Assignment 3 answer

June 17, 2021

```
In [2]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
```

1 Exercise 6.8

(a) Describe a nonexperimental research situation—real or contrived—in which failure to control statistically for an omitted variable induces a correlation between the error and an explanatory variable, producing erroneous conclusions. (For example: An educational researcher discovers that university students who study more get lower grades on average; the researcher concludes that studying has an adverse effect on students' grades.)

If X_1 causes both X_2 and Y, then regressing $Y = A + BX_2 + E$ will induce a correlation between X_2 and E. The point is that variation in X_1 is part of the variation in E.

In the example given in the problem, X_1 could be "finding a subject more difficult" and X_2 is "studying more" and Y is "grades on tests". Finding a subject more difficult causes you to both study more and to get worse grades.

(b) Describe an experiment—real or contrived—in which faulty experimental practice induces an explanatory variable to become correlated with the error, compromising the validity of the results produced by the experiment. (For example: In an experimental study of a promising new therapy for depression, doctors administering the treatments tend to use the new therapy with patients for whom more traditional approaches have failed; it is discovered that subjects receiving the new treatment tend to do worse, on average, than those receiving older treatments or a placebo; the researcher concludes that the new treatment is not effective.)

The question gives a good example.

(c) Is it fair to conclude that a researcher is never able absolutely to rule out the possibility that an explanatory variable of interest is correlated with the error? Is experimental research no better than observational research in this respect? Explain your answer.

I would say yes: You can never know all of the explanatory variables. Experiment is better in this respect because, in an experiment you can at least control for the variablees you know about, which is something you can't do with observational data.

2 Exercise D5.2

Analyze Sahlinsís data, given in sahlins.txt, by regressing Acres/Gardener on Consumers/Gardener. In a society characterized by ëprimitive communism,íthe social product of the village would be redistributed according to need, while each household would work in proportion to its capacity, implying a regression slope of zero. In contrast, in a society in which redistribution is purely through the market, each household should have to work in proportion to its consumption needs, suggesting a positive regression slope and an intercept of zero. Interpret the results of the regression in light of these observations. Examine and interpret the values of A;B;SE; and r (or r2). Do the results change if the fourth household is deleted? Plot the regression lines calculated with and without the fourth household on a scatterplot of the data. Does either regression do a good job of summarizing the relationship between Acres/Gardener and Consumers/Gardener? (Cf., Exercise D2.1.)

```
In [41]: df = pd.read_csv('https://socialsciences.mcmaster.ca/jfox/Books/Applied-Regression-3E/
                                sep = '\s', names = ['','consumers','acres'], index_col = 0, ski

         df
```

Out[41]:

	consumers	acres
1	1.00	1.71
2	1.08	1.52
3	1.15	1.29
4	1.15	3.09
5	1.20	2.21
6	1.30	2.26
7	1.37	2.40
8	1.37	2.10
9	1.43	1.96
10	1.46	2.09
11	1.52	2.02
12	1.57	1.31
13	1.65	2.17
14	1.65	2.28
15	1.65	2.41
16	1.66	2.23
17	1.87	3.04
18	2.03	2.06
19	2.05	2.73
20	2.30	2.36

** Examine and interpret the values of A;B;SE; and r (or r2) **

```
In [77]: # "Examine"

         def examine(df):
             n = len(df)
             y = df['acres']
```

```
    x = df['consumers']
    k = 1

    X = np.concatenate((
        np.ones(shape=x.shape)[:,None],
        x.values[:,None]
    ), axis=1)

    B = np.linalg.inv(X.T @ X) @ (X.T @ y)
    A = B[0]
    B = B[1]

    yhat = A + B*x
    E = y-yhat

    SSE = (E**2).sum()
    RegSS = ((yhat - y.mean())**2).sum()
    TSS = SSE + RegSS

    # SE is estimate of sigma_eps, standard error of yhat
    SE = np.sqrt( SSE / (n - k - 1) )

    R2 = RegSS/TSS
    corr = np.sqrt(R2)

    print (f"n = {n}")
    print (f"A = {A:.2f}  B = {B:.2f}")
    print (f"SE = {SE:.4f}")
    print (f"R2 = {R2:.4f}")
    print (f"corr = {corr:.4f}")
```

```
In [78]: examine(df)
```

```
n = 20
A = 1.38  B = 0.52
SE = 0.4543
R2 = 0.1411
corr = 0.3757
```

Your interpretation of the above regression analysis should be expressed in the units of the problem (consumers, acres), and explain how the linear model and the quality of its fit relate to the hypothesis under consideration (Is this community, in practice – by the data – one that is "primitive communist" or subject to market forces?)

The problem says that a purely communist village would have B=0, meaning that the amount of work or output (acres) done by a household would be independent of the number of people (consumers) in the household. If this were the case, the intercept, A, would need to be non-zero if, in fact, any output were created at all.

In a market economy a household with more members (consumers) would have to work more, with the amount of work generally proportional to the number of consumers. In other words, B would be positive. There would also be no obligation to produce more than your needs, so A would be zero.

The data seems to tell neither story exactly. Both A and B are positive, indicating that this village is a mix of communist and market economies. R2 is low, so perhaps we can say it is mostly communist.

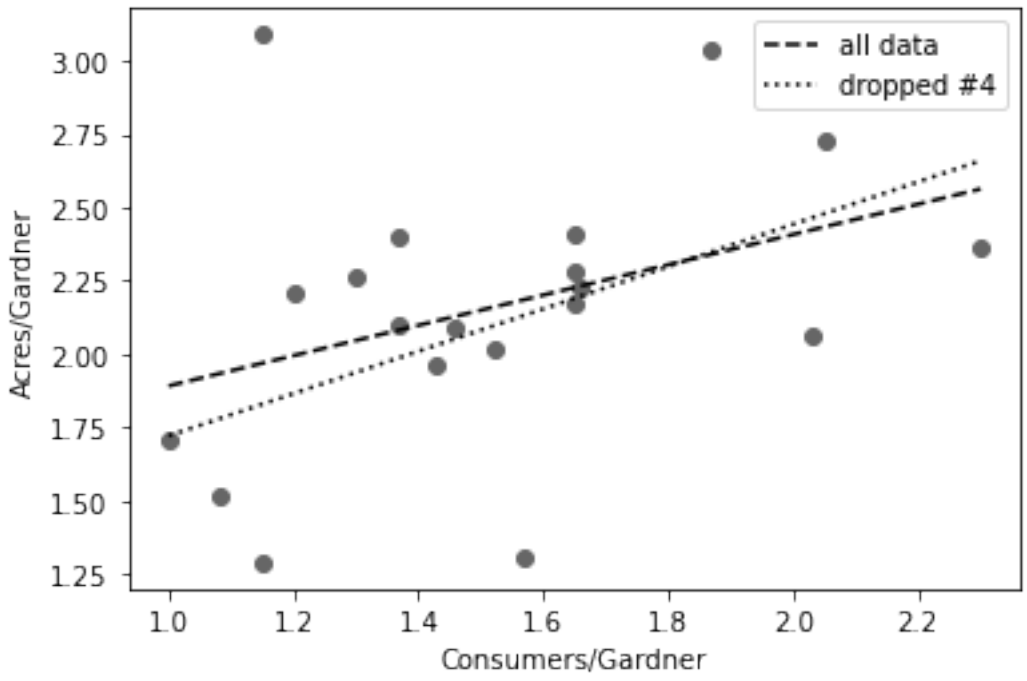
```
In [79]: examine(df.drop(4))
```

```
n = 19
A = 1.00  B = 0.72
SE = 0.3681
R2 = 0.3264
corr = 0.5713
```

The linear relationship is much stronger without household #4, indicating a more market-like economy for most households.

```
In [53]: plt.plot(df['consumers'], df['acres'], 'o', color="#666666");
         xx = np.array([1, 2.3])
         plt.plot(xx, A + B*xx, '--k')
         plt.plot(xx, A1 + B1*xx, ':k')
         plt.legend(['__nolegend__', 'all data', 'dropped #4'])
         plt.xlabel('Consumers/Gardner')
         plt.ylabel('Acres/Gardner')
         plt.show()
```

```
/Users/dsweet2/env_python3/lib/python3.6/site-packages/ipykernel_launcher.py:5: UserWarning: Tl
  """
```



The regression lines are similar and tell the same qualitative story: The economy is mixed.

2.1 Exercise D6.2

In Exercise D5.2 you calculated a simple regression of acres/gardener on con- sumers/gardener for the 20 households of Mazulu village. Find the standard errors of the least-squares intercept and slope. Can we conclude that the population slope is greater than zero? Can we conclude that the intercept is greater than zero? Repeat these computations omitting the fourth household.

```
In [74]: def examine_6_2(df):
             n = len(df)
             y = df['acres']
             x = df['consumers']
             k = 1

             X = np.concatenate((
                 np.ones(shape=x.shape)[:,None],
                 x.values[:,None]
             ), axis=1)

             B = np.linalg.inv(X.T @ X) @ (X.T @ y)
             A = B[0]
             B = B[1]

             yhat = A + B*x
             E = y-yhat
```

```
    SSE = (E**2).sum()
    RegSS = ((yhat - y.mean())**2).sum()
    TSS = SSE + RegSS

    # SE is estimate of sigma_eps, standard error of yhat
    SE = np.sqrt( SSE / (n - k - 1) )
    SSX = ((X[:,1]-X[:,1].mean())**2).sum()
    SE_B = np.sqrt(SE**2 / SSX)
    SE_A = SE_B * np.sqrt((X[:,1]**2).sum())
    t_A = A / SE_A
    t_B = B / SE_B

    R2 = RegSS/TSS
    corr = np.sqrt(R2)

    print (f"n = {n}")
    print (f"A = {A:.2f}  B = {B:.2f}")
    print (f"SE = {SE:.4f}")
    print (f"SE_A = {SE_A:.4f}")
    print (f"SE_B = {SE_B:.4f}")
    print (f"t_A = {t_A:.4f}")
    print (f"t_B = {t_B:.4f}")
    print (f"R2 = {R2:.4f}")
    print (f"corr = {corr:.4f}")
```

```
In [75]: examine_6_2(df)
```

```
n = 20
A = 1.38  B = 0.52
SE = 0.4543
SE_A = 2.0948
SE_B = 0.3002
t_A = 0.6567
t_B = 1.7197
R2 = 0.1411
corr = 0.3757
```

The t values suggest that the intercept (A) is not significantly different from 0 at the p=.05 level and that the slope (B) is (just barely).

```
In [76]: examine_6_2(df.drop(4))
```

```
n = 19
A = 1.00  B = 0.72
SE = 0.3681
SE_A = 1.7302
SE_B = 0.2514
```