#3

a) Numerical statistics of both models initially below:

> summary(pdata.lm1)

Call:

lm(formula = y ~ X1 + X2 + X3 + X4, data = pdata)

Residuals:

Min 1Q Median 3Q Max

-66.648 -10.246 -2.180 4.843 175.897

Coefficients:

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

(Intercept) -48.74779 39.09656 -1.247 0.2156

X1 2.78319 0.47290 5.885 6.39e-08 \*\*\*

X2 0.03239 0.07236 0.448 0.6555

X3 -0.28014 0.45730 -0.613 0.5417

X4 10.09513 5.17913 1.949 0.0543 .

---

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

Residual standard error: 31.88 on 92 degrees of freedom

Multiple R-squared: 0.4144, Adjusted R-squared: 0.3889

F-statistic: 16.28 on 4 and 92 DF, p-value: 4.092e-10

> summary(pdata.lm2)

Call:

lm(formula = y ~ Z1 + Z2 + Z3 + Z4, data = pdata)

Residuals:

Min 1Q Median 3Q Max

-62.044 -9.035 0.341 6.065 170.231

Coefficients:

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

(Intercept) -40.769 33.244 -1.226 0.223199

Z1 2.028 0.584 3.473 0.000787 \*\*\*

Z2 17.857 10.751 1.661 0.100111

Z3 1.104 1.325 0.833 0.407104

Z4 6.393 5.025 1.272 0.206518

---

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

Residual standard error: 30.99 on 92 degrees of freedom

Multiple R-squared: 0.4467, Adjusted R-squared: 0.4226

F-statistic: 18.57 on 4 and 92 DF, p-value: 3.246e-11

To correct for violations in assumptions, outliers must be checked and removed if they are significant

> outlierTest(pdata.lm1)

rstudent unadjusted p-value Bonferroni p

96 6.922725 6.0796e-10 5.8972e-08

97 6.457904 5.1050e-09 4.9519e-07

> outlierTest(pdata.lm2)

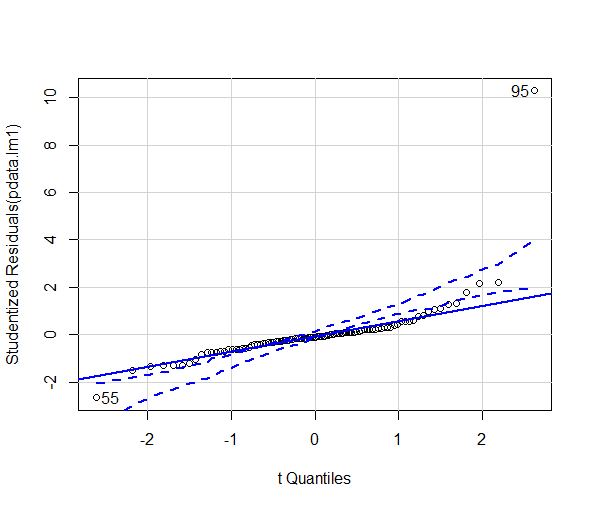
rstudent unadjusted p-value Bonferroni p

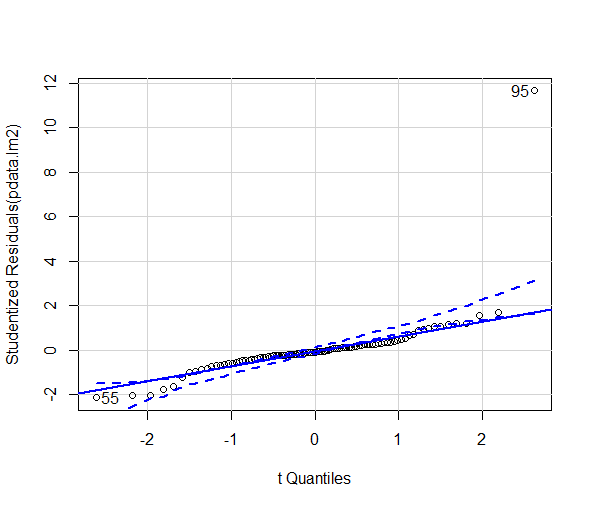
96 7.046140 3.4313e-10 3.3283e-08

97 6.745516 1.3755e-09 1.3342e-07

We see that 96 and 97 are shown to be outliers in both models so they must be removed, as well as check for multicollinearity, normality, and residual fits.

Studentized resids for both our models below





We see that the studentized resids fit well apart from potential outlier at 95, however since it passed our outlier test earlier it is not signifigant

Summary data of models without 96 and 97:

> summary(pdata.lm1)

Call:

lm(formula = y ~ X1 + X2 + X3 + X4, data = pdata)

Residuals:

Min 1Q Median 3Q Max

-43.574 -7.128 -1.895 3.597 117.585

Coefficients:

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

(Intercept) -15.43444 21.69049 -0.712 0.47857

X1 1.72515 0.27392 6.298 1.08e-08 \*\*\*

X2 0.03259 0.03989 0.817 0.41605

X3 -0.48456 0.25250 -1.919 0.05815 .

X4 7.63972 2.86704 2.665 0.00913 \*\*

---

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

Residual standard error: 17.57 on 90 degrees of freedom

Multiple R-squared: 0.4707, Adjusted R-squared: 0.4472

F-statistic: 20.01 on 4 and 90 DF, p-value: 8.171e-12

> summary(pdata.lm2)

Call:

lm(formula = y ~ Z1 + Z2 + Z3 + Z4, data = pdata)

Residuals:

Min 1Q Median 3Q Max

-33.142 -5.783 -1.393 4.330 120.455

Coefficients:

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

(Intercept) -27.9840 18.5454 -1.509 0.1348

Z1 1.5093 0.3263 4.625 1.25e-05 \*\*\*

Z2 16.8182 6.1261 2.745 0.0073 \*\*

Z3 -0.5842 0.7888 -0.741 0.4609

Z4 5.0698 2.8047 1.808 0.0740 .

---

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

Residual standard error: 17.21 on 90 degrees of freedom

Multiple R-squared: 0.4926, Adjusted R-squared: 0.4701

F-statistic: 21.84 on 4 and 90 DF, p-value: 1.275e-12

(ii)

After reviewing both models, I decided to go with model 2 because model 2 has a higher adjusted R-squared, as well as a lower P-value in both the T score and the F statistic.

(b) Run forward stepwise subset selection

> app.step.fw=step(fit2, direction="forward", scope=list(upper=fit1, lower=fit2))

Start: AIC=601.93

y ~ 1

Df Sum of Sq RSS AIC

+ Z1 1 21938 30578 552.55

+ Z2 1 15851 36666 569.79

+ Z3 1 12210 40306 578.79

+ Z4 1 10354 42162 583.06

<none> 52516 601.93

Step: AIC=552.55

y ~ Z1

Df Sum of Sq RSS AIC

+ Z2 1 2891.57 27687 545.11

+ Z4 1 1571.07 29007 549.53

<none> 30578 552.55

+ Z3 1 386.86 30191 553.34

Step: AIC=545.11

y ~ Z1 + Z2

Df Sum of Sq RSS AIC

+ Z4 1 877.43 26809 544.05

<none> 27687 545.11

+ Z3 1 72.42 27614 546.86

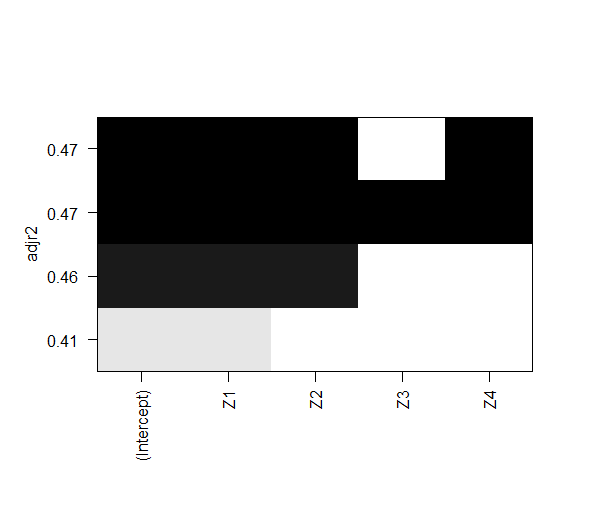
Step: AIC=544.05

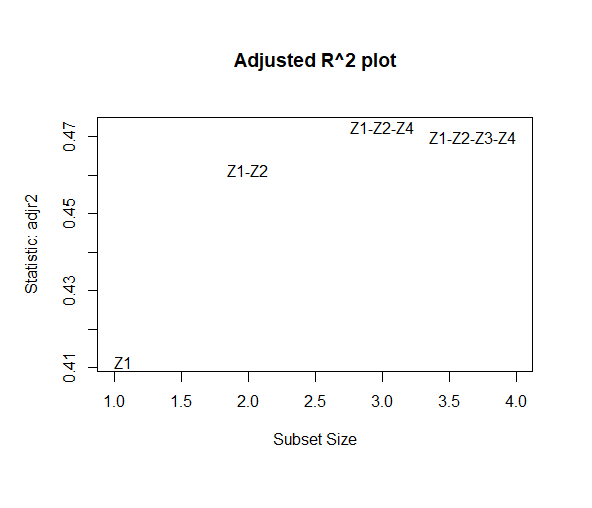
y ~ Z1 + Z2 + Z4

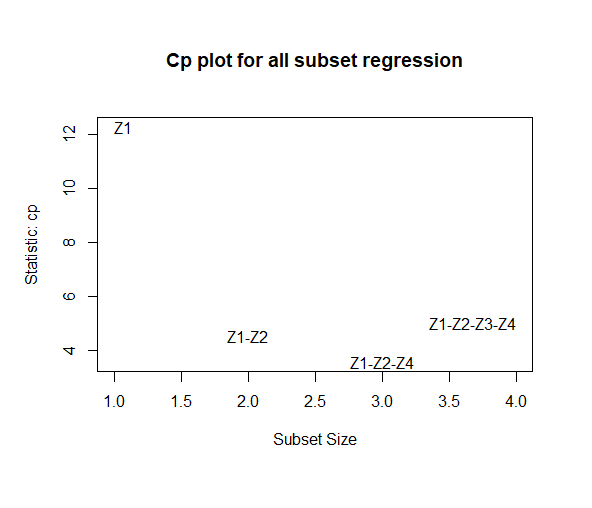
Df Sum of Sq RSS AIC

<none> 26809 544.05

+ Z3 1 162.39 26647 545.47







> res.sum

Subset selection object

Call: regsubsets.formula(y ~ Z1 + Z2 + Z3 + Z4, data = pdata, nvmax = 4,

nbest = 4)

4 Variables (and intercept)

Forced in Forced out

Z1 FALSE FALSE

Z2 FALSE FALSE

Z3 FALSE FALSE

Z4 FALSE FALSE

4 subsets of each size up to 4

Selection Algorithm: exhaustive

Z1 Z2 Z3 Z4

1 ( 1 ) "\*" " " " " " "

1 ( 2 ) " " "\*" " " " "

1 ( 3 ) " " " " "\*" " "

1 ( 4 ) " " " " " " "\*"

2 ( 1 ) "\*" "\*" " " " "

2 ( 2 ) "\*" " " " " "\*"

2 ( 3 ) "\*" " " "\*" " "

2 ( 4 ) " " "\*" " " "\*"

3 ( 1 ) "\*" "\*" " " "\*"

3 ( 2 ) "\*" "\*" "\*" " "

3 ( 3 ) "\*" " " "\*" "\*"

3 ( 4 ) " " "\*" "\*" "\*"

4 ( 1 ) "\*" "\*" "\*" "\*"

> cbind(res.sum$which, res.sum$adjr2,res.sum$cp, res.sum$bic)

(Intercept) Z1 Z2 Z3 Z4

1 1 1 0 0 0 0.4114810 12.278381 -42.27217

1 1 0 1 0 0 0.2943177 32.839195 -25.02436

1 1 0 0 1 0 0.2242462 45.135946 -16.03070

1 1 0 0 0 1 0.1885315 51.403470 -11.75472

2 1 1 1 0 0 0.4613413 4.512029 -47.15540

2 1 1 0 0 1 0.4356500 8.972071 -42.72912

2 1 1 0 1 0 0.4126106 12.971754 -38.92785

2 1 0 1 0 1 0.3497744 23.880213 -29.27287

3 1 1 1 0 1 0.4726805 3.548490 -45.66096

3 1 1 1 1 0 0.4568464 6.267437 -42.85033

3 1 1 0 1 1 0.4319828 10.536871 -38.59819

3 1 0 1 1 1 0.3512784 24.394996 -25.97725

4 1 1 1 1 1 0.4700510 5.000000 -41.68428

From the statistics above, we see that the stepwise forward subset selection matched the highest R^2a subset in the R2adjusted graph. To further verify, we see this in the adjr plot as well and our cbind statistic reflect the same.

Our chosen subset model is: Y~Z1+Z2+Z4

> summary(pdata.lm3)

Call:

lm(formula = y ~ Z1 + Z2 + Z4, data = pdata)

Residuals:

Min 1Q Median 3Q Max

-35.350 -5.260 -0.895 4.315 118.029

Coefficients:

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

(Intercept) -26.1289 18.3298 -1.425 0.15744

Z1 1.4123 0.2981 4.737 7.94e-06 \*\*\*

Z2 14.6382 5.3592 2.731 0.00757 \*\*

Z4 4.7816 2.7707 1.726 0.08778 .

---

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

Residual standard error: 17.16 on 91 degrees of freedom

Multiple R-squared: 0.4895, Adjusted R-squared: 0.4727

F-statistic: 29.09 on 3 and 91 DF, p-value: 2.806e-13

outlierTest(pdata.lm3)

rstudent unadjusted p-value Bonferroni p

95 10.57034 1.8803e-17 1.7863e-15

We run an outlier test on the subset model, and find that 95 is an outlier for our model, so to correct the assumption we remove the point.

summary(pdata.lm3)

Call:

lm(formula = y ~ Z1 + Z2 + Z4, data = pdata2)

Residuals:

Min 1Q Median 3Q Max

-27.808 -4.930 -0.763 3.877 37.613

Coefficients:

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

(Intercept) -15.6662 12.3507 -1.268 0.20791

Z1 1.3353 0.2004 6.665 2.07e-09 \*\*\*

Z2 9.8644 3.6277 2.719 0.00785 \*\*

Z4 3.2752 1.8663 1.755 0.08268 .

---

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

Residual standard error: 11.53 on 90 degrees of freedom

Multiple R-squared: 0.5908, Adjusted R-squared: 0.5771

F-statistic: 43.31 on 3 and 90 DF, p-value: < 2.2e-16

Our final statistics for the best subset, we achieve a lower p value for both F-Statistic and T scores.

c)

> summary(p1data1)

Call:

lm(formula = y1 ~ X1 + X2 + X3 + X4)

Residuals:

Min 1Q Median 3Q Max

-45.863 -18.668 0.304 14.630 79.321

Coefficients:

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

(Intercept) 27.9160 123.2886 0.226 0.8241

X1 0.5777 0.8936 0.646 0.5285

X2 -0.1854 0.4279 -0.433 0.6714

X3 -2.2036 1.0705 -2.059 0.0587 .

X4 21.6275 13.4687 1.606 0.1306

---

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

Residual standard error: 32.14 on 14 degrees of freedom

Multiple R-squared: 0.4826, Adjusted R-squared: 0.3348

F-statistic: 3.265 on 4 and 14 DF, p-value: 0.04346

> summary(p1data2)

Call:

lm(formula = y1 ~ Z1 + Z2 + Z3 + Z4)

Residuals:

Min 1Q Median 3Q Max

-38.892 -15.798 -1.619 6.507 112.942

Coefficients: (1 not defined because of singularities)

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

(Intercept) -127.4250 105.7442 -1.205 0.247

Z1 1.6187 0.9962 1.625 0.125

Z2 NA NA NA NA

Z3 -0.9888 2.1419 -0.462 0.651

Z4 20.8823 14.6007 1.430 0.173

Residual standard error: 36.07 on 15 degrees of freedom

Multiple R-squared: 0.3014, Adjusted R-squared: 0.1617

F-statistic: 2.158 on 3 and 15 DF, p-value: 0.1357

> summary(p0data1)

Call:

lm(formula = y11 ~ X11 + X22 + X33 + X44)

Residuals:

Min 1Q Median 3Q Max

-45.863 -18.668 0.304 14.630 79.321

Coefficients:

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

(Intercept) 27.9160 123.2886 0.226 0.8241

X11 0.5777 0.8936 0.646 0.5285

X22 -0.1854 0.4279 -0.433 0.6714

X33 -2.2036 1.0705 -2.059 0.0587 .

X44 21.6275 13.4687 1.606 0.1306

---

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

Residual standard error: 32.14 on 14 degrees of freedom

Multiple R-squared: 0.4826, Adjusted R-squared: 0.3348

F-statistic: 3.265 on 4 and 14 DF, p-value: 0.04346

> summary(p0data2)

Call:

lm(formula = y11 ~ Z11 + Z22 + Z33 + Z44)

Residuals:

Min 1Q Median 3Q Max

-38.892 -15.798 -1.619 6.507 112.942

Coefficients: (1 not defined because of singularities)

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

(Intercept) -127.4250 105.7442 -1.205 0.247

Z11 1.6187 0.9962 1.625 0.125

Z22 NA NA NA NA

Z33 -0.9888 2.1419 -0.462 0.651

Z44 20.8823 14.6007 1.430 0.173

Residual standard error: 36.07 on 15 degrees of freedom

Multiple R-squared: 0.3014, Adjusted R-squared: 0.1617

F-statistic: 2.158 on 3 and 15 DF, p-value: 0.1357

For both subdata sets I would go with Model 1 because the F statistic has a signigicant P value, whereas In Model 2 this is not achieved