





R version 3.1.1 (2014-07-10) -- "Sock it to Me"

Copyright (C) 2014 The R Foundation for Statistical Computing

Platform: x86\_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.

Type 'contributors()' for more information and

'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or

'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

> risk<-read.csv(choose.files(),header=TRUE)

> attach(risk)

> summary(SOPH)

Min. 1st Qu. Median Mean 3rd Qu. Max.

5.00 18.00 23.00 21.19 25.00 31.00

> SOPHlow=(1\*(SOPH<=10))

> SOPHmd=(1\*(SOPH>10&SOPH<=20))

> SOPHhigh=(1\*(SOPH>20))

> SOPHmed=(1\*(SOPH>10&SOPH<=20))

> check=data.frame(SOPH<SOPHlow,SOPHmed,SOPHhigh)

> fix(check)

>

> #this way you have different variables and can choose which one to omit

> #and make the baseline

>

> categories=cut(SOPH,br=c(5,10,20,31))

> table(categories)

categories

(5,10] (10,20] (20,31]

1 29 42

> model1=lm(FIRMCOST~ASSUME+CAP+SIZELOG+INDCOST+SOPH+CENTRAL)

> lnASSUME=log(ASSUME+0.000000000000000000001)

> lnFIRMCOST=log(FIRMCOST)

> model2=lm(lnFIRMCOST~lnASSUME+CAP+SIZELOG+INDCOST+SOPH+CENTRAL,subset=-c(67,32,10,34,15)

+ )

> layout(matrix(c(1,2,3,4,5,6,7,8,9,10,11,12),byrow=TRUE,ncol=6))

> layout(matrix(c(1,2,3,4,5,6,7,8,9,10,11,12),byrow=TRUE,ncol=6))

> plot.new()

> hist(ASSUME)

> hist(SIZELOG)

> hist(INDCOST)

> hist(CENTRAL)

> hist(SOPH)

> hist(FIRMCOST)

> plot(ASSUME,FIRMCOST)

> text(ASSUME,FIRMCOST,labels=row.names(risk),pos=1)

> plot(SIZELOG,FIRMCOST)

> plot(INDCOST,FIRMCOST)

> plot(CENTRAL,FIRMCOST)

> plot(SOPH,FIRMCOST)

> #pos=1 means that put the number under the points

> model2=lm(lnFIRMCOST~lnASSUME+CAP+SIZELOG+INDCOST+SOPH+CENTRAL,subset=-c(67,32,10,34,15)

+ )

> risk1=risk[-c(72,15,57,16,73),]

> #before the comma is the row and after the comma is the column

> #when we leave blank in front of the comma means that we want all the columns

> ##this will withdraw those 5 rows, but still give us all columns since nothing is after the comma

> #we are doing this to see if after we would withdraw thse points, would we still

> #need to log ASSUME and FIRMCOST or would that not be a problem anymore

> risk1

FIRMCOST ASSUME CAP SIZELOG INDCOST CENTRAL SOPH

1 3.29 0.29 1 9.55 0.32 1 25

2 9.31 0.89 0 8.04 0.33 2 24

3 4.07 1.67 0 7.90 0.34 2 15

4 6.94 1.21 0 8.10 0.34 1 16

5 5.35 0.28 0 7.74 0.09 3 18

6 28.86 14.86 1 8.16 0.42 3 25

7 2.79 0.85 0 9.47 0.41 1 11

8 15.00 1.47 0 8.13 0.32 1 22

9 3.89 3.47 0 7.50 0.53 2 18

10 4.07 5.86 0 8.55 1.22 1 27

11 4.34 0.39 0 8.08 0.26 5 26

12 4.33 2.08 1 8.01 0.34 4 11

13 5.29 1.14 0 7.59 0.42 3 20

14 7.90 0.24 0 8.52 0.33 2 22

17 8.51 0.23 0 8.39 0.63 2 23

18 4.31 0.50 0 8.86 0.42 1 19

19 12.45 1.02 0 8.30 0.32 1 11

20 2.55 0.44 0 10.12 0.34 1 17

21 8.18 3.86 0 7.00 0.42 3 30

22 11.29 2.98 0 7.65 0.34 1 24

23 1.25 0.01 0 8.52 0.41 5 24

24 11.92 1.68 0 6.29 0.34 1 25

25 0.65 0.09 0 8.70 0.10 2 23

26 3.15 0.21 1 8.59 0.50 2 19

27 19.38 0.34 1 7.38 0.33 1 27

28 3.75 3.28 0 8.07 0.33 2 23

29 13.33 0.83 0 8.01 0.53 2 18

30 4.58 0.44 1 8.48 0.34 3 18

31 13.96 0.02 1 8.49 0.50 3 29

32 0.28 0.00 0 9.86 0.10 1 5

33 6.08 0.50 1 7.60 0.33 3 22

34 0.94 0.01 1 8.16 0.33 1 18

35 2.97 0.17 0 7.45 0.10 3 25

36 4.11 0.17 1 8.41 0.33 4 25

37 1.22 0.50 1 10.60 0.34 1 13

38 6.29 0.01 0 9.04 0.86 2 20

39 6.12 0.05 1 8.26 0.33 4 23

40 3.51 0.33 0 9.15 0.32 1 27

41 2.16 0.90 1 8.96 0.34 2 19

42 0.36 0.07 0 9.95 0.34 4 22

43 7.83 1.67 0 8.01 0.26 2 25

44 5.09 0.50 0 7.00 0.33 4 26

45 0.20 0.07 1 10.27 0.10 1 25

46 8.85 1.35 0 8.27 0.33 1 24

47 0.76 0.02 0 10.02 0.50 3 19

48 5.71 1.53 0 8.16 0.26 3 24

49 17.53 0.13 0 7.31 0.86 1 18

50 14.00 0.63 0 8.99 0.33 4 23

51 2.06 0.33 1 9.10 0.33 1 24

52 0.93 0.79 1 9.55 0.34 2 28

53 10.00 0.71 0 8.16 0.50 1 27

54 5.82 0.29 1 8.55 0.50 2 9

55 9.13 2.50 0 8.29 0.32 5 31

56 9.00 2.03 0 7.60 0.50 4 24

58 2.15 0.13 0 9.01 0.32 4 23

59 22.22 0.42 0 7.50 0.33 1 18

60 12.71 3.42 0 7.09 0.32 2 27

61 15.97 0.79 0 8.95 0.62 3 28

62 4.32 1.90 0 8.58 0.32 3 27

63 8.49 0.29 0 8.58 0.32 1 15

64 5.25 6.46 1 8.29 0.42 1 11

65 18.33 0.29 0 8.70 0.86 1 25

66 21.72 0.92 0 7.50 0.41 4 18

67 0.40 0.00 1 10.36 0.41 5 21

68 3.70 0.43 1 8.17 0.34 1 17

69 15.00 2.92 0 7.50 0.34 1 14

70 18.00 2.30 1 9.21 0.86 3 18

71 29.12 0.79 0 7.96 0.33 1 19

> attach(risk1)

The following objects are masked from risk:

ASSUME, CAP, CENTRAL, FIRMCOST, INDCOST, SIZELOG, SOPH

> par(mfrow=c(1,2)

+ )

> hist(FIRMCOST)

> hist(ASSUME)

> #thse are both better, but still right skewed, so maybe we should still stick with the logs

>

> attach(risk)

The following objects are masked from risk1:

ASSUME, CAP, CENTRAL, FIRMCOST, INDCOST, SIZELOG, SOPH

The following objects are masked from risk (pos = 4):

ASSUME, CAP, CENTRAL, FIRMCOST, INDCOST, SIZELOG, SOPH

> model2=lm(lnFIRMCOST~lnASSUME+CAP+SIZELOG+INDCOST+SOPH+CENTRAL,subset=-c(67,32,10,34,15)

+ )

> summary(model2)

Call:

lm(formula = lnFIRMCOST ~ lnASSUME + CAP + SIZELOG + INDCOST +

SOPH + CENTRAL, subset = -c(67, 32, 10, 34, 15))

Residuals:

Min 1Q Median 3Q Max

-1.44056 -0.43247 -0.00076 0.43228 1.53867

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.794997 1.057506 6.425 2.23e-08 \*\*\*

lnASSUME 0.095035 0.060128 1.581 0.119

CAP -0.177675 0.192151 -0.925 0.359

SIZELOG -0.672300 0.107350 -6.263 4.21e-08 \*\*\*

INDCOST 1.957882 0.454386 4.309 6.07e-05 \*\*\*

SOPH 0.002572 0.018123 0.142 0.888

CENTRAL -0.077998 0.075617 -1.031 0.306

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7082 on 61 degrees of freedom

Multiple R-squared: 0.6164, Adjusted R-squared: 0.5786

F-statistic: 16.33 on 6 and 61 DF, p-value: 4.125e-11

> #Let's start backwards regression

> model3=lm(lnFIRMCOST~lnASSUME+CAP+SIZELOG+INDCOST+CENTRAL,subset=-c(67,32,10,34,15)

+ )

> summary(model3)

Call:

lm(formula = lnFIRMCOST ~ lnASSUME + CAP + SIZELOG + INDCOST +

CENTRAL, subset = -c(67, 32, 10, 34, 15))

Residuals:

Min 1Q Median 3Q Max

-1.44973 -0.43795 0.00297 0.44570 1.53889

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.85984 0.94618 7.250 7.93e-10 \*\*\*

lnASSUME 0.09457 0.05956 1.588 0.117

CAP -0.17898 0.19041 -0.940 0.351

SIZELOG -0.67415 0.10571 -6.378 2.54e-08 \*\*\*

INDCOST 1.95741 0.45077 4.342 5.31e-05 \*\*\*

CENTRAL -0.07532 0.07264 -1.037 0.304

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7026 on 62 degrees of freedom

Multiple R-squared: 0.6162, Adjusted R-squared: 0.5853

F-statistic: 19.91 on 5 and 62 DF, p-value: 8.754e-12

> #we take the worse p value

> model4=lm(lnFIRMCOST~lnASSUME+SIZELOG+INDCOST+CENTRAL,subset=-c(67,32,10,34,15)

+ )

> summary(model4)

Call:

lm(formula = lnFIRMCOST ~ lnASSUME + SIZELOG + INDCOST + CENTRAL,

subset = -c(67, 32, 10, 34, 15))

Residuals:

Min 1Q Median 3Q Max

-1.38661 -0.51652 0.05331 0.48126 1.61111

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.02903 0.92804 7.574 1.98e-10 \*\*\*

lnASSUME 0.08612 0.05883 1.464 0.148

SIZELOG -0.69784 0.10256 -6.804 4.38e-09 \*\*\*

INDCOST 1.88085 0.44294 4.246 7.29e-05 \*\*\*

CENTRAL -0.07709 0.07255 -1.063 0.292

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.702 on 63 degrees of freedom

Multiple R-squared: 0.6108, Adjusted R-squared: 0.5861

F-statistic: 24.71 on 4 and 63 DF, p-value: 2.498e-12

> model5=lm(lnFIRMCOST~lnASSUME+SIZELOG+INDCOST,subset=-c(67,32,10,34,15))

> summary(model5)

Call:

lm(formula = lnFIRMCOST ~ lnASSUME + SIZELOG + INDCOST, subset = -c(67,

32, 10, 34, 15))

Residuals:

Min 1Q Median 3Q Max

-1.3765 -0.5114 0.0439 0.4551 1.4780

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.64765 0.85668 7.760 8.55e-11 \*\*\*

lnASSUME 0.10037 0.05733 1.751 0.0848 .

SIZELOG -0.67370 0.10011 -6.729 5.54e-09 \*\*\*

INDCOST 1.91523 0.44220 4.331 5.33e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7027 on 64 degrees of freedom

Multiple R-squared: 0.6038, Adjusted R-squared: 0.5852

F-statistic: 32.51 on 3 and 64 DF, p-value: 6.89e-13

> model6=lm(lnFIRMCOST~SIZELOG+INDCOST,subset=-c(67,32,10,34,15))

> summary(model6)

Call:

lm(formula = lnFIRMCOST ~ SIZELOG + INDCOST, subset = -c(67,

32, 10, 34, 15))

Residuals:

Min 1Q Median 3Q Max

-1.44998 -0.45180 0.04349 0.51332 1.51761

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.14900 0.82012 8.717 1.56e-12 \*\*\*

SIZELOG -0.74241 0.09355 -7.936 3.79e-11 \*\*\*

INDCOST 1.95979 0.44843 4.370 4.57e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7137 on 65 degrees of freedom

Multiple R-squared: 0.5848, Adjusted R-squared: 0.572

F-statistic: 45.78 on 2 and 65 DF, p-value: 3.914e-13

> #we wanna see if by removing it we lose a lot of R2 and R2a

> #there is no right or wrong answer..so you can take out the last variable or leave it in

>

> risk2=data.frame(SIZELOG,INDCOST,lnFIRMCOST)

> pairs(risk2,upper.panel=NULL)

> #for INDCOST, we can see that the plot first seems steep, then it curve as it suggest quadratic

> #so we squared them

> IND2=INDCOST\*INDCOST

> model7=lm(lnFIRMCOST~SIZELOG+IND2,subset=-c(67,32,10,34,15))

> summary(model7)

Call:

lm(formula = lnFIRMCOST ~ SIZELOG + IND2, subset = -c(67, 32,

10, 34, 15))

Residuals:

Min 1Q Median 3Q Max

-1.66530 -0.46091 0.03448 0.49002 1.50999

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.64576 0.82364 9.283 1.58e-13 \*\*\*

SIZELOG -0.74041 0.09725 -7.614 1.42e-10 \*\*\*

IND2 1.40831 0.38595 3.649 0.000526 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7396 on 65 degrees of freedom

Multiple R-squared: 0.5542, Adjusted R-squared: 0.5404

F-statistic: 40.39 on 2 and 65 DF, p-value: 3.968e-12

> model7=lm(lnFIRMCOST~SIZELOG+INDCOST+IND2,subset=-c(67,32,10,34,15))

> summary(model7)

Call:

lm(formula = lnFIRMCOST ~ SIZELOG + INDCOST + IND2, subset = -c(67,

32, 10, 34, 15))

Residuals:

Min 1Q Median 3Q Max

-1.47554 -0.46097 0.03016 0.51759 1.51891

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.73559 0.85767 7.853 5.85e-11 \*\*\*

SIZELOG -0.75676 0.09315 -8.124 1.95e-11 \*\*\*

INDCOST 4.25445 1.59099 2.674 0.0095 \*\*

IND2 -1.98475 1.32140 -1.502 0.1380

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7069 on 64 degrees of freedom

Multiple R-squared: 0.599, Adjusted R-squared: 0.5802

F-statistic: 31.86 on 3 and 64 DF, p-value: 1.012e-12

> #those removed points were outliers or high leverage points based on a n old model,

> #maybe now that we have removed some of the variablse, maybe the variables were the

> #problem and some of these points may be fine now.

> model8=lm(lnFIRMCOST~SIZELOG+INDCOST+IND2)

> summary(model8)

Call:

lm(formula = lnFIRMCOST ~ SIZELOG + INDCOST + IND2)

Residuals:

Min 1Q Median 3Q Max

-1.78663 -0.57616 0.00556 0.55472 2.91533

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.3525 0.9530 6.666 5.33e-09 \*\*\*

SIZELOG -0.7726 0.1013 -7.626 9.64e-11 \*\*\*

INDCOST 6.2635 1.6104 3.889 0.000229 \*\*\*

IND2 -3.5850 1.2654 -2.833 0.006039 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8235 on 69 degrees of freedom

Multiple R-squared: 0.5655, Adjusted R-squared: 0.5466

F-statistic: 29.94 on 3 and 69 DF, p-value: 1.645e-12

> #the squared term is now significant, but the s and R2 are a little worse

>

> rstandard=rstandard(model8)

> rstandard[order(rstandard)]

34 9 23 42 67 47

-2.187890970 -1.875406132 -1.853620830 -1.749986768 -1.556735151 -1.412741527

44 73 21 10 13 26

-1.234466211 -1.125058650 -1.045709742 -1.022350769 -1.010993559 -0.990716615

52 24 25 45 32 57

-0.942846950 -0.924674723 -0.844265462 -0.823943135 -0.788013677 -0.737135759

49 3 56 28 68 12

-0.730647953 -0.687327940 -0.641836420 -0.578865161 -0.547663695 -0.506850954

41 33 64 51 38 58

-0.460450927 -0.433392542 -0.352848776 -0.339019678 -0.335821362 -0.323349265

17 54 36 35 18 11

-0.313378619 -0.272373960 -0.144677512 -0.128670621 -0.054952156 -0.034827326

30 53 7 60 62 29

0.006802684 0.023244393 0.030352464 0.037580622 0.125331425 0.139864340

4 39 22 48 37 40

0.156238136 0.200864726 0.327531462 0.378728070 0.408310926 0.412176483

2 69 72 43 46 66

0.506777007 0.535275805 0.603807868 0.625041448 0.661896078 0.683569163

65 1 31 14 55 27

0.685691528 0.717066577 0.748158750 0.759400974 0.767282672 0.785616185

20 63 5 61 59 16

0.861368307 0.952898603 1.005824158 1.030542443 1.066806676 1.080203446

19 70 8 6 71 50

1.156530714 1.168890005 1.224383804 1.613690806 1.829044334 1.909837392

15

3.648540517

> #34 and 15 are outliers outside of |2|

> leverages=hatvalues(model8)

> leverages[order(leverages)]

30 36 46 39 68 14 19

0.01592652 0.01607942 0.01622150 0.01625435 0.01635631 0.01638396 0.01666519

55 34 4 62 63 28 8

0.01668977 0.01674930 0.01685590 0.01720600 0.01720600 0.01745354 0.01749456

12 2 23 71 64 3 6

0.01771616 0.01774276 0.01848436 0.01864720 0.01890487 0.01910053 0.01938348

41 50 58 18 22 51 33

0.02099963 0.02181053 0.02265675 0.02284457 0.02360881 0.02404610 0.02511433

40 48 11 43 69 13 59

0.02563904 0.02574079 0.02639635 0.02712885 0.02722179 0.02752023 0.02760689

53 31 54 26 66 27 29

0.02793847 0.02800603 0.02837244 0.02867724 0.02897753 0.03099748 0.03262620

56 52 7 1 9 60 17

0.03536632 0.03678905 0.03703348 0.03742874 0.04113908 0.04161216 0.04350302

44 21 61 57 42 15 20

0.04460971 0.04629978 0.04918419 0.05301346 0.05348655 0.05857997 0.06204943

47 65 67 24 49 73 38

0.07138592 0.07750345 0.07919350 0.08127000 0.08309945 0.08492790 0.08503612

70 37 25 35 5 32 72

0.09011452 0.09094859 0.11983662 0.13280966 0.13646035 0.15010457 0.16259037

45 10 16

0.17054415 0.40604031 0.42058786

> #cutoff=3(k+1)/n=3(3+1)/73

> 12/73

[1] 0.1643836

> #high leverage points are 45,10, and 16

> model9=lm(lnFIRMCOST~SIZELOG+INDCOST+IND2,subset=-c(34,15,45,10,16))

> summary(model9)

Call:

lm(formula = lnFIRMCOST ~ SIZELOG + INDCOST + IND2, subset = -c(34,

15, 45, 10, 16))

Residuals:

Min 1Q Median 3Q Max

-1.46593 -0.45479 0.00515 0.53313 1.54529

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.85892 0.91750 7.476 2.71e-10 \*\*\*

SIZELOG -0.78452 0.09356 -8.386 6.75e-12 \*\*\*

INDCOST 4.76311 2.09308 2.276 0.0262 \*

IND2 -2.60927 2.00293 -1.303 0.1973

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7203 on 64 degrees of freedom

Multiple R-squared: 0.5804, Adjusted R-squared: 0.5607

F-statistic: 29.51 on 3 and 64 DF, p-value: 4.244e-12

> #to pick the best model there is no right or wrong answer, but you need to explain why you pick those model

> #for example you can say that the p value is now significant, or the R2 is now significant

> residuals6=residuals(model6)

> plot(INDCOST,residuals8,xlab="INDCOST",ylab="RESIDUALS")

Error in xy.coords(x, y, xlabel, ylabel, log) :

object 'residuals8' not found

> #snafu I'll figure out later