p8130_hw5_xj2249

xj2249

12/2/2019

Problem1

```
state_df <-
    state.x77 %>%
    as.data.frame() %>%
    janitor::clean_names()
```

a) Descriptive statistics

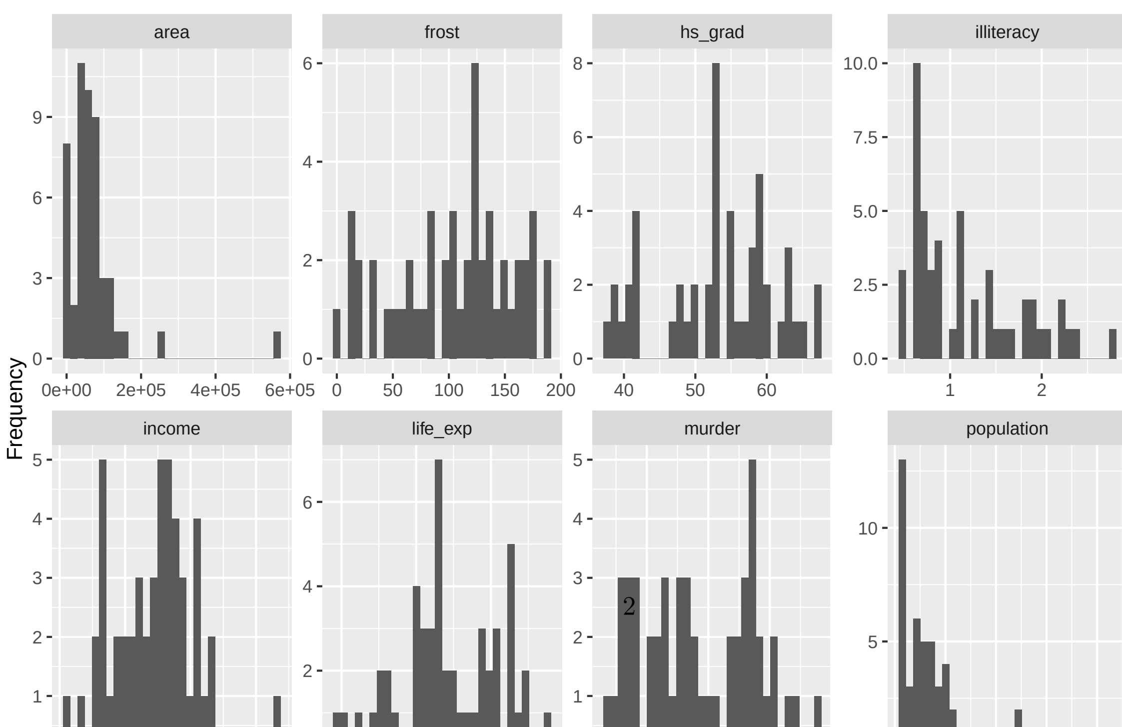
```
# descriptive statistics for variables of interest
control_table <- tableby.control(
        total = FALSE,
        test = FALSE,
        numeric.stats = c("meansd","medianq1q3","range"),
        stats.labels = list(meansd = "Mean (SD)",
                            medianq1q3 = "Median (Q1, Q3)",
                            range = "Min - Max"),
        digits = 2
        )


state_df %>%
        tableby(~.,
                data = .,
                control = control_table) %>%
        summary(text = TRUE) %>%
        kableExtra::kable(caption = "Characcteristics of patients") %>%
        kableExtra::kable_styling(latex_options = "hold_position")
```

b) Exploratory plots

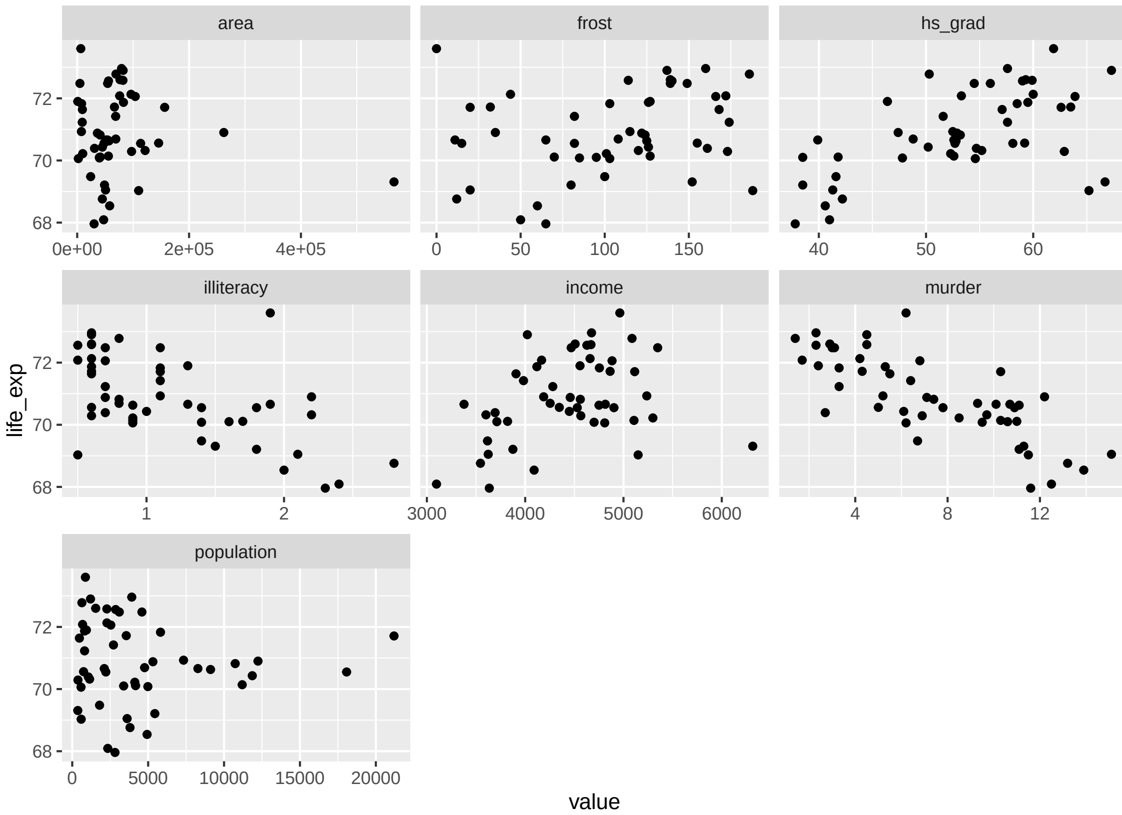
```
plot_histogram(state_df)
```

Table 1: Characcteristics of patients

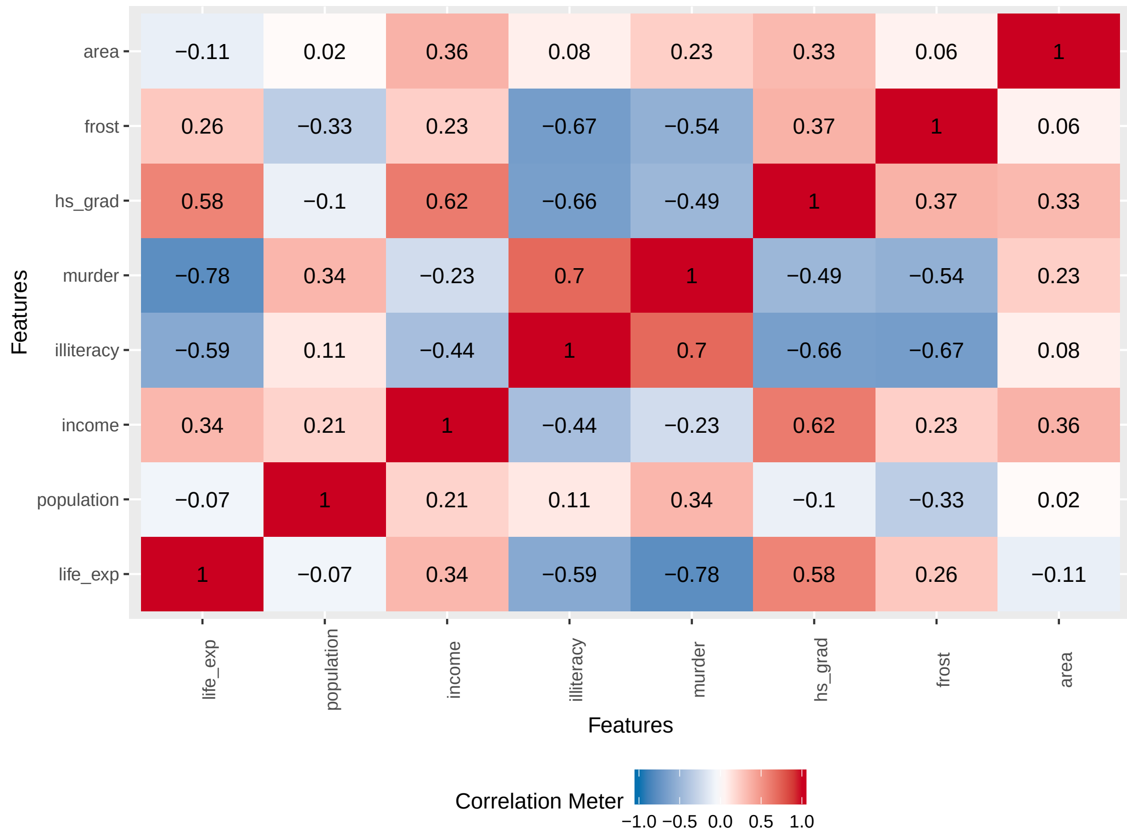
	Overall (N=50)
population	
- Mean (SD)	4246.42 (4464.49)
- Median (Q1, Q3)	2838.50 (1079.50, 4968.50)
- Min - Max	365.00 - 21198.00
income	
- Mean (SD)	4435.80 (614.47)
- Median (Q1, Q3)	4519.00 (3992.75, 4813.50)
- Min - Max	3098.00 - 6315.00
illiteracy	
- Mean (SD)	1.17 (0.61)
- Median (Q1, Q3)	0.95 (0.62, 1.58)
- Min - Max	0.50 - 2.80
life_exp	
- Mean (SD)	70.88 (1.34)
- Median (Q1, Q3)	70.67 (70.12, 71.89)
- Min - Max	67.96 - 73.60
murder	
- Mean (SD)	7.38 (3.69)
- Median (Q1, Q3)	6.85 (4.35, 10.67)
- Min - Max	1.40 - 15.10
hs_grad	
- Mean (SD)	53.11 (8.08)
- Median (Q1, Q3)	53.25 (48.05, 59.15)
- Min - Max	37.80 - 67.30
frost	
- Mean (SD)	104.46 (51.98)
- Median (Q1, Q3)	114.50 (66.25, 139.75)
- Min - Max	0.00 - 188.00
area	
- Mean (SD)	70735.88 (85327.30)
- Median (Q1, Q3)	54277.00 (36985.25, 81162.50)
- Min - Max	1049.00 - 566432.00



```r
plot_scatterplot(state_df,by = "life_exp")
```




```r
plot_correlation(state_df %>% dplyr::select(life_exp,everything()))
```



c) Automatic procedure

Backwards elimination

```
full <- lm(life_exp~.,data = state_df)
summary(full)
```

```
##
## Call:
## lm(formula = life_exp ~ ., data = state_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## population   5.180e-05  2.919e-05   1.775   0.0832 .
## income      -2.180e-05  2.444e-04  -0.089   0.9293
## illiteracy   3.382e-02  3.663e-01   0.092   0.9269
## murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## hs_grad      4.893e-02  2.332e-02   2.098   0.0420 *
## frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
```

```
## area        -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

```r
# No area
step1 <- update(full, . ~ . -area)
summary(step1)
```

```
##
## Call:
## lm(formula = life_exp ~ population + income + illiteracy + murder +
##     hs_grad + frost, data = state_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49047 -0.52533 -0.02546  0.57160  1.50374
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
## population   5.188e-05  2.879e-05   1.802   0.0785 .
## income      -2.444e-05  2.343e-04  -0.104   0.9174
## illiteracy   2.846e-02  3.416e-01   0.083   0.9340
## murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
## hs_grad      4.847e-02  2.067e-02   2.345   0.0237 *
## frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7361 on 43 degrees of freedom
## Multiple R-squared:  0.7361, Adjusted R-squared:  0.6993
## F-statistic: 19.99 on 6 and 43 DF,  p-value: 5.362e-11
```

```r
# No illiteracy
step2 <- update(step1, . ~ . -illiteracy)
summary(step2)
```

```
##
## Call:
## lm(formula = life_exp ~ population + income + murder + hs_grad +
##     frost, data = state_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4892 -0.5122 -0.0329  0.5645  1.5166
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
```

```
## population   5.115e-05  2.709e-05   1.888   0.0657 .
## income      -2.477e-05  2.316e-04  -0.107   0.9153
## murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
## hs_grad      4.776e-02  1.859e-02   2.569   0.0137 *
## frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7277 on 44 degrees of freedom
## Multiple R-squared:  0.7361, Adjusted R-squared:  0.7061
## F-statistic: 24.55 on 5 and 44 DF,  p-value: 1.019e-11
```

```r
# No income
step3 <- update(step2, . ~ . -income)
summary(step3)
```

```
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##     data = state_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## population   5.014e-05  2.512e-05   1.996  0.05201 .
## murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## hs_grad      4.658e-02  1.483e-02   3.142  0.00297 **
## frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

```r
# No population
step4 <- update(step3, . ~ . -population)
summary(step4)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost, data = state_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## murder      -0.283065   0.036731  -7.706 8.04e-10 ***
## hs_grad      0.049949   0.015201   3.286  0.00195 **
## frost       -0.006912   0.002447  -2.824  0.00699 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

The "best subset" is `murder + hs_grad + frost` for backward elimination.

Forward elimination

```
null = lm( life_exp ~ 1, data = state_df )
addterm( null, scope = full, test = "F" )
```

```
## Single term additions
##
## Model:
## life_exp ~ 1
##            Df Sum of Sq    RSS     AIC F Value     Pr(F)
## <none>                  88.299  30.435
## population  1     0.409 87.890  32.203   0.223   0.63866
## income      1    10.223 78.076  26.283   6.285   0.01562 *
## illiteracy  1    30.578 57.721  11.179  25.429 6.969e-06 ***
## murder      1    53.838 34.461 -14.609  74.989 2.260e-11 ***
## hs_grad     1    29.931 58.368  11.737  24.615 9.196e-06 ***
## frost       1     6.064 82.235  28.878   3.540   0.06599 .
## area        1     1.017 87.282  31.856   0.559   0.45815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# add murder
step1 = update(null,.~.+murder)
addterm( step1, scope = full, test = "F" )
```

```
## Single term additions
##
## Model:
## life_exp ~ murder
##            Df Sum of Sq    RSS     AIC F Value    Pr(F)
## <none>                  34.461 -14.609
## population  1    4.0161 30.445 -18.805  6.1999 0.016369 *
## income      1    2.4047 32.057 -16.226  3.5257 0.066636 .
## illiteracy  1    0.2732 34.188 -13.007  0.3756 0.542910
## hs_grad     1    4.6910 29.770 -19.925  7.4059 0.009088 **
## frost       1    3.1346 31.327 -17.378  4.7029 0.035205 *
## area        1    0.4697 33.992 -13.295  0.6494 0.424375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# add hs_grad
step2 = update(step1,.~.+ hs_grad)
addterm( step2, scope = full, test = "F" )
```

```
## Single term additions
##
## Model:
## life_exp ~ murder + hs_grad
##            Df Sum of Sq    RSS     AIC F Value    Pr(F)
## <none>                  29.770 -19.925
## population  1    3.3405 26.430 -23.877  5.8141 0.019949 *
## income      1    0.1022 29.668 -18.097  0.1585 0.692418
## illiteracy  1    0.4419 29.328 -18.673  0.6931 0.409421
## frost       1    4.3987 25.372 -25.920  7.9751 0.006988 **
## area        1    0.2775 29.493 -18.394  0.4329 0.513863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# add frost
step3 = update(step2,.~.+ frost)
addterm( step3, scope = full, test = "F" )
```

```
## Single term additions
##
## Model:
## life_exp ~ murder + hs_grad + frost
##            Df Sum of Sq    RSS     AIC F Value   Pr(F)
## <none>                  25.372 -25.920
## population  1   2.06358 23.308 -28.161  3.9841 0.05201 .
## income      1   0.18232 25.189 -24.280  0.3257 0.57103
## illiteracy  1   0.17184 25.200 -24.259  0.3069 0.58236
## area        1   0.02573 25.346 -23.970  0.0457 0.83173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(step3)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost, data = state_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## murder      -0.283065   0.036731  -7.706 8.04e-10 ***
## hs_grad      0.049949   0.015201   3.286  0.00195 **
## frost       -0.006912   0.002447  -2.824  0.00699 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

The "best subset" is `murder + hs_grad + frost` for forward elimination.

Stepwise selection

```r
step(full, direction = 'both')
```

```
## Start:  AIC=-22.18
## life_exp ~ population + income + illiteracy + murder + hs_grad +
##     frost + area
##
##              Df Sum of Sq    RSS     AIC
## - area        1    0.0011 23.298 -24.182
## - income      1    0.0044 23.302 -24.175
## - illiteracy  1    0.0047 23.302 -24.174
## <none>                    23.297 -22.185
## - population  1    1.7472 25.044 -20.569
## - frost       1    1.8466 25.144 -20.371
## - hs_grad     1    2.4413 25.738 -19.202
## - murder      1   23.1411 46.438  10.305
##
## Step:  AIC=-24.18
## life_exp ~ population + income + illiteracy + murder + hs_grad +
##     frost
##
##              Df Sum of Sq    RSS     AIC
## - illiteracy  1    0.0038 23.302 -26.174
## - income      1    0.0059 23.304 -26.170
## <none>                    23.298 -24.182
## - population  1    1.7599 25.058 -22.541
## + area        1    0.0011 23.297 -22.185
## - frost       1    2.0488 25.347 -21.968
## - hs_grad     1    2.9804 26.279 -20.163
## - murder      1   26.2721 49.570  11.569
##
## Step:  AIC=-26.17
## life_exp ~ population + income + murder + hs_grad + frost
##
##              Df Sum of Sq    RSS     AIC
## - income      1     0.006 23.308 -28.161
## <none>                    23.302 -26.174
## - population  1     1.887 25.189 -24.280
## + illiteracy  1     0.004 23.298 -24.182
## + area        1     0.000 23.302 -24.174
## - frost       1     3.037 26.339 -22.048
## - hs_grad     1     3.495 26.797 -21.187
## - murder      1    34.739 58.041  17.456
```

```
##
## Step:  AIC=-28.16
## life_exp ~ population + murder + hs_grad + frost
##
##              Df Sum of Sq    RSS     AIC
## <none>                     23.308 -28.161
## + income      1     0.006 23.302 -26.174
## + illiteracy  1     0.004 23.304 -26.170
## + area        1     0.001 23.307 -26.163
## - population  1     2.064 25.372 -25.920
## - frost       1     3.122 26.430 -23.877
## - hs_grad     1     5.112 28.420 -20.246
## - murder      1    34.816 58.124  15.528
```

```
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##     data = state_df)
##
## Coefficients:
## (Intercept)   population       murder      hs_grad        frost
##   7.103e+01    5.014e-05   -3.001e-01    4.658e-02   -5.943e-03
```

The “best subset” is `population + murder + hs_grad + frost` for stepwise selection.

```r
model_back <-  lm(life_exp~murder + hs_grad + frost,data = state_df)
summary(model_back)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost, data = state_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## murder      -0.283065   0.036731  -7.706 8.04e-10 ***
## hs_grad      0.049949   0.015201   3.286  0.00195 **
## frost       -0.006912   0.002447  -2.824  0.00699 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

```r
model_step <-  lm(life_exp~murder + hs_grad + frost + population,data = state_df)
summary(model_step)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost + population,
##     data = state_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## hs_grad      4.658e-02  1.483e-02   3.142  0.00297 **
## frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
## population   5.014e-05  2.512e-05   1.996  0.05201 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

- The automatic procedures do not necessarily generate the same model. In this case, backwards and forward elimination generate the same model, whereas stepwise selection generate a different one.
- The variable `population` is a close call and I decide to keep it. After adding `population`, the adjusted R-squared increase from 0.6939 to 0.7126. The larger model have a better predictive ability, and because our goal is a predictive model, it's better to keep `population` in the model.

```r
cor(state_df[,3],state_df[,6])
```

```
## [1] -0.6571886
```

- The is a moderate correlation between `Illiteracy` and `HS graduation rate`. Only `HS graduation rate` is contained in the subset.

d) criterion-based procedures

```r
best <- function(model, ...)
{
  subsets <- regsubsets(formula(model), model.frame(model), ...)
  subsets <- with(summary(subsets),
                  cbind(p = as.numeric(rownames(which)), which, rss, rsq, adjr2, cp, bic))
  return(subsets)
}

best(full) %>% kableExtra::kable() %>% kableExtra::kable_styling(latex_options = "scale_down")
```

p	(Intercept)	population	income	illiteracy	murder	hs_grad	frost	area	rss	rsq	adjr2	cp	bic
1	1	0	0	0	1	0	0	0	34.46133	0.6097201	0.6015893	16.126760	-39.22051
2	1	0	0	0	1	1	0	0	29.77036	0.6628461	0.6484991	9.669894	-42.62472
3	1	0	0	0	1	1	1	0	25.37162	0.7126624	0.6939230	3.739878	-46.70678
4	1	1	0	0	1	1	1	0	23.30804	0.7360328	0.7125690	2.019659	-47.03640
5	1	1	1	0	1	1	1	0	23.30198	0.7361014	0.7061129	4.008737	-43.13738
6	1	1	1	1	1	1	1	0	23.29822	0.7361440	0.6993268	6.001959	-39.23342
7	1	1	1	1	1	1	1	1	23.29714	0.7361563	0.6921823	8.000000	-35.32373

The "best subset" is `population + murder + hs_grad + frost` for stepwise selection.

e) criterion-based procedures

Actually, the prefered model from c) and d) is the same, and the model comparison is in c). The final model is `life_exp = murder + hs_grad + frost + population`. ### leverage & influential points

```r
final_model <- lm(life_exp~murder + hs_grad + frost + population,data = state_df)
influence <- influence.measures(final_model)
summary(influence)
```

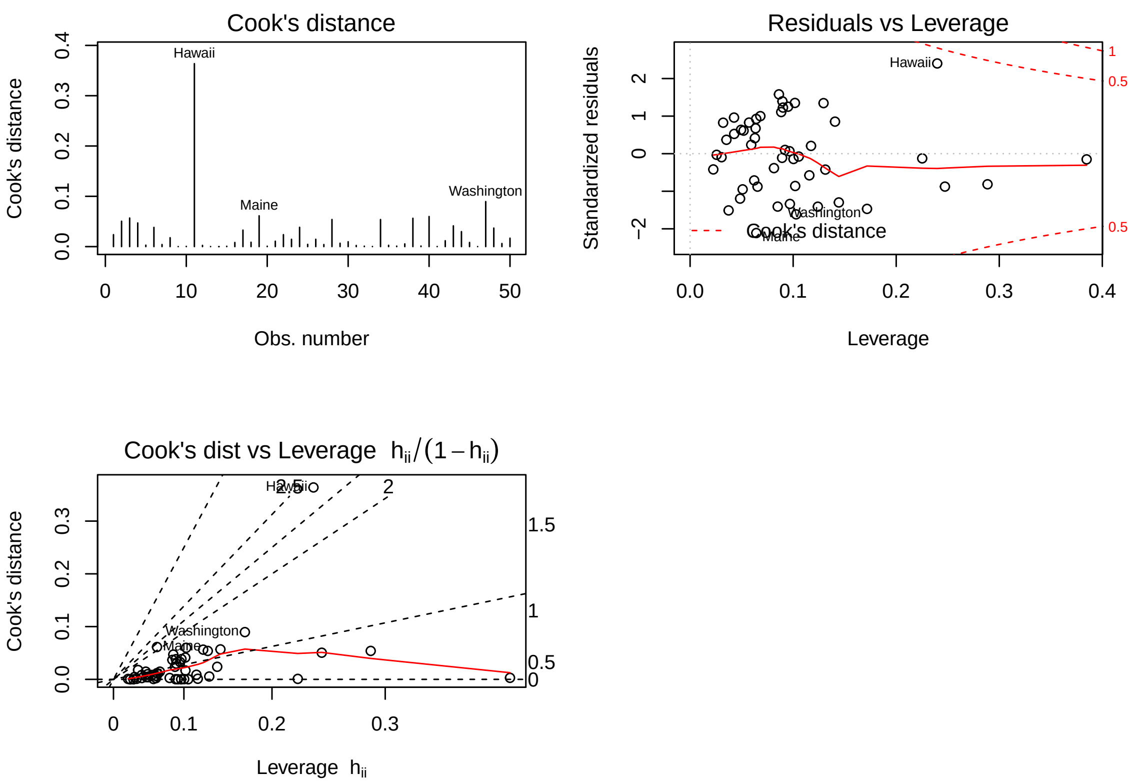
```
## Potentially influential observations of
##   lm(formula = life_exp ~ murder + hs_grad + frost + population,      data = state_df) :
##
##            dfb.1_ dfb.mrdr dfb.hs_g dfb.frst dfb.pplt dffit   cov.r
## Alaska      0.41  -0.40    -0.35    -0.16     0.18    -0.50    1.36_*
## California  0.04   0.00    -0.04     0.03    -0.09    -0.12    1.81_*
## Hawaii     -0.03  -0.28     0.66    -1.24_*  -0.57     1.43_*  0.74
## Nevada      0.40  -0.42    -0.29    -0.28     0.14    -0.52    1.46_*
## New York    0.01   0.00     0.00    -0.01    -0.06    -0.07    1.44_*
##            cook.d hat
## Alaska      0.05   0.25
## California  0.00   0.38_*
## Hawaii      0.36   0.24
## Nevada      0.05   0.29
## New York    0.00   0.23
```

```r
hatvalues(final_model)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##     0.14061825     0.24727915     0.14434012     0.08623296     0.38475924
##       Colorado    Connecticut       Delaware        Florida        Georgia
##     0.08960146     0.04944598     0.03735911     0.09648760     0.10033898
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##     0.23979244     0.04280306     0.10541465     0.02574946     0.05932553
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##     0.04264019     0.09506497     0.11572004     0.06424817     0.02251734
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##     0.06542733     0.08844258     0.06818938     0.09685602     0.03207145
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##     0.04851763     0.05189556     0.28860921     0.06221607     0.05097477
##     New Mexico       New York North Carolina   North Dakota           Ohio
##     0.06286777     0.22522744     0.08927508     0.12949804     0.08138412
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
```

```
##      0.03526037     0.13125063     0.12395238     0.11735640     0.10289140
##    South Dakota      Tennessee          Texas           Utah        Vermont
##      0.09208789     0.06417731     0.10172016     0.09012184     0.05722013
##        Virginia     Washington  West Virginia      Wisconsin        Wyoming
##      0.03054924     0.17168830     0.08498652     0.06355888     0.10198735
```

```r
par(mfrow = c(2, 2))
plot(final_model,c(4,5,6))
```



Moderate leverages are: Alaska, California, Hawaii, Nevada and New York. Hawaii could be a influential point, withh dffit > 1 but cook's distance < 0.5. Therefore, we can fit the model with and without Hawaii, and see the change.

```r
model_no_hawaii <- lm(life_exp~murder + hs_grad + frost + population,
                      data = state_df[(row.names(state_df) != "Hawaii"), ])
summary(model_no_hawaii)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost + population,
##     data = state_df[(row.names(state_df) != "Hawaii"), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48967 -0.50158  0.01999  0.54355  1.11810
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  7.106e+01  8.998e-01  78.966  < 2e-16 ***
## murder      -2.906e-01  3.477e-02  -8.357 1.24e-10 ***
## hs_grad      3.728e-02  1.447e-02   2.576   0.0134 *
## frost       -3.099e-03  2.545e-03  -1.218   0.2297
## population   6.363e-05  2.431e-05   2.618   0.0121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6796 on 44 degrees of freedom
## Multiple R-squared:  0.7483, Adjusted R-squared:  0.7254
## F-statistic: 32.71 on 4 and 44 DF,  p-value: 1.15e-12
```

```r
summary(final_model)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost + population,
##     data = state_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## hs_grad      4.658e-02  1.483e-02   3.142  0.00297 **
## frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
## population   5.014e-05  2.512e-05   1.996  0.05201 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

```r
(model_no_hawaii$coefficients-final_model$coefficients)/final_model$coefficients
```

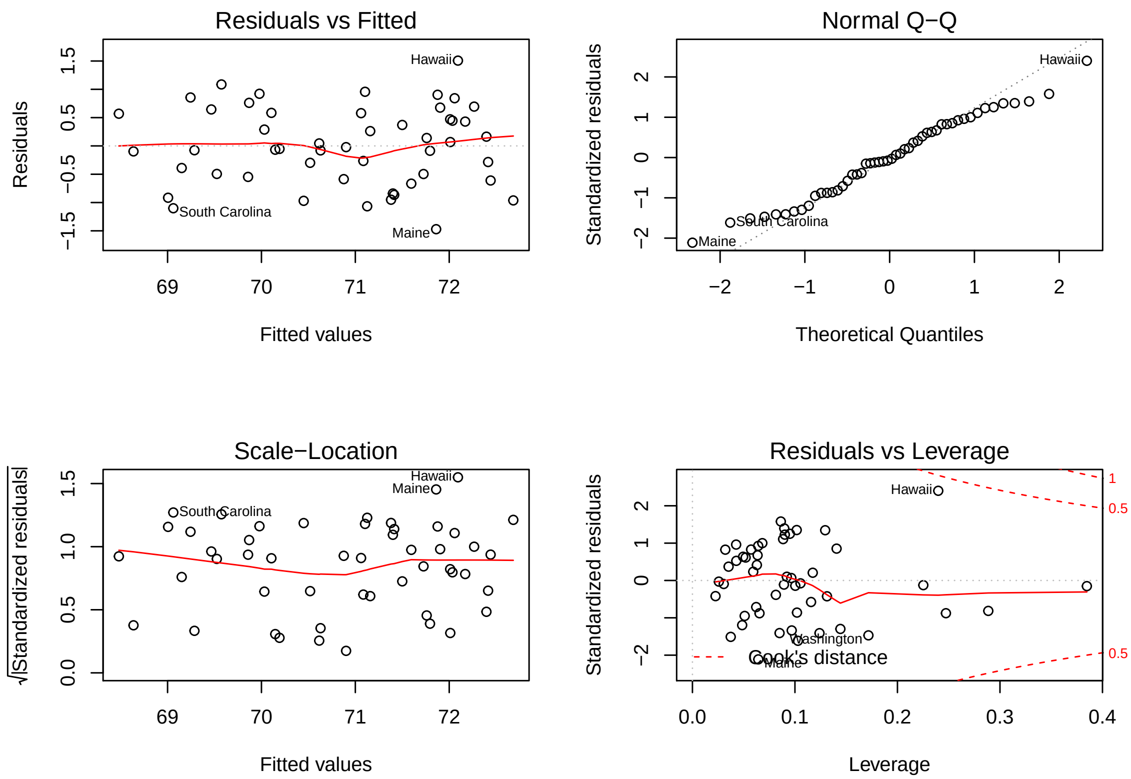
```
##   (Intercept)        murder       hs_grad         frost    population
##  0.0004181555 -0.0317796965 -0.1997080752 -0.4784953062  0.2689780916
```

As we can see, after removal of "Hawaii" some coefficients change greatly in magnitude, including `hs_grad`,`frost` and `population`(up to 20% and more).

Since we have no way to know if the data for "Hawaii" is reliable, we can not just remove casually. Therefore, we may report the results with and without "Hawaii" in the model.

Model assumptions

```r
par(mfrow = c(2, 2))
plot(final_model)
```



- Constant variance: the "residual vs fitted" and "scale-location" plots suggest a constant variance.
- Normality: Points fall along a line in the middle of the graph, but curve off at two ends.

Cross validation

Test the model predictive ability using a 10-fold cross-validation (10 repeats).

```r
train_ctr <- trainControl(method = "repeatedcv", number = 10, repeats = 10)

# Fit the 4-variables model that we discussed in previous lectures
model_cv <- train(life_exp ~ murder + hs_grad + frost + population,
                  data = state_df,
                  trControl = train_ctr,
                  method = 'lm')
model_cv$results
```

```
##   intercept      RMSE  Rsquared       MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 0.7404084 0.7420079 0.6313099 0.1998243  0.1939375 0.1826762
```

The R-squared is 0.77 and RMSE is 0.75, which indicates the model has a good predictive ability.

f) Summary

In summary, the model with predictor `population`, `murder` ,`hs_grad` and `frost` is our final model and it has a good predictive ability overall.

Problem2

```
com_df <-
    read_csv("./hw5/CommercialProperties.csv") %>%
    janitor::clean_names()

com_df %>% view()
```

a) Model with all variables

```
full_model <- lm(rental_rate ~.,data = com_df)
summary(full_model)
```

```
##
## Call:
## lm(formula = rental_rate ~ ., data = com_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.220e+01  5.780e-01  21.110  < 2e-16 ***
## age          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## taxes         2.820e-01  6.317e-02   4.464 2.75e-05 ***
## vacancy_rate  6.193e-01  1.087e+00   0.570     0.57
## sq_footage    7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

```
full_model$terms
```

```
## rental_rate ~ age + taxes + vacancy_rate + sq_footage
## attr(,"variables")
## list(rental_rate, age, taxes, vacancy_rate, sq_footage)
## attr(,"factors")
##              age taxes vacancy_rate sq_footage
## rental_rate    0     0            0          0
```

```
## age            1     0            0           0
## taxes          0     1            0           0
## vacancy_rate   0     0            1           0
## sq_footage     0     0            0           1
## attr(,"term.labels")
## [1] "age"          "taxes"        "vacancy_rate" "sq_footage"
## attr(,"order")
## [1] 1 1 1 1
## attr(,"intercept")
## [1] 1
## attr(,"response")
## [1] 1
## attr(,".Environment")
## <environment: R_GlobalEnv>
## attr(,"predvars")
## list(rental_rate, age, taxes, vacancy_rate, sq_footage)
## attr(,"dataClasses")
##  rental_rate          age        taxes vacancy_rate   sq_footage
##    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"
```

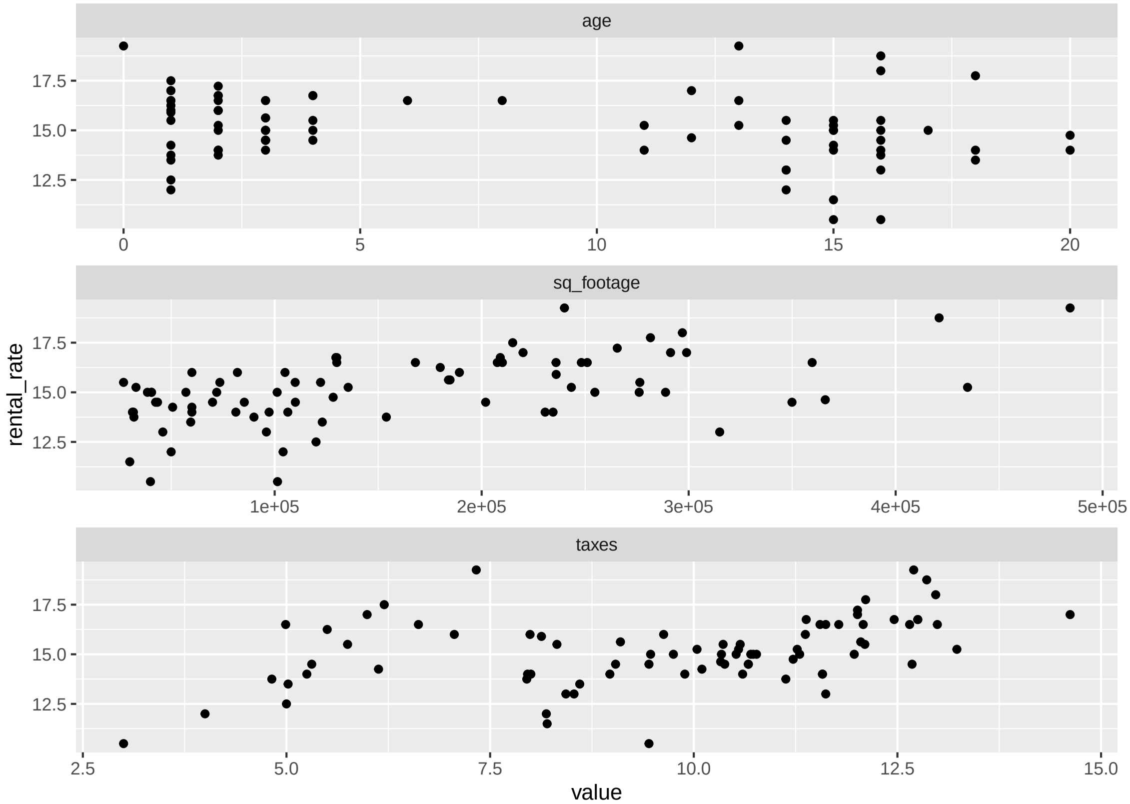
- `age`, `taxes`, and `sq_footage` are significant predictors whereas `vacancy_rate` is a non-significant predictor.
- According to overall F test, p-value< 0.001, at a significance level of 0.05, we reject H_0 and conclude that there is a linear relationship between rental rate and the set of all variables.
- The R-squared is 0.5847, suggesting the a poor performance of overall fit.

b) Scatter plot

```r
dev.off
```

```
## function (which = dev.cur())
## {
##     if (which == 1)
##         stop("cannot shut down device 1 (the null device)")
##     .External(C_devoff, as.integer(which))
##     dev.cur()
## }
## <bytecode: 0x7f9c9c5b5000>
## <environment: namespace:grDevices>
```

```r
plot_scatterplot(data = com_df[,-4], by = "rental_rate", ncol = 1)
```



comment???

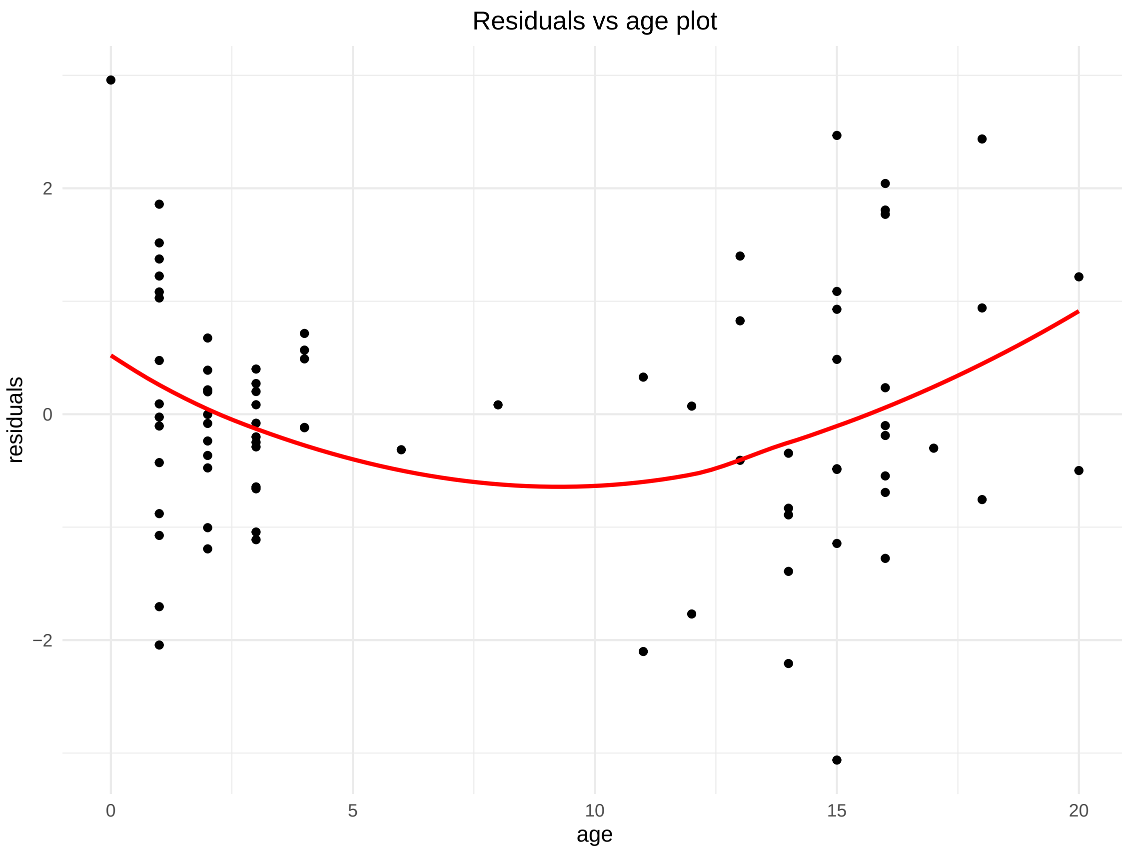
c) Model with significant predictors

```
sig_model <- lm(rental_rate ~.,data =  com_df[,-4])
```

d) Model with significant predictors

Higher order term

```
com_df %>%
    mutate(residuals = residuals(sig_model)) %>%
    ggplot(aes(y = residuals, x = age)) +
    geom_point() +
    geom_smooth(aes(y = residuals),se = F,color = "red") +
    labs(title = "Residuals vs age plot")
```



The residuals vs age plots shows a concave curve so we may use fit age with a quadratic term.

```
quartfit_age <- lm(rental_rate ~age + I(age^2) + taxes + sq_footage , data = com_df)
vif(quartfit_age)
```

```
##        age   I(age^2)      taxes sq_footage
##  34.673257  32.956178   1.532560   1.268814
```

```
summary(quartfit_age)
```

```
##
## Call:
## lm(formula = rental_rate ~ age + I(age^2) + taxes + sq_footage,
##     data = com_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89596 -0.62547 -0.08907  0.62793  2.68309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.249e+01  4.805e-01  26.000  < 2e-16 ***
## age         -4.043e-01  1.089e-01  -3.712  0.00039 ***
## I(age^2)     1.415e-02  5.821e-03   2.431  0.01743 *
## taxes        3.140e-01  5.880e-02   5.340 9.33e-07 ***
## sq_footage   8.046e-06  1.267e-06   6.351 1.42e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 76 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.5927
## F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

The vif of age and age^2 is very large so we should center age.

Let's fit the model with centerd age.

```r
center_df = mutate(com_df, center_age = age-mean(age))
quartfit_centerage <- lm(rental_rate ~ center_age + I(center_age^2)+ taxes + sq_footage , data = center_

vif(quartfit_centerage)
```

```
##      center_age I(center_age^2)           taxes      sq_footage
##        1.901945        1.608797        1.532560        1.268814
```

```r
summary(quartfit_centerage )
```

```
##
## Call:
## lm(formula = rental_rate ~ center_age + I(center_age^2) + taxes +
##     sq_footage, data = center_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89596 -0.62547 -0.08907  0.62793  2.68309
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.019e+01  6.709e-01  15.188  < 2e-16 ***
## center_age      -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
## I(center_age^2)  1.415e-02  5.821e-03   2.431   0.0174 *
## taxes            3.140e-01  5.880e-02   5.340 9.33e-07 ***
## sq_footage       8.046e-06  1.267e-06   6.351 1.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 76 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.5927
## F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

Piecewise linear model

```r
com_df_nonlin <-
    com_df %>%
    mutate(knot = (age - 10)*(age >= 10))
piecewise_age <- lm(rental_rate ~ age + knot + taxes + sq_footage , data = com_df_nonlin)
```

I choose age=10 as the knot, because it seems to be a truning point. When age<10, with the increase of age, y has a increasing trend, while after age >10, y has a decreasing trend.

Model comparison

```r
summary(quartfit_centerage)
```

```
##
## Call:
## lm(formula = rental_rate ~ center_age + I(center_age^2) + taxes +
##     sq_footage, data = center_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89596 -0.62547 -0.08907  0.62793  2.68309
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.019e+01  6.709e-01  15.188  < 2e-16 ***
## center_age      -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
## I(center_age^2)  1.415e-02  5.821e-03   2.431   0.0174 *
## taxes            3.140e-01  5.880e-02   5.340 9.33e-07 ***
## sq_footage       8.046e-06  1.267e-06   6.351 1.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 76 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.5927
## F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

```r
summary(piecewise_age)
```

```
##
## Call:
## lm(formula = rental_rate ~ age + knot + taxes + sq_footage, data = com_df_nonlin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9321 -0.6387 -0.0901  0.6188  2.6443
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.238e+01  4.787e-01  25.866  < 2e-16 ***
## age         -2.865e-01  6.330e-02  -4.526 2.18e-05 ***
## knot         3.261e-01  1.374e-01   2.374   0.0201 *
## taxes        3.036e-01  5.772e-02   5.260 1.29e-06 ***
## sq_footage   8.373e-06  1.270e-06   6.591 5.13e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.099 on 76 degrees of freedom
## Multiple R-squared:  0.6118, Adjusted R-squared:  0.5913
## F-statistic: 29.94 on 4 and 76 DF,  p-value: 5.89e-15
```

The two models have very similar R^2 and adjusted R^2. And piecewise model is much easier to interpret so I would recommend the piecewise model.

e) Model comparision

```r
rbind(broom::glance(sig_model),broom::glance(piecewise_age)) %>%
    mutate(model = c("non-piecewise model","piecewise model")) %>%
    dplyr::select(model,everything(),-c(sigma,logLik,deviance,df.residual)) %>%
    kableExtra::kable(digits = 3)
```

model	r.squared	adj.r.squared	statistic	p.value	df	AIC	BIC
non-piecewise model	0.583	0.567	35.88	0	4	255.836	267.808
piecewise model	0.612	0.591	29.94	0	5	252.041	266.408

```r
# try cross validation
non_piecewiese_cv <-
    train( rental_rate ~ ., data = com_df[, -4],
           trControl = train_ctr,
           method = 'lm')

piecewiese_cv <-
    train(rental_rate ~ age + knot + taxes + sq_footage, data = com_df_nonlin,
          trControl = train_ctr,
          method = 'lm')
```