The Logic of Multiple Regression

based on Chapter 8 of Judd & McClelland (1989)

DATA

The data are the Industrial Production (IP) index and the number of Unemployed (UN) workers (in millions) for the 10 years from 1950 to 1959. For convenience, the years have been relabeled from 1 to 10. Here are the data:

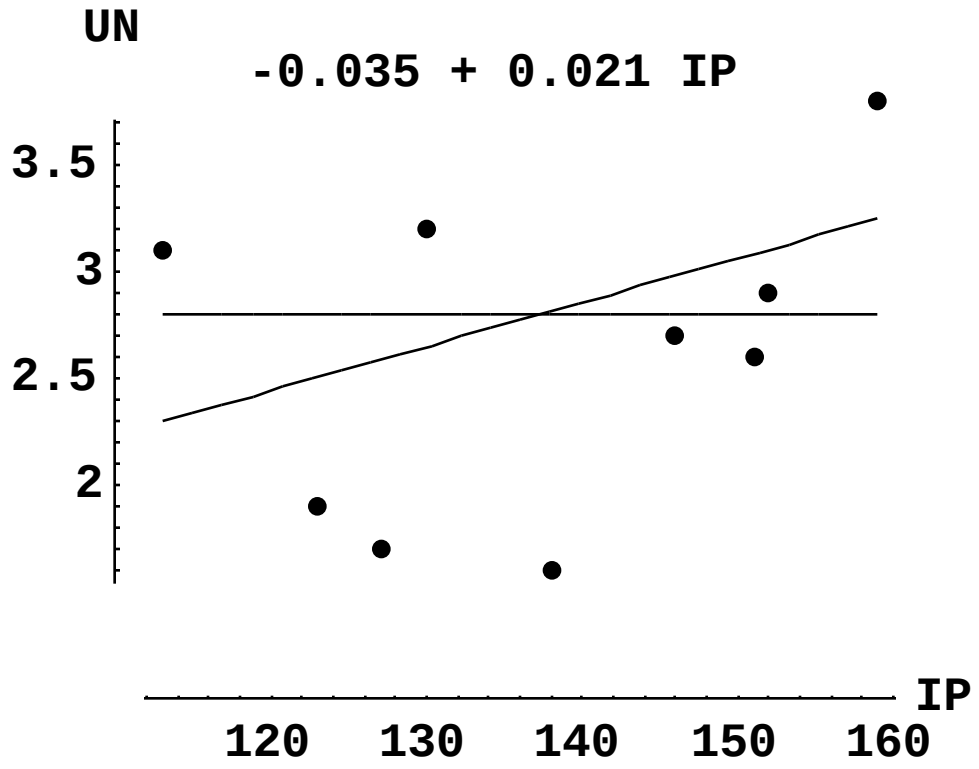
YR	IP	UN
1	113	3.1
2	123	1.9
3	127	1.7
4	138	1.6
5	130	3.2
6	146	2.7
7	151	2.6
8	152	2.9
9	141	4.7
10	159	3.8

Predicting UN with IP

We will first try to predict UN with IP because it seems reasonable that with higher Industrial Production there would be more jobs and hence less Unemployment.

A: UN^ = -0.035 + 0.021 IP

C: UN^ = 2.8



Doesn't look too good! Let's look at the predictions and the errors. We will use the special notation UN.0.IP to represent the errors in order to indicate that these are the errors from a model predicting UN using a constant (X0) and one predictor (IP).

YR	IP	UN	UN^	UN.0,IP	UN.0,IP^2
1	113	3.1	2.3	0.8	0.64
2	123	1.9	2.5	-0.61	0.37
3	127	1.7	2.6	-0.89	0.8
4	138	1.6	2.8	-1.2	1.5
5	130	3.2	2.7	0.55	0.3
6	146	2.7	3.	-0.29	0.082
7	151	2.6	3.1	-0.49	0.24
8	152	2.9	3.1	-0.21	0.044
9	141	4.7	2.9	1.8	3.3
10	159	3.8	3.3	0.55	0.3

SSE(C) = 8.38, SSE(A) = 7.56, SSR = 0.82

PRE = 0.098, F*(1,8) = 0.87, p = 0.38

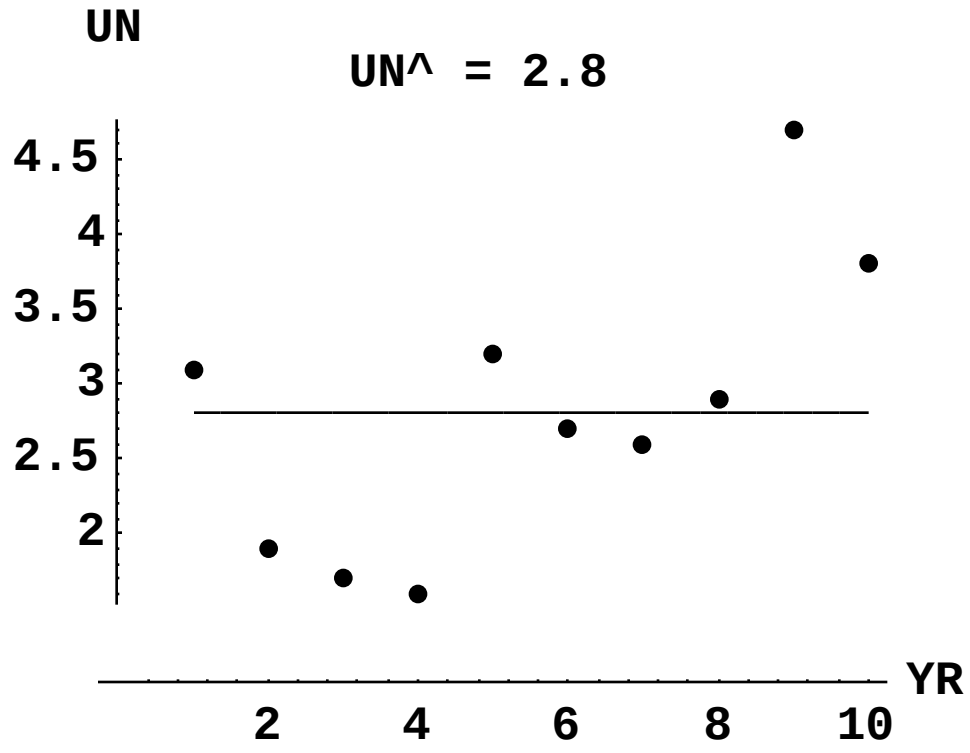
That is, using IP reduces error in predicting UN (over a simple model) by only about 10%, which is not surprising. So, do NOT reject MODEL C! IP, by itself, is NOT a useful predictor of UN.

Predicting UN with YR

Let's try predicting UN with YR. Perhaps there are consistent yearly changes in unemployment. But we wil do this step-by-step, slowing taking UN apart into its component pieces. We will start with modeling UN with its mean.

A: UN^ = 2.8

C: UN^ = 0

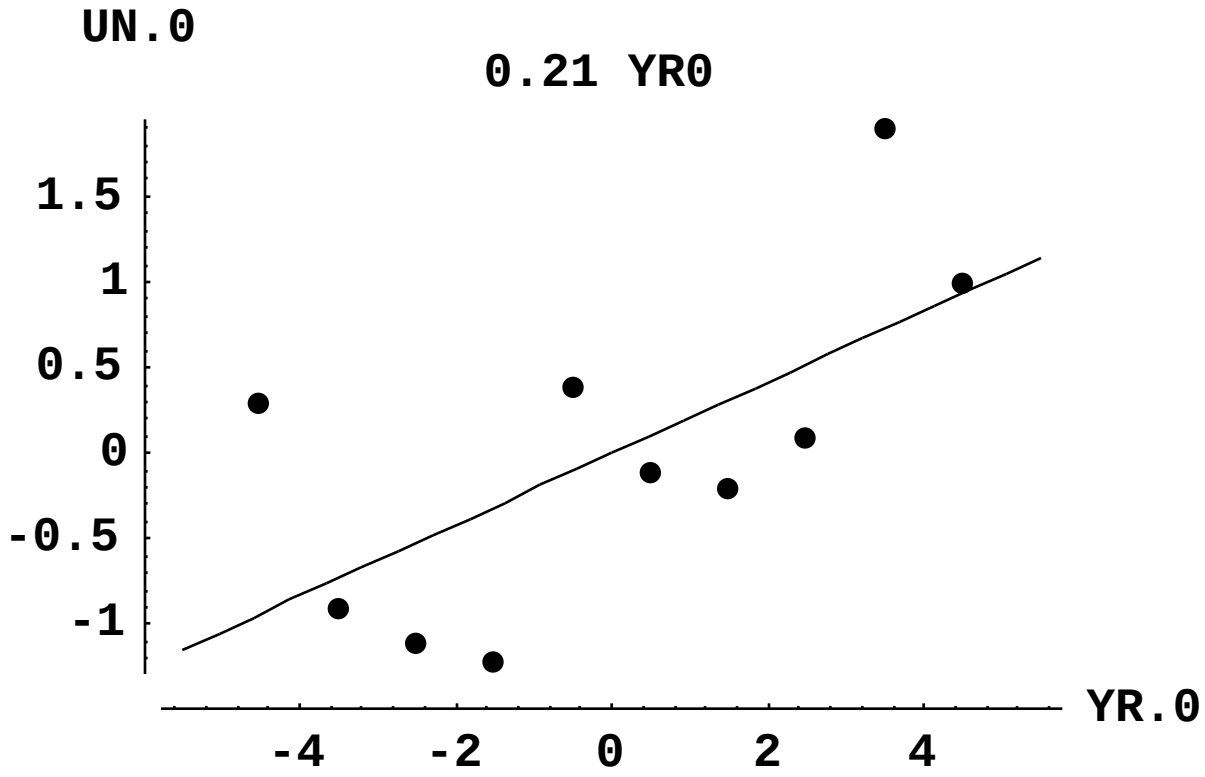


YR	UN	UN^	UN.0	YR.0
1	3.1	2.82	0.28	-4.5
2	1.9	2.82	-0.92	-3.5
3	1.7	2.82	-1.12	-2.5
4	1.6	2.82	-1.22	-1.5
5	3.2	2.82	0.38	-0.5
6	2.7	2.82	-0.12	0.5
7	2.6	2.82	-0.22	1.5
8	2.9	2.82	0.08	2.5
9	4.7	2.82	1.88	3.5
10	3.8	2.82	0.98	4.5

UN.0 is interpreted as the amount by which unemployment was unexpectedly (relative to the mean) high or low in a given year. For example, the value of UN.0 = 0.28 for YR = 1 means unemployment was unexpectedly high by .28 million = 280,000 workers in the first year of these data. YR.0 has a similar interpretation, even though it may seem at first. The value of YR.0 = -4.5 means that YR is "unexpectedly" low (relative to the mean for YR) in the first year. The question now becomes whether knowing whether YR.0 is unexpectedly high or low (i.e., early or late) can predict when UN.0 is unexpectedly high or low.

A: UN.0^ = 0.21 YR0

C: UN.0^ = 0



YR	UN	UN.0	UN.0^	UN.0,YR
1	3.1	0.28	-0.94	1.2
2	1.9	-0.92	-0.73	-0.19
3	1.7	-1.12	-0.52	-0.6
4	1.6	-1.22	-0.31	-0.91
5	3.2	0.38	-0.1	0.48
6	2.7	-0.12	0.1	-0.22
7	2.6	-0.22	0.31	-0.53
8	2.9	0.08	0.52	-0.44
9	4.7	1.88	0.73	1.2
10	3.8	0.98	0.94	0.042

What do the values of UN.0,YR mean? UN.0,YR=1.2 for YR 1 means that relative to our model of UN based on yearly changes, there were 1.2 million more unemployed workers than expected. UN.0,YR = -.19 for YR 2 means that relative to the model there were 190,000 fewer workers unemployed than expected. Etc.

SSE(C) = 8.38, SSE(A) = 4.79, SSR = 3.6

PRE = 0.43, F*(1,8) = 5.99, p = 0.04

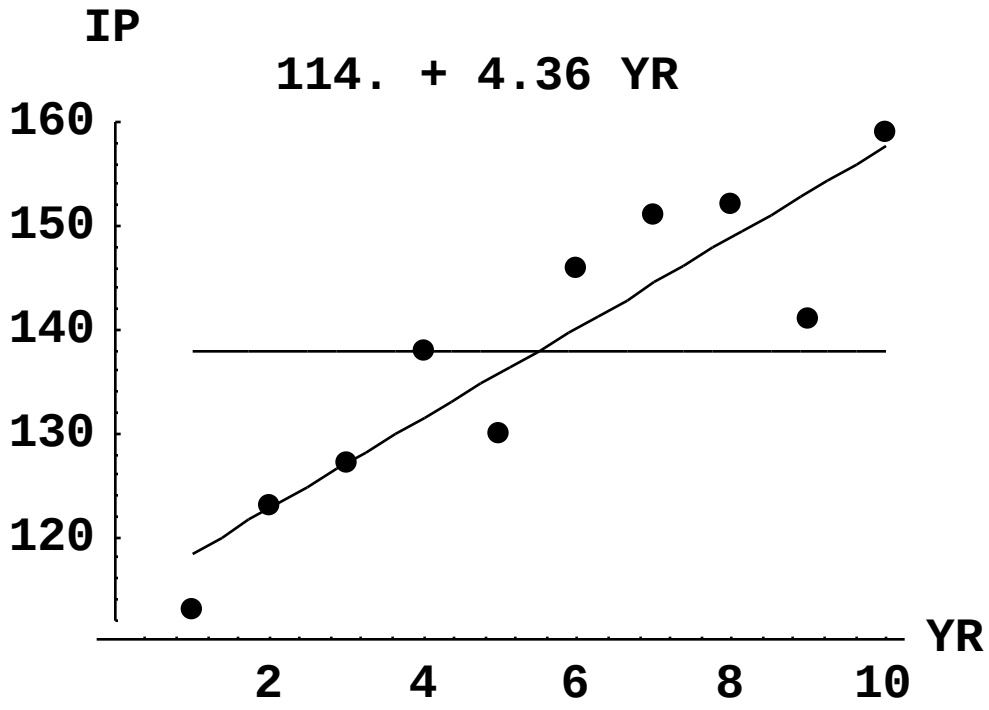
That is, using YR reduces the error in our predictions (relative to a simple model) of UN by 43% and a reduction of that magnitude is statistically reliable ($p < .04$). Reject MODEL C in favor of MODEL A! Note that UN.0 has been divided into two parts: UN.0^ which represents 43% of the original UN.0 variable and UN.0.YR which represents the other 57%.

How might we improve on our model of UN? We can only make progress if we can find something that will predict when UN is unexpectedly low or high. The expected part related to YR we've already taken care of. Thus, we want something that will predict UN.0.YR, the remaining 57% of the original UN variable that we don't understand!

An obvious candidate is to try IP again. However, there may also be yearly changes in IP. So first we want to get that part of IP that isn't related to YR. In other words, we want to know when IP is unexpectedly high or low relative to a model of consistent yearly changes in IP.

Predicting IP with YR

```
A:  IP^ = 114. + 4.36 YR
C:  IP^ = 138.
```



On average, Industrial Production increases each year by about 4.36 units.

YR	IP	IP^	IP.0,YR
1	113	118.4	-5.4
2	123	122.7	0.27
3	127	127.1	-0.091
4	138	131.5	6.5
5	130	135.8	-5.8
6	146	140.2	5.8
7	151	144.5	6.5
8	152	148.9	3.1
9	141	153.3	-12.
10	159	157.6	1.4

The values of IP.0,YR have similar interpretations to those for UN.0,YR. That is, IP.0,YR = -5.4 for YR 1 means that relative to a model of consistent yearly changes, Industrial Production was lower than expected by 5.4 units. The next year it was marginally higher than expected by .27 units. Note that in year 9, industrial production was a whopping 12 units lower than expected.

SSE(C) = 1914., SSE(A) = 343.1, SSR = 1571.
PRE = 0.82, F*(1,8) = 36.6, p = 0.0003

Tol = 1 - PRE = 0.18

Yes, there are reliable yearly changes in Industrial Production. That is, IP and YR are highly related (PRE=.82) or, in other words, only 1-.82 = .18 of IP is not related to YR.

Predicting UN.0,YR with IP.0,YR

The key question becomes whether Unemployment is unexpectedly high (low) when Industrial Production is unexpectedly low (high) relative to a model of consistent yearly changes. In other words, can IP.0,YR predict UN.0,YR? Let's look at the data:

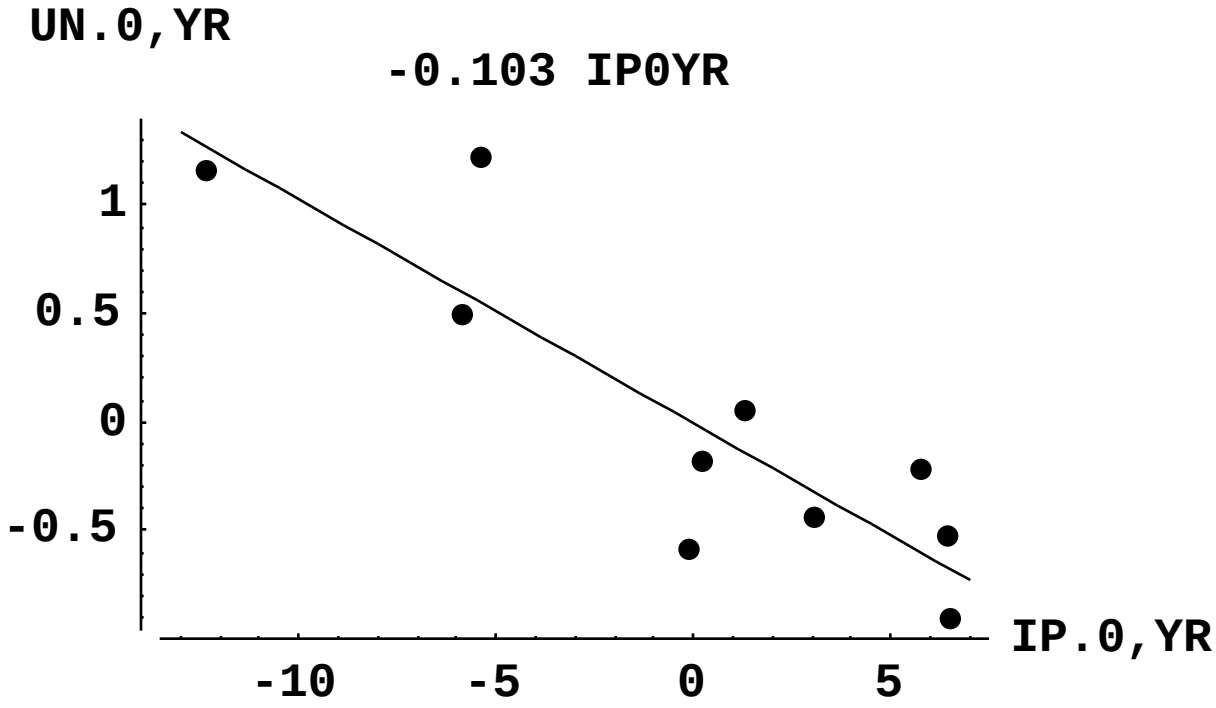
YR	UN.0,YR	IP.0,YR
1	1.22	-5.36
2	-0.19	0.273
3	-0.599	-0.0909
4	-0.907	6.55
5	0.484	-5.82
6	-0.224	5.82
7	-0.533	6.45
8	-0.441	3.09
9	1.15	-12.3
10	0.0418	1.36

In YR 1 unemployment is higher than expected (by 1.22 million) and industrial production is lower than expected (by 5.36 units). In YR 2, unemployment is a little lower than expected and industrial production is a little higher than expected. But in YR 3, both unemployment and industrial production are a little lower than expected. But then again, in YR 9 unemployment is higher than expected and industrial production is a lot lower than expected. So let's ask whether there is a relationship on average, using the only tool we know:

A: UN.0,YR^ = -0.103 IP0YR
C: UN.0,YR^ = 0

Note that we did not estimate a constant because we KNOW that the mean of UN.0,YR = 0 and we KNOW that IP.0,YR must be in mean deviation form. Thus, even before doing the regression, we KNOW that the constant must be 0. We therefore do not waste estimating a parameter whose value we already know.

The coefficient for IP.0,YR means that whenever industrial production is unexpectedly high by 1 unit, we predict that unemployment will be unexpectedly high by .103 million = 103,000 workers where our expectations are based on a model of consistent yearly changes.



Note that the regression line must go through the point {0,0} because 0 is the mean of both variables.

YR	UN.0,YR	UN.0,YR^	UN.0,YR,IP
1	1.2	0.55	0.66
2	-0.19	-0.028	-0.16
3	-0.6	0.0094	-0.61
4	-0.91	-0.68	-0.23
5	0.48	0.6	-0.12
6	-0.22	-0.6	0.38
7	-0.53	-0.67	0.13
8	-0.44	-0.32	-0.12
9	1.2	1.3	-0.12
10	0.042	-0.14	0.18

UN.0,YR,IP has the usual interpretation. UN.0,YR,IP = 0.66 for YR 1 means that relative to a model using both YR and IP, unemployment was unexpectedly high in YR 1 by 660,000 workers. In YR 2 unemployment was unexpectedly low by 160,000 workers. Etc. Notice that on average the size of the unexpected unemployment has become smaller. Note that we have used a total of 3 parameters to get to UN.0,YR,IP so we have n-PA = 10-3 = 7 potential parameters left to use.

SSE(C) = 4.79, SSE(A) = 1.13, SSR = 3.66

PRE = 0.76, F*(1,7) = 22.8, p = 0.002

t*(7) = 4.77

Predicting UN with YR & IP

Now we are ready for the punchline. Above we did a lot of work to estimate the coefficient for IP.0,YR. The punchline is that this is exactly the same coefficient we get when we do the multiple regression using both IP and YR to predict UN. Thus, the interpretation of the coefficient we developed above is the appropriate interpretation for the IP coefficient in multiple regression.

UN^ = 13.5 - 0.103 IP + 0.659 YR

Note how different our conclusion about the usefulness of IP is in the context of this model also including YR than it was when we asked about IP by itself as a predictor of UN. Even the sign is different! Different questions can have very different answers in multiple regression.