STAT 252 R3

LAB 4

DONG, Boyuan

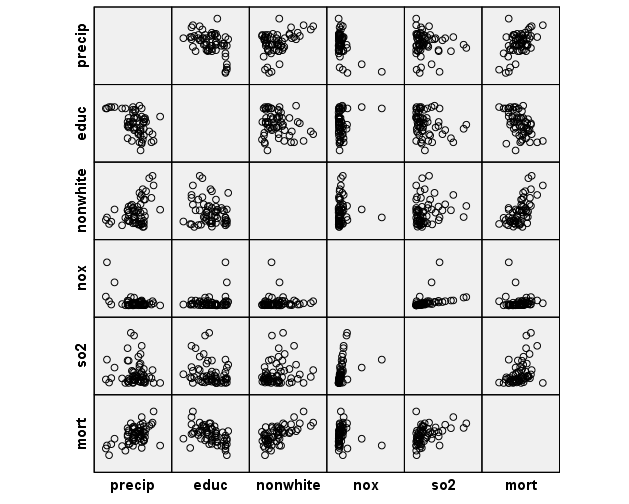
STAT 252 Lab4

1.

The random sample & selecting are presented to imply population inferences.

No. because association does not imply the cause-and-effect conclusions.

2.



(a)

Mort & PRECIP: strong positive linear relationship, with one outlier

Mort & EDUC: strong negative linear relationship, without outliers

Mort & NONWHITE: strong positive linear relationship, without outliers

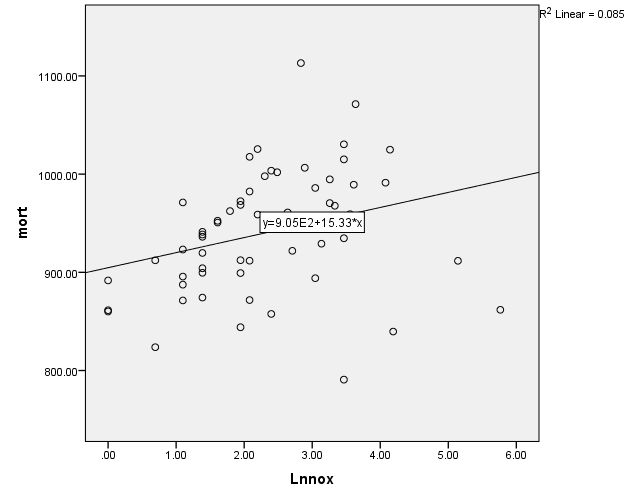
Mort & NOX: nonlinearity with sever outliers

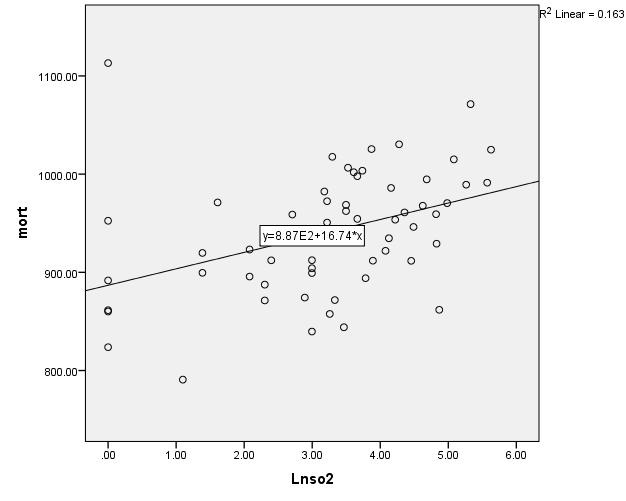
Mort & SO2: weak positive linear relationship with outliers

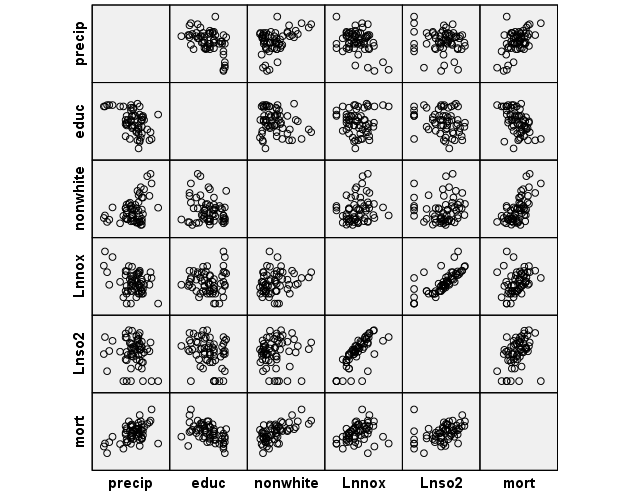
(b)

Transform the nox and so2 into Lnnox and Lnso2

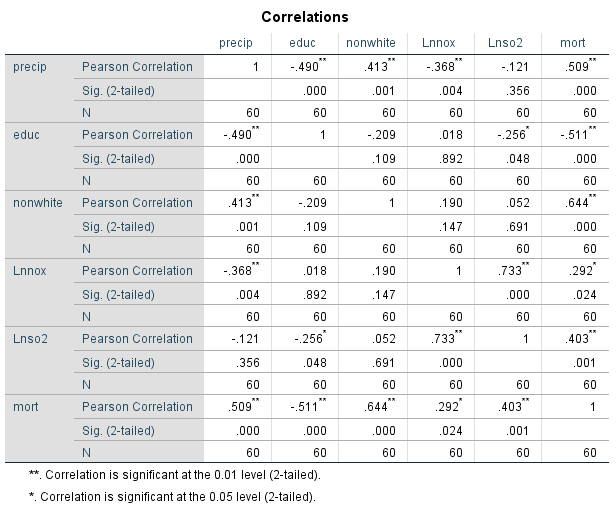
Because both the two variables have week or no linear relationship with the response variable mort, so transform them into ln form to obtain a stronger linear relationship with the mort. R^2 would be strong and the association would be stronger.







3.



(a)

 The explanatory variable with the highest correlation with the response: 3 points

I would choose the nonwhite variable with the highest correlation with the response mortality. It’s 0.644.

(b)

I expected the correlations between each pair of explanatory variables are all not significant.

The correlation between Lnso2 and Lnnox is 0.733, so there is a strong association between these two explanatory variables. So, multicollinearity is a problem in this case.

I would delete one of the variable and reserve the other variable.

4.

Mort = β0 + β1 \* precip + β2 \* educ + β3 \* nonwhite + β4 \* Lnnox + β5 \* Lnso2

Or, equivalently in terms of the mean as:

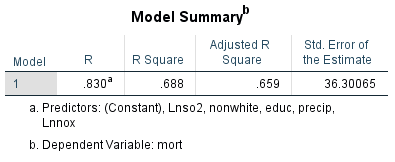
µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = β0 + β1 \* precip + β2 \* educ + β3 \* nonwhite + β4 \* Lnnox + β5 \* Lnso2

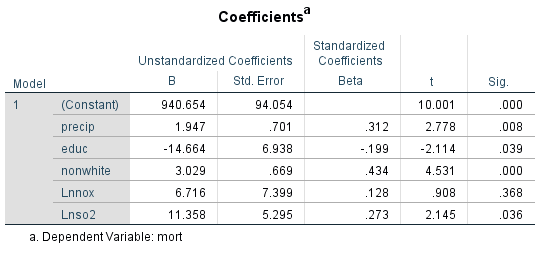
Assumptions:

It is assumed, for any (X1 = x1, ..., Xp = xp), that the errors in the response are independent and that ε ~ N(με = 0, σε2 ), where σε is independent of (X1, ..., Xp).

1. Linearity (με = 0): The mean response is a linear function of (X1, ..., Xp).
2. Constant Variance (σε is independent of any X): The variability in the response for any fixed (X1 = x1, ..., Xp = xp) is constant.
3. Normality: For any fixed (X1 = x1, ..., Xp = xp), the response is normally distributed.
4. Independence (errors are independent): The responses are all independent.

5.





(a)

Mort = β0 + β1 \* precip + β2 \* educ + β3 \* nonwhite + β4 \* Lnnox + β5 \* Lnso2

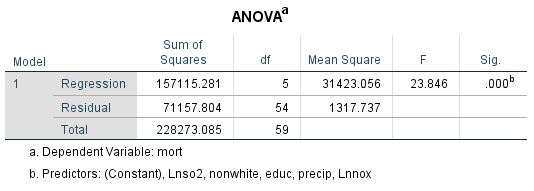
Mort = 940.654 + 1.947 \* precip + (-14.664) \* educ + 3.029 \* nonwhite + 6.716 \* Lnnox + 11.358 \* Lnso2

Or, equivalently in terms of the mean as:

µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = β0 + β1 \* precip + β2 \* educ + β3 \* nonwhite + β4 \* Lnnox + β5 \* Lnso2

µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = 940.654 + 1.947 \* precip + (-14.664) \* educ + 3.029 \* nonwhite + 6.716 \* Lnnox + 11.358 \* Lnso2

(b)



H0 : µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = β0

H0 : β1 = β2= β3= β4= β5=0

HA : µ (mort | precip, educ, nonwhite, Lnnox, Lnso2)= β0 + β1 \* precip + β2 \* educ + β3 \* nonwhite + β4 \* Lnnox + β5 \* Lnso2

HA : at least one βi ≠ 0, i=1,2,3,4,5

SSR(r)=228273.085 df(r)=59

SSR(f)=71157.804 df(f)=54

F0= ( (SSR(r) – SSR(f)) / (df(r) – df(f)) ) / (SSR(f) / df(f) ) =23.846 ~ F(5,54)

P-value = P(F(5,54) >23.846) =0.000 (approx.)

Follows a distribution F (5,54)

Conclusion:

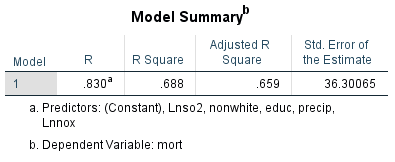
0<P-value <0.01

🡪Strong to convincing evidence against H0

🡪Reject H0

🡪We have enough evidence to suggest that at least one explanatory variable is useful.

(c)

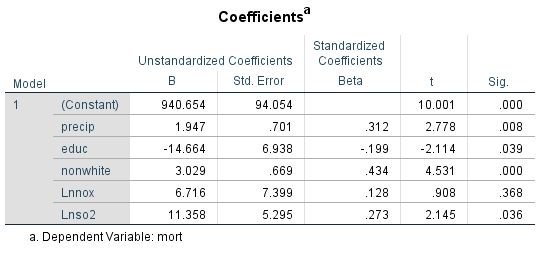


68.8% of the variation in mortality is explained be the five explanatory variables.

(d)

H0: βi=0 vs. HA: βi≠0 i=1,2,3,4,5

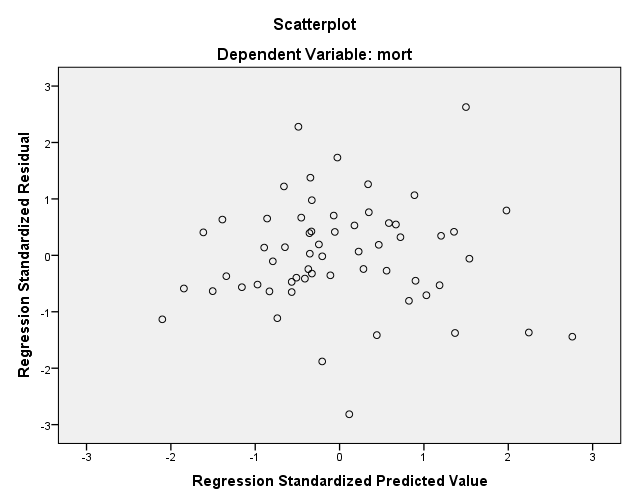
1. precip: t0=2.778, P-value=0.008 (2-tail), very significant, strong to convincing against β1=0, reject β1=0 🡪 have enough evidence to suggest β1≠0
2. educ: t0= - 2.114, P-value=0.039 (2-tail), very significant, strong to convincing against β2=0, reject β2=0 🡪 have enough evidence to suggest β2≠0
3. nonwhite: t0=4.531, P-value=0.000 (2-tail), very significant, strong to convincing against β3=0, reject β3=0 🡪 have enough evidence to suggest β3≠0
4. Lnnox: t0=0.908, P-value=0.368 (2-tail), not significant, week to no evidence against β4=0, do not reject β4=0 🡪have insufficient evidence to suggest β4≠0
5. Lnso2: t0=2.145, P-value=0.036 (2-tail), moderate significant, moderate evidence against β5=0, reject β5=0 🡪have enough evidence to suggest β5≠0



6.

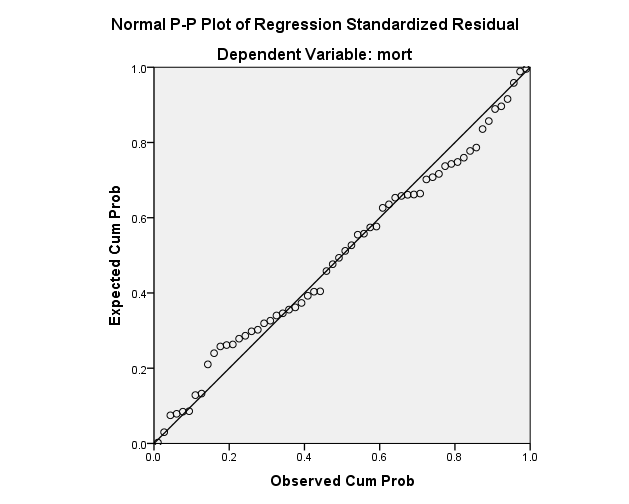
The graph shows approximately constant variance, the spread of the responses about the line is approximately the same at all levels of the explanatory variable.

There’s insufficient evidence to show that the variance of the residuals increases with increasing fitted values, with few outliers.



7.

The points approximately fall on or close to the line, showing increasing linear relationship. The assumption of normality is not violated.



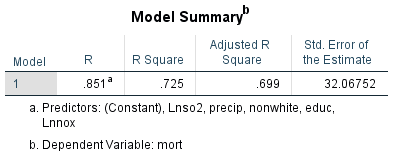
8.

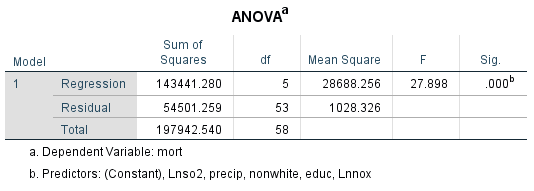
There’s one case (# 37 observation) with a large studentized residual of 3.55531. this is the only one with absolute value greater than 3, this is also has the only one Cook’s distance greater than 1 (about 1.75027).

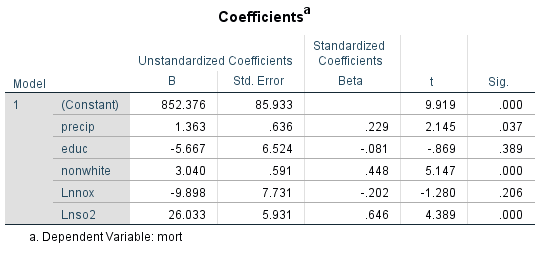
p=6 n=60, 2p/n= 2\*6/60=0.2, #29 with 0.21778 #37 with 0.43713 and #49 with 0.23878 all greater than 2p/n.

Only the #37 observation violates all the cases above, so the #37 observation is the outlier and influential case.

9.







(a)

Mort = 852.376 + 1.363 \* precip + (-5.667) \* educ + 3.040 \* nonwhite + (-9.898) \* Lnnox + (26.033) \* Lnso2

Or, equivalently in terms of the mean as:

µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = 852.376 + 1.363 \* precip + (-5.667) \* educ + 3.040 \* nonwhite + (-9.898) \* Lnnox + 26.033 \* Lnso2

(b)

Q5(c). 68.8% of the variation in mortality is explained be the five explanatory variables.

72.5% of the variation in mortality is explained be the five explanatory variables.

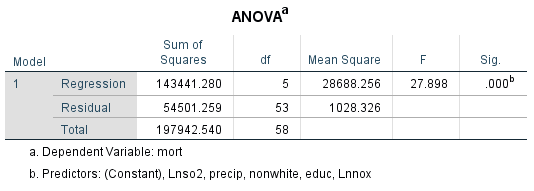
The percentage is larger than that in Q5

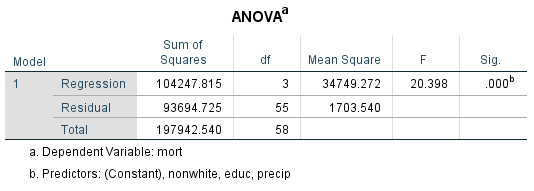
(c)

H0: βi=0 vs. HA: βi≠0 i=1,2,3,4,5

1. precip: t0=2.145, P-value=0.037 (2-tail), moderate significant, moderate evidence against β1=0, reject β1=0 🡪 have enough evidence to suggest β1≠0
2. educ: t0= - 0.869, P-value=0.389 (2-tail), not significant, week to no evidence against β2=0, do not reject β2=0 🡪have insufficient evidence to suggest β2≠0
3. nonwhite: t0=5.147, P-value=0.000 (2-tail), very significant, strong to convincing against β3=0, reject β3=0 🡪 have enough evidence to suggest β3≠0
4. Lnnox: t0= -1.280, P-value=0.206 (2-tail), not significant, week to no evidence against β4=0, do not reject β4=0 🡪have insufficient evidence to suggest β4≠0
5. Lnso2: t0=4.389, P-value=0.000 (2-tail), very significant, strong to convincing evidence against β5=0, reject β5=0 🡪have enough evidence to suggest β5≠0

10.





H0 : µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = β0 + β4\*Lnnox + β5\*Lnso2

H0 : β4= β5=0

HA : µ (mort | precip, educ, nonwhite, Lnnox, Lnso2)= β0 + β1 \* precip + β2 \* educ + β3 \* nonwhite + β4 \* Lnnox + β5 \* Lnso2

HA : at least one βi ≠ 0, i=4,5

SSR(r)=93694.725 df(r)=55

SSR(f)=54501.259 df(f)=53

F0= ( (SSR(r) – SSR(f)) / (df(r) – df(f)) ) / (SSR(f) / df(f) ) =( 93694.725 – 54501.259) / (55 – 53) ) / (54501.259/ 53 )=19.057 ~ F(2,53)

P-value = P(F(2,53) >19.057) = P(F(2,50) >19.057) (approx.) 🡺 (0,0.001)

Follows a distribution F (2,53)

Conclusion:

0<P-value <0.01

🡪Strong to convincing evidence against H0

🡪Reject H0

🡪We have enough evidence to suggest that at least one explanatory variable is useful.

11.

(a)

µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = 852.376 + 1.363 \* precip + (-5.667) \* educ + 3.040 \* nonwhite + (-9.898) \* Lnnox + 26.033 \* Lnso2

µ (mort | precip, educ, nonwhite, Lnnox, Lnso2) = 852.376 + 1.363 \* (35) + (-5.667) \* (11) + 3.040 \* (3.5) + (-9.898) \* (Ln10) + 26.033 \* (Ln39)=920.89938

(b)

Confidence interval for the mean mortality: (908.64802, 933.30210)

Prediction interval: (855.48515, 986.46496)

The predicted interval is wider than the confidence interval for the mean mortality.

Because the SE for the confidence interval of the mean mortality is:

S.E.(μˆY|x\*) = σˆ

while for the prediction interval:

S.E.( yˆ ) = σˆ

S.E.(μˆY|x\*) < S.E.( yˆ ) and also

For CI: μˆY|x\*± tα/2, n – 2 ×S.E.(μˆY|x\*) for PI: ŷ ± t α / 2 , n – 2 × S.E. ( yˆ )

So the PI is wider than CI.