.059
## ==============================================================================
## Omnibus:                      124.288   Durbin-Watson:                   0.945
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              244.026
## Skew:                           1.362   Prob(JB):                     1.02e-53
## Kurtosis:                       5.038   Cond. No.                         201.
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## """
```

The estimates are

$$\widehat{\beta}_0 = 33.22; \qquad \widehat{\beta}_1 = -1.03; \qquad \widehat{\beta}_2 = 0 - 03$$

We get for the model

$$\texttt{medv} = 33.22 - 1.03 \cdot \texttt{lstat} + 0.03 \cdot \texttt{age}$$

Interpretation of the estimates:

- $\widehat{\beta}_0 = 33.22$

  In neighborhoods where there is no population of lower status and no units build before 1940, the medium value of houses is \$ 33 220.

- $\widehat{\beta}_1 = -1.03$

  For each additional percent of population of lower status, the medium value decreases by \$ 1030.

- $\widehat{\beta}_2 = 0.03$

  For each additional percent of units build before 1949, the medium value increases by \$ 30.

- All p-values are significant (below the significance level of 5 %), so all estimates individually contribute significantly to the model.

- The R^2 value is 0.5513, therefore about 55 % of the variation is explained by the model.

- The p-value of the F value is below the significance level and therefore significant. The null hypothesis H_0

  $$\beta_1 = \beta_2 = 0$$

  is rejected. One of β's is signigficantly different from 0. At least one variables contributes significantly to the model.

```
all_columns = "+".join(df.columns.drop("medv"))
all_columns
```

b)

```
## 'crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+black+lstat'
```

```
fit = ols("medv~" + all_columns , data=df).fit()

fit.summary()
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   medv   R-squared:                       0.741
## Model:                            OLS   Adj. R-squared:                  0.734
## Method:                 Least Squares   F-statistic:                     108.1
## Date:                Mon, 11 May 2020   Prob (F-statistic):          6.72e-135
## Time:                        10:18:15   Log-Likelihood:                -1498.8
## No. Observations:                 506   AIC:                             3026.
## Df Residuals:                     492   BIC:                             3085.
```

```
## Df Model:                          13
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept     36.4595      5.103      7.144      0.000      26.432      46.487
## crim          -0.1080      0.033     -3.287      0.001      -0.173      -0.043
## zn             0.0464      0.014      3.382      0.001       0.019       0.073
## indus          0.0206      0.061      0.334      0.738      -0.100       0.141
## chas           2.6867      0.862      3.118      0.002       0.994       4.380
## nox          -17.7666      3.820     -4.651      0.000     -25.272     -10.262
## rm             3.8099      0.418      9.116      0.000       2.989       4.631
## age            0.0007      0.013      0.052      0.958      -0.025       0.027
## dis           -1.4756      0.199     -7.398      0.000      -1.867      -1.084
## rad            0.3060      0.066      4.613      0.000       0.176       0.436
## tax           -0.0123      0.004     -3.280      0.001      -0.020      -0.005
## ptratio       -0.9527      0.131     -7.283      0.000      -1.210      -0.696
## black          0.0093      0.003      3.467      0.001       0.004       0.015
## lstat         -0.5248      0.051    -10.347      0.000      -0.624      -0.425
## ==============================================================================
## Omnibus:                      178.041   Durbin-Watson:                   1.078
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              783.126
## Skew:                           1.521   Prob(JB):                    8.84e-171
## Kurtosis:                       8.281   Cond. No.                     1.51e+04
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## [2] The condition number is large, 1.51e+04. This might indicate that there are
## strong multicollinearity or other numerical problems.
## """
```

The p-value is almost 1, so not significant at all. But in a), the p-value is 0.005, which is significant. That means that the variable **age** must correlate strongly with other variables (see d)).

c) The more variables you have the bigger the R^2 value. That means that the R^2 is not a good indicator to compare different models.

d) Model:

$$\texttt{medv} = \beta_0 + \beta_1 \cdot \texttt{lstat} + \beta_2 \cdot \texttt{age} + \beta_{12} \cdot \texttt{lstat*age}$$

Remark: $*$ in **lstat*age** does *not* signify multiplication, it just means interaction.

```
fit = ols("medv~lstat*age", data=df).fit()

fit.summary()

## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   medv   R-squared:                       0.556
## Model:                            OLS   Adj. R-squared:                  0.553
## Method:                 Least Squares   F-statistic:                     209.3
## Date:                Mon, 11 May 2020   Prob (F-statistic):           4.86e-88
## Time:                        10:18:15   Log-Likelihood:                -1635.0
## No. Observations:                 506   AIC:                             3278.
## Df Residuals:                     502   BIC:                             3295.
```

```
## Df Model:                           3
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept     36.0885      1.470     24.553      0.000      33.201      38.976
## lstat        -1.3921      0.167     -8.313      0.000      -1.721      -1.063
## age          -0.0007      0.020     -0.036      0.971      -0.040       0.038
## lstat:age     0.0042      0.002      2.244      0.025       0.001       0.008
## ==============================================================================
## Omnibus:                     135.601   Durbin-Watson:                   0.965
## Prob(Omnibus):                 0.000   Jarque-Bera (JB):              296.955
## Skew:                          1.417   Prob(JB):                     3.29e-65
## Kurtosis:                      5.461   Cond. No.                     6.88e+03
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## [2] The condition number is large, 6.88e+03. This might indicate that there are
## strong multicollinearity or other numerical problems.
## """
```

The estimates are

$$\widehat{\beta}_0 = 36.10; \qquad \widehat{\beta}_1 = -1.39; \qquad \widehat{\beta}_2 = -0.0007; \qquad \widehat{\beta}_{12} = 0.004$$

We get for the model

$$\texttt{medv} = 36.10 - 1.39 \cdot \texttt{lstat} - 0.00072 \cdot \texttt{age} + 0.0041 \cdot \texttt{lstat*age}$$

Interpretation of the estimates:

- $\widehat{\beta}_0 = 36.10$

  In neighborhoods where there is no population of lower status and no units build before 1940, the medium value of houses is \$ 36 100.

- $\widehat{\beta}_1 = -1.39$

  For each additional percent of population of lower status, the medium value decreases by \$ 1930.

- $\widehat{\beta}_2 = -0.00072$

  For each additional percent of units build before 1949, the medium value decreases by \$ 0.27.

  As you can imagine, this value is not significant, as you can see from the output.

- $\widehat{\beta}_{12} = 0.004$

  This coefficient is somewhat difficult to interpret and we did't do it in class.

- Not all p-values are significant (below the significance level of 5 %) anymore.

  The p value for **age** ist 0.97, so this not significance anymore, whereas without interaction it was. What is the reason for this?

  The p-value of the interaction term is 0.0252 which is below the significance level of 5 %. The null hypothesis H_0, that there is no interaction, is rejected. There is statistically significant interaction.

  Now, let's take a look at the correlation coefficient of the two explanatory variables **lstat** and **age**.

```
df.corr()
```

```
##                crim        zn     indus  ...     black     lstat      medv
## crim     1.000000 -0.200469  0.406583  ... -0.385064  0.455621 -0.388305
## zn      -0.200469  1.000000 -0.533828  ...  0.175520 -0.412995  0.360445
## indus    0.406583 -0.533828  1.000000  ... -0.356977  0.603800 -0.483725
## chas    -0.055892 -0.042697  0.062938  ...  0.048788 -0.053929  0.175260
## nox      0.420972 -0.516604  0.763651  ... -0.380051  0.590879 -0.427321
## rm      -0.219247  0.311991 -0.391676  ...  0.128069 -0.613808  0.695360
## age      0.352734 -0.569537  0.644779  ... -0.273534  0.602339 -0.376955
## dis     -0.379670  0.664408 -0.708027  ...  0.291512 -0.496996  0.249929
## rad      0.625505 -0.311948  0.595129  ... -0.444413  0.488676 -0.381626
## tax      0.582764 -0.314563  0.720760  ... -0.441808  0.543993 -0.468536
## ptratio  0.289946 -0.391679  0.383248  ... -0.177383  0.374044 -0.507787
## black   -0.385064  0.175520 -0.356977  ...  1.000000 -0.366087  0.333461
## lstat    0.455621 -0.412995  0.603800  ... -0.366087  1.000000 -0.737663
## medv    -0.388305  0.360445 -0.483725  ...  0.333461 -0.737663  1.000000
##
## [14 rows x 14 columns]
```

  This value is quite high. An explanation *could* be that in the poorer neighborhoods, people didn't have the money to build new houses, so there are more houses built before 1940.

- The R^2 value is 0.56, therefore about 56 % of the variation is explained by the model.

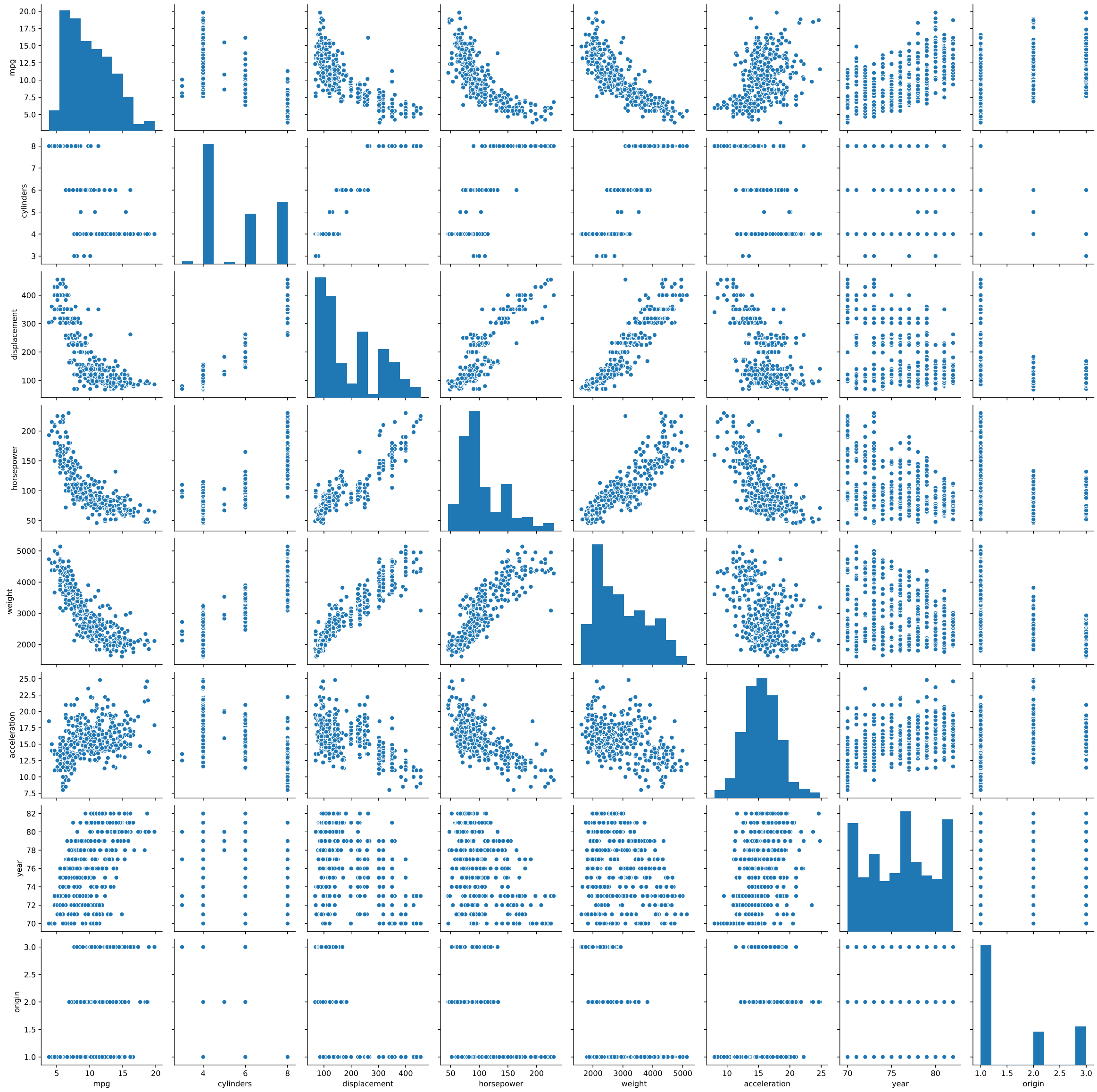
- The p-value of the F value is below the significance level and therefore significant. The null hypothesis H_0

$$\beta_1 = \beta_2 = \beta_{12} = 0$$

  is rejected. One of β's is signigficantly different from 0. At least one variables contributes significantly to the model.

Lösung 12.2

a) Streudiagramme:



b) Korrelationsmatrix

```
df.corr()

##                    mpg  cylinders  ...      year    origin
## mpg           1.000000  -0.777618  ...  0.580541  0.565209
## cylinders    -0.777618   1.000000  ... -0.345647 -0.568932
## displacement -0.805127   0.950823  ... -0.369855 -0.614535
## horsepower   -0.778427   0.842983  ... -0.416361 -0.455171
## weight       -0.832244   0.897527  ... -0.309120 -0.585005
## acceleration  0.423329  -0.504683  ...  0.290316  0.212746
## year          0.580541  -0.345647  ...  1.000000  0.181528
```

```
## origin         0.565209  -0.568932  ...  0.181528  1.000000
##
## [8 rows x 8 columns]
```

```
df.corr().loc["horsepower", "displacement"]
```

```
## 0.897257001843467
```

c) Output

```
all_columns = "+".join(df.columns.drop("mpg"))
all_columns
```

```
## 'cylinders+displacement+horsepower+weight+acceleration+year+origin'
```

```
fit = ols("mpg~" + all_columns , data=df).fit()
```

```
fit.summary()
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                    mpg   R-squared:                       0.821
## Model:                            OLS   Adj. R-squared:                  0.818
## Method:                 Least Squares   F-statistic:                     252.4
## Date:                Mon, 11 May 2020   Prob (F-statistic):          2.04e-139
## Time:                        10:18:15   Log-Likelihood:                -688.05
## No. Observations:                 392   AIC:                             1392.
## Df Residuals:                     384   BIC:                             1424.
## Df Model:                           7
## Covariance Type:            nonrobust
## ================================================================================
##                    coef    std err          t      P>|t|      [0.025      0.975]
## --------------------------------------------------------------------------------
## Intercept       -7.3178      1.974     -3.707      0.000     -11.199      -3.437
## cylinders       -0.2097      0.137     -1.526      0.128      -0.480       0.060
## displacement     0.0085      0.003      2.647      0.008       0.002       0.015
## horsepower      -0.0072      0.006     -1.230      0.220      -0.019       0.004
## weight          -0.0028      0.000     -9.929      0.000      -0.003      -0.002
## acceleration     0.0342      0.042      0.815      0.415      -0.048       0.117
## year             0.3191      0.022     14.729      0.000       0.276       0.362
## origin           0.6061      0.118      5.127      0.000       0.374       0.839
## ==============================================================================
## Omnibus:                       31.906   Durbin-Watson:                   1.309
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):               53.100
## Skew:                           0.529   Prob(JB):                     2.95e-12
## Kurtosis:                       4.460   Cond. No.                     8.59e+04
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## [2] The condition number is large, 8.59e+04. This might indicate that there are
## strong multicollinearity or other numerical problems.
## """
```

i) Der p-Wert zum zugehörigen F-Wert ist praktisch 0 und somit besteht ein statistisch signifikanter Zusammenhang zwischen Zielvariable und den Prädiktoren.

ii) Dies sind die Koeffizienten (**`displacement`**, **`weight`**, **`year`** und **`origin`**).

iii) Der Koeffizient für **year** ist positiv. Das heisst, man mit jüngeren Autos weiter pro Gallone Benzin kommt. Die neueren Autos sind als im Allgemeinen sparsamer.

d) Output:

```
fit = ols("mpg~weight*year" , data=df).fit()

fit.summary()

## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                    mpg   R-squared:                       0.834
## Model:                            OLS   Adj. R-squared:                  0.833
## Method:                 Least Squares   F-statistic:                     649.3
## Date:                Mon, 11 May 2020   Prob (F-statistic):          8.06e-151
## Time:                        10:18:15   Log-Likelihood:                -673.91
## No. Observations:                 392   AIC:                             1356.
## Df Residuals:                     388   BIC:                             1372.
## Df Model:                           3
## Covariance Type:            nonrobust
## ===============================================================================
##                   coef    std err          t      P>|t|      [0.025      0.975]
## -------------------------------------------------------------------------------
## Intercept     -46.9421      5.502     -8.531      0.000     -57.760     -36.124
## weight          0.0117      0.002      6.242      0.000       0.008       0.015
## year            0.8672      0.073     11.876      0.000       0.724       1.011
## weight:year    -0.0002   2.51e-05     -7.752      0.000      -0.000      -0.000
## ==============================================================================
## Omnibus:                       48.151   Durbin-Watson:                   1.345
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):               86.841
## Skew:                           0.722   Prob(JB):                     1.39e-19
## Kurtosis:                       4.798   Cond. No.                     1.87e+07
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## [2] The condition number is large, 1.87e+07. This might indicate that there are
## strong multicollinearity or other numerical problems.
## """

fit.pvalues

## Intercept      3.295213e-16
## weight         1.136624e-09
## year           5.881890e-28
## weight:year    8.015486e-14
## dtype: float64
```

Der p-Wert des Interaktionsterm ist von der Grössenordnung 10^{-14}, also sehr nahe bei 0. Die Nullhypothese, dass keine Interaktion vorliegt, wird also verworfen.

Dies lässt sich damit erklären, dass das Gewicht mit den jüngeren Autos immer kleiner geworden ist.