1.

a)

> ts<- read.csv(file.choose(), header=TRUE)

> attach(ts)

> plot(Time,Apple)



b)

> model=lm(Apple~Time)

> summary(model)

Call:

lm(formula = Apple ~ Time)

Residuals:

Min 1Q Median 3Q Max

-8.8357 -1.8845 -0.1322 2.3075 8.7195

Coefficients:

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

(Intercept) 68.827212 0.553083 124.44 <2e-16 \*\*\*

Time 0.183267 0.005711 32.09 <2e-16 \*\*\*

---

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

Residual standard error: 3.558 on 165 degrees of freedom

Multiple R-squared: 0.8619, Adjusted R-squared: 0.8611

F-statistic: 1030 on 1 and 165 DF, p-value: < 2.2e-16

#we could read the model is significant since F-pvalue of the model is small <alpha=.05

#also the R sqr and R sqr adj are both high and differ very little, thus the model is well explained by the variable listed even after a penalty taken account for the number of X variables.

Also, the Time variable is significant since p-value is < alpha =.05

c)

> T2=Time\*Time

> model1=lm(Apple~Time+T2)

> summary(model1)

Call:

lm(formula = Apple ~ Time + T2)

Residuals:

Min 1Q Median 3Q Max

-6.6616 -1.9688 0.0426 2.0819 6.4198

Coefficients:

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

(Intercept) 7.350e+01 6.830e-01 107.616 < 2e-16 \*\*\*

Time 1.747e-02 1.877e-02 0.931 0.353

T2 9.869e-04 1.082e-04 9.120 2.61e-16 \*\*\*

---

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

Residual standard error: 2.907 on 164 degrees of freedom

Multiple R-squared: 0.9084, Adjusted R-squared: 0.9073

F-statistic: 813 on 2 and 164 DF, p-value: < 2.2e-16

#we could read the model is significant since F-pvalue of the model is small <alpha=.05 still ,

#also the R sqr=.9084 and R sqr=.9073 adj are both high and differ very little, and the model1 has been improved from model (R sqr=.8619 and R sqr=.8611 ) Also, the T2 variable is significant since p-value is < alpha =.05,yet Time becomes not significant.

This is expectable since T2 could conform to the pattern of the plot we did.

d)

Thus Sep 3 is with 171 in the time period

1.747e-02 \*171 +9.869e-04 \*171\*171

We have 31.84531 as the predicted stock price

e)

No, the overall increase that looks like the linear trend shows that this is

#not stationary in the mean since the values of APPLE are increasing

#with time

#The overall spread in the points is also increasing so it is not stationary

#in the variation, It has heterocedasticity

f)

> Apple

[1] 102.50 102.25 102.13 100.89 101.54 101.32 100.58 100.57 100.53 99.16

[11] 97.98 97.50 97.24 95.97 95.99 94.74 94.48 94.49 94.65 95.12

[21] 95.65 95.13 97.66 97.89 98.53 97.19 96.55 96.71 94.25 93.48

[31] 93.96 92.63 94.31 94.85 95.97 94.75 94.57 94.92 94.88 95.50

[41] 93.56 93.02 93.06 92.47 91.52 90.45 89.91 89.83 90.38 90.46

[51] 91.41 91.72 91.62 91.74 90.83 91.83 93.40 93.78 93.24 91.77

[61] 92.02 91.66 90.63 89.36 89.98 90.32 88.70 88.93 87.30 86.32

[71] 86.19 85.96 85.94 84.94 83.70 84.42 84.40 84.27 83.23 83.58

[81] 83.73 84.03 84.95 83.77 83.61 83.42 83.73 83.98 80.85 80.26

[91] 74.18 75.16 75.09 74.21 73.37 73.22 73.74 73.45 74.00 74.97

[101] 73.99 74.00 75.18 76.16 76.70 76.57 75.87 75.89 75.98 76.30

[111] 77.04 76.22 75.33 74.74 75.10 75.12 74.46 74.17 75.01 75.86

[121] 75.78 75.05 74.98 75.03 75.25 75.10 74.60 74.39 74.59 73.13

[131] 73.80 74.57 74.25 75.08 75.96 77.18 76.90 76.96 75.76 75.76

[141] 74.78 73.46 72.45 72.03 71.49 70.47 70.34 70.23 70.37 71.17

[151] 77.36 76.73 78.15 77.50 77.15 75.97 77.88 78.32 76.78 75.28

[161] 74.89 75.39 76.37 75.89 76.43 76.02 77.73

> Apple1=Apple[-1]

> diff = Apple1-Apple[1:(length(Apple)-1)]

> diff

[1] -0.25 -0.12 -1.24 0.65 -0.22 -0.74 -0.01 -0.04 -1.37 -1.18 -0.48 -0.26

[13] -1.27 0.02 -1.25 -0.26 0.01 0.16 0.47 0.53 -0.52 2.53 0.23 0.64

[25] -1.34 -0.64 0.16 -2.46 -0.77 0.48 -1.33 1.68 0.54 1.12 -1.22 -0.18

[37] 0.35 -0.04 0.62 -1.94 -0.54 0.04 -0.59 -0.95 -1.07 -0.54 -0.08 0.55

[49] 0.08 0.95 0.31 -0.10 0.12 -0.91 1.00 1.57 0.38 -0.54 -1.47 0.25

[61] -0.36 -1.03 -1.27 0.62 0.34 -1.62 0.23 -1.63 -0.98 -0.13 -0.23 -0.02

[73] -1.00 -1.24 0.72 -0.02 -0.13 -1.04 0.35 0.15 0.30 0.92 -1.18 -0.16

[85] -0.19 0.31 0.25 -3.13 -0.59 -6.08 0.98 -0.07 -0.88 -0.84 -0.15 0.52

[97] -0.29 0.55 0.97 -0.98 0.01 1.18 0.98 0.54 -0.13 -0.70 0.02 0.09

[109] 0.32 0.74 -0.82 -0.89 -0.59 0.36 0.02 -0.66 -0.29 0.84 0.85 -0.08

[121] -0.73 -0.07 0.05 0.22 -0.15 -0.50 -0.21 0.20 -1.46 0.67 0.77 -0.32

[133] 0.83 0.88 1.22 -0.28 0.06 -1.20 0.00 -0.98 -1.32 -1.01 -0.42 -0.54

[145] -1.02 -0.13 -0.11 0.14 0.80 6.19 -0.63 1.42 -0.65 -0.35 -1.18 1.91

[157] 0.44 -1.54 -1.50 -0.39 0.50 0.98 -0.48 0.54 -0.41 1.71

plot(Time[1: 166],diff)



Yes, it is stationary in the mean as it all hovers around 0 or nearby and the variance is stationary in general,except for the outliers we read,as the y spread is pretty much the same over time

2.

a)We first build a full model:

> model=lm(claims ~ geog + area + age + age65 + age85 + trav + inc + employ + pop +hs + cd + married +dens + loc)

> summary(model)

Call:

lm(formula = claims ~ geog + area + age + age65 + age85 + trav +

inc + employ + pop + hs + cd + married + dens + loc)

Residuals:

Min 1Q Median 3Q Max

-174.079 -50.952 -9.652 65.005 202.783

Coefficients:

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

(Intercept) 1.638e+03 9.638e+02 1.700 0.098289 .

geognorth -1.342e+02 7.785e+01 -1.724 0.093861 .

geogsouth 3.218e+01 7.369e+01 0.437 0.665072

geogwest -7.811e+01 6.964e+01 -1.122 0.269909

area 3.540e-04 3.437e-04 1.030 0.310294

age -3.854e+01 1.428e+01 -2.699 0.010745 \*

age65 5.650e+01 3.024e+01 1.869 0.070299 .

age85 2.433e+01 1.303e+02 0.187 0.852977

trav 5.867e+00 1.220e+01 0.481 0.633693

inc 2.087e-02 4.889e-03 4.269 0.000149 \*\*\*

employ -1.605e+01 9.799e+00 -1.638 0.110693

pop -1.206e-06 3.780e-06 -0.319 0.751689

hs 1.590e+01 1.104e+01 1.441 0.158700

cd -9.240e+00 8.490e+00 -1.088 0.284142

married -2.337e+01 8.262e+00 -2.828 0.007791 \*\*

dens -2.143e-02 2.613e-02 -0.820 0.417716

loc -2.929e+00 2.626e+01 -0.112 0.911844

---

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

Residual standard error: 99.05 on 34 degrees of freedom

Multiple R-squared: 0.7806, Adjusted R-squared: 0.6774

F-statistic: 7.562 on 16 and 34 DF, p-value: 4.32e-07

As we found,

F-statistic is significant with such a low p-value 4.32e-07 < alpha = 0.05, which is good

s= 99.05 is hard to judge but seems to be high

R² =0.7806, medium high, this is good. This model is explaining only about 78.06 % of the variation in LOSS.

R²adj =.6774, much low, this is bad since it is less a lot after taking into account the number of X’s variables.

b)

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

plot.new()

hist(area)

hist(age)

hist(age65)

hist(age85)

hist(trav)

hist(inc)

hist(employ)

hist(pop)

hist(hs)

hist(cd)

hist(married)

hist(dens)

#I can’t get the result.

> hist(area)

#I suspect the following variables need adjustments



This is too high on the left so we don’t use log

plot(area,claims)



It shows model inequality though as Xs are not well spread

hist(age85)

> plot(age85,claims)





The above shows age85 need a log since it could be normalized





It works fine in inc,so I Ignore it





It works fine in employ too,so I Ignore it





Pop seems reasonable,since it could spread unevenly





We also need a log in hs to fix the normality





It seems work fine with the married





Dense shows a model inequality,but we could not make it too good





As we see trav need a sqr term

c)

to sum up, we need log on age85,employ,hs,and an added sqr term on trav

lnage85=log(age85)

lnemploy=log(employ)

lnhs=log(hs)

t2=trav\*trav

model1=lm(claims ~ geog + area + age + age65 + lnage85 + trav + t2+ inc + lnemploy + pop +lnhs + cd + married +dens + loc)

> summary(model1)

Call:

lm(formula = claims ~ geog + area + age + age65 + lnage85 + trav +

t2 + inc + lnemploy + pop + lnhs + cd + married + dens +

loc)

Residuals:

Min 1Q Median 3Q Max

-172.532 -51.247 -8.667 65.852 201.422

Coefficients:

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

(Intercept) 6.473e+02 4.850e+03 0.133 0.894643

geognorth -1.309e+02 8.197e+01 -1.597 0.119812

geogsouth 3.186e+01 7.882e+01 0.404 0.688704

geogwest -7.639e+01 7.321e+01 -1.043 0.304364

area 3.580e-04 3.682e-04 0.972 0.337987

age -3.854e+01 1.454e+01 -2.650 0.012254 \*

age65 5.783e+01 3.385e+01 1.708 0.097013 .

lnage85 2.068e+01 2.086e+02 0.099 0.921615

trav 1.514e+01 6.208e+01 0.244 0.808775

t2 -2.099e-01 1.206e+00 -0.174 0.862903

inc 2.102e-02 5.014e-03 4.193 0.000194 \*\*\*

lnemploy -1.173e+03 7.940e+02 -1.477 0.149176

pop -1.154e-06 3.995e-06 -0.289 0.774531

lnhs 1.373e+03 9.598e+02 1.430 0.161991

cd -8.831e+00 9.327e+00 -0.947 0.350628

married -2.341e+01 8.676e+00 -2.698 0.010903 \*

dens -2.153e-02 2.647e-02 -0.813 0.421985

loc -4.768e+00 3.068e+01 -0.155 0.877455

---

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

Residual standard error: 100.1 on 33 degrees of freedom

Multiple R-squared: 0.7825, Adjusted R-squared: 0.6704

F-statistic: 6.983 on 17 and 33 DF, p-value: 1.087e-06

#it seems changes little but F-pvalue has improved a little

e)

rstudent = rstudent(model1)

rstudent[order(rstudent)]

9 12 47 34 2 21

-3.27997343 -2.36893787 -1.97921067 -1.54579311 -1.48670618 -1.38992146

37 32 5 20 14 1

-1.09507835 -1.03891492 -1.03411390 -0.98477771 -0.85849770 -0.82586932

43 50 25 28 38 39

-0.76483888 -0.74071123 -0.59284514 -0.58862613 -0.58204505 -0.53778221

36 16 41 4 7 15

-0.52394711 -0.45611789 -0.43661122 -0.35118183 -0.16646656 -0.15685469

11 13 42 17 48 51

-0.14665087 -0.09721081 -0.07362296 -0.03197064 -0.01493347 0.12156393

3 30 35 46 10 26

0.18230037 0.20957397 0.27290541 0.34594052 0.35625341 0.41680135

18 22 45 6 29 24

0.75132909 0.84367628 0.86273850 0.96301817 0.96432042 1.02578885

8 33 44 49 27 23

1.07553346 1.15788328 1.33467446 1.36781529 1.65301980 1.72192052

31 40 19

1.77342182 1.86960356 2.46906258

> #we need to remove point 19, 9, and 12 since their absolute value is bigger than 2

> #they are outliers

> leverages = hatvalues(model1)

> leverages[order(leverages)]

43 41 1 4 16 50 15

0.1186433 0.1225331 0.1232585 0.1613045 0.1701246 0.1812276 0.1858922

36 17 32 26 47 13 19

0.2075555 0.2151250 0.2236021 0.2270922 0.2279638 0.2307046 0.2335178

22 48 28 18 42 24 23

0.2359965 0.2382780 0.2439789 0.2441450 0.2523642 0.2551294 0.2748031

14 25 31 37 39 30 8

0.2765416 0.2835921 0.2955874 0.2979161 0.3076383 0.3119938 0.3150100

7 11 34 38 35 6 46

0.3189641 0.3216887 0.3227759 0.3423136 0.3470090 0.3483737 0.3652484

20 21 12 27 40 51 49

0.3696887 0.3966979 0.3968428 0.3991427 0.4447908 0.4726516 0.4776213

3 29 33 44 10 45 5

0.4993648 0.5354731 0.5643485 0.5869298 0.6769886 0.7115469 0.7124318

2 9

0.9369373 0.9906509

K=15

N=51

> dim(dz)

[1] 51 16

> 3\*(16)/51

[1] 0.9411765

Thus, we cast out #9

f)

> model1=lm(claims ~ geog + area + age + age65 + lnage85 + trav + t2+ inc + lnemploy + pop +lnhs + cd + married +dens + loc,subset=-c(19,9,12))

> summary(model1)

Call:

lm(formula = claims ~ geog + area + age + age65 + lnage85 + trav +

t2 + inc + lnemploy + pop + lnhs + cd + married + dens +

loc, subset = -c(19, 9, 12))

Residuals:

Min 1Q Median 3Q Max

-112.199 -46.190 -5.715 43.856 154.769

Coefficients:

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

(Intercept) -1.879e+03 4.171e+03 -0.450 0.6556

geognorth -3.320e+01 6.897e+01 -0.481 0.6338

geogsouth 7.203e+01 6.323e+01 1.139 0.2636

geogwest 4.902e+01 6.471e+01 0.758 0.4547

area 6.680e-04 2.987e-04 2.237 0.0329 \*

age -2.403e+01 1.212e+01 -1.983 0.0566 .

age65 5.803e+01 2.690e+01 2.158 0.0391 \*

lnage85 -5.558e+01 1.722e+02 -0.323 0.7491

trav 5.184e+01 4.995e+01 1.038 0.3076

t2 -7.831e-01 9.619e-01 -0.814 0.4220

inc 1.329e-02 5.991e-03 2.218 0.0342 \*

lnemploy -7.170e+02 6.633e+02 -1.081 0.2883

pop -2.977e-06 3.182e-06 -0.936 0.3570

lnhs 1.208e+03 7.915e+02 1.526 0.1374

cd -5.692e+00 7.420e+00 -0.767 0.4490

married -1.520e+01 7.180e+00 -2.118 0.0426 \*

dens 2.492e-01 1.007e-01 2.475 0.0192 \*

loc -7.948e+00 2.441e+01 -0.326 0.7470

---

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

Residual standard error: 78.58 on 30 degrees of freedom

Multiple R-squared: 0.8559, Adjusted R-squared: 0.7743

F-statistic: 10.48 on 17 and 30 DF, p-value: 2.361e-08

We could see the model1 has improved a lot by F-pvalue,R sqr and R sqr adj changes

g)

backward regression:

we cast out variable via the largest p value’s variable one step at a time

> model2=lm(claims ~ geog + area + age + age65 + lnage85 + trav + t2+ inc + lnemploy + pop +lnhs + cd + married +dens,subset=-c(19,9,12))

> summary(model2)

Call:

lm(formula = claims ~ geog + area + age + age65 + lnage85 + trav +

t2 + inc + lnemploy + pop + lnhs + cd + married + dens, subset = -c(19,

9, 12))

Residuals:

Min 1Q Median 3Q Max

-109.689 -50.179 -9.104 44.277 152.211

Coefficients:

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

(Intercept) -1.404e+03 3.851e+03 -0.365 0.7179

geognorth -4.376e+01 5.999e+01 -0.729 0.4712

geogsouth 6.288e+01 5.581e+01 1.127 0.2686

geogwest 3.843e+01 5.514e+01 0.697 0.4910

area 6.791e-04 2.924e-04 2.322 0.0269 \*

age -2.518e+01 1.143e+01 -2.204 0.0351 \*

age65 6.102e+01 2.492e+01 2.449 0.0202 \*

lnage85 -6.036e+01 1.691e+02 -0.357 0.7235

trav 4.332e+01 4.192e+01 1.033 0.3095

t2 -6.237e-01 8.160e-01 -0.764 0.4505

inc 1.372e-02 5.759e-03 2.382 0.0235 \*

lnemploy -7.906e+02 6.146e+02 -1.286 0.2078

pop -2.592e-06 2.911e-06 -0.890 0.3801

lnhs 1.181e+03 7.757e+02 1.523 0.1380

cd -5.260e+00 7.195e+00 -0.731 0.4702

married -1.458e+01 6.817e+00 -2.139 0.0405 \*

dens 2.465e-01 9.890e-02 2.493 0.0182 \*

---

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

Residual standard error: 77.44 on 31 degrees of freedom

Multiple R-squared: 0.8554, Adjusted R-squared: 0.7808

F-statistic: 11.46 on 16 and 31 DF, p-value: 7.096e-09

> model3=lm(claims ~ geog + area + age + age65 + trav + t2+ inc + lnemploy + pop +lnhs + cd + married +dens,subset=-c(19,9,12))

> summary(model3)

Call:

lm(formula = claims ~ geog + area + age + age65 + trav + t2 +

inc + lnemploy + pop + lnhs + cd + married + dens, subset = -c(19,

9, 12))

Residuals:

Min 1Q Median 3Q Max

-105.248 -48.941 -4.803 44.150 153.805

Coefficients:

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

(Intercept) -1.626e+03 3.748e+03 -0.434 0.66736

geognorth -5.464e+01 5.095e+01 -1.072 0.29156

geogsouth 6.600e+01 5.437e+01 1.214 0.23361

geogwest 3.312e+01 5.237e+01 0.633 0.53153

area 7.081e-04 2.771e-04 2.556 0.01555 \*

age -2.587e+01 1.111e+01 -2.329 0.02631 \*

age65 5.456e+01 1.691e+01 3.226 0.00289 \*\*

trav 4.728e+01 3.987e+01 1.186 0.24446

t2 -6.919e-01 7.824e-01 -0.884 0.38316

inc 1.414e-02 5.562e-03 2.542 0.01607 \*

lnemploy -7.317e+02 5.839e+02 -1.253 0.21921

pop -2.818e-06 2.802e-06 -1.006 0.32196

lnhs 1.195e+03 7.640e+02 1.564 0.12767

cd -6.629e+00 6.006e+00 -1.104 0.27793

married -1.526e+01 6.451e+00 -2.366 0.02420 \*

dens 2.358e-01 9.295e-02 2.537 0.01625 \*

---

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

Residual standard error: 76.37 on 32 degrees of freedom

Multiple R-squared: 0.8548, Adjusted R-squared: 0.7868

F-statistic: 12.56 on 15 and 32 DF, p-value: 2.057e-09

> model4=lm(claims ~ area + age + age65 + trav + t2+ inc + lnemploy + pop +lnhs + cd + married +dens,subset=-c(19,9,12))

> summary(model4)

Call:

lm(formula = claims ~ area + age + age65 + trav + t2 + inc +

lnemploy + pop + lnhs + cd + married + dens, subset = -c(19,

9, 12))

Residuals:

Min 1Q Median 3Q Max

-142.630 -53.240 -0.902 64.839 134.198

Coefficients:

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

(Intercept) 2.926e+03 3.471e+03 0.843 0.40501

area 7.151e-04 2.832e-04 2.525 0.01626 \*

age -2.563e+01 1.008e+01 -2.542 0.01562 \*

age65 5.845e+01 1.777e+01 3.290 0.00229 \*\*

trav 4.466e+01 4.189e+01 1.066 0.29371

t2 -6.464e-01 8.312e-01 -0.778 0.44195

inc 1.296e-02 5.526e-03 2.345 0.02480 \*

lnemploy -1.115e+03 5.882e+02 -1.895 0.06637 .

pop -4.557e-06 2.651e-06 -1.719 0.09447 .

lnhs 4.618e+02 6.673e+02 0.692 0.49353

cd -4.037e-01 5.591e+00 -0.072 0.94285

married -1.054e+01 6.295e+00 -1.674 0.10304

dens 2.005e-01 9.270e-02 2.163 0.03748 \*

---

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

Residual standard error: 81.45 on 35 degrees of freedom

Multiple R-squared: 0.8194, Adjusted R-squared: 0.7575

F-statistic: 13.23 on 12 and 35 DF, p-value: 1.151e-09

> model5=lm(claims ~ area + age + age65 + trav + t2+ inc + lnemploy + pop +lnhs + married +dens,subset=-c(19,9,12))

> summary(model5)

Call:

lm(formula = claims ~ area + age + age65 + trav + t2 + inc +

lnemploy + pop + lnhs + married + dens, subset = -c(19, 9,

12))

Residuals:

Min 1Q Median 3Q Max

-143.226 -53.077 -0.903 64.883 136.455

Coefficients:

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

(Intercept) 3.041e+03 3.042e+03 0.999 0.32424

area 7.163e-04 2.787e-04 2.570 0.01446 \*

age -2.564e+01 9.943e+00 -2.579 0.01414 \*

age65 5.876e+01 1.701e+01 3.454 0.00143 \*\*

trav 4.444e+01 4.121e+01 1.079 0.28795

t2 -6.455e-01 8.195e-01 -0.788 0.43605

inc 1.292e-02 5.425e-03 2.382 0.02261 \*

lnemploy -1.125e+03 5.606e+02 -2.008 0.05223 .

pop -4.586e-06 2.583e-06 -1.775 0.08434 .

lnhs 4.434e+02 6.083e+02 0.729 0.47079

married -1.042e+01 5.998e+00 -1.737 0.09090 .

dens 1.989e-01 8.891e-02 2.237 0.03154 \*

---

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

Residual standard error: 80.32 on 36 degrees of freedom

Multiple R-squared: 0.8194, Adjusted R-squared: 0.7642

F-statistic: 14.84 on 11 and 36 DF, p-value: 2.885e-10

> model6=lm(claims ~ area + age + age65 + trav + t2+ inc + lnemploy + pop + married +dens,subset=-c(19,9,12))

> summary(model6)

Call:

lm(formula = claims ~ area + age + age65 + trav + t2 + inc +

lnemploy + pop + married + dens, subset = -c(19, 9, 12))

Residuals:

Min 1Q Median 3Q Max

-153.338 -53.044 -8.836 56.535 135.335

Coefficients:

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

(Intercept) 4.289e+03 2.498e+03 1.717 0.09436 .

area 7.541e-04 2.722e-04 2.771 0.00870 \*\*

age -2.444e+01 9.743e+00 -2.509 0.01662 \*

age65 6.073e+01 1.669e+01 3.639 0.00083 \*\*\*

trav 3.997e+01 4.049e+01 0.987 0.32997

t2 -5.567e-01 8.052e-01 -0.691 0.49364

inc 1.437e-02 5.015e-03 2.865 0.00683 \*\*

lnemploy -9.880e+02 5.246e+02 -1.884 0.06751 .

pop -5.123e-06 2.461e-06 -2.082 0.04433 \*

married -9.164e+00 5.709e+00 -1.605 0.11694

dens 1.777e-01 8.348e-02 2.129 0.03999 \*

---

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

Residual standard error: 79.81 on 37 degrees of freedom

Multiple R-squared: 0.8167, Adjusted R-squared: 0.7672

F-statistic: 16.49 on 10 and 37 DF, p-value: 8.775e-11

> model7=lm(claims ~ area + age + age65 + trav + inc + lnemploy + pop + married +dens,subset=-c(19,9,12))

> summary(model7)

Call:

lm(formula = claims ~ area + age + age65 + trav + inc + lnemploy +

pop + married + dens, subset = -c(19, 9, 12))

Residuals:

Min 1Q Median 3Q Max

-155.38 -47.18 -6.09 53.38 124.63

Coefficients:

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

(Intercept) 5.175e+03 2.130e+03 2.430 0.01995 \*

area 6.956e-04 2.569e-04 2.708 0.01009 \*

age -2.473e+01 9.668e+00 -2.558 0.01464 \*

age65 6.035e+01 1.656e+01 3.644 0.00080 \*\*\*

trav 1.268e+01 8.941e+00 1.418 0.16442

inc 1.497e-02 4.905e-03 3.053 0.00412 \*\*

lnemploy -1.118e+03 4.863e+02 -2.299 0.02707 \*

pop -5.119e-06 2.444e-06 -2.095 0.04290 \*

married -9.381e+00 5.661e+00 -1.657 0.10575

dens 1.677e-01 8.164e-02 2.054 0.04691 \*

---

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

Residual standard error: 79.26 on 38 degrees of freedom

Multiple R-squared: 0.8143, Adjusted R-squared: 0.7704

F-statistic: 18.52 on 9 and 38 DF, p-value: 2.454e-11

> model8=lm(claims ~ area + age + age65 + inc + lnemploy + pop + married +dens,subset=-c(19,9,12))

> summary(model8)

Call:

lm(formula = claims ~ area + age + age65 + inc + lnemploy + pop +

married + dens, subset = -c(19, 9, 12))

Residuals:

Min 1Q Median 3Q Max

-152.01 -56.71 -6.96 66.68 135.77

Coefficients:

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

(Intercept) 7.595e+03 1.290e+03 5.886 7.45e-07 \*\*\*

area 4.343e-04 1.813e-04 2.396 0.02146 \*

age -2.519e+01 9.786e+00 -2.574 0.01398 \*

age65 5.958e+01 1.677e+01 3.553 0.00101 \*\*

inc 1.963e-02 3.691e-03 5.318 4.57e-06 \*\*\*

lnemploy -1.626e+03 3.328e+02 -4.886 1.79e-05 \*\*\*

pop -3.380e-06 2.141e-06 -1.579 0.12240

married -1.207e+01 5.403e+00 -2.233 0.03135 \*

dens 1.461e-01 8.124e-02 1.798 0.07986 .

---

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

Residual standard error: 80.28 on 39 degrees of freedom

Multiple R-squared: 0.8045, Adjusted R-squared: 0.7644

F-statistic: 20.06 on 8 and 39 DF, p-value: 1.341e-11

> model9=lm(claims ~ area + age + age65 + inc + lnemploy + married +dens,subset=-c(19,9,12))

> summary(model9)

Call:

lm(formula = claims ~ area + age + age65 + inc + lnemploy + married +

dens, subset = -c(19, 9, 12))

Residuals:

Min 1Q Median 3Q Max

-166.333 -53.177 -3.056 67.070 144.800

Coefficients:

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

(Intercept) 6.643e+03 1.162e+03 5.717 1.18e-06 \*\*\*

area 4.135e-04 1.841e-04 2.246 0.03030 \*

age -1.946e+01 9.257e+00 -2.102 0.04192 \*

age65 5.089e+01 1.613e+01 3.155 0.00305 \*\*

inc 1.751e-02 3.501e-03 5.000 1.18e-05 \*\*\*

lnemploy -1.429e+03 3.142e+02 -4.548 4.93e-05 \*\*\*

married -1.050e+01 5.409e+00 -1.940 0.05941 .

dens 1.652e-01 8.182e-02 2.019 0.05020 .

---

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

Residual standard error: 81.76 on 40 degrees of freedom

Multiple R-squared: 0.792, Adjusted R-squared: 0.7556

F-statistic: 21.76 on 7 and 40 DF, p-value: 8.767e-12

> model10=lm(claims ~ area + age + age65 + inc + lnemploy +dens,subset=-c(19,9,12))

> summary(model10)

Call:

lm(formula = claims ~ area + age + age65 + inc + lnemploy + dens,

subset = -c(19, 9, 12))

Residuals:

Min 1Q Median 3Q Max

-163.64 -55.13 -14.15 62.71 188.13

Coefficients:

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

(Intercept) 6.863e+03 1.195e+03 5.744 1.00e-06 \*\*\*

area 4.596e-04 1.886e-04 2.437 0.01926 \*

age -1.568e+01 9.350e+00 -1.677 0.10111

age65 5.207e+01 1.665e+01 3.127 0.00325 \*\*

inc 1.778e-02 3.614e-03 4.919 1.45e-05 \*\*\*

lnemploy -1.646e+03 3.034e+02 -5.426 2.83e-06 \*\*\*

dens 1.901e-01 8.349e-02 2.277 0.02804 \*

---

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

Residual standard error: 84.47 on 41 degrees of freedom

Multiple R-squared: 0.7724, Adjusted R-squared: 0.7391

F-statistic: 23.19 on 6 and 41 DF, p-value: 9.817e-12

> model11=lm(claims ~ area + age65 + inc + lnemploy +dens,subset=-c(19,9,12))

> summary(model11)

Call:

lm(formula = claims ~ area + age65 + inc + lnemploy + dens, subset = -c(19,

9, 12))

Residuals:

Min 1Q Median 3Q Max

-165.71 -58.49 -10.62 56.15 207.11

Coefficients:

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

(Intercept) 6.457e+03 1.195e+03 5.403 2.85e-06 \*\*\*

area 4.296e-04 1.918e-04 2.240 0.03045 \*

age65 3.132e+01 1.139e+01 2.750 0.00875 \*\*

inc 1.575e-02 3.479e-03 4.528 4.85e-05 \*\*\*

lnemploy -1.606e+03 3.089e+02 -5.200 5.55e-06 \*\*\*

dens 2.138e-01 8.404e-02 2.544 0.01473 \*

---

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

Residual standard error: 86.28 on 42 degrees of freedom

Multiple R-squared: 0.7568, Adjusted R-squared: 0.7279

F-statistic: 26.14 on 5 and 42 DF, p-value: 6.758e-12

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 11 |  | Model 2 |  |
| R² | .7568 |  | .8554 | B |
| R² adj | .7279 |  | .7808 | B |
| s | 86.28 |  | 77.44 | B |
| F p-value | 6.758e-12 | B | 7.096e-09 |  |

B = better

I could pick model 2 upon so far, though from R2 to R 2 adj has been gapped larger there than model11

h)  
 > cor(cbind(claims , geog , area , age, age65 , age85 , trav , inc ,employ , pop , hs , cd , married , dens, loc),use="pairwise.complete.obs")

claims geog area age age65

claims 1.000000000 -0.25377961 -0.03001100 -0.010618572 -0.03812652

geog -0.253779608 1.00000000 0.26838868 -0.405631319 -0.20148551

area -0.030011001 0.26838868 1.00000000 -0.409792982 -0.58841122

age -0.010618572 -0.40563132 -0.40979298 1.000000000 0.76955315

age65 -0.038126525 -0.20148551 -0.58841122 0.769553148 1.00000000

age85 -0.007349446 -0.27744746 -0.47112919 0.662183712 0.82463132

trav 0.679712055 -0.31843917 -0.32667935 -0.015296579 -0.04753646

inc 0.489305142 -0.36414475 0.01959007 -0.002954480 -0.27704351

employ -0.284333240 -0.25432855 -0.11669010 0.187741986 0.08728170

pop 0.222651570 0.02570838 0.12688786 -0.206145037 -0.10714285

hs -0.177245319 -0.13071675 0.10783987 0.200679164 -0.02037537

cd 0.459444087 -0.42212502 -0.11257946 -0.001287920 -0.18151542

married -0.499393965 0.27285194 0.11746372 -0.065833930 -0.07977184

dens 0.430433150 -0.28043756 -0.17904795 -0.026624466 0.02215157

loc -0.400457933 0.36587478 -0.05487633 -0.001836416 -0.01646195

age85 trav inc employ pop

claims -0.007349446 0.67971205 0.48930514 -0.28433324 0.22265157

geog -0.277447455 -0.31843917 -0.36414475 -0.25432855 0.02570838

area -0.471129191 -0.32667935 0.01959007 -0.11669010 0.12688786

age 0.662183712 -0.01529658 -0.00295448 0.18774199 -0.20614504

age65 0.824631322 -0.04753646 -0.27704351 0.08728170 -0.10714285

age85 1.000000000 -0.05387221 0.03001476 0.31381860 -0.05234635

trav -0.053872207 1.00000000 0.43551739 -0.39110224 0.48245271

inc 0.030014760 0.43551739 1.00000000 0.40750763 0.13098212

employ 0.313818596 -0.39110224 0.40750763 1.00000000 -0.29298185

pop -0.052346346 0.48245271 0.13098212 -0.29298185 1.00000000

hs 0.271600526 -0.37172627 0.46178058 0.79015464 -0.36499680

cd 0.115536088 0.43860417 0.72866538 0.40355171 0.06571017

married -0.035788812 -0.41763633 0.01724437 0.25358756 -0.04439305

dens 0.087986478 0.34284535 0.14024149 -0.04142778 -0.07757156

loc -0.097301228 -0.35317650 -0.31887167 0.03375986 -0.36762823

hs cd married dens loc

claims -0.17724532 0.45944409 -0.49939396 0.43043315 -0.400457933

geog -0.13071675 -0.42212502 0.27285194 -0.28043756 0.365874783

area 0.10783987 -0.11257946 0.11746372 -0.17904795 -0.054876329

age 0.20067916 -0.00128792 -0.06583393 -0.02662447 -0.001836416

age65 -0.02037537 -0.18151542 -0.07977184 0.02215157 -0.016461950

age85 0.27160053 0.11553609 -0.03578881 0.08798648 -0.097301228

trav -0.37172627 0.43860417 -0.41763633 0.34284535 -0.353176502

inc 0.46178058 0.72866538 0.01724437 0.14024149 -0.318871666

employ 0.79015464 0.40355171 0.25358756 -0.04142778 0.033759863

pop -0.36499680 0.06571017 -0.04439305 -0.07757156 -0.367628234

hs 1.00000000 0.40159564 0.31180933 -0.08652042 0.070443398

cd 0.40159564 1.00000000 -0.41657735 0.57187024 -0.408847966

married 0.31180933 -0.41657735 1.00000000 -0.84457985 0.200510171

dens -0.08652042 0.57187024 -0.84457985 1.00000000 -0.289775527

loc 0.07044340 -0.40884797 0.20051017 -0.28977553 1.000000000

#as we read there may be some variables has high inflation factor

i)

> cor(cbind(claims , geog , area , age, age65 , age85 , trav , inc ,employ , pop , hs , cd , married , dens, loc),use="pairwise.complete.obs")

claims geog area age age65

claims 1.000000000 -0.25377961 -0.03001100 -0.010618572 -0.03812652

geog -0.253779608 1.00000000 0.26838868 -0.405631319 -0.20148551

area -0.030011001 0.26838868 1.00000000 -0.409792982 -0.58841122

age -0.010618572 -0.40563132 -0.40979298 1.000000000 0.76955315

age65 -0.038126525 -0.20148551 -0.58841122 0.769553148 1.00000000

age85 -0.007349446 -0.27744746 -0.47112919 0.662183712 0.82463132

trav 0.679712055 -0.31843917 -0.32667935 -0.015296579 -0.04753646

inc 0.489305142 -0.36414475 0.01959007 -0.002954480 -0.27704351

employ -0.284333240 -0.25432855 -0.11669010 0.187741986 0.08728170

pop 0.222651570 0.02570838 0.12688786 -0.206145037 -0.10714285

hs -0.177245319 -0.13071675 0.10783987 0.200679164 -0.02037537

cd 0.459444087 -0.42212502 -0.11257946 -0.001287920 -0.18151542

married -0.499393965 0.27285194 0.11746372 -0.065833930 -0.07977184

dens 0.430433150 -0.28043756 -0.17904795 -0.026624466 0.02215157

loc -0.400457933 0.36587478 -0.05487633 -0.001836416 -0.01646195

age85 trav inc employ pop

claims -0.007349446 0.67971205 0.48930514 -0.28433324 0.22265157

geog -0.277447455 -0.31843917 -0.36414475 -0.25432855 0.02570838

area -0.471129191 -0.32667935 0.01959007 -0.11669010 0.12688786

age 0.662183712 -0.01529658 -0.00295448 0.18774199 -0.20614504

age65 0.824631322 -0.04753646 -0.27704351 0.08728170 -0.10714285

age85 1.000000000 -0.05387221 0.03001476 0.31381860 -0.05234635

trav -0.053872207 1.00000000 0.43551739 -0.39110224 0.48245271

inc 0.030014760 0.43551739 1.00000000 0.40750763 0.13098212

employ 0.313818596 -0.39110224 0.40750763 1.00000000 -0.29298185

pop -0.052346346 0.48245271 0.13098212 -0.29298185 1.00000000

hs 0.271600526 -0.37172627 0.46178058 0.79015464 -0.36499680

cd 0.115536088 0.43860417 0.72866538 0.40355171 0.06571017

married -0.035788812 -0.41763633 0.01724437 0.25358756 -0.04439305

dens 0.087986478 0.34284535 0.14024149 -0.04142778 -0.07757156

loc -0.097301228 -0.35317650 -0.31887167 0.03375986 -0.36762823

hs cd married dens loc

claims -0.17724532 0.45944409 -0.49939396 0.43043315 -0.400457933

geog -0.13071675 -0.42212502 0.27285194 -0.28043756 0.365874783

area 0.10783987 -0.11257946 0.11746372 -0.17904795 -0.054876329

age 0.20067916 -0.00128792 -0.06583393 -0.02662447 -0.001836416

age65 -0.02037537 -0.18151542 -0.07977184 0.02215157 -0.016461950

age85 0.27160053 0.11553609 -0.03578881 0.08798648 -0.097301228

trav -0.37172627 0.43860417 -0.41763633 0.34284535 -0.353176502

inc 0.46178058 0.72866538 0.01724437 0.14024149 -0.318871666

employ 0.79015464 0.40355171 0.25358756 -0.04142778 0.033759863

pop -0.36499680 0.06571017 -0.04439305 -0.07757156 -0.367628234

hs 1.00000000 0.40159564 0.31180933 -0.08652042 0.070443398

cd 0.40159564 1.00000000 -0.41657735 0.57187024 -0.408847966

married 0.31180933 -0.41657735 1.00000000 -0.84457985 0.200510171

dens -0.08652042 0.57187024 -0.84457985 1.00000000 -0.289775527

loc 0.07044340 -0.40884797 0.20051017 -0.28977553 1.000000000

#there may be high inflation factors in the model2

> library(Rcmdr)

Loading required package: splines

Loading required package: RcmdrMisc

Loading required package: car

Loading required package: sandwich

Rcmdr Version 2.1-4

> vif(model2)

GVIF Df GVIF^(1/(2\*Df))

geog 21.993150 3 1.673842

area 5.093877 1 2.256962

age 5.184230 1 2.276891

age65 13.685086 1 3.699336

lnage85 12.446787 1 3.528000

trav 170.451480 1 13.055707

t2 141.101281 1 11.878606

inc 13.666305 1 3.696797

lnemploy 8.563757 1 2.926390

pop 2.972113 1 1.723982

lnhs 8.537268 1 2.921860

cd 9.023241 1 3.003871

married 2.550634 1 1.597070

dens 5.285861 1 2.299100

#as 10 is the cutoff

We will get rid of the following

age65 13.685086 1 3.699336

lnage85 12.446787 1 3.528000

trav 170.451480 1 13.055707

t2 141.101281 1 11.878606

inc 13.666305 1 3.696797

model12=lm(claims ~ geog + area + age + lnemploy + pop +lnhs + cd + married +dens,subset=-c(19,9,12))

> summary(model12)

Call:

lm(formula = claims ~ geog + area + age + lnemploy + pop + lnhs +

cd + married + dens, subset = -c(19, 9, 12))

Residuals:

Min 1Q Median 3Q Max

-174.38 -57.25 -19.33 32.72 237.60

Coefficients:

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

(Intercept) -9.332e+02 3.682e+03 -0.253 0.8013

geognorth -7.864e+01 6.131e+01 -1.283 0.2078

geogsouth 6.628e+00 6.440e+01 0.103 0.9186

geogwest 9.610e+00 6.064e+01 0.158 0.8750

area 3.212e-04 2.145e-04 1.497 0.1431

age -8.924e+00 1.029e+01 -0.867 0.3917

lnemploy -1.083e+03 4.984e+02 -2.172 0.0365 \*

pop 2.440e-06 2.869e-06 0.851 0.4006

lnhs 1.671e+03 9.445e+02 1.769 0.0853 .

cd -4.026e-01 6.558e+00 -0.061 0.9514

married -1.724e+01 7.647e+00 -2.254 0.0304 \*

dens 4.476e-01 8.115e-02 5.515 3.09e-06 \*\*\*

---

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

Residual standard error: 96.35 on 36 degrees of freedom

Multiple R-squared: 0.7401, Adjusted R-squared: 0.6606

F-statistic: 9.317 on 11 and 36 DF, p-value: 1.32e-07

#I will apply model13 as the best model after eliminating all the high inflation factors

Though the R sqr adj gets worse from the model2 and many individual variables become insignificant based on alpha=.05 and s gets larger because this model best predict the future data than being actifical too badly