## Regression

In unserer Analyse haben wir ausschließlich lineare Regressionsmodelle verwendet, um die Daten zu modellieren. Dabei wurden die Variablen Age, Height, BMI, Weight, FoodConsumption und Activity berücksichtigt. Andere verfügbare Daten, wie beispielsweise jene, die lediglich aus den Datensätzen (1) und (2) bestanden, wurden nicht einbezogen, da sie für die Modellierung als nicht aussagekräftig eingeschätzt wurden.

**Training eines LR-Modells**

Wir haben in diesem Abschnitt das Ziel, die Daten zu modellieren und dadurch zu analysieren.

1.py

| | Age | Height | Weight | BMI | FoodConsumption | Activity |
|---|---|---|---|---|---|---|
| count | 1610.000000 | 1610.000000 | 1610.000000 | 1610.000000 | 1610.000000 | 1610.000000 |
| mean | 33.115528 | 167.741615 | 73.504745 | 25.990435 | 5.824845 | 6.614907 |
| std | 9.835076 | 7.979873 | 17.096973 | 5.149009 | 1.707041 | 1.881387 |
| min | 18.000000 | 150.000000 | 38.760000 | 17.000000 | 1.000000 | 1.000000 |
| 25% | 25.000000 | 161.000000 | 59.800000 | 21.700000 | 5.000000 | 5.000000 |
| 50% | 32.000000 | 168.000000 | 70.400000 | 27.500000 | 6.000000 | 7.000000 |
| 75% | 41.000000 | 174.000000 | 85.180000 | 27.500000 | 7.000000 | 8.000000 |
| max | 54.000000 | 193.000000 | 130.370000 | 35.000000 | 11.000000 | 12.000000 |

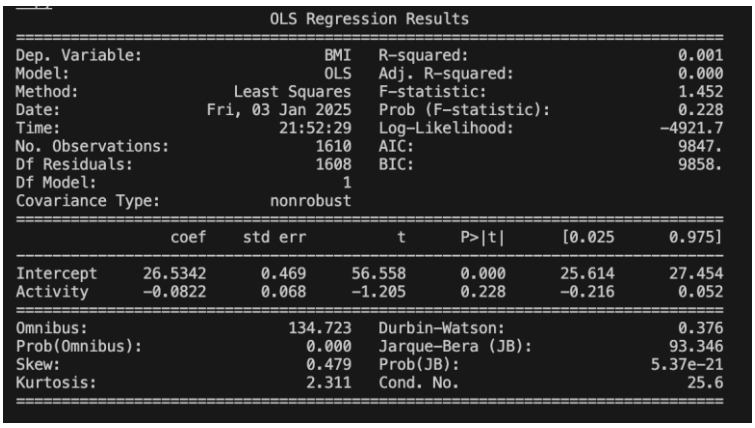
BMI: Ein BMI über 24,9 zählt als übergewichtig. Somit lässt sich ablesen, dass mehr als die Hälfte übergewichtig sind. Außerdem liegt der Durchschnittswert bei 25,99 und somit im Bereich des Übergewichts.

Activtiy: Bei Activity steht eine hohe Zahl für mehr Aktivität. Somit lässt sich sagen, dass mehr als die hälfte Aktiv tätig sind.

FoodConsumption: Bei FoodConsumption steht eine niedrige Zahl für „gut“. Somit lässt sich sagen, dass Durchschnittlich mehr als die Hälfte zu viel Essen Konsumiert.

**Interpretation eines LR-Modells**

Aus der Analysesicht kann man entweder mit einem Modell beginnen, das alle Features enthält, und es dann schrittweise verfeinern, oder man **beginnt mit einem künstlich einfachen Modell als Vergleichspunkt. Annahme: Geringere Aktivität führt zu höherem BMI.**



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | BMI | R-squared: | 0.001 |
| Model: | OLS | Adj. R-squared: | 0.000 |
| Method: | Least Squares | F-statistic: | 1.452 |
| Date: | Fri, 03 Jan 2025 | Prob (F-statistic): | 0.228 |
| Time: | 21:52:29 | Log-Likelihood: | -4921.7 |
| No. Observations: | 1610 | AIC: | 9847. |
| Df Residuals: | 1608 | BIC: | 9858. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

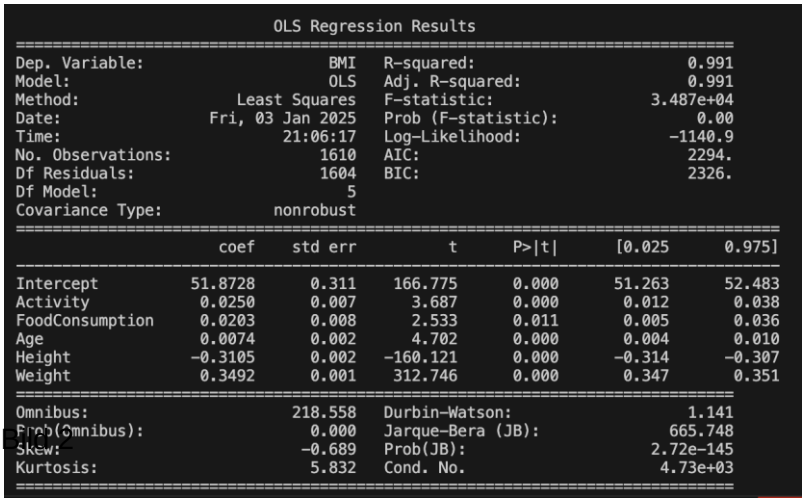
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 26.5342 | 0.469 | 56.558 | 0.000 | 25.614 | 27.454 |
| Activity | -0.0822 | 0.068 | -1.205 | 0.228 | -0.216 | 0.052 |

| | | | |
|---|---|---|---|
| Omnibus: | 134.723 | Durbin-Watson: | 0.376 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 93.346 |
| Skew: | 0.479 | Prob(JB): | 5.37e-21 |
| Kurtosis: | 2.311 | Cond. No. | 25.6 |

R-squared-Wert: Korrelationen gelten ab r=0.6 als “stark”. Somit ist R-squared von 0.001 schwach.

P-Wert: Der p-Wert liegt mit 0,23 deutlich über 0,05. Somit ist die Nullhypothese, dass der Acitivty-Wert keine Auswirkung auf den BMI hat, nicht zu verwerfen.

Koeffizienten: Das Modell sagt voraus, dass für jeden Activity-Wert, der BMI um 0,08 sinkt.



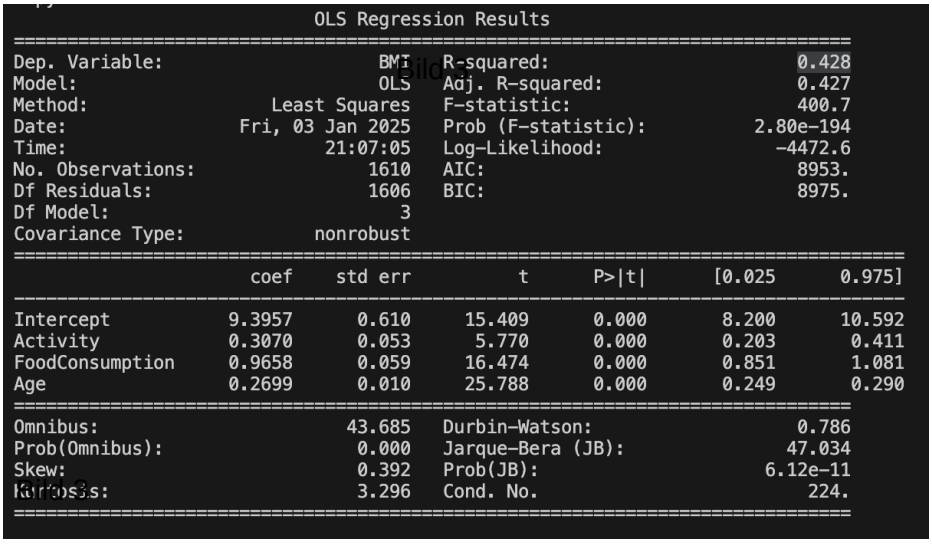
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | BMI | R-squared: | 0.991 |
| Model: | OLS | Adj. R-squared: | 0.991 |
| Method: | Least Squares | F-statistic: | 3.487e+04 |
| Date: | Fri, 03 Jan 2025 | Prob (F-statistic): | 0.00 |
| Time: | 21:06:17 | Log-Likelihood: | -1140.9 |
| No. Observations: | 1610 | AIC: | 2294. |
| Df Residuals: | 1604 | BIC: | 2326. |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 51.8728 | 0.311 | 166.775 | 0.000 | 51.263 | 52.483 |
| Activity | 0.0250 | 0.007 | 3.687 | 0.000 | 0.012 | 0.038 |
| FoodConsumption | 0.0203 | 0.008 | 2.533 | 0.011 | 0.005 | 0.036 |
| Age | 0.0074 | 0.002 | 4.702 | 0.000 | 0.004 | 0.010 |
| Height | -0.3105 | 0.002 | -160.121 | 0.000 | -0.314 | -0.307 |
| Weight | 0.3492 | 0.001 | 312.746 | 0.000 | 0.347 | 0.351 |

| | | | |
|---|---|---|---|
| Omnibus: | 218.558 | Durbin-Watson: | 1.141 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 665.748 |
| Skew: | -0.689 | Prob(JB): | 2.72e-145 |
| Kurtosis: | 5.832 | Cond. No. | 4.73e+03 |

Jeder Wert ist relevant, da der p-Wert jeweils unter 0.05 liegt. Der R-squared liegt bei 0.991, somit ist die Korrelation „stark“. Dies könnte an „Height“ und „Weight“ liegen, da der BMI-Wert daraus resultiert. Als nächstes werden diese Werte außer Acht gelassen.



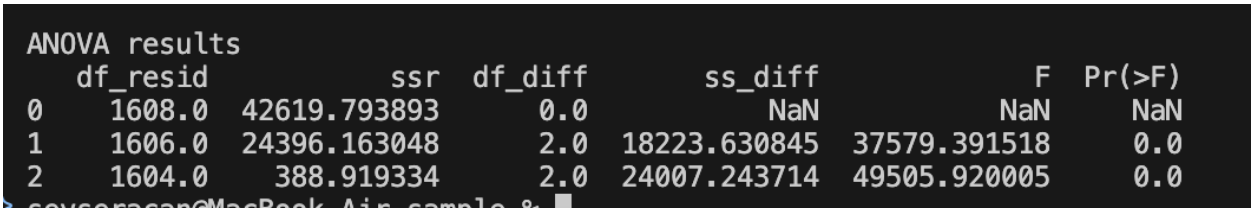
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | BMI | R-squared: | 0.428 |
| Model: | OLS | Adj. R-squared: | 0.427 |
| Method: | Least Squares | F-statistic: | 400.7 |
| Date: | Fri, 03 Jan 2025 | Prob (F-statistic): | 2.80e-194 |
| Time: | 21:07:05 | Log-Likelihood: | -4472.6 |
| No. Observations: | 1610 | AIC: | 8953. |
| Df Residuals: | 1606 | BIC: | 8975. |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 9.3957 | 0.610 | 15.409 | 0.000 | 8.200 | 10.592 |
| Activity | 0.3070 | 0.053 | 5.770 | 0.000 | 0.203 | 0.411 |
| FoodConsumption | 0.9658 | 0.059 | 16.474 | 0.000 | 0.851 | 1.081 |
| Age | 0.2699 | 0.010 | 25.788 | 0.000 | 0.249 | 0.290 |

| | | | |
|---|---|---|---|
| Omnibus: | 43.685 | Durbin-Watson: | 0.786 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 47.034 |
| Skew: | 0.392 | Prob(JB): | 6.12e-11 |
| Kurtosis: | 3.296 | Cond. No. | 224. |

In diesem Modell werden nur „Activity“, „FoodConsumption“ und „Age“ betrachtet. Hierbei bleibt jeder p-Wert relevant. Jedoch hat sich R-squared deutlich verschlechtert, aber dennoch ist die Korrelation „stark“. Gemessen an R-squared ist also Bild 2 > Bild 3 > Bild 1.



```
ANOVA results
   df_resid           ssr  df_diff       ss_diff             F  Pr(>F)
0    1608.0  42619.793893      0.0           NaN           NaN     NaN
1    1606.0  24396.163048      2.0  18223.630845  37579.391518     0.0
2    1604.0    388.919334      2.0  24007.243714  49505.920005     0.0
```

**Modellvergleich mit ANOVA**

Vergleich von Bild 3 mit Bild 1:

Der Vergleich in der zweiten Zeile zeigt einen signifikanten p-Wert (Pr(>F) = 0.0), was bedeutet, dass Bild 3 eine signifikante Verbesserung gegenüber Bild 1 darstellt.

Vergleich von Bild 2 mit Bild 3:

Der Vergleich in der dritten Zeile hat ebenfalls einen signifikanten p-Wert (Pr(>F) = 0.0), was darauf hinweist, dass Bild 2 eine signifikante Verbesserung gegenüber Bild 2 darstellt.

**Schlussfolgerung:**
Da sowohl der Vergleich von Bild 3 mit Bild 1 als auch der Vergleich von Bild2 mit Bild3 signifikant ist, können wir sagen: Bild 2 ist besser als Bild 2 somit auch besser als Bild1.