Multiple linear regression

The ASTA team

Contents

1 Multiple regression model

1.1 Multiple regression model

- We look at data from `Table 9.15` in Agresti. The data are measurements in the 67 counties of Florida.
- Our focus is on
  - The response y: `Crime` which is the crime rate
  - The predictor x_1: `Education` which is proportion of the population with high school exam
  - The predictor x_2: `Urbanisation` which is proportion of the population living in urban areas

1.2 Example

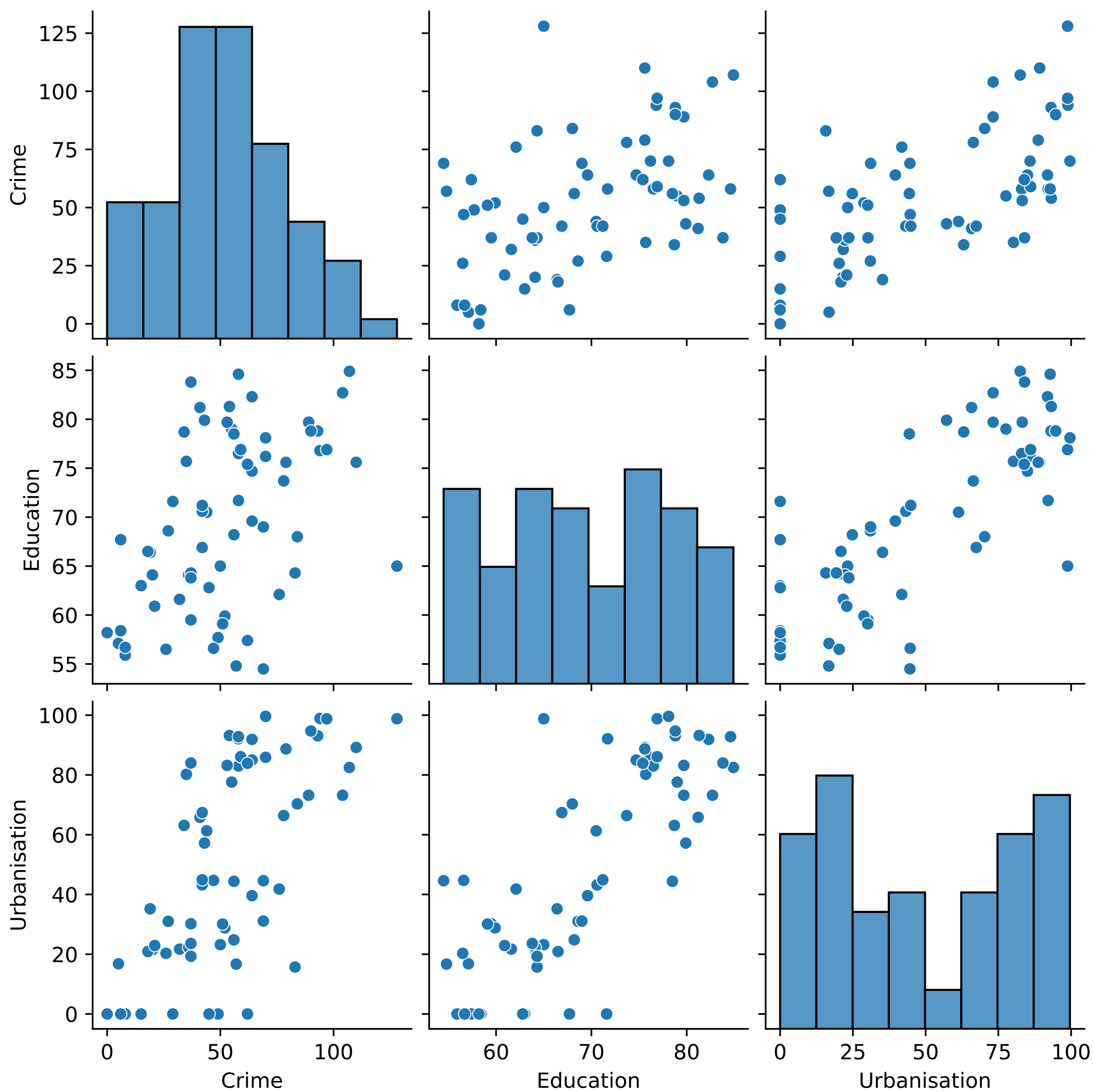
```
import pandas as pd
```

```python
FL = pd.read_csv("https://asta.math.aau.dk/datasets?file=fl-crime.txt", sep='\t')
FL.head(n = 3)
```

```
##    Crime  Education  Urbanisation
## 0    104       82.7          73.2
## 1     20       64.1          21.5
## 2     64       74.7          85.0
```

```python
import seaborn as sns
import matplotlib.pyplot as plt

p = sns.pairplot(FL)
```



1.3 Correlations

- There is significant (p$\approx 7 \times 10^{-5}$) positive correlation (r=0.47) between `Crime` and `Education`
- Then there is also significant positive correlation (r=0.68) between `Crime` and `Urbanisation`

```
FL.corr()
```

```
##                  Crime  Education  Urbanisation
## Crime         1.000000   0.466912      0.677368
## Education     0.466912   1.000000      0.790719
## Urbanisation  0.677368   0.790719      1.000000
```

```
import pingouin as pg

pg.corr(FL['Crime'], FL['Education'], method='pearson')
```

```
##           n         r         CI95%     p-val     BF10     power
## pearson  67  0.466912  [0.26, 0.64]  0.000068  357.897  0.982794
```

1.4 Several predictors

- Both `Education` (x_1) and `Urbanisation` (x_2) are pretty good predictors for `Crime` (y).
- We therefore want to consider the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- The errors ϵ are random noise with mean zero and standard deviation $\sigma_{y|x}$.
- The graph for the mean response is in other words a 2-dimensional plane in a 3-dimensional space.
- We determine estimates (a, b_1, b_2) for $(\alpha, \beta_1, \beta_2)$ via the least squares method, i.e deviations from the plane.

1.5 Example

```
import numpy as np
import statsmodels.formula.api as smf

model = smf.ols('Crime ~ Education + Urbanisation', data=FL).fit()
model.summary(slim = True)
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                  Crime   R-squared:                       0.471
## Model:                            OLS   Adj. R-squared:                  0.455
## No. Observations:                  67   F-statistic:                     28.54
## Covariance Type:            nonrobust   Prob (F-statistic):           1.38e-09
## ================================================================================
##                    coef    std err          t      P>|t|      [0.025      0.975]
## --------------------------------------------------------------------------------
## Intercept       59.1181     28.365      2.084      0.041       2.452     115.784
## Education       -0.5834      0.472     -1.235      0.221      -1.527       0.360
## Urbanisation     0.6825      0.123      5.539      0.000       0.436       0.929
## ================================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
## """
```

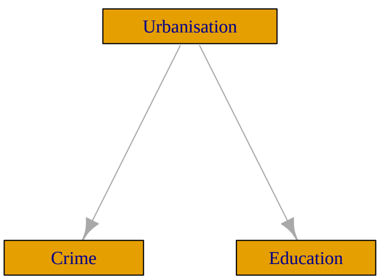
- From the output we find the prediction equation

$$\widehat{y} = 59 - 0.58x_1 + 0.68x_2$$

- Not exactly what we expected based on the correlation.
- Now there appears to be a negative association between y and x_1 (Simpsons Paradox)!
- We can also find the standard error (0.4725) and the corresponding t-score (-1.235) for the the slope of `Education`
- This yields a p-value of 22%, i.e. the slope is not significantly different from zero.

1.6 Simpsons paradox

- The example illustrates **Simpson's paradox**.
- When considered alone `Education` is a good predictor for `Crime` (with positive correlation).
- When we add `Urbanisation`, then `Education` has a negative effect on `Crime` (but not significant).



- A possible explanation is illustrated by the graph above.
  - `Urbanisation` has positive effect on both `Education` and `Crime`.
  - For a given level of `urbanisation` there is a (non-significant) negative association between `Education` and `Crime`.
  - Viewed alone `Education` is a good predictor for `Crime`. If `Education` has a large value, then this indicates a large value of `Urbanisation` and thereby a large value of `Crime`.

2 The general model

2.1 Regression model

- We have a sample of size n, where we have measured
  - the response y.
  - k potential predictors $x_1, x_2, \dots, x_k$.
- Multiple regression model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon.$$

- The errors ϵ are a sample from a population with mean zero and standard deviation $\sigma_{y|x}$.
- The **systematic** part of the model, i.e. when all errors are zero, says that **the mean response** is a linear function of the predictors:

$$E(y|x_1, x_2, \dots, x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- The symbol E is used here to denote expectation, i.e., mean value.

2.2 Interpretation of parameters

- In the multiple linear regression model

$$E(y|x_1, x_2, \dots, x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- The parameter α is the `Intercept`, corresponding to the mean response, when all predictors are equal to zero.
- The parameters $(\beta_1, \beta_2, \ldots, \beta_k)$ are called **partial regression coefficients**.

- Imagine that all predictors but x_1 are held fixed. Then $y = \tilde{\alpha} + \beta_1 x_1$ is a line with slope β_1, which describes the rate of change in the mean response, when x_1 is changed one unit. Here

$$\tilde{\alpha} = \alpha + \beta_2 x_2 + \cdots + \beta_k x_k$$

is a constant number since we assumed all predictors but x_1 was held fixed.
- The rate of change β_1 does not depend on the value of the remaining predictors. In this case we say that there is **no interaction** between the effects of the predictors on the response.
- The above holds similarly for the other partial regression coefficients.
- An example of a model with interaction is

$$E(y|x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \alpha + \beta_2 x_2 + (\beta_1 + \beta_3 x_2) x_1$$

- When we fix x_2 the line has slope $\beta_1 + \beta_3 x_2$, which depends on the chosen value of x_2.

3 Estimation

3.1 Estimation of model

- The estimate $(a, b_1, b_2, \ldots, b_k)$ for $(\alpha, \beta_1, \beta_2, \ldots, \beta_k)$ is determined by minimizing the sum of squared errors.
- Based on this estimate we write the prediction equation as

$$\widehat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

- The distance between model and data is measured by the sum of squared erros

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

- We estimate $\sigma_{y|x}$ by the quantity

$$s_{y|x} = \sqrt{\frac{SSE}{n-k-1}}.$$

- Rather than n we divide SSE by **the degrees of freedom** $df = n - k - 1$. Theory shows, that this is reasonable.
- The degrees of freedom df are determined by the sample size minus the number of parameters in the regression equation.
- Currently we have $k + 1$ parameters: 1 intercept and k slopes.

4 Multiple R-squared

4.1 Multiple R^2

- We want to compare two models to predict the response y. Analogous to simple linear regression we have the following setup:
- Model 1: We do not use the predictors, and use $\bar{y}$ to predict any y-measurement. The corresponding prediction error is
  - $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ and is called the **Total Sum of Squares**.
- Model 2: We use the multiple prediction equation to predict y, i.e. the prediction error is
  - $SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ and is called **Sum of Squared Errors**.

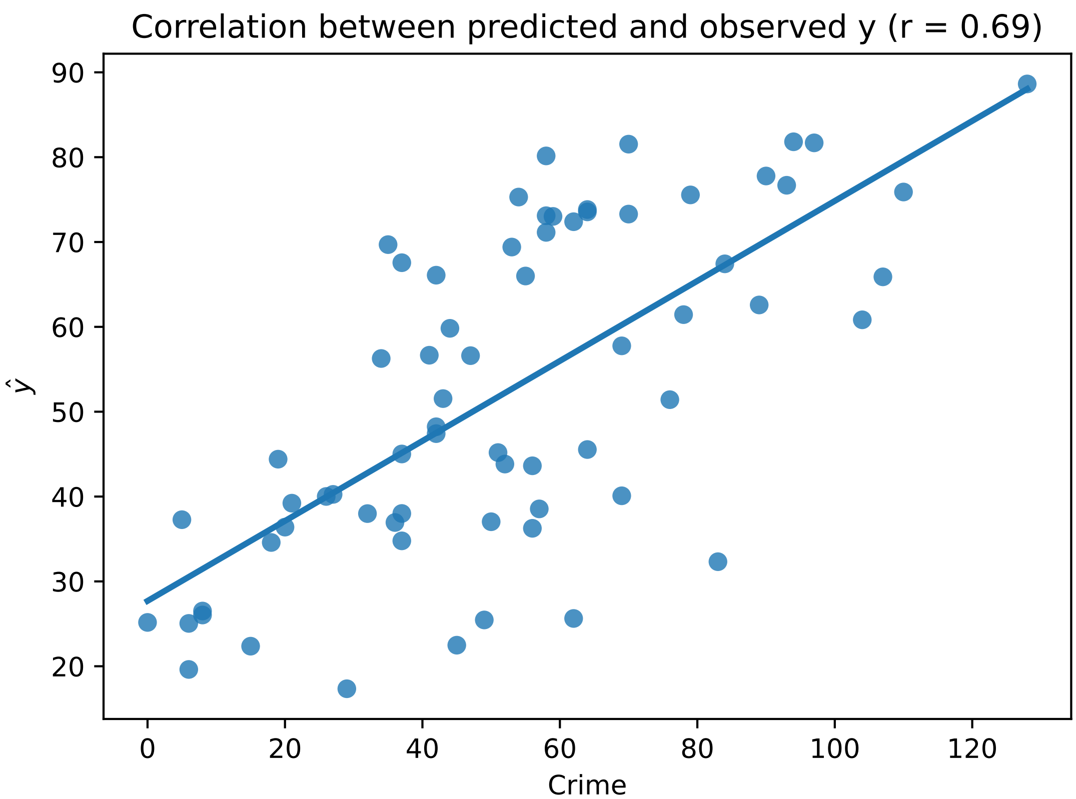
- We then define **the multiple coefficient of determination**

$$R^2 = \frac{TSS - SSE}{TSS}.$$

- Thus, R^2 is the relative reduction in prediction error, when we use $x_1, x_2, \ldots, x_k$ as explanatory variables.
- It can be shown that the **multiple correlation** $R = \sqrt{R^2}$ is the correlation between y and $\widehat{y}$.

```
FL['CrimePred'] = model.fittedvalues
r_val = model.rsquared**0.5
sns.regplot(x='Crime', y='CrimePred', data=FL, ci=None)

# Labels and title
plt.title(f"Correlation between predicted and observed y (r = {r_val:.2f})")
plt.xlabel("Crime")
plt.ylabel(r'$\hat{y}$') #r for raw string, no need to escape backslash in math
```



4.2 Example

```
model.summary(slim = True)
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                                 OLS Regression Results
## ==============================================================================
## Dep. Variable:                  Crime   R-squared:                       0.471
## Model:                            OLS   Adj. R-squared:                  0.455
```

```
## No. Observations:                  67   F-statistic:                     28.54
## Covariance Type:            nonrobust   Prob (F-statistic):           1.38e-09
## ================================================================================
##                    coef    std err          t      P>|t|      [0.025      0.975]
## --------------------------------------------------------------------------------
## Intercept       59.1181     28.365      2.084      0.041       2.452     115.784
## Education       -0.5834      0.472     -1.235      0.221      -1.527       0.360
## Urbanisation     0.6825      0.123      5.539      0.000       0.436       0.929
## ================================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## """
```

```
np.sqrt(model.mse_resid)
```

```
## np.float64(20.81558240979863)
```

```
model.df_resid
```

```
## np.float64(64.0)
```

- The prediction equation is $\widehat{y} = 59 - 0.58x_1 + 0.68x_2$
- The estimate for $\sigma_{y|x}$ is $s_{y|x} = 20.82$ (`Residual standard error`) with $df = 67 - 3 = 64$ degrees of freedom.
- Multiple $R^2 = 0.4714$, i.e. 47% of the variation in the response is explained by including the predictors in the model.
- The estimate $b_1 = -0.5834$ has standard error (`Std. Error`) $se = 0.4725$ with corresponding t-score (`t value`) $t_{\text{obs}} = \frac{-0.5834}{0.4725} = -1.235$.
- The hypothesis $H_0 : \beta_1 = 0$ has the t-score $t_{\text{obs}} = -1.235$, which means that b_1 isn't significantly different from zero, since the p-value (`Pr(>|t|)`) is 22%. That means that we should exclude `Education` as a predictor.

4.3 Example

- Our final model is then a simple linear regression:

```
model2 = smf.ols('Crime ~ Urbanisation', data=FL).fit()
model2.summary(slim = True)
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                  Crime   R-squared:                       0.459
## Model:                            OLS   Adj. R-squared:                  0.451
## No. Observations:                  67   F-statistic:                     55.11
## Covariance Type:            nonrobust   Prob (F-statistic):           3.08e-10
## ================================================================================
##                    coef    std err          t      P>|t|      [0.025      0.975]
## --------------------------------------------------------------------------------
## Intercept       24.5412      4.539      5.406      0.000      15.476      33.607
## Urbanisation     0.5622      0.076      7.424      0.000       0.411       0.713
## ================================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
## """
```

- The coefficient of determination always decreases, when the model is simpler. Now we have $R^2 = 46\%$, where before we had 47%. But the decrease is not significant.

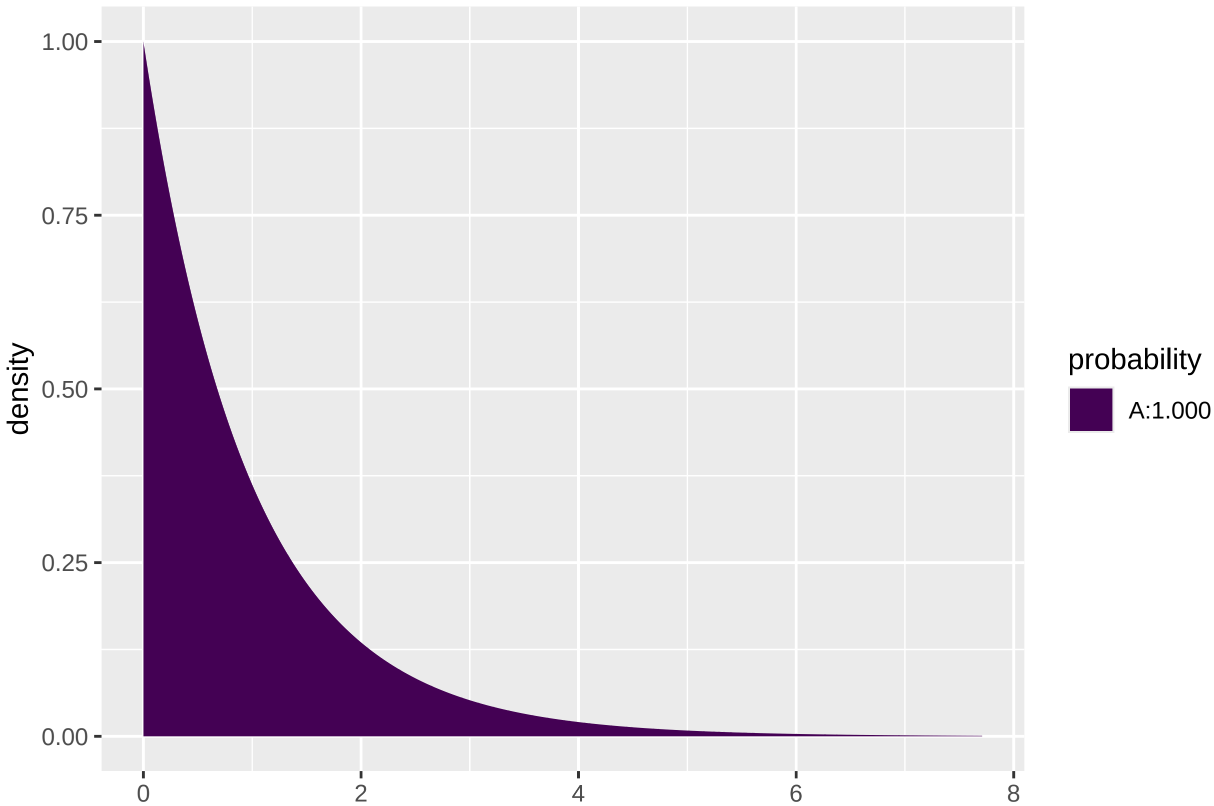
5 F-test for effect of predictors

5.1 F-test

- We consider the hypothesis
$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
against the alternative, that at least one of these are non-zero.
- As test statistic we use
$$F_{obs} = \frac{(n-k-1)R^2}{k(1-R^2)}$$
- Large values of R^2 implies large values of F, which points to the alternative hypothesis.
- I.e. when we have calculated the observed value F_{obs}, then we have to find the probability that a new experiment would result in a larger value.
- It can be shown that the reference distribution is (can be approximated by) a so-called **F-distribution** with **degrees of freedom** $df_1 = k$ and $df_2 = n - k - 1$.

5.2 Example

- We return to `Crime` and the prediction equation $\widehat{y} = 59 - 0.58x_1 + 0.68x_2$, where $n = 67$ and $R^2 = 0.4714$. We have
  - $df_1 = k = 2$ since we have 2 predictors.
  - $df_2 = n - k - 1 = 67 - 2 - 1 = 64$.
  - Then we can calculate $F_{obs} = \frac{(n-k-1)R^2}{k(1-R^2)} = 28.54$
- To evaluate the value 28.54 in the relevant F-distribution:




```
from scipy.stats import f

1 - f.cdf(28.54, 2, 64)
```

```
## np.float64(1.3786117802894182e-09)
```

- So p-value$=1.38 \times 10^{-9}$ (notice we don't multiply by 2 since this is a one-sided test; only large values point more towards the alternative than the null hypothesis).
- All this can be found in the summary output we already have:

```
model.summary(slim = True)
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                  Crime   R-squared:                       0.471
## Model:                            OLS   Adj. R-squared:                  0.455
## No. Observations:                  67   F-statistic:                     28.54
## Covariance Type:            nonrobust   Prob (F-statistic):           1.38e-09
## ================================================================================
##                    coef    std err          t      P>|t|      [0.025      0.975]
## --------------------------------------------------------------------------------
## Intercept       59.1181     28.365      2.084      0.041       2.452     115.784
## Education       -0.5834      0.472     -1.235      0.221      -1.527       0.360
## Urbanisation     0.6825      0.123      5.539      0.000       0.436       0.929
## ================================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
## """
```

6 Test for interaction

6.1 Interaction between effects of predictors

- Could it be possible that a combination of `Education` and `Urbanisation` is good for prediction? We want to investigate this using the model

$$E(y|x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

where we have extended with a possible effect of the product $x_1 x_2$:

```
model3 = smf.ols('Crime ~ Education * Urbanisation', data=FL).fit()
model3.summary(slim = True)
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                  Crime   R-squared:                       0.479
## Model:                            OLS   Adj. R-squared:                  0.454
## No. Observations:                  67   F-statistic:                     19.32
## Covariance Type:            nonrobust   Prob (F-statistic):           5.37e-09
## ==========================================================================================
##                              coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------------------
## Intercept                 19.3175     49.959      0.387      0.700     -80.517     119.152
## Education                  0.0340      0.794      0.043      0.966      -1.552       1.620
## Urbanisation               1.5143      0.868      1.744      0.086      -0.220       3.249
## Education:Urbanisation    -0.0120      0.012     -0.968      0.337      -0.037       0.013
## ==========================================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## [2] The condition number is large, 8.97e+04. This might indicate that there are
## strong multicollinearity or other numerical problems.
## """
```

- When we look at the p-values in the table nothing is significant at the 5% level!
- But the F-statistic tells us that the predictors collectively have a significant prediction ability.
- Why has the highly significant effect of x_2 disappeared? Because the predictors x_1 and $x_1 x_2$ are able to explain the same as x_2.
- Previously we only had x_1 as alternative explanation to x_2 - and that wasn't enough.
- The phenomenon is called **multicollinearity** and illustrates that we can have different models with equally good predictive properties.
- In this case we will choose the model with x_2 since it is simpler.
- However, in general it can be difficult to choose between models. For example, if both height and weight are good predictors of some response, but one of them can be left out, which one do we choose?