Inferență statistică în ML

Cap 8. Proprietățile regresiei multi-variabilă. Variabile categoriale.

May 14, 2019

1 Proprietățile regresiei multi-variabilă

2 Variabile categoriale (factor)

3 Regresia cu variabile continue și categoriale

Swiss dataset

- măsura standardizată a fertilității și indicatorii socio-economici pentru 47 de provincii vorbitoare de limba franceză pentru Elveția anului 1888
- dataframe cu 47 de observații pentru 6 variabile (features), fiecare variabilă exprimată în procente, adică în intervalul $[0, 100]$
  1. **Fertility**, măsură standardizată a fertilității
  2. **Agriculture**, % bărbați implicați în agricultură ca ocupație
  3. **Examination**, % recruți perfect sănătoși, examenul medical al armatei
  4. **Education**, % educație peste școala primară al celor recrutați în armată
  5. **Catholic**, % catolici din regiune (două religii, catolică vs. protestantă)
  6. **Infant.Mortality**, % născuți vii care trăiesc mai puțin de 1 an
- toate variabilele în afară de Fertility sunt definite ca proporții din populație
- ce variabilă/variabile explică Fertility în aceste provincii?

Swiss: correlations

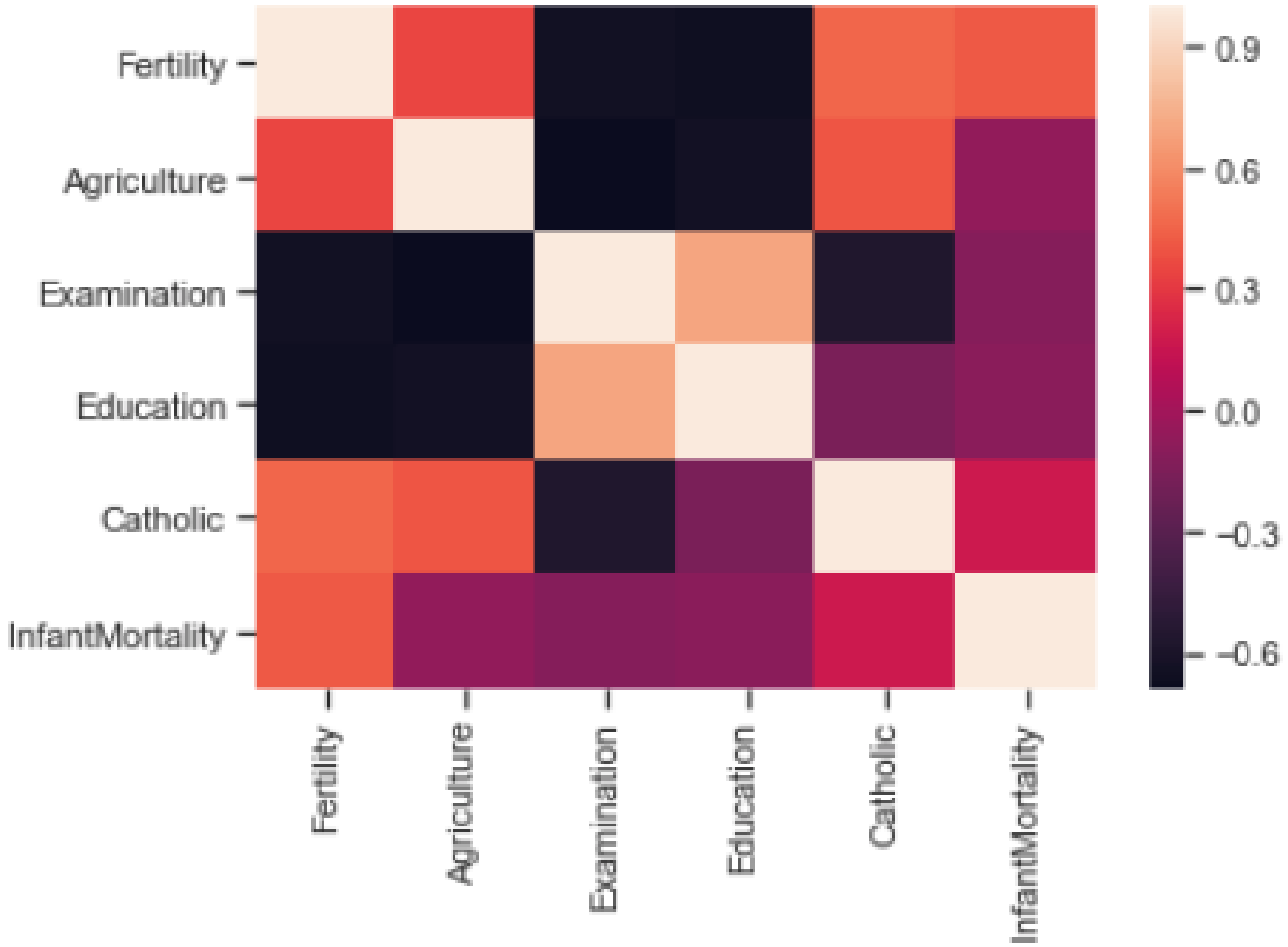
	District	Fertility	Agriculture	Examination	Education	Catholic	InfantMortality
0	Courtelary	80.2	17.0	15	12	9.96	22.2
1	Delemont	83.1	45.1	6	9	84.84	22.2
2	Franches-Mnt	92.5	39.7	5	5	93.40	20.2
3	Moutier	85.8	36.5	12	7	33.77	20.3
4	Neuveville	76.9	43.5	17	15	5.16	20.6

```
swiss.corr()
```

	Fertility	Agriculture	Examination	Education	Catholic	InfantMortality
Fertility	1.000000	0.353079	-0.645883	-0.663789	0.463685	0.416556
Agriculture	0.353079	1.000000	-0.686542	-0.639523	0.401095	-0.060859
Examination	-0.645883	-0.686542	1.000000	0.698415	-0.572742	-0.114022
Education	-0.663789	-0.639523	0.698415	1.000000	-0.153859	-0.099322
Catholic	0.463685	0.401095	-0.572742	-0.153859	1.000000	0.175496
InfantMortality	0.416556	-0.060859	-0.114022	-0.099322	0.175496	1.000000

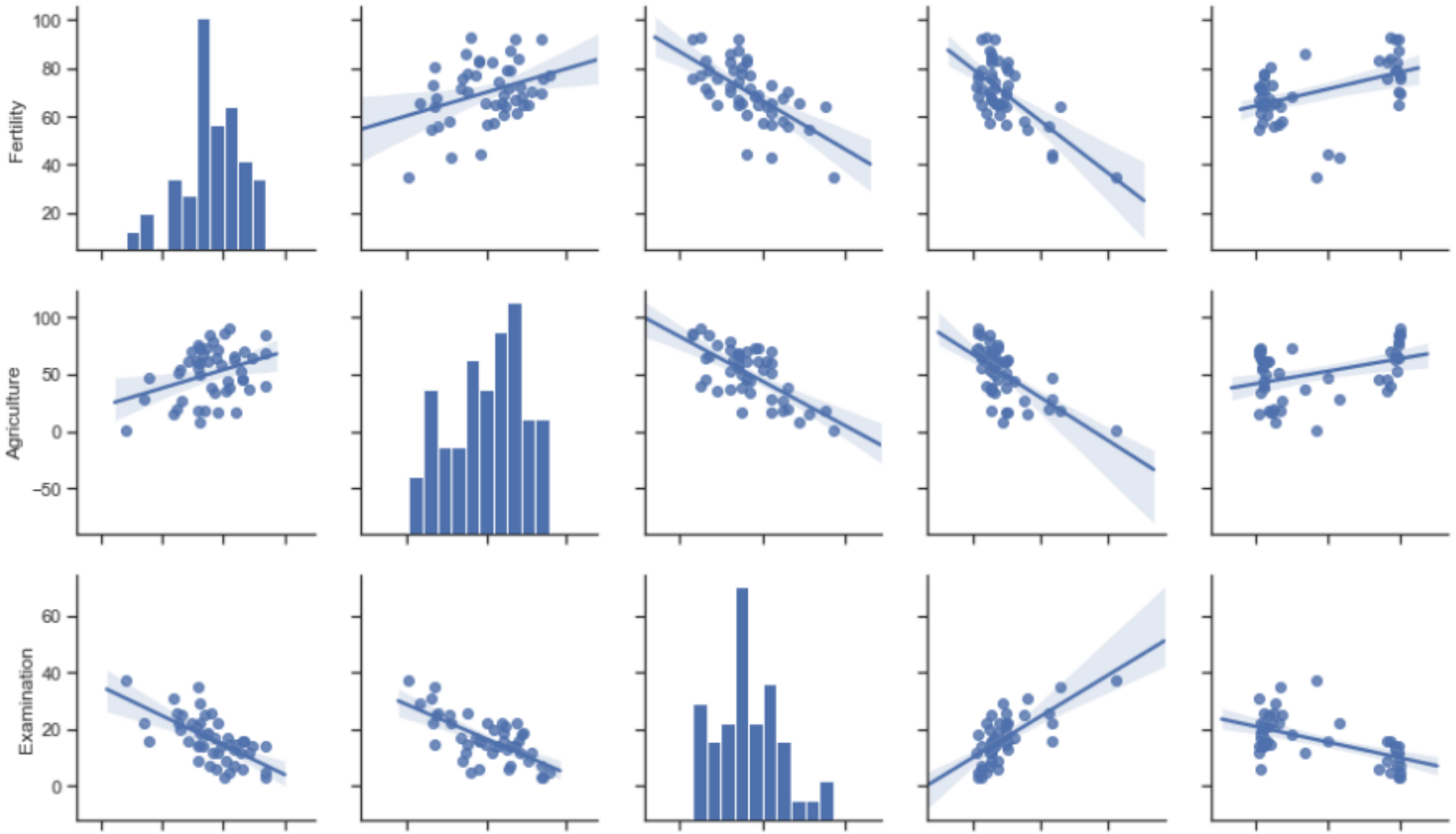
Swiss: correlations heatmap

```
sns.heatmap(swiss.corr())
plt.show()
```

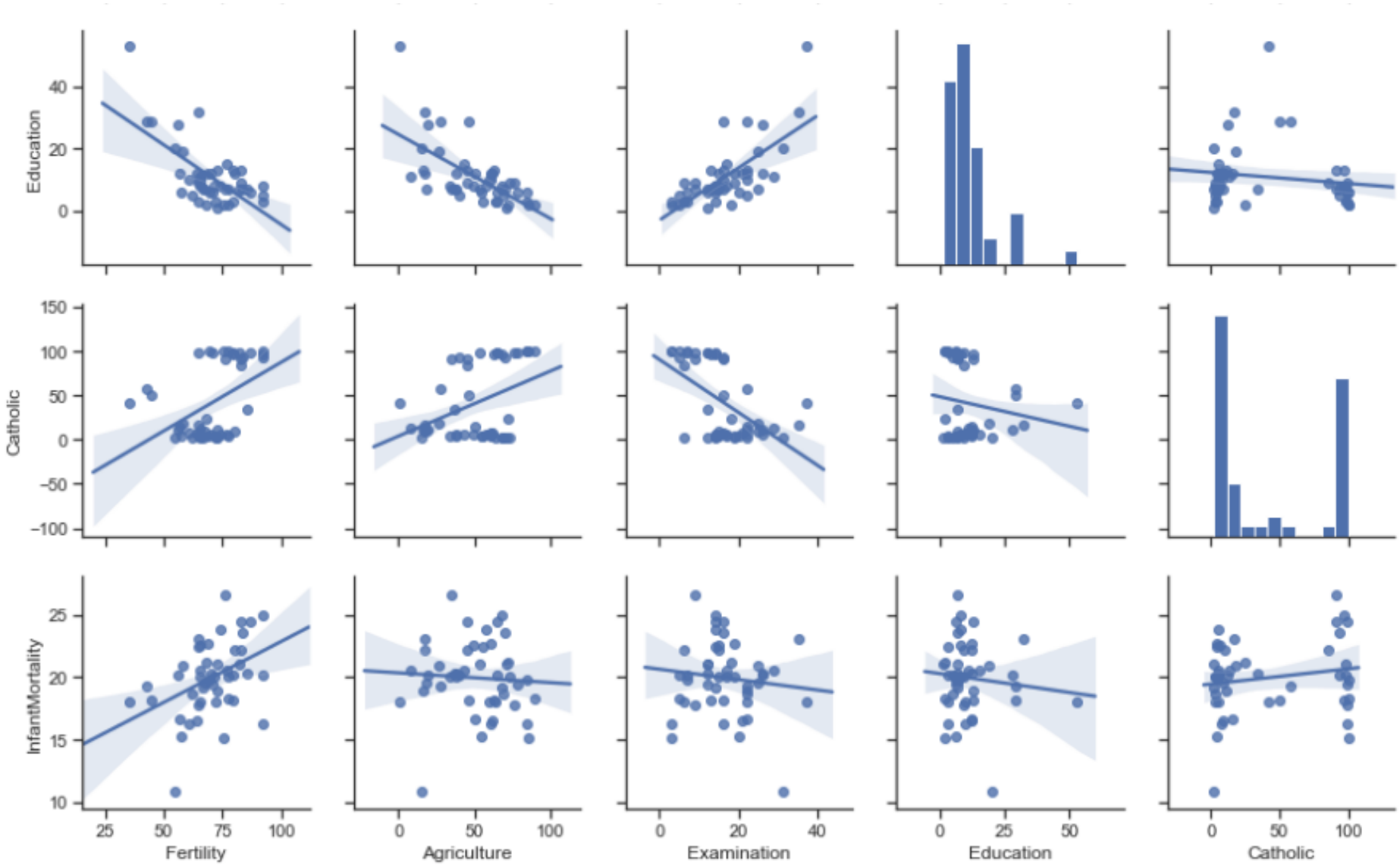


Swiss: pair plot

```
sns.set(style="ticks")
sns.pairplot(swiss, kind='reg')
plt.show()
```

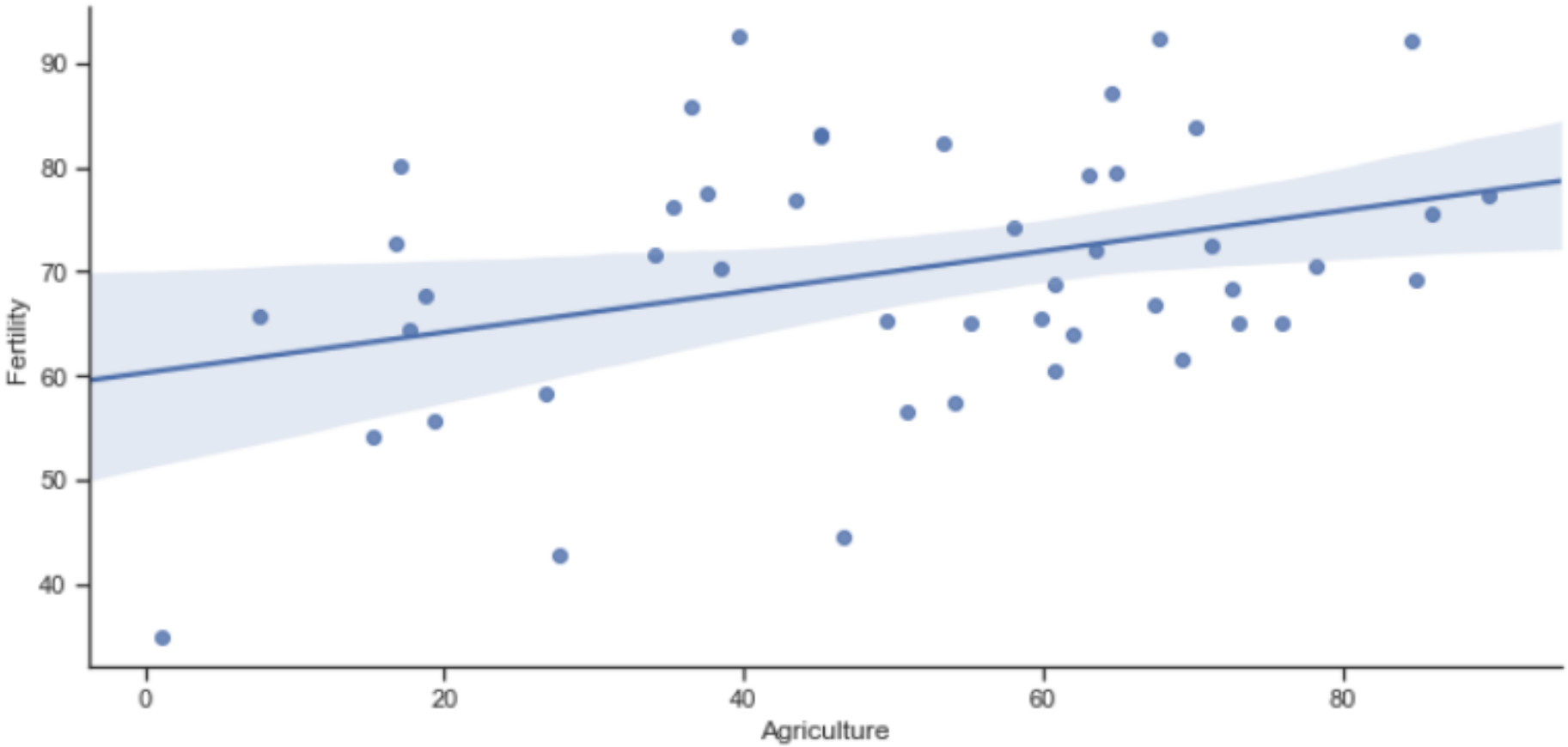


Swiss: pair plot (2)



Swiss: Fertility vs. Agriculture

```
sns.lmplot(x='Agriculture', y='Fertility', data=swiss, aspect=2)
plt.show()
print(swiss.corr()['Agriculture']['Fertility'])
```




```
0.35307918361997476
```

Swiss: Fertility Linear Regression

```
# https://stackoverflow.com/questions/19991445/run-an-ols-regression-with-pandas-data-frame
lm = smf.ols(formula='Fertility ~ Agriculture + Examination + Education + Catholic + InfantMortality',
             data=swiss).fit()
print(lm.params)
lm.summary()
```

```
Intercept          66.915182
Agriculture        -0.172114
Examination        -0.258008
Education          -0.870940
Catholic            0.104115
InfantMortality     1.077048
dtype: float64
```

Interpretare: Fertility

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	66.9152	10.706	6.250	0.000	45.294	88.536
Agriculture	-0.1721	0.070	-2.448	0.019	-0.314	-0.030
Examination	-0.2580	0.254	-1.016	0.315	-0.771	0.255
Education	-0.8709	0.183	-4.758	0.000	-1.241	-0.501
Catholic	0.1041	0.035	2.953	0.005	0.033	0.175
InfantMortality	1.0770	0.382	2.822	0.007	0.306	1.848

- predictorul Agriculture este exprimat în procente
- coeficientul său este estimat la -0.1721
- așteptarea este de o scădere de 0.17% în fertilitate pentru fiecare creștere de 1% în procentul de bărbați angajați în agricultură, dacă celelalte variabile sunt menținute constante
- Student T test pentru $H_0 : \beta_{Agri} = 0$ vs. $H_a : \beta_{Agri} \neq 0$ $(\frac{\beta_{Agri} - 0}{stderr_{Agri}})$, este semnificativ statistic (pValue=0.018)

Fertility în funcție doar de Agriculture

```
lm = smf.ols(formula='Fertility ~ Agriculture',
  data=swiss).fit()
lm.summary()
```

Fertility (Agriculture)

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	60.3044	4.251	14.185	0.000	51.742	68.867
Agriculture	0.1942	0.077	2.532	0.015	0.040	0.349

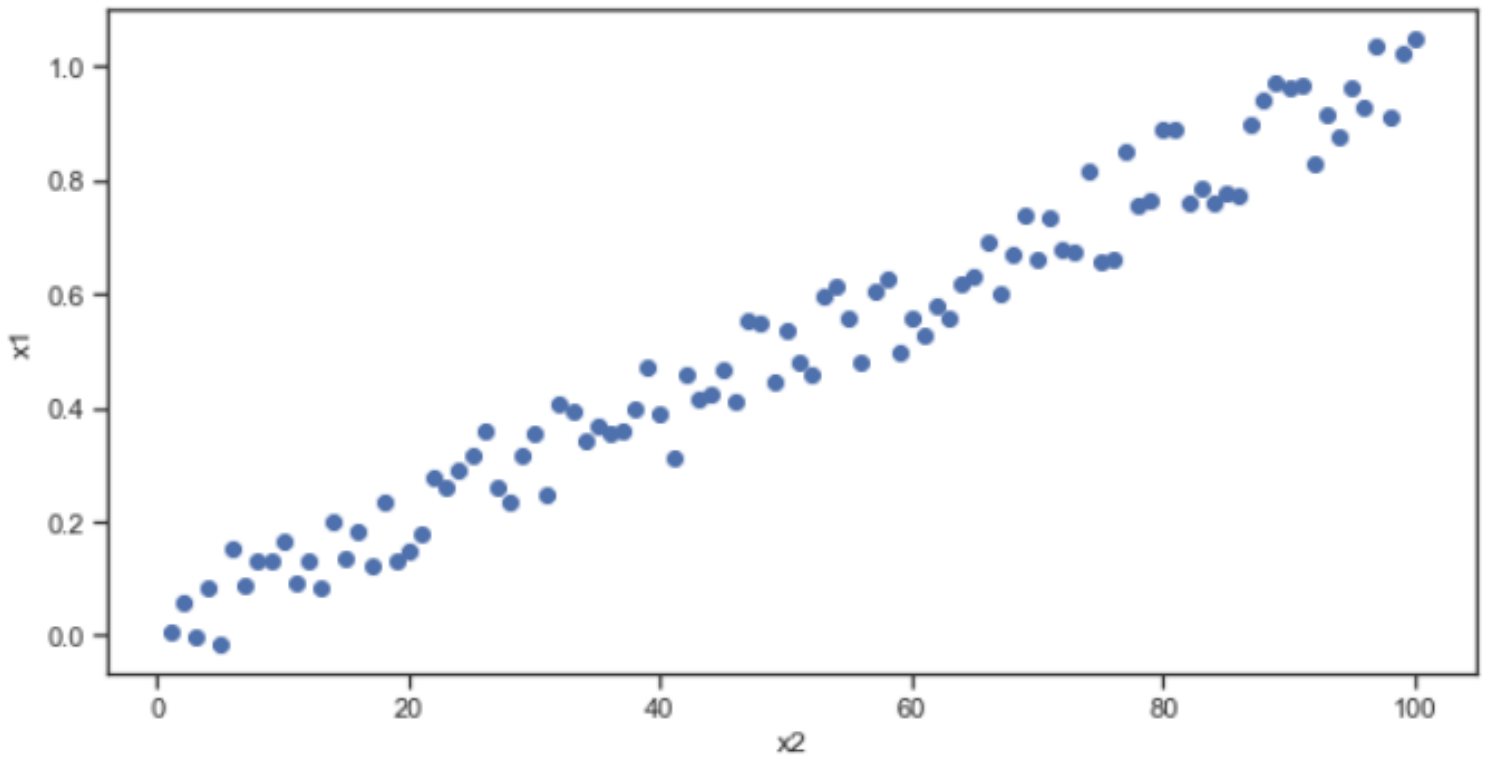
Fertility (all)

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	66.9152	10.706	6.250	0.000	45.294	88.536
Agriculture	-0.1721	0.070	-2.448	0.019	-0.314	-0.030
Examination	-0.2580	0.254	-1.016	0.315	-0.771	0.255
Education	-0.8709	0.183	-4.758	0.000	-1.241	-0.501
Catholic	0.1041	0.035	2.953	0.005	0.033	0.175
InfantMortality	1.0770	0.382	2.822	0.007	0.306	1.848

- coeficientul β_{Agri} are acum un efect pozitiv asupra Fertilității (v. https://en.wikipedia.org/wiki/Simpson%27s_paradox)
- în ambele cazuri coeficientul este statistic significant
- va trebui să aducem argumente despre ce predictori e necesar să includem

Exemplu: variabile corelate

```
n = 100 ; x2 = np.array(range(1, n+1))
x1 = .01 * x2 + np.random.uniform(-.1, .1, size=n)
y = -x1 + x2 + np.random.randn(n)*.01
```



- y este asociat negativ cu x1 (scade cu x1) și crește cu x2

Exemplu: y ∼ x1

```
y = -x1 + x2 + np.random.randn(n)*.01
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.6444	1.208	-0.533	0.595	-3.042	1.753
x1	98.7365	2.062	47.888	0.000	94.645	102.828

- pentru y ∼ x1, x1 are coeficientul 98.74, nu -1 cum ne-am fi așteptat
- efectul lui x2 este capturat în x1

Exemplu: y ∼ x1 + x2

```
y = -x1 + x2 + np.random.randn(n)*.01
```

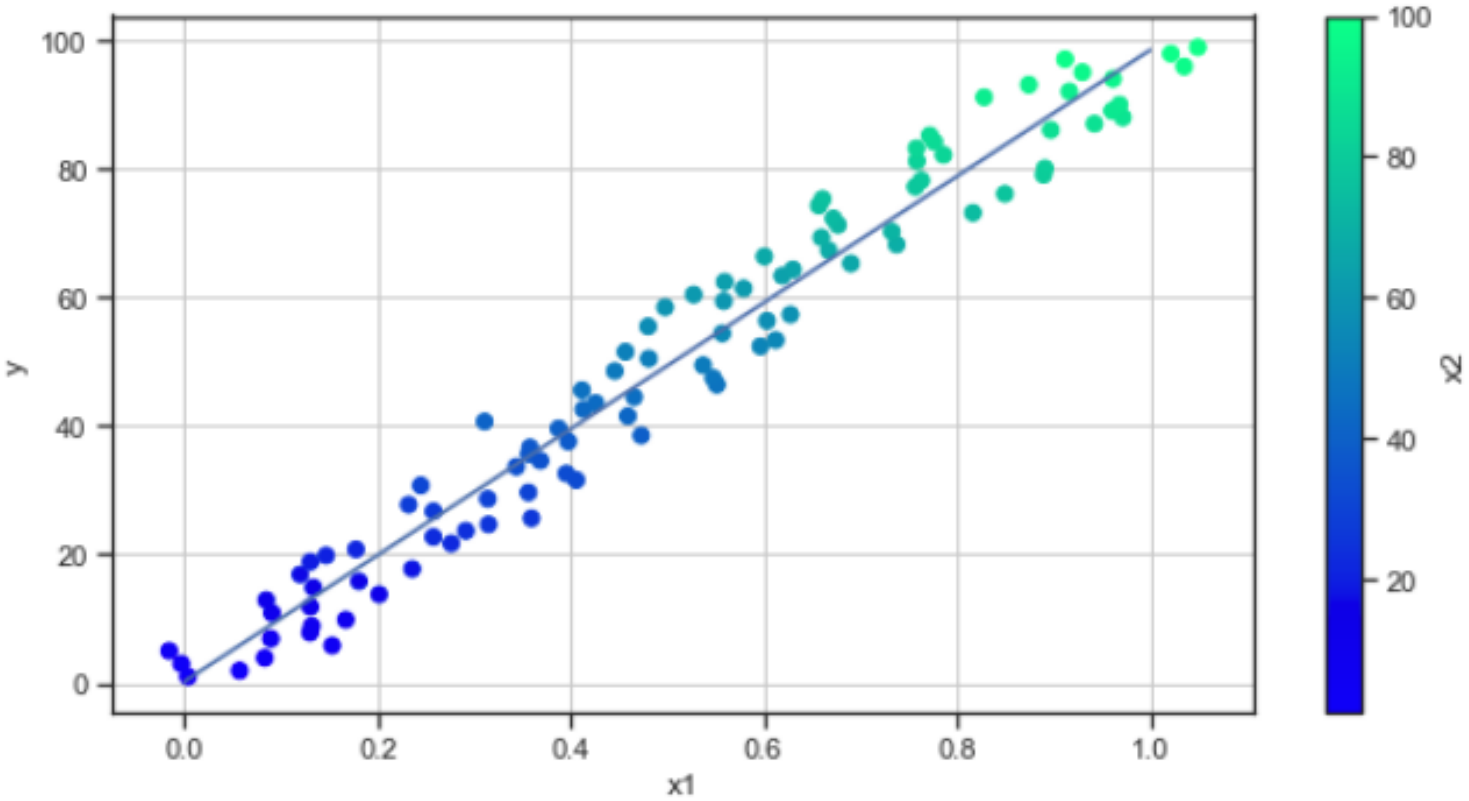
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0032	0.002	-1.593	0.115	-0.007	0.001
x1	-1.0022	0.017	-58.521	0.000	-1.036	-0.968
x2	1.0000	0.000	5944.564	0.000	1.000	1.000

- se obțin corect cei doi coeficienți
- regresia ia x1 și șterge efectul lui x2

Exemplu: y ∼ x1

```
newdf = pd.DataFrame({'y': y, 'x1': x1, 'x2': x2,
    'ey': smf.ols(formula='y ~ x2', data=df).fit().resid,
    'ex1': smf.ols(formula='x1 ~ x2', data=df).fit().resid})
params = smf.ols(formula='y ~ x1', data=df).fit().params
x = np.linspace(0, 1, 10)
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
cm = plt.cm.get_cmap('winter')
sc = ax.scatter(x=x1, y=y, c=x2, cmap=cm, data=newdf)
ax.plot(x, params[0] + x * params[1])
ax.set_xlabel('x1') ; ax.set_ylabel('y')
fig.colorbar(sc, ax=ax).set_label('x2')
ax.grid() ; plt.show()
```

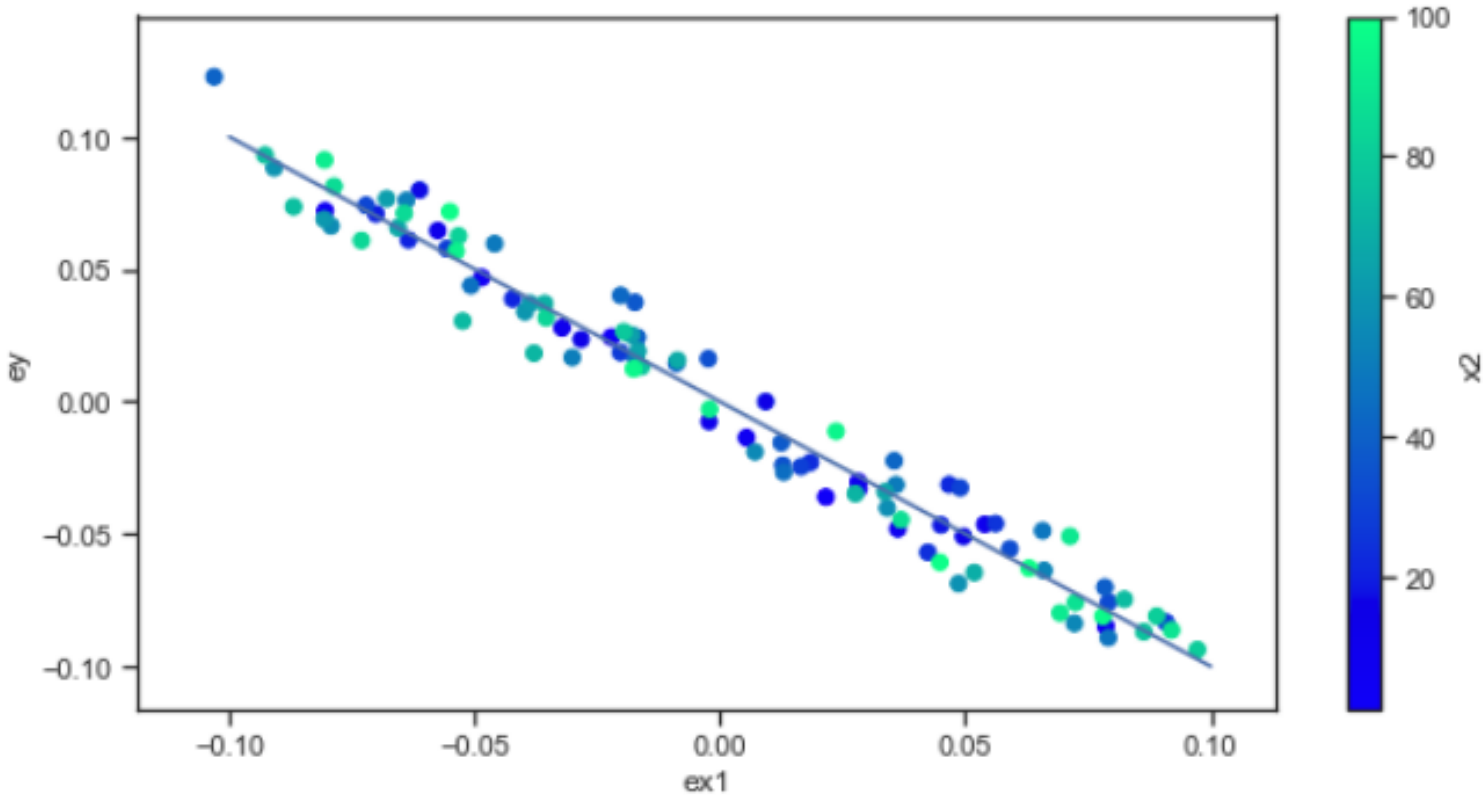
Exemplu: y ∼ x1



- regresia doar în funcție de x1: y crește pe măsură ce x1 crește (coeficientul 98)
- apare acest **confounding effect**[1], se vede cum variabila x1 este influențată de x2

[1] https://en.wikipedia.org/wiki/Confounding

Exemplu: ey ∼ ex1



- panta negativă
- reziduul ex1 nu este influențat de variabila x2
- modelul liniar scoate infuența lui x2 din x1 și din y
- asta nu înseamnă că modul corect este să folosim toate variabilele

Swiss dataset din nou

	Fertility	Agriculture	Examination
Fertility	1.000000	0.353079	-0.645883
Agriculture	0.353079	1.000000	-0.686542
Examination	-0.645883	-0.686542	1.000000
Education	-0.663789	-0.639523	0.698415

- am văzut că semnul coeficientului pentru variabila Agriculture se inversează dacă introducem celelalte variabile (de ex. Examination, Education)
- procentul de bărbați ocupați în Agriculture este corelat negativ cu Education (-0.639), iar Education și Examination (corr = 0.698) par să explice același lucru pentru Fertility
- nu putem susține o asociere directă între Fertility și Agriculture, pentru că dacă scoatem variabile, trendul se inversează

Includerea unei variabile combinație liniară

```
swiss2 = swiss.copy()
swiss2['z'] = swiss2['Agriculture'] + swiss2['Education']
lm = smf.ols(formula='Fertility ~ Agriculture +
  Examination + Education + Catholic +
  InfantMortality + z', data=swiss2).fit()
lm.summary()
```

InfantMortality	1.0770	0.382	2.822	0.007	0.306	1.848
z	-0.3477	0.073	-4.760	0.000	-0.495	-0.200

Omnibus:	0.058	**Durbin-Watson:**	1.454
Prob(Omnibus):	0.971	**Jarque-Bera (JB):**	0.155
Skew:	-0.077	**Prob(JB):**	0.925
Kurtosis:	2.764	**Cond. No.**	2.36e+16

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 8.5e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

1 Proprietățile regresiei multi-variabilă

2 Variabile categoriale (factor)

3 Regresia cu variabile continue și categoriale

Variabilă de tip factor, binară

- dacă pe post de regresori se folosesc variabile factor (categorii), se ajunge la covariance analysis
- considerăm un singur regresor factor, de tip binar, $X_{i1} \in \{0, 1\}$ (în grupul de control $= 0$, tratați $= 1$)

$$Y_i = \beta_0 + X_{i1}\beta_1 + \epsilon_i$$

- pentru cei care au primit tratamentul, media este $E[Y_i] = \beta_0 + \beta_1$
- pentru cei din grupul de control (covariate $= 0$), media lor este $E[Y_i] = \beta_0$
- modelul LS potrivește coeficienții astfel ca $\hat{\beta}_0 + \hat{\beta}_1$ să fie media celor din grupul tratat, iar $\hat{\beta}_0$ media pentru grupul de control
- β_1 este interpretat ca și creșterea sau scăderea răspunsului Y a celor care au fost tratați
- statistica pentru β_1 este exact statistica pentru two-group A/B testing

Variabilă factor cu mai multe valori

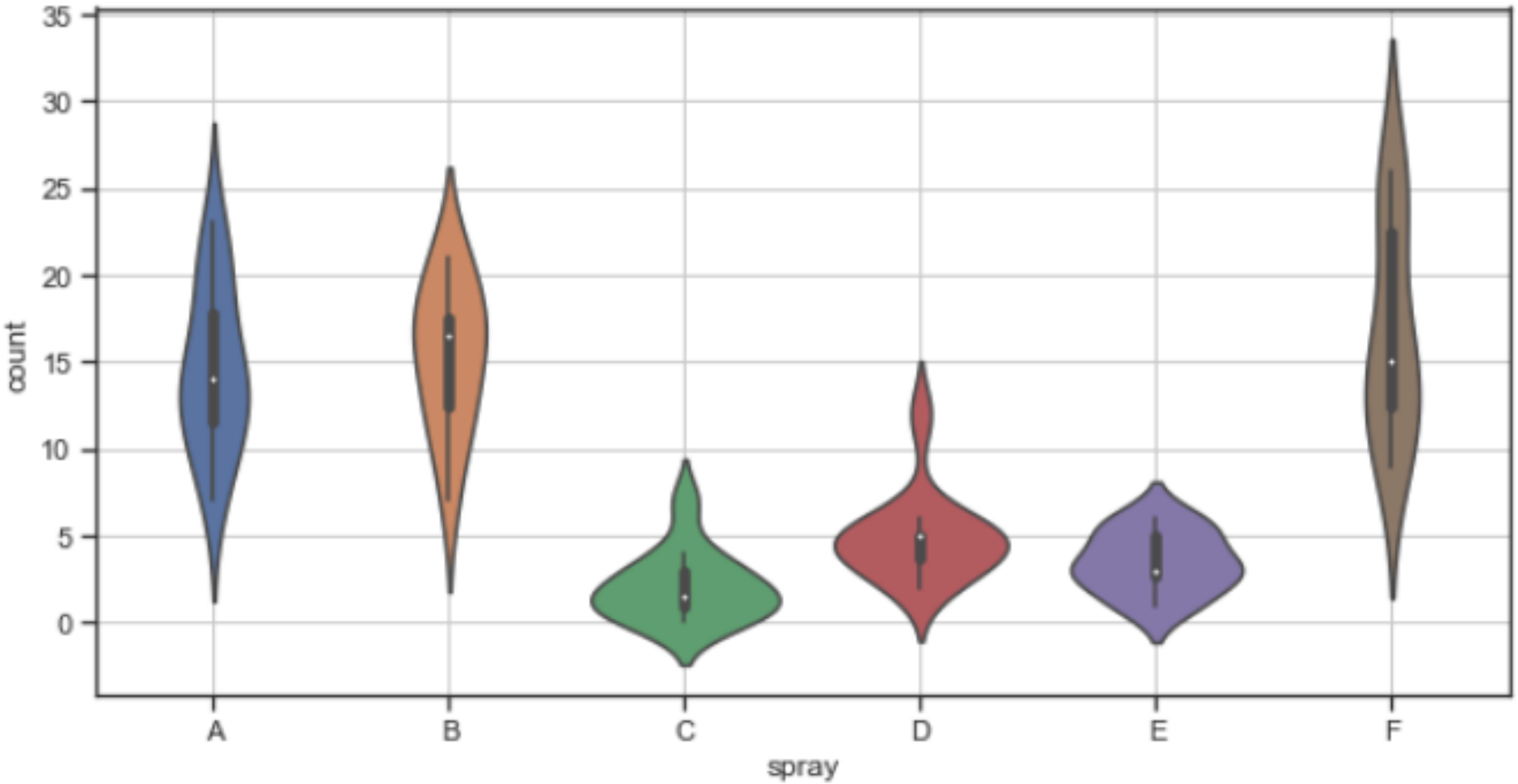
- considerăm o variabilă factor cu mai multe nivele (nivelele nu sunt comparabile)
- exemplu de 3 nivele: afiliere politică: Republican, Democrat, Independent

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i$$

- $X_{i1} = 1$ pentru Republicani, 0 în rest
- $X_{i2} = 1$ pentru Democrați, 0 în rest
- nu este nevoie de o a treia variabilă, modelul ar fi redundant
- dacă i este Republican, $E[Y_i] = \beta_0 + \beta_1$
- dacă i este Democrat, $E[Y_i] = \beta_0 + \beta_2$
- dacă i este Independent, $E[Y_i] = \beta_0$
- β_1 compară Republicanii cu Independ., iar β_2 Democrații cu Independ.
- $\beta_1 - \beta_2$ compară Republicanii cu Democrații
- ce se alege ca nivel de referință are impact asupra explicării modelului

Insectsprays

```
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax = sns.violinplot(x='spray', y='count', data=insectsprays, ax=ax)
ax.grid()
plt.show()
```



Regresia cu variabilă multi-nivel

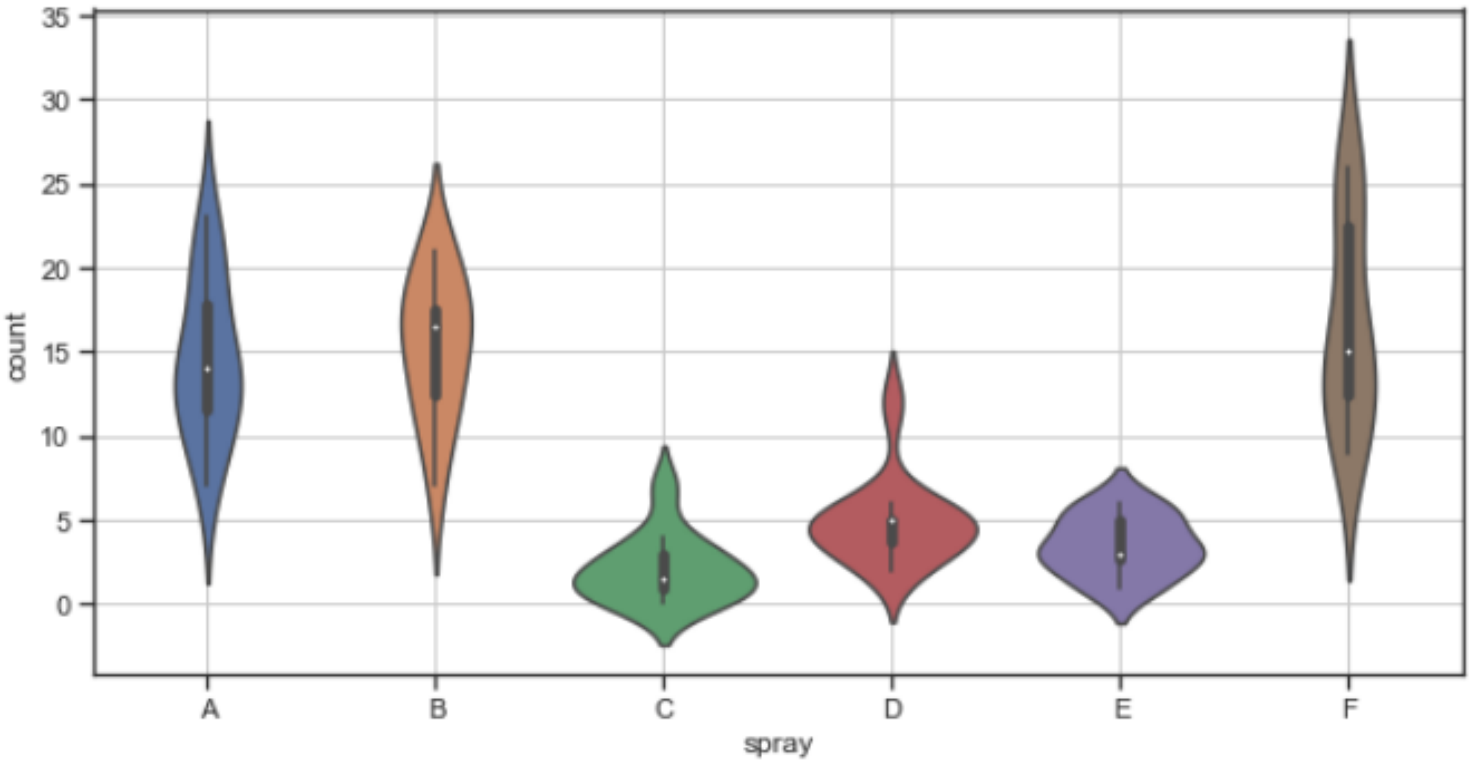
```
lm = smf.ols(formula='count ~ spray', data=insectsprays).fit()
lm.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	14.5000	1.132	12.807	0.000	12.240	16.760
spray[T.B]	0.8333	1.601	0.520	0.604	-2.363	4.030
spray[T.C]	-12.4167	1.601	-7.755	0.000	-15.613	-9.220
spray[T.D]	-9.5833	1.601	-5.985	0.000	-12.780	-6.387
spray[T.E]	-11.0000	1.601	-6.870	0.000	-14.197	-7.803
spray[T.F]	2.1667	1.601	1.353	0.181	-1.030	5.363

- totul se compară cu spray A: $\bar{X}_A = 14.5$, $\hat{\beta}_B = 0.833$ e schimbarea în medie față de spray A

Regresia pe variabile factor

	coef
Intercept	14.5000
spray[T.B]	0.8333
spray[T.C]	-12.4167
spray[T.D]	-9.5833
spray[T.E]	-11.0000
spray[T.F]	2.1667



- de observat diferențele în medii față de spray A

Realizarea manuală a modelului (alegem A ca referință)

```
is2 = insectsprays.copy()
for x in ['A', 'B', 'C', 'D', 'E', 'F']:
    is2[x] = 1*(is2.spray == x) # is2[x] = is2[x].astype(float)
print(is2.head())

lm = smf.ols(formula='count ~ B + C + D + E + F', data=is2).fit()
print(lm.summary())

# vezi condition number cand introducem A (redundant)
```

	Unnamed: 0	count	spray	A	B	C	D	E	F
0	1	10	A	1	0	0	0	0	0
1	2	7	A	1	0	0	0	0	0
2	3	20	A	1	0	0	0	0	0
3	4	14	A	1	0	0	0	0	0
4	5	14	A	1	0	0	0	0	0

Excluderea intercept-ului

```
lm = smf.ols(formula='count ~ spray - 1', data=insectsprays).fit()
lm.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
spray[A]	14.5000	1.132	12.807	0.000	12.240	16.760
spray[B]	15.3333	1.132	13.543	0.000	13.073	17.594
spray[C]	2.0833	1.132	1.840	0.070	-0.177	4.344
spray[D]	4.9167	1.132	4.343	0.000	2.656	7.177
spray[E]	3.5000	1.132	3.091	0.003	1.240	5.760
spray[F]	16.6667	1.132	14.721	0.000	14.406	18.927

- nu mai există o referință, referința e media 0 acum
- de regulă dorim o referință, pentru că testele t vor măsura confidența cu care spray-urile sunt diferite față de referință
- aici, p-Values măsoară dacă mediile sunt diferite de zero (a omorât ceva?)

Sumar

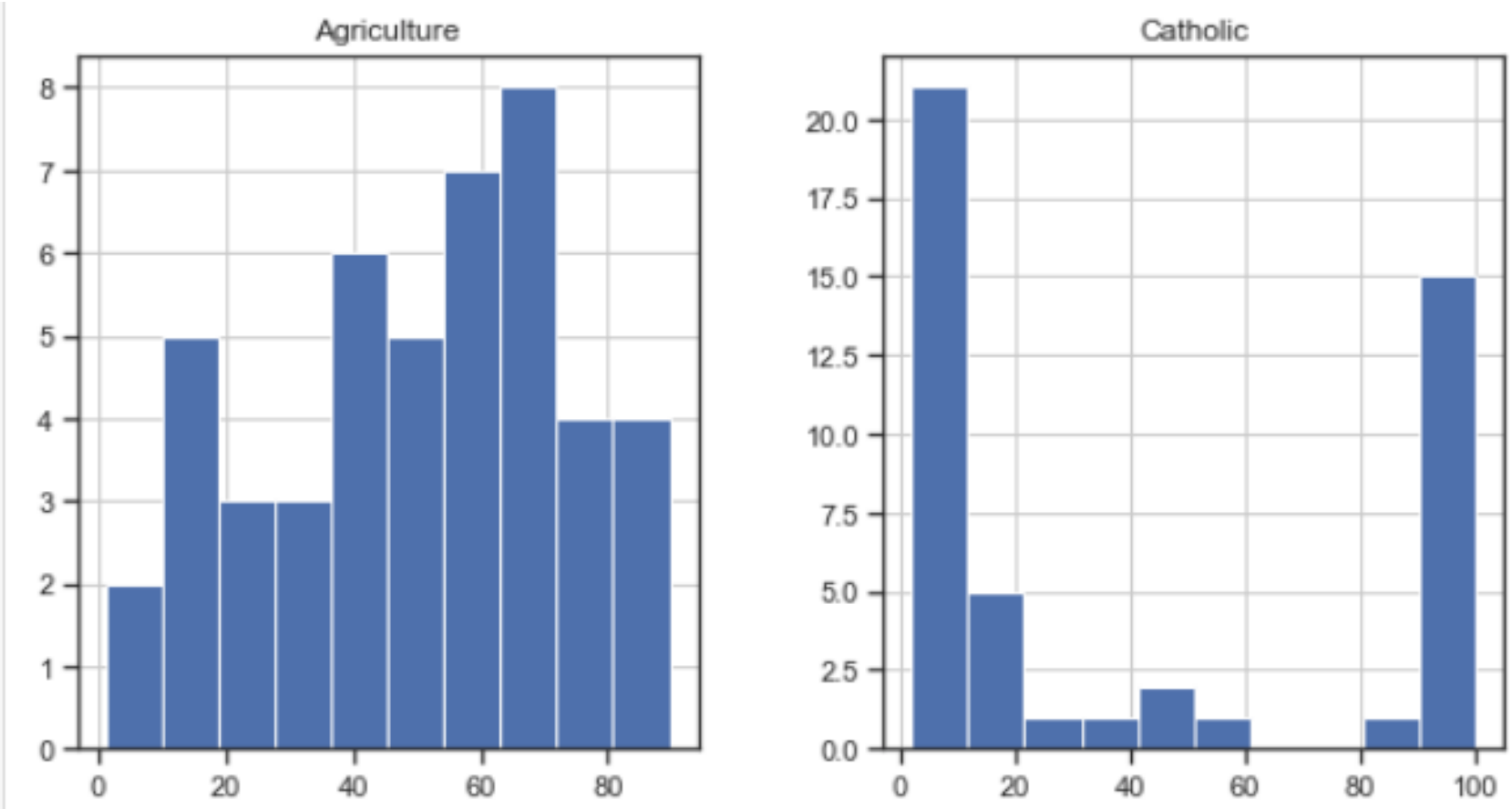
- dacă folosim spray ca factor, pachetul statmodels folosește primul (alfabetic) factor ca referință
- toate testele se referă la comparația cu spray A
- sample mean pentru A este intercept-ul (acel β_0)
- celelalte medii sunt egale cu intercept + coeficientul lor
- dacă se omite intercept-ul, testele sunt pentru media $\neq 0$
- probleme cu modelul:
  - modelarea se face mai degrabă cu distribuții Poisson
  - variația față de medie nu este constantă precum asumă regresia (heteroscedasticity)

1 Proprietățile regresiei multi-variabilă

2 Variabile categoriale (factor)

3 Regresia cu variabile continue și categoriale

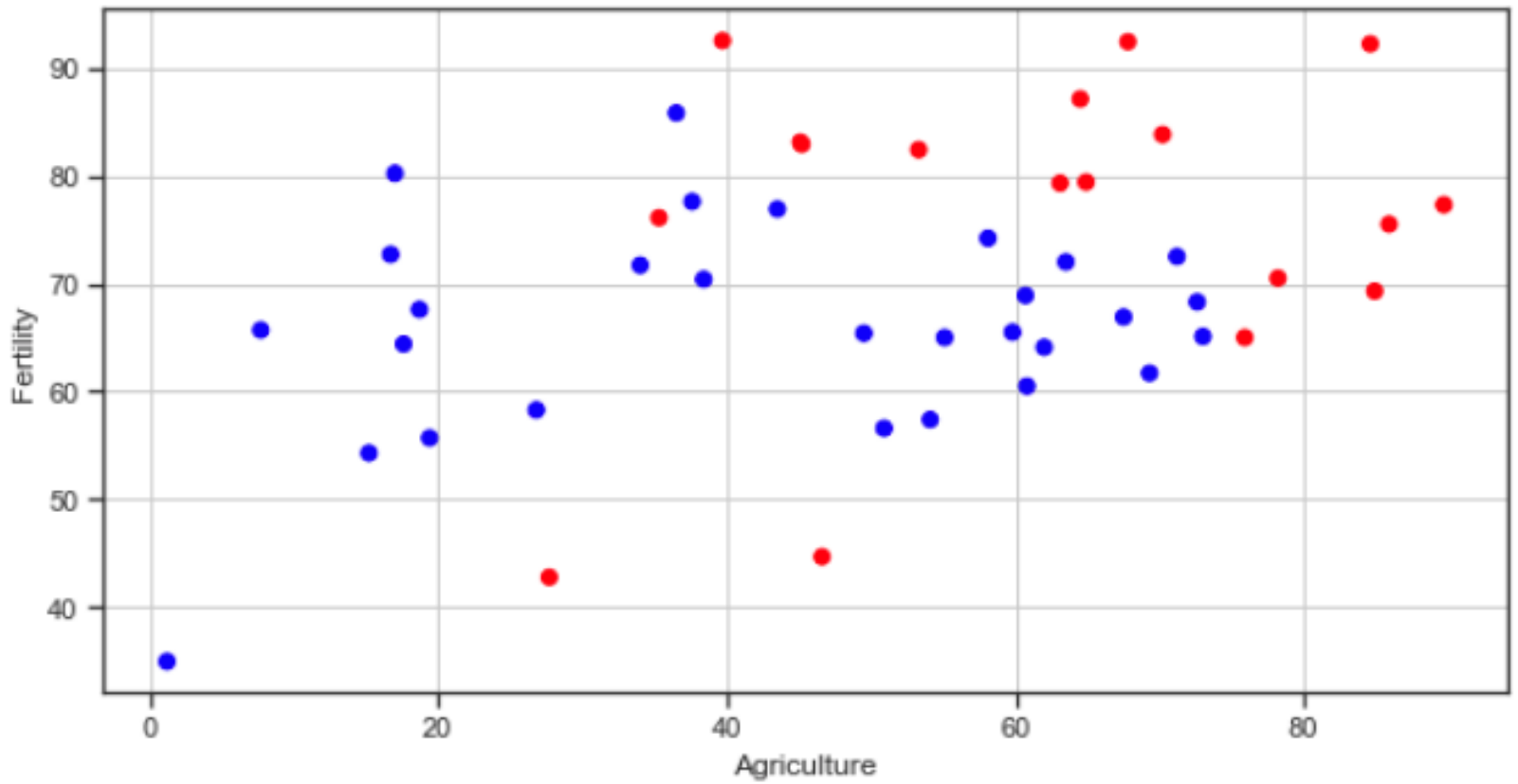
Distribuție (foarte) bimodală ca și categorie



- majoritatea regiunilor au fie un procent $> 80\%$ catolici, fie $< 20\%$

Catholics (red) vs. NonCatholic (blue)

```
swiss3 = swiss.copy() ; swiss3['CatholicBin'] = 1 * (swiss.Catholic > 50)
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
plt.scatter(swiss3.Agriculture, swiss3.Fertility,
            c=['Red' if x else 'Blue' for x in swiss3.CatholicBin.values])
ax.set_xlabel('Agriculture') ; ax.set_ylabel('Fertility')
plt.grid() ; plt.show()
```



Regresia Fertility $\sim$ Agriculture, C(CatholicBin)

- avem următoarele variabile (răspuns și regresori):

  $Y = Fertility$

  $X_1 = Agriculture$

  $X_2 = 1$ **dacă** $> 50\%$ Catholic, **altfel** 0

- putem aborda mai multe modele:
  1. o linie care nu ține seama de religia provinciei:

     $$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1$$

  2. fără interacțiune între variabile, obținem două regresii cu aceeași pantă:

     $$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

     $$\begin{aligned} &\text{dacă } X_2 = 0 \\ &= \beta_0 + \beta_1 X_1 \\ &\text{dacă } X_2 = 1 \\ &= \beta_0 + \beta_2 + \beta_1 X_1 \end{aligned}$$

Regresia Fertility $\sim$ Agriculture, C(CatholicBin) (2)

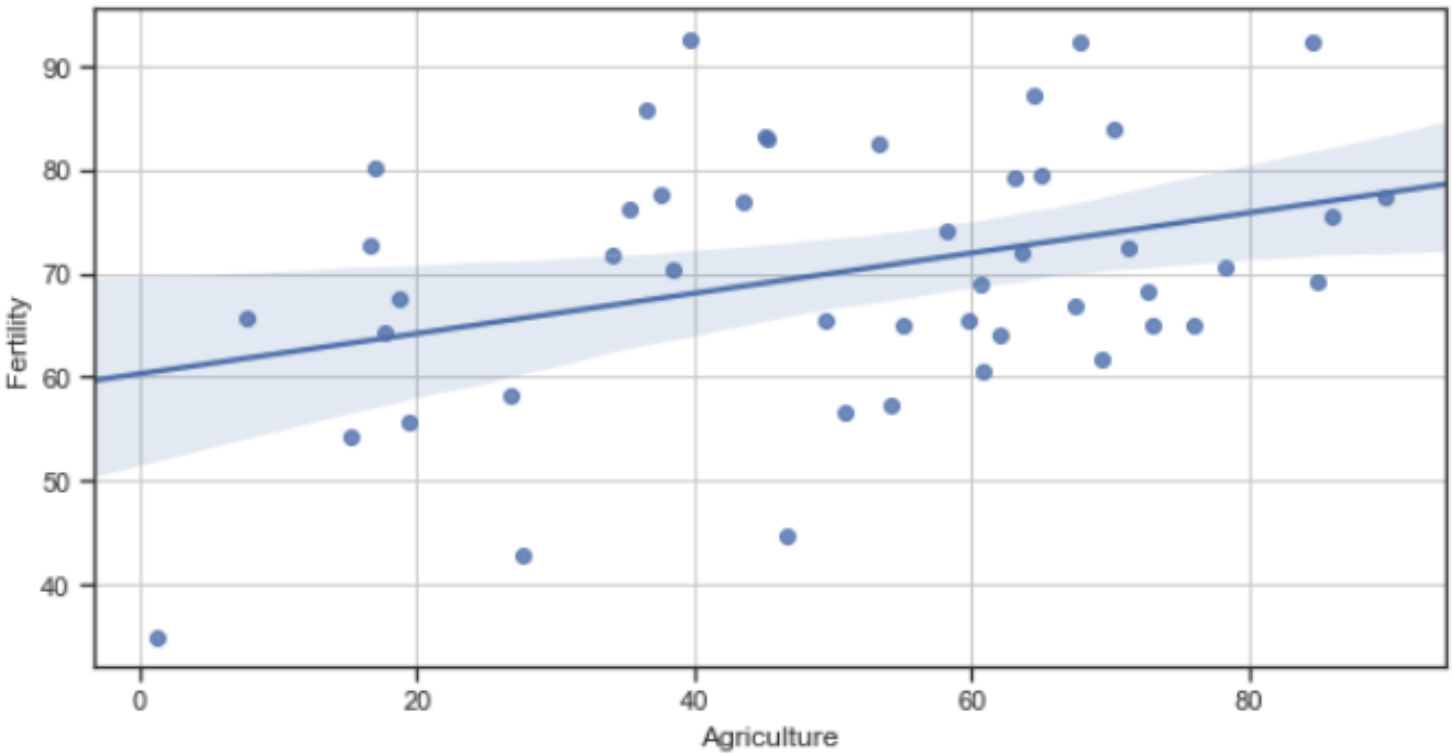
4. cu interacțiune între regresori, obținem două linii de regresie, intercepts și slopes diferite:

$$
\begin{aligned}
E[Y|X_1, X_2] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \\
&\text{dacă } X_2 = 0 \\
&= \beta_0 + \beta_1 X_1 \\
&\text{dacă } X_2 = 1 \\
&= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1
\end{aligned}
$$

- coeficientul β_2 asociat Catholics va fi schimbarea de intercept de la Protestant la Catholic
- coeficientul β_3 va fi asociat cu schimbarea de pantă la trecerea de la Protestant la Catholic

Fertility ∼ Agriculture

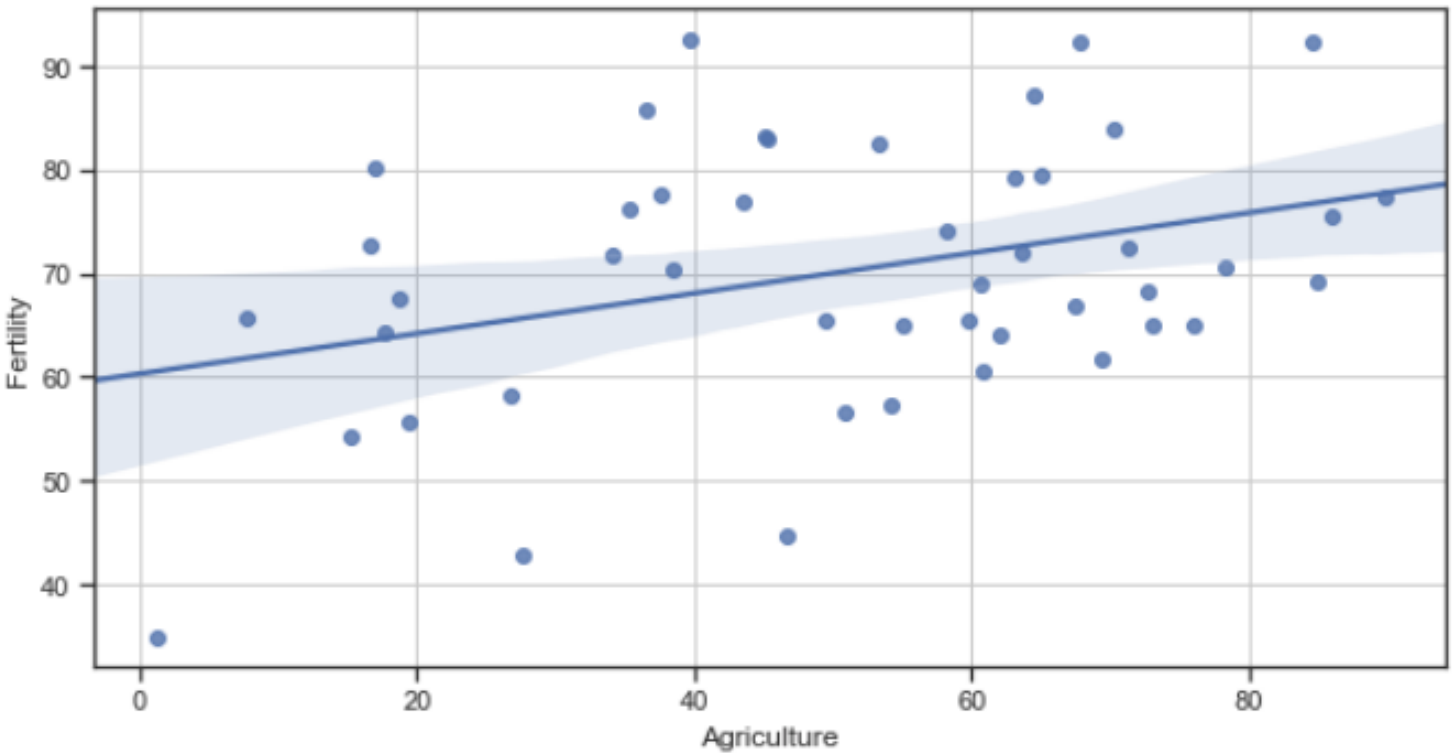
```
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
sns.regplot(y='Fertility', x='Agriculture', data=swiss3, ax=ax)
plt.grid() ; plt.show()
```



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	60.3044	4.251	14.185	0.000	51.742	68.867
Agriculture	0.1942	0.077	2.532	0.015	0.040	0.349

Fertility $\sim$ Agriculture

```
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
sns.regplot(y='Fertility', x='Agriculture', data=swiss3, ax=ax)
plt.grid() ; plt.show()
```



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	60.3044	4.251	14.185	0.000	51.742	68.867
Agriculture	0.1942	0.077	2.532	0.015	0.040	0.349

Fertility ∼ Agriculture + C(CatholicBin)

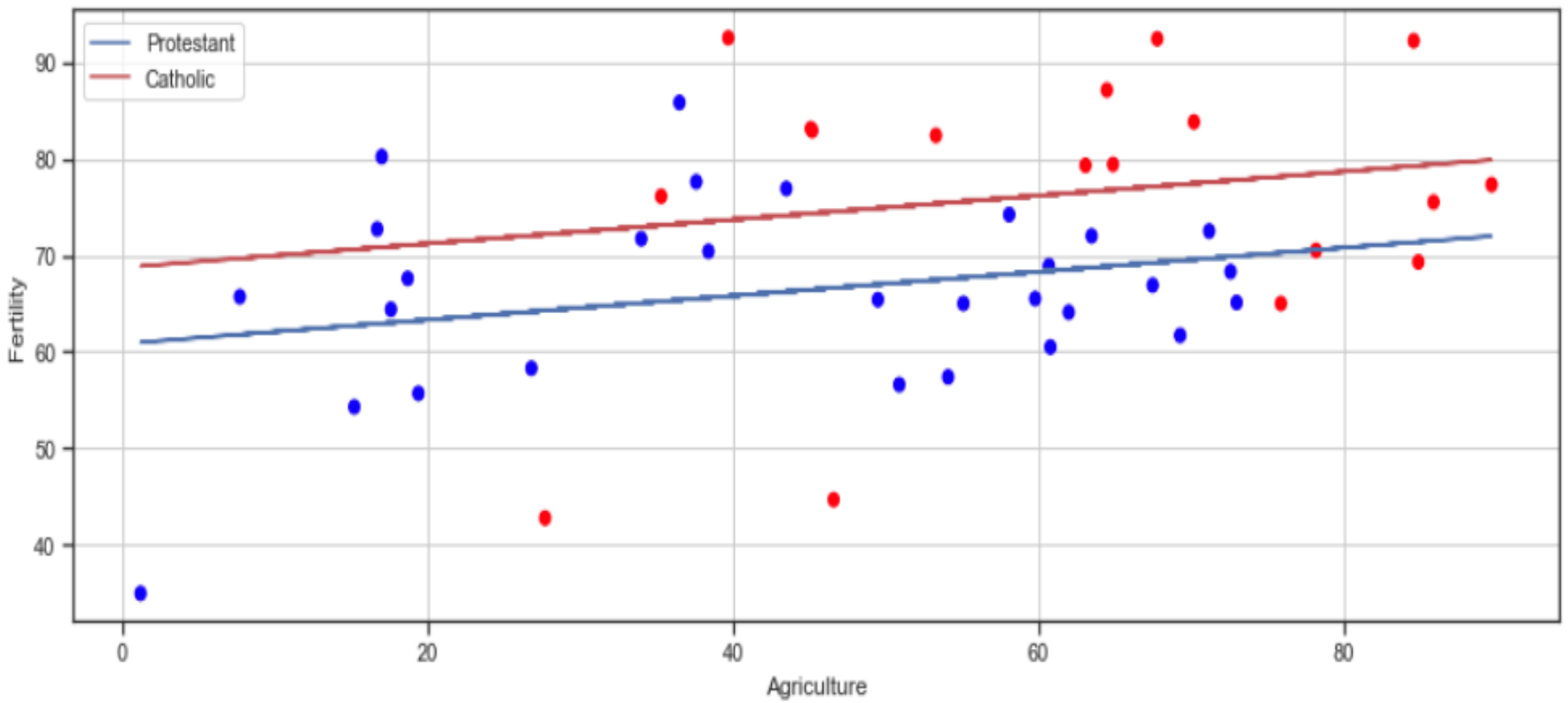
```
lm = smf.ols(formula='Fertility ~ Agriculture + C(CatholicBin)', data=swiss3).fit()
lm.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	60.8322	4.106	14.816	0.000	52.557	69.107
C(CatholicBin)[T.1]	7.8843	3.748	2.103	0.041	0.330	15.439
Agriculture	0.1242	0.081	1.531	0.133	-0.039	0.288

- intercept-ul este 60.83
- panta este aceeași indiferent de religie
- schimbarea de intercept este 7.88 pentru Catholics

Fertility ∼ Agriculture + C(CatholicBin) (2)

```
params = smf.ols(formula='Fertility ~ Agriculture + C(CatholicBin)', data=swiss3).fit().params
x = swiss3.Agriculture.values ; fig, ax = plt.subplots(1, 1, figsize=(15, 5))
ax.scatter(swiss3.Agriculture, swiss3.Fertility, c=['Red' if x else 'Blue' for x in swiss3.CatholicBin.values])
ax.plot(x, params[0] + params[2] * x, 'b')
ax.plot(x, params[0] + params[1] + params[2] * x, 'r')
ax.set_xlabel('Agriculture') ; ax.set_ylabel('Fertility')
ax.legend(['Protestant', 'Catholic']) ; plt.grid() ; plt.show()
```



Fertility ~ Agriculture * C(CatholicBin)

```
lm = smf.ols(formula='Fertility ~ Agriculture * C(CatholicBin)', data=swiss3).fit()
print(lm.summary())
```

- steluța * din formulă determină statmodels să considere regresorul de tip interacțiune dintre cele două variabile și introduce și efectele liniare ale celor două

```
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                     62.0499      4.789     12.956      0.000      52.392      71.708
C(CatholicBin)[T.1]            2.8577     10.626      0.269      0.789     -18.573      24.288
Agriculture                    0.0961      0.099      0.973      0.336      -0.103       0.295
Agriculture:C(CatholicBin)[T.1]  0.0891      0.176      0.506      0.615      -0.266       0.444
```

Fertility ∼ Agriculture * C(CatholicBin) (2)

```
x = swiss3.Agriculture.values ; fig, ax = plt.subplots(1, 1, figsize=(15, 5))
ax.scatter(swiss3.Agriculture, swiss3.Fertility, c=['Red' if x else 'Blue' for x in swiss3.CatholicBin.values])
ax.plot(x, lm.params[0] + lm.params[2] * x, 'b')
ax.plot(x, lm.params[0] + lm.params[1] + (lm.params[2] + lm.params[3]) * x, 'r')
ax.set_xlabel('Agriculture') ; ax.set_ylabel('Fertility')
ax.legend(['Protestant', 'Catholic']) ; plt.grid() ; plt.show()
```