model selection

Eleanor Zhang

12/16/2018

read data and select variables to import

```
cancer_data <- read_csv("./data/Cancer_Registry.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   binnedInc = col_character(),
##   Geography = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
cancer_county <- cancer_data %>%
  janitor::clean_names() %>%
  separate(geography, into = c("county", "state"), sep = ", ") %>%
  dplyr::select(target_death_rate, incidence_rate, med_income, poverty_percent, median_age:median_age_f
  dplyr::select(-pct_hs25_over, -pct_bach_deg25_over, -pct_employed16_over, -percent_married) %>%
  mutate(pct_upto_hs18_24 = pct_no_hs18_24 + pct_hs18_24,
         pct_above_hs18_24 = 100 - pct_upto_hs18_24,
         pct_with_coverage = pct_private_coverage + pct_public_coverage_alone,
         income_cat = ifelse(med_income < 35000, 0, 1)) %>%
  dplyr::select(-(pct_no_hs18_24:pct_bach_deg18_24), -pct_above_hs18_24, -(pct_private_coverage:pct_publ
  na.omit

dim(cancer_county)
```

```
## [1] 3047   17
```

```
anyNA(cancer_county)
```

```
## [1] FALSE
```

Data description:

- **target_death_rate:** mean per capita (100,000) cancer mortalities (a)
- **avg_ann_count:** mean number of reported cases of cancer diagnosed annually (a)
- **avg_deaths_per_year:** mean number of reported mortalities due to cancer (a)
- **incidence_rate:** mean per capita (100,000) cancer diagnoses (a)
- **med_income:** median income per county (b)
- **pop_est2015:** population of county (b)
- **poverty_percent:** percent of population in poverty (b)
- **study_per_cap** per capita number of cancer-related clinical trials per county (a)
- **binned_inc:** median income per capita binned by decile (b)
- **median_age:** median age of county residents (b)
- **median_age_male:** median age of male county residents (b)
- **median_age_female:** median age of female county residents (b)
- **geography:** county name (b)

- **avg_household_size:** mean household size of county (b)
- **percent_married:** percent of county residents who are married (b)
- **pct_no_hs18_24:** percent of county residents ages 18-24 highest education attained: less than high school (b)
- **pct_hs18_24:** percent of county residents ages 18-24 highest education attained: high school diploma (b)
- **pct_some_col18_24:** percent of county residents ages 18-24 highest education attained: some college (b)
- **pct_bach_deg18_24:** percent of county residents ages 18-24 highest education attained: bachelor's degree (b)
- **pct_hs25_over:** percent of county residents ages 25 and over highest education attained: high school diploma (b)

- **pct_bach_deg25_over:** percent of county residents ages 25 and over highest education attained: bachelor's degree (b)
- **pct_employed16_over:** percent of county residents ages 16 and over employed (b)

- **pct_unemployed16_over:** percent of county residents ages 16 and over unemployed (b)

- **pct_private_coverage:** percent of county residents with private health coverage (b)
- **pct_private_coverage_alone:** percent of county residents with private health coverage alone (no public assistance) (b)
- **pct_emp_priv_coverage:** percent of county residents with employee-provided private health coverage (b)

- **pct_public_coverage:** percent of county residents with government-provided health coverage (b)
- **pct_public_coverage_alone:** percent of county residents with government-provided health coverage alone (b)

- **pct_white:** percent of county residents who identify as White (b)

- **pct_black:** percent of county residents who identify as Black (b)
- **pct_asian:** percent of county residents who identify as Asian (b)

- **pct_other_race:** percent of county residents who identify in a category which is not White, Black, or Asian (b)

- **pct_married_households:** percent of married households (b)
- **birth_rate:** number of live births relative to number of women in county (b)

Separate entire dataset into two groups: low income and high income:

```
income_low_data <- cancer_county %>% filter(income_cat == 0) %>% dplyr::select(-income_cat)
income_high_data <- cancer_county %>% filter(income_cat == 1) %>% dplyr::select(-income_cat)
```

Description

```
summary(income_low_data)
```

```
##  target_death_rate incidence_rate    med_income    poverty_percent
##  Min.   : 66.3     Min.   :211.1   Min.   :22640   Min.   :17.60
##  1st Qu.:181.6     1st Qu.:404.7   1st Qu.:30467   1st Qu.:23.50
##  Median :202.3     Median :453.9   Median :32458   Median :26.50
##  Mean   :201.0     Mean   :446.7   Mean   :31818   Mean   :27.68
##  3rd Qu.:224.0     3rd Qu.:492.4   3rd Qu.:33948   3rd Qu.:30.77
##  Max.   :292.5     Max.   :651.3   Max.   :34991   Max.   :47.40
```

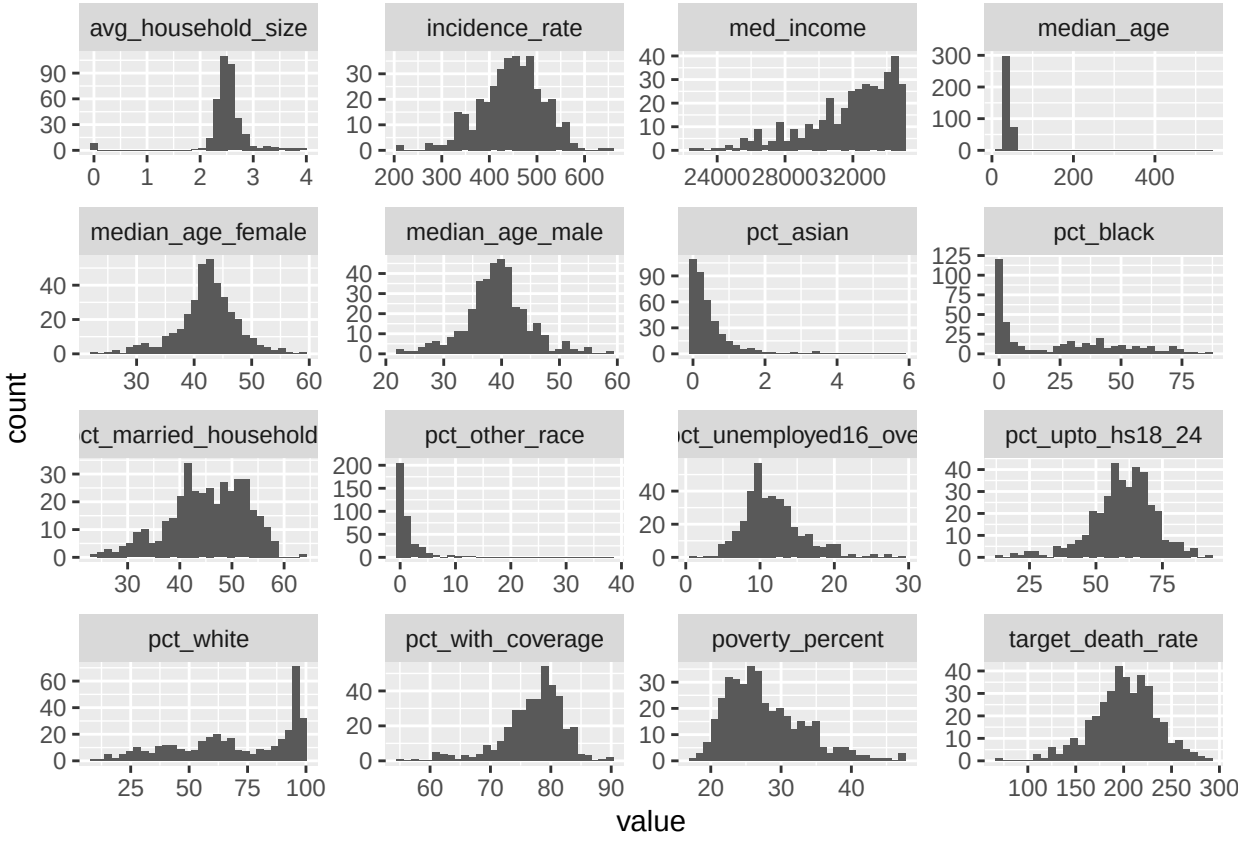
```
##    median_age     median_age_male median_age_female avg_household_size
##  Min.   : 22.30   Min.   :22.4    Min.   :22.30     Min.   :0.0243
##  1st Qu.: 38.40   1st Qu.:36.1    1st Qu.:40.00     1st Qu.:2.4000
##  Median : 40.75   Median :39.2    Median :42.60     Median :2.5100
##  Mean   : 44.03   Mean   :39.2    Mean   :42.29     Mean   :2.5138
##  3rd Qu.: 43.40   3rd Qu.:42.0    3rd Qu.:45.10     3rd Qu.:2.6275
##  Max.   :536.40   Max.   :58.6    Max.   :58.70     Max.   :3.9700
##  pct_unemployed16_over   pct_white       pct_black         pct_asian
##  Min.   : 1.200        Min.   :10.20   Min.   : 0.0000   Min.   :0.00000
##  1st Qu.: 9.125        1st Qu.:50.64   1st Qu.: 0.8307   1st Qu.:0.07866
##  Median :11.100        Median :69.97   Median :12.8157   Median :0.26847
##  Mean   :11.747        Mean   :69.60   Mean   :23.3729   Mean   :0.47431
##  3rd Qu.:13.800        3rd Qu.:94.89   3rd Qu.:41.8741   3rd Qu.:0.56603
##  Max.   :29.400        Max.   :99.85   Max.   :85.9478   Max.   :5.81482
##  pct_other_race    pct_married_households pct_upto_hs18_24
##  Min.   : 0.0000   Min.   :22.99          Min.   :14.20
##  1st Qu.: 0.1602   1st Qu.:40.74          1st Qu.:53.65
##  Median : 0.5722   Median :45.69          Median :60.65
##  Mean   : 1.8305   Mean   :45.24          Mean   :60.16
##  3rd Qu.: 1.6440   3rd Qu.:51.08          3rd Qu.:67.70
##  Max.   :37.8590   Max.   :63.16          Max.   :93.00
##  pct_with_coverage
##  Min.   :54.60
##  1st Qu.:74.60
##  Median :78.10
##  Mean   :77.14
##  3rd Qu.:80.58
##  Max.   :89.30
```

```
summary(income_high_data)
```

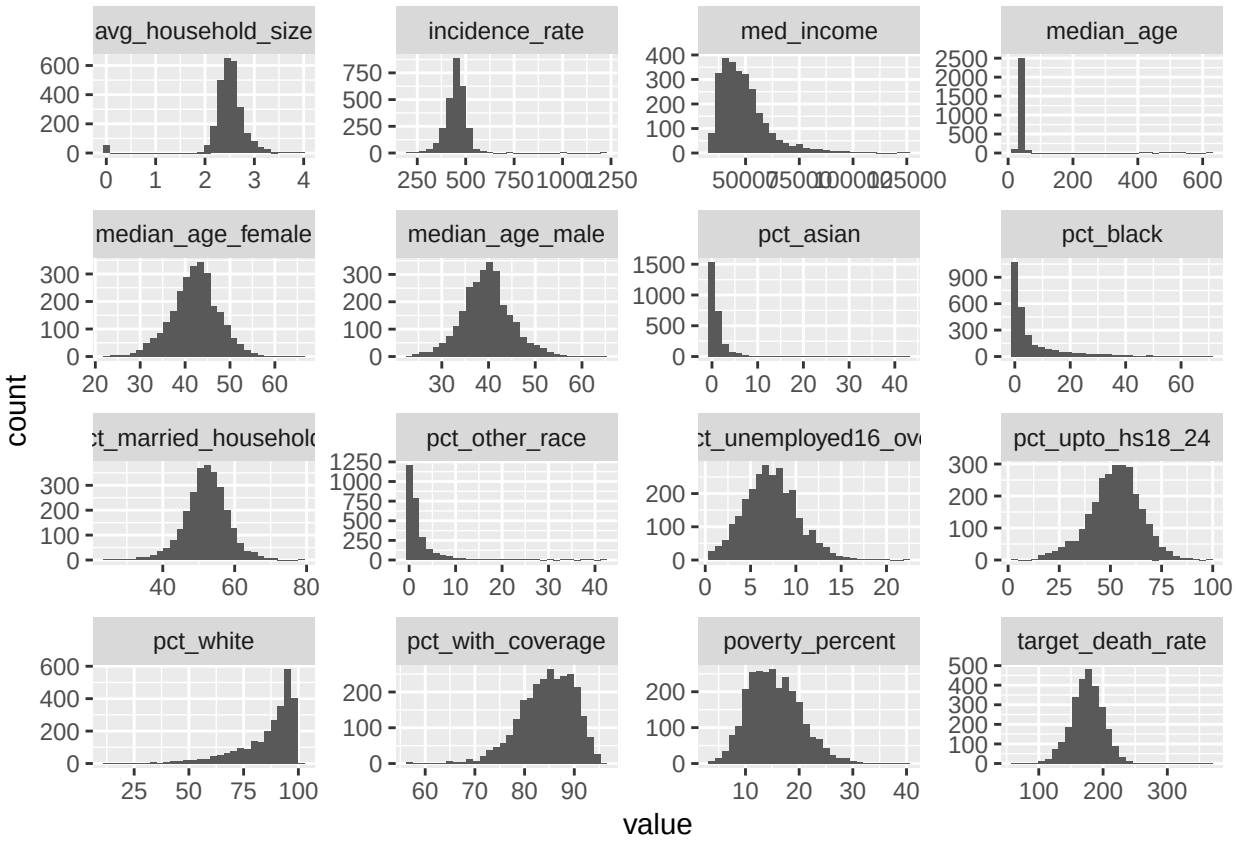
```
##  target_death_rate incidence_rate     med_income     poverty_percent
##  Min.   : 59.7     Min.   : 201.3   Min.   : 35002   Min.   : 3.20
##  1st Qu.:159.8     1st Qu.: 422.3   1st Qu.: 41346   1st Qu.:11.70
##  Median :176.0     Median : 453.5   Median : 46895   Median :14.90
##  Mean   :175.5     Mean   : 448.5   Mean   : 49222   Mean   :15.35
##  3rd Qu.:191.3     3rd Qu.: 479.3   3rd Qu.: 53739   3rd Qu.:18.60
##  Max.   :362.8     Max.   :1206.9   Max.   :125635   Max.   :39.50
##    median_age     median_age_male median_age_female avg_household_size
##  Min.   : 23.20   Min.   :23.00   Min.   :22.30     Min.   :0.0221
##  1st Qu.: 37.60   1st Qu.:36.40   1st Qu.:38.90     1st Qu.:2.3600
##  Median : 41.00   Median :39.60   Median :42.30     Median :2.4900
##  Mean   : 45.45   Mean   :39.62   Mean   :42.12     Mean   :2.4748
##  3rd Qu.: 44.10   3rd Qu.:42.60   3rd Qu.:45.40     3rd Qu.:2.6300
##  Max.   :624.00   Max.   :64.70   Max.   :65.70     Max.   :3.9700
##  pct_unemployed16_over   pct_white        pct_black
##  Min.   : 0.400        Min.   : 11.01   Min.   : 0.0000
##  1st Qu.: 5.300        1st Qu.: 80.04   1st Qu.: 0.6121
##  Median : 7.200        Median : 90.61   Median : 2.0426
##  Mean   : 7.301        Mean   : 85.64   Mean   : 7.0877
##  3rd Qu.: 9.200        3rd Qu.: 95.52   3rd Qu.: 8.4456
##  Max.   :21.900        Max.   :100.00   Max.   :70.3080
##    pct_asian       pct_other_race    pct_married_households
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :23.89
##  1st Qu.: 0.2934   1st Qu.: 0.3219   1st Qu.:48.67
```

```
##  Median : 0.5970   Median : 0.8681   Median :52.11
##  Mean   : 1.3644   Mean   : 2.0052   Mean   :52.09
##  3rd Qu.: 1.3150   3rd Qu.: 2.2368   3rd Qu.:55.74
##  Max.   :42.6194   Max.   :41.9303   Max.   :78.08
##  pct_upto_hs18_24 pct_with_coverage
##  Min.   :  4.80   Min.   :56.50
##  1st Qu.: 44.80   1st Qu.:81.10
##  Median : 52.80   Median :85.00
##  Mean   : 52.24   Mean   :84.51
##  3rd Qu.: 60.40   3rd Qu.:88.70
##  Max.   :100.00   Max.   :95.70
```

```
income_low_data %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins = 30)
```




```
income_high_data %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins = 30)
```




```
cor(income_low_data) %>%
  knitr::kable()
```

	target_death_rate	incidence_rate	med_income	poverty_percent	median_age	median
target_death_rate	1.0000000	0.6486077	-0.2240703	0.1361154	0.0013132	
incidence_rate	0.6486077	1.0000000	-0.1169041	0.1135944	0.0200734	
med_income	-0.2240703	-0.1169041	1.0000000	-0.7061875	0.0744970	
poverty_percent	0.1361154	0.1135944	-0.7061875	1.0000000	-0.0880504	
median_age	0.0013132	0.0200734	0.0744970	-0.0880504	1.0000000	
median_age_male	-0.0243540	-0.0820241	0.1468815	-0.5574503	0.1095884	
median_age_female	0.0138467	-0.0834793	0.0507732	-0.4606824	0.1084074	
avg_household_size	-0.0602017	-0.0705208	-0.0295925	0.2537305	-0.0022481	
pct_unemployed16_over	0.1213959	0.1463679	-0.3729463	0.5136861	0.0659541	
pct_white	0.0287757	-0.0366273	0.2348080	-0.4816891	0.0180313	
pct_black	0.0319415	0.1285239	-0.2826310	0.4134618	-0.0618720	
pct_asian	-0.1608526	-0.0689404	0.0760789	0.0498449	-0.0546289	
pct_other_race	-0.2730811	-0.2077668	0.0375456	0.0117471	0.0244835	
pct_married_households	-0.0701690	-0.1378267	0.3886596	-0.5581316	0.0527556	
pct_upto_hs18_24	0.0750697	-0.0150144	-0.0949035	0.0743566	0.0743857	
pct_with_coverage	0.2579686	0.2703036	0.1835471	-0.2733050	-0.0939198	

```
cor(income_high_data) %>%
  knitr::kable()
```

	target_death_rate	incidence_rate	med_income	poverty_percent	median_age	median
target_death_rate	1.0000000	0.4344595	-0.3699982	0.3617096	0.0090007	

	target_death_rate	incidence_rate	med_income	poverty_percent	median_age	median
incidence_rate	0.4344595	1.0000000	-0.0031009	0.0019127	0.0178751	
med_income	-0.3699982	-0.0031009	1.0000000	-0.7553018	-0.0238311	
poverty_percent	0.3617096	0.0019127	-0.7553018	1.0000000	-0.0217394	
median_age	0.0090007	0.0178751	-0.0238311	-0.0217394	1.0000000	
median_age_male	-0.0127602	-0.0027654	-0.1327490	-0.2053428	0.1315390	
median_age_female	0.0086043	0.0039883	-0.1821019	-0.1595703	0.1268465	
avg_household_size	-0.0459818	-0.1276909	0.1572489	0.0393694	-0.0355821	
pct_unemployed16_over	0.3327304	0.1081013	-0.3395022	0.5615274	0.0192614	
pct_white	-0.1291861	-0.0136929	0.0053752	-0.4079393	0.0386841	
pct_black	0.2249300	0.1315126	-0.1300678	0.3991478	-0.0034839	
pct_asian	-0.1727010	-0.0074408	0.4290119	-0.1242369	-0.0403319	
pct_other_race	-0.1762481	-0.2092576	0.0938002	0.0867307	-0.0386233	
pct_married_households	-0.2440334	-0.1724538	0.3629736	-0.5281702	0.0057737	
pct_upto_hs18_24	0.2179310	-0.1066694	-0.2739013	0.1735028	0.0397649	
pct_with_coverage	-0.1885411	0.2450192	0.4746496	-0.6003613	0.0112912	

Model selection

full model

```
# low income
full_model_low <- lm(target_death_rate ~., data = income_low_data)
summary(full_model_low)
```

```
##
## Call:
## lm(formula = target_death_rate ~ ., data = income_low_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.735 -14.007   0.099  13.636  71.783
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.141e+02  4.829e+01   2.363 0.018652 *
## incidence_rate          2.805e-01  2.046e-02  13.707  < 2e-16 ***
## med_income             -2.733e-03  7.686e-04  -3.555 0.000428 ***
## poverty_percent        -3.712e-01  4.402e-01  -0.843 0.399743
## median_age              6.391e-03  3.281e-02   0.195 0.845649
## median_age_male        -1.824e+00  6.001e-01  -3.040 0.002538 **
## median_age_female       1.439e+00  5.919e-01   2.431 0.015532 *
## avg_household_size     -1.732e-01  3.020e+00  -0.057 0.954290
## pct_unemployed16_over  -1.168e-01  3.826e-01  -0.305 0.760374
## pct_white              -1.952e-01  1.370e-01  -1.425 0.155101
## pct_black              -4.154e-01  1.199e-01  -3.464 0.000597 ***
## pct_asian              -4.461e+00  2.012e+00  -2.217 0.027230 *
## pct_other_race         -1.167e+00  3.296e-01  -3.541 0.000451 ***
## pct_married_households -3.193e-01  3.326e-01  -0.960 0.337682
## pct_upto_hs18_24        2.457e-01  1.154e-01   2.128 0.033981 *
## pct_with_coverage       1.273e+00  3.000e-01   4.245 2.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 23.59 on 362 degrees of freedom
## Multiple R-squared:  0.5168, Adjusted R-squared:  0.4967
## F-statistic: 25.81 on 15 and 362 DF,  p-value: < 2.2e-16
```

```
# high income
full_model_high <- lm(target_death_rate ~., data = income_high_data)
summary(full_model_high)
```

```
##
## Call:
## lm(formula = target_death_rate ~ ., data = income_high_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.922  -10.984    0.228   10.995  125.851
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.259e+02  1.412e+01   8.921  < 2e-16 ***
## incidence_rate          1.991e-01  7.900e-03  25.199  < 2e-16 ***
## med_income             -4.466e-04  6.992e-05  -6.387 1.99e-10 ***
## poverty_percent         1.503e-01  1.771e-01   0.848   0.3963
## median_age             -3.203e-03  8.297e-03  -0.386   0.6995
## median_age_male        -2.155e-01  2.219e-01  -0.971   0.3316
## median_age_female      -2.175e-01  2.197e-01  -0.990   0.3224
## avg_household_size      3.374e-01  1.040e+00   0.324   0.7457
## pct_unemployed16_over   1.068e+00  1.646e-01   6.490 1.02e-10 ***
## pct_white              -3.829e-02  6.698e-02  -0.572   0.5676
## pct_black               7.910e-02  6.709e-02   1.179   0.2385
## pct_asian              -3.477e-01  1.883e-01  -1.846   0.0650 .
## pct_other_race         -1.004e+00  1.331e-01  -7.546 6.15e-14 ***
## pct_married_households -1.525e-01  9.028e-02  -1.690   0.0912 .
## pct_upto_hs18_24        3.919e-01  3.931e-02   9.971  < 2e-16 ***
## pct_with_coverage      -2.146e-01  1.083e-01  -1.981   0.0477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.6 on 2653 degrees of freedom
## Multiple R-squared:  0.4051, Adjusted R-squared:  0.4017
## F-statistic: 120.4 on 15 and 2653 DF,  p-value: < 2.2e-16
```

step wise procedure

```
# low income
step(full_model_low, direction = "backward")
```

```
## Start:  AIC=2405.37
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age + median_age_male + median_age_female + avg_household_size +
##     pct_unemployed16_over + pct_white + pct_black + pct_asian +
##     pct_other_race + pct_married_households + pct_upto_hs18_24 +
##     pct_with_coverage
##
##                          Df Sum of Sq    RSS    AIC
```

```
## - avg_household_size      1         2 201524 2403.4
## - median_age              1        21 201543 2403.4
## - pct_unemployed16_over   1        52 201574 2403.5
## - poverty_percent         1       396 201918 2404.1
## - pct_married_households  1       513 202035 2404.3
## <none>                              201522 2405.4
## - pct_white               1      1130 202652 2405.5
## - pct_upto_hs18_24        1      2522 204044 2408.1
## - pct_asian               1      2737 204259 2408.5
## - median_age_female       1      3291 204813 2409.5
## - median_age_male         1      5145 206667 2412.9
## - pct_black               1      6679 208201 2415.7
## - pct_other_race          1      6980 208503 2416.2
## - med_income              1      7036 208559 2416.3
## - pct_with_coverage       1     10033 211555 2421.7
## - incidence_rate          1    104587 306109 2561.4
##
## Step:  AIC=2403.37
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age + median_age_male + median_age_female + pct_unemployed16_over +
##     pct_white + pct_black + pct_asian + pct_other_race + pct_married_households +
##     pct_upto_hs18_24 + pct_with_coverage
##
##                          Df Sum of Sq    RSS    AIC
## - median_age              1        21 201545 2401.4
## - pct_unemployed16_over   1        53 201577 2401.5
## - poverty_percent         1       415 201939 2402.2
## - pct_married_households  1       524 202048 2402.4
## <none>                              201524 2403.4
## - pct_white               1      1151 202675 2403.5
## - pct_upto_hs18_24        1      2547 204071 2406.1
## - pct_asian               1      2782 204306 2406.6
## - median_age_female       1      3352 204876 2407.6
## - median_age_male         1      5149 206673 2410.9
## - pct_black               1      6896 208420 2414.1
## - pct_other_race          1      7037 208561 2414.3
## - med_income              1      7101 208625 2414.5
## - pct_with_coverage       1     10036 211560 2419.8
## - incidence_rate          1    104910 306434 2559.8
##
## Step:  AIC=2401.41
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age_male + median_age_female + pct_unemployed16_over +
##     pct_white + pct_black + pct_asian + pct_other_race + pct_married_households +
##     pct_upto_hs18_24 + pct_with_coverage
##
##                          Df Sum of Sq    RSS    AIC
## - pct_unemployed16_over   1        46 201591 2399.5
## - poverty_percent         1       418 201963 2400.2
## - pct_married_households  1       524 202069 2400.4
## <none>                              201545 2401.4
## - pct_white               1      1169 202714 2401.6
## - pct_upto_hs18_24        1      2546 204091 2404.2
## - pct_asian               1      2783 204328 2404.6
```

```
## - median_age_female       1      3393 204938 2405.7
## - median_age_male         1      5136 206681 2408.9
## - pct_black               1      6984 208530 2412.3
## - pct_other_race          1      7020 208565 2412.4
## - med_income              1      7082 208627 2412.5
## - pct_with_coverage       1     10071 211616 2417.8
## - incidence_rate          1    105980 307525 2559.1
##
## Step:  AIC=2399.5
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age_male + median_age_female + pct_white + pct_black +
##     pct_asian + pct_other_race + pct_married_households + pct_upto_hs18_24 +
##     pct_with_coverage
##
##                          Df Sum of Sq    RSS    AIC
## - poverty_percent         1       478 202069 2398.4
## - pct_married_households  1       506 202097 2398.4
## <none>                                201591 2399.5
## - pct_white               1      1122 202714 2399.6
## - pct_upto_hs18_24        1      2553 204144 2402.3
## - pct_asian               1      2745 204336 2402.6
## - median_age_female       1      3472 205063 2404.0
## - median_age_male         1      5229 206820 2407.2
## - pct_black               1      6942 208533 2410.3
## - med_income              1      7037 208628 2410.5
## - pct_other_race          1      7082 208673 2410.6
## - pct_with_coverage       1     10181 211772 2416.1
## - incidence_rate          1    106374 307966 2557.7
##
## Step:  AIC=2398.4
## target_death_rate ~ incidence_rate + med_income + median_age_male +
##     median_age_female + pct_white + pct_black + pct_asian + pct_other_race +
##     pct_married_households + pct_upto_hs18_24 + pct_with_coverage
##
##                          Df Sum of Sq    RSS    AIC
## - pct_married_households  1       412 202481 2397.2
## - pct_white               1       990 203059 2398.2
## <none>                                202069 2398.4
## - pct_upto_hs18_24        1      2285 204354 2400.7
## - pct_asian               1      2730 204799 2401.5
## - median_age_female       1      3778 205848 2403.4
## - median_age_male         1      4798 206867 2405.3
## - pct_black               1      6762 208831 2408.8
## - pct_other_race          1      6899 208968 2409.1
## - med_income              1      9367 211436 2413.5
## - pct_with_coverage       1     10041 212110 2414.7
## - incidence_rate          1    106769 308838 2556.8
##
## Step:  AIC=2397.17
## target_death_rate ~ incidence_rate + med_income + median_age_male +
##     median_age_female + pct_white + pct_black + pct_asian + pct_other_race +
##     pct_upto_hs18_24 + pct_with_coverage
##
##                     Df Sum of Sq    RSS    AIC
```

```
## <none>                           202481 2397.2
## - pct_upto_hs18_24   1      2084 204565 2399.0
## - pct_white          1      2125 204606 2399.1
## - pct_asian          1      2390 204872 2399.6
## - median_age_female  1      3765 206247 2402.1
## - median_age_male    1      4913 207394 2404.2
## - pct_black          1      6480 208961 2407.1
## - pct_other_race     1      6545 209026 2407.2
## - pct_with_coverage  1     10589 213071 2414.4
## - med_income         1     12809 215290 2418.3
## - incidence_rate     1    110663 313144 2560.0
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_male + median_age_female + pct_white + pct_black +
##     pct_asian + pct_other_race + pct_upto_hs18_24 + pct_with_coverage,
##     data = income_low_data)
##
## Coefficients:
##       (Intercept)     incidence_rate         med_income
##         70.929330           0.283192           -0.002389
##   median_age_male  median_age_female          pct_white
##         -1.736992           1.514622           -0.227054
##         pct_black          pct_asian     pct_other_race
##         -0.395269          -3.996507           -1.088218
##  pct_upto_hs18_24  pct_with_coverage
##          0.218499           1.288771
```

```
backward_model_low <- lm(target_death_rate ~ incidence_rate + med_income +
    median_age_male + median_age_female + pct_white + pct_black +
    pct_asian + pct_other_race + pct_upto_hs18_24 + pct_with_coverage,
    data = income_low_data)
summary(backward_model_low)
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_male + median_age_female + pct_white + pct_black +
##     pct_asian + pct_other_race + pct_upto_hs18_24 + pct_with_coverage,
##     data = income_low_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.155 -13.088   0.046  13.593  75.734
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       70.9293300 27.2514067   2.603 0.009622 **
## incidence_rate     0.2831921  0.0199959  14.163  < 2e-16 ***
## med_income        -0.0023887  0.0004958  -4.818 2.12e-06 ***
## median_age_male   -1.7369925  0.5820663  -2.984 0.003034 **
## median_age_female  1.5146221  0.5797669   2.612 0.009358 **
## pct_white         -0.2270544  0.1156907  -1.963 0.050448 .
## pct_black         -0.3952686  0.1153337  -3.427 0.000679 ***
```

```
## pct_asian         -3.9965072  1.9200022  -2.082 0.038080 *
## pct_other_race    -1.0882180  0.3159474  -3.444 0.000639 ***
## pct_upto_hs18_24   0.2184992  0.1124250   1.944 0.052719 .
## pct_with_coverage  1.2887706  0.2941692   4.381 1.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.49 on 367 degrees of freedom
## Multiple R-squared:  0.5145, Adjusted R-squared:  0.5012
## F-statistic: 38.89 on 10 and 367 DF,  p-value: < 2.2e-16
```

```
# high income
step(full_model_high, direction = "backward")
```

```
## Start:  AIC=15899.64
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age + median_age_male + median_age_female + avg_household_size +
##     pct_unemployed16_over + pct_white + pct_black + pct_asian +
##     pct_other_race + pct_married_households + pct_upto_hs18_24 +
##     pct_with_coverage
##
##                          Df Sum of Sq     RSS   AIC
## - avg_household_size      1        40 1019337 15898
## - median_age              1        57 1019354 15898
## - pct_white               1       126 1019422 15898
## - poverty_percent         1       277 1019573 15898
## - median_age_male         1       362 1019659 15899
## - median_age_female       1       376 1019673 15899
## - pct_black               1       534 1019831 15899
## <none>                                1019297 15900
## - pct_married_households  1      1097 1020394 15900
## - pct_asian               1      1309 1020606 15901
## - pct_with_coverage       1      1508 1020805 15902
## - med_income              1     15674 1034971 15938
## - pct_unemployed16_over   1     16182 1035479 15940
## - pct_other_race          1     21876 1041172 15954
## - pct_upto_hs18_24        1     38196 1057493 15996
## - incidence_rate          1    243974 1263271 16470
##
## Step:  AIC=15897.75
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age + median_age_male + median_age_female + pct_unemployed16_over +
##     pct_white + pct_black + pct_asian + pct_other_race + pct_married_households +
##     pct_upto_hs18_24 + pct_with_coverage
##
##                          Df Sum of Sq     RSS   AIC
## - median_age              1        56 1019393 15896
## - pct_white               1       136 1019473 15896
## - poverty_percent         1       283 1019620 15896
## - median_age_male         1       371 1019708 15897
## - median_age_female       1       400 1019737 15897
## - pct_black               1       521 1019858 15897
## <none>                                1019337 15898
## - pct_married_households  1      1056 1020394 15898
## - pct_asian               1      1303 1020640 15899
```

```
## - pct_with_coverage       1      1566 1020903 15900
## - med_income              1     15655 1034992 15936
## - pct_unemployed16_over   1     16534 1035871 15939
## - pct_other_race          1     21836 1041173 15952
## - pct_upto_hs18_24        1     38484 1057821 15995
## - incidence_rate          1    244207 1263544 16469
##
## Step:  AIC=15895.9
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age_male + median_age_female + pct_unemployed16_over +
##     pct_white + pct_black + pct_asian + pct_other_race + pct_married_households +
##     pct_upto_hs18_24 + pct_with_coverage
##
##                          Df Sum of Sq     RSS   AIC
## - pct_white               1       138 1019531 15894
## - poverty_percent         1       289 1019682 15895
## - median_age_male         1       382 1019775 15895
## - median_age_female       1       402 1019795 15895
## - pct_black               1       513 1019906 15895
## <none>                                1019393 15896
## - pct_married_households  1      1050 1020444 15897
## - pct_asian               1      1301 1020694 15897
## - pct_with_coverage       1      1561 1020954 15898
## - med_income              1     15646 1035039 15934
## - pct_unemployed16_over   1     16489 1035882 15937
## - pct_other_race          1     21858 1041251 15950
## - pct_upto_hs18_24        1     38477 1057871 15993
## - incidence_rate          1    244157 1263550 16467
##
## Step:  AIC=15894.26
## target_death_rate ~ incidence_rate + med_income + poverty_percent +
##     median_age_male + median_age_female + pct_unemployed16_over +
##     pct_black + pct_asian + pct_other_race + pct_married_households +
##     pct_upto_hs18_24 + pct_with_coverage
##
##                          Df Sum of Sq     RSS   AIC
## - poverty_percent         1       335 1019866 15893
## - median_age_male         1       391 1019922 15893
## - median_age_female       1       427 1019957 15893
## <none>                                1019531 15894
## - pct_asian               1      1179 1020710 15895
## - pct_married_households  1      1285 1020815 15896
## - pct_with_coverage       1      1833 1021364 15897
## - pct_black               1      2556 1022086 15899
## - med_income              1     15649 1035180 15933
## - pct_unemployed16_over   1     16795 1036326 15936
## - pct_other_race          1     23028 1042559 15952
## - pct_upto_hs18_24        1     39981 1059512 15995
## - incidence_rate          1    244120 1263651 16465
##
## Step:  AIC=15893.13
## target_death_rate ~ incidence_rate + med_income + median_age_male +
##     median_age_female + pct_unemployed16_over + pct_black + pct_asian +
##     pct_other_race + pct_married_households + pct_upto_hs18_24 +
```

```
##     pct_with_coverage
##
##                           Df Sum of Sq     RSS   AIC
## - median_age_female        1       445 1020311 15892
## - median_age_male          1       471 1020338 15892
## <none>                                 1019866 15893
## - pct_asian                1      1020 1020886 15894
## - pct_married_households   1      1579 1021445 15895
## - pct_with_coverage        1      2488 1022354 15898
## - pct_black                1      3163 1023029 15899
## - pct_unemployed16_over    1     20351 1040217 15944
## - pct_other_race           1     22956 1042822 15950
## - med_income               1     35558 1055424 15983
## - pct_upto_hs18_24         1     39818 1059684 15993
## - incidence_rate           1    243891 1263757 16463
##
## Step:  AIC=15892.3
## target_death_rate ~ incidence_rate + med_income + median_age_male +
##     pct_unemployed16_over + pct_black + pct_asian + pct_other_race +
##     pct_married_households + pct_upto_hs18_24 + pct_with_coverage
##
##                           Df Sum of Sq     RSS   AIC
## <none>                                 1020311 15892
## - pct_asian                1      1021 1021332 15893
## - pct_married_households   1      1553 1021864 15894
## - pct_black                1      2941 1023253 15898
## - pct_with_coverage        1      2994 1023305 15898
## - median_age_male          1     11288 1031599 15920
## - pct_unemployed16_over    1     20723 1041034 15944
## - pct_other_race           1     22915 1043226 15950
## - med_income               1     35284 1055596 15981
## - pct_upto_hs18_24         1     39388 1059700 15991
## - incidence_rate           1    245198 1265510 16465

##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_male + pct_unemployed16_over + pct_black + pct_asian +
##     pct_other_race + pct_married_households + pct_upto_hs18_24 +
##     pct_with_coverage, data = income_high_data)
##
## Coefficients:
##            (Intercept)          incidence_rate              med_income
##              1.331e+02               1.989e-01              -4.707e-04
##        median_age_male   pct_unemployed16_over               pct_black
##             -4.630e-01               1.135e+00               1.127e-01
##              pct_asian          pct_other_race  pct_married_households
##             -2.747e-01              -9.756e-01              -1.709e-01
##       pct_upto_hs18_24       pct_with_coverage
##              3.899e-01              -2.783e-01
```

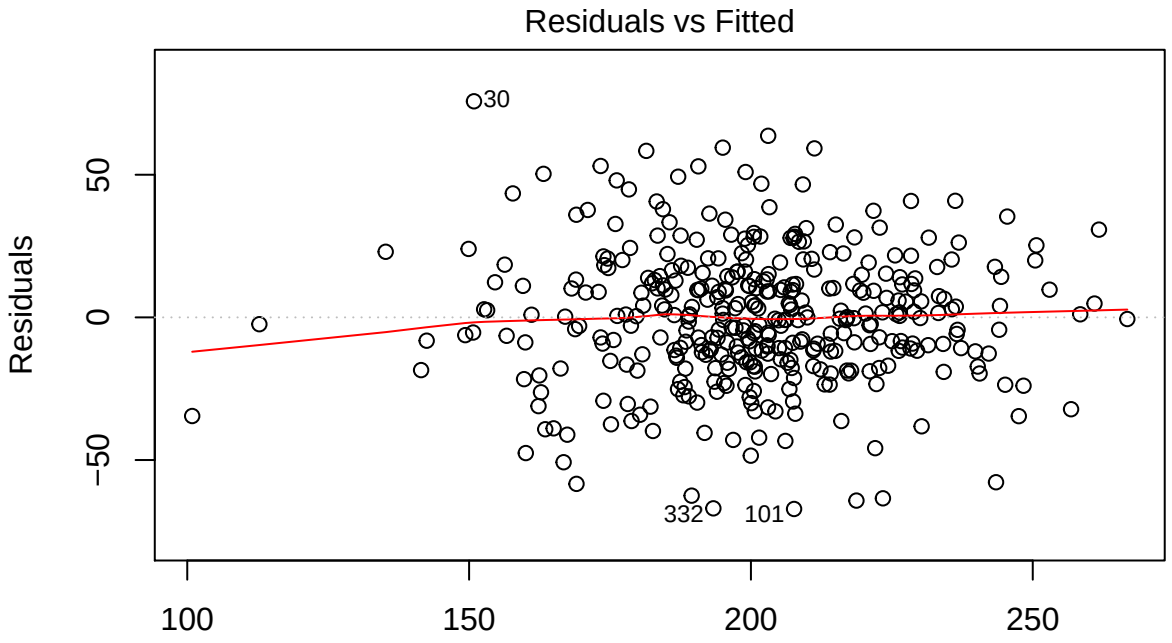
```
backward_model_high <- lm(target_death_rate ~ incidence_rate + med_income +
    median_age_male + pct_unemployed16_over + pct_black + pct_asian +
    pct_other_race + pct_married_households + pct_upto_hs18_24 +
    pct_with_coverage, data = income_high_data)
```

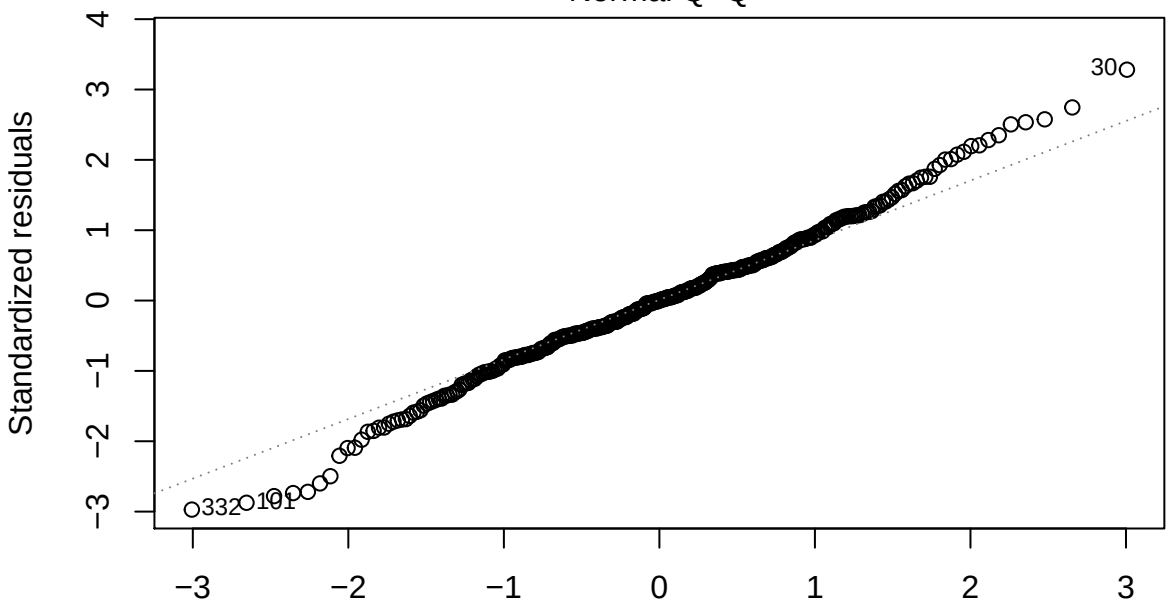
```r
summary(backward_model_high)
```

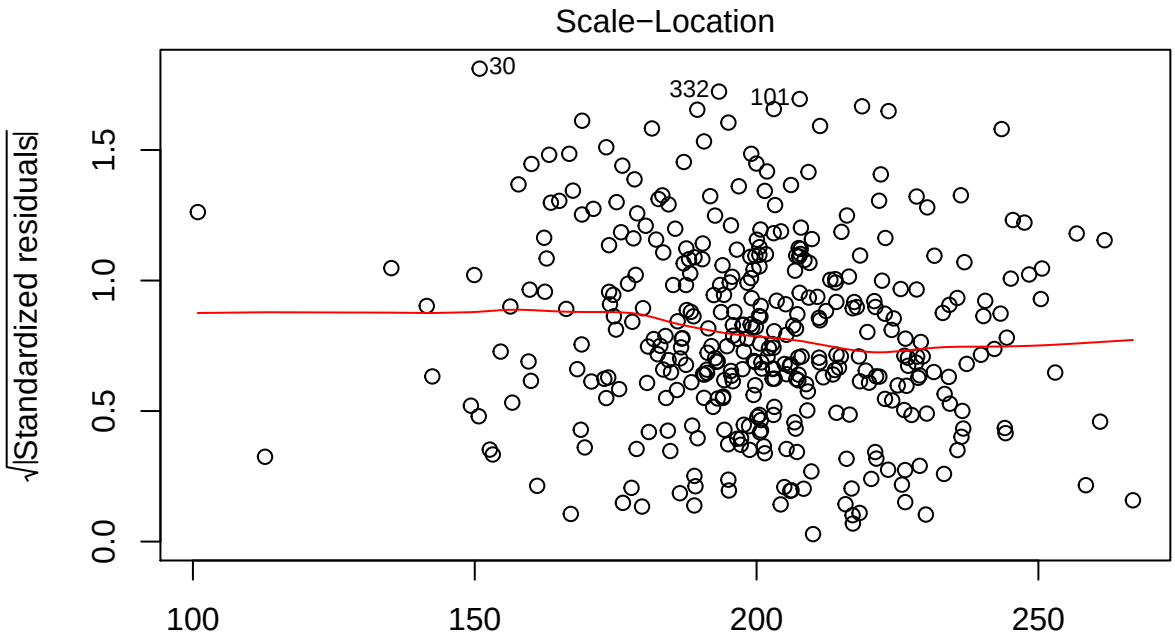
```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_male + pct_unemployed16_over + pct_black + pct_asian +
##     pct_other_race + pct_married_households + pct_upto_hs18_24 +
##     pct_with_coverage, data = income_high_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.718  -10.910    0.098   10.903  126.259
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.331e+02  9.495e+00  14.016  < 2e-16 ***
## incidence_rate           1.989e-01  7.871e-03  25.274  < 2e-16 ***
## med_income              -4.706e-04  4.909e-05  -9.587  < 2e-16 ***
## median_age_male         -4.630e-01  8.539e-02  -5.423 6.40e-08 ***
## pct_unemployed16_over    1.135e+00  1.545e-01   7.347 2.67e-13 ***
## pct_black                1.127e-01  4.073e-02   2.768  0.00568 **
## pct_asian               -2.747e-01  1.685e-01  -1.631  0.10309
## pct_other_race          -9.756e-01  1.263e-01  -7.726 1.56e-14 ***
## pct_married_households -1.709e-01  8.495e-02  -2.011  0.04441 *
## pct_upto_hs18_24         3.899e-01  3.849e-02  10.130  < 2e-16 ***
## pct_with_coverage       -2.783e-01  9.966e-02  -2.793  0.00526 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.59 on 2658 degrees of freedom
## Multiple R-squared:  0.4045, Adjusted R-squared:  0.4023
## F-statistic: 180.6 on 10 and 2658 DF,  p-value: < 2.2e-16
```

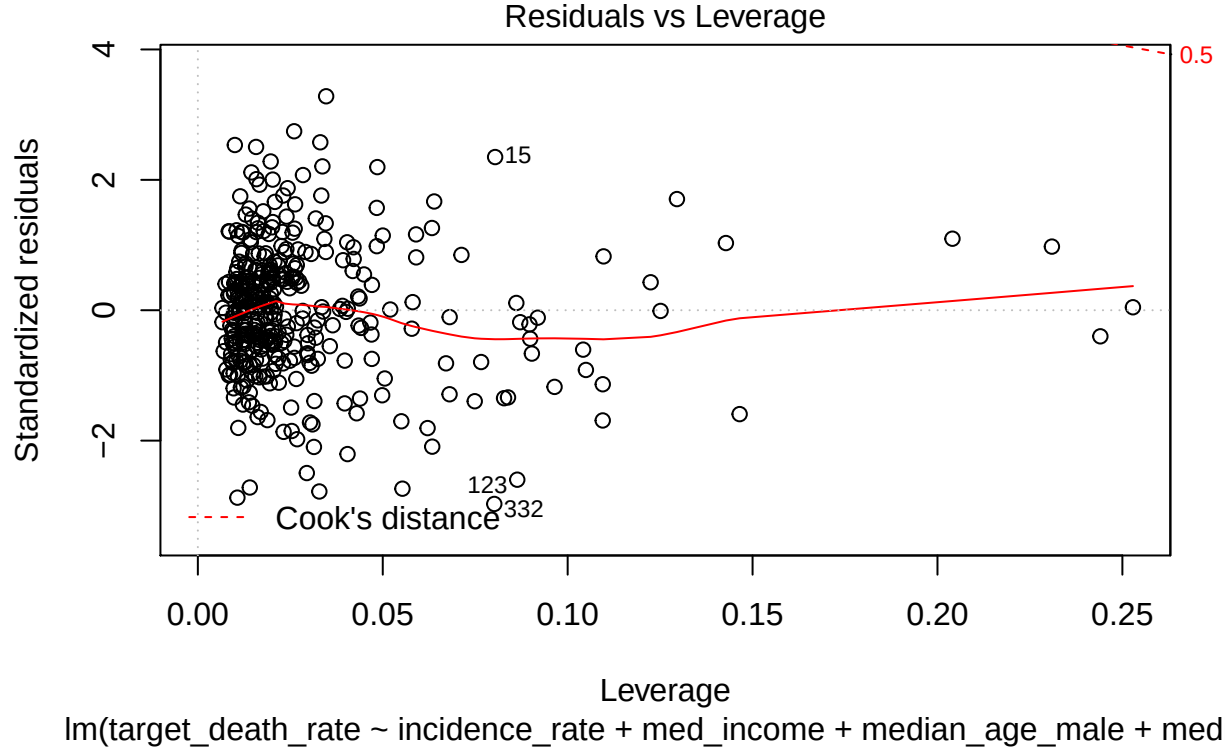
Check assumption

```r
plot(backward_model_low)
```

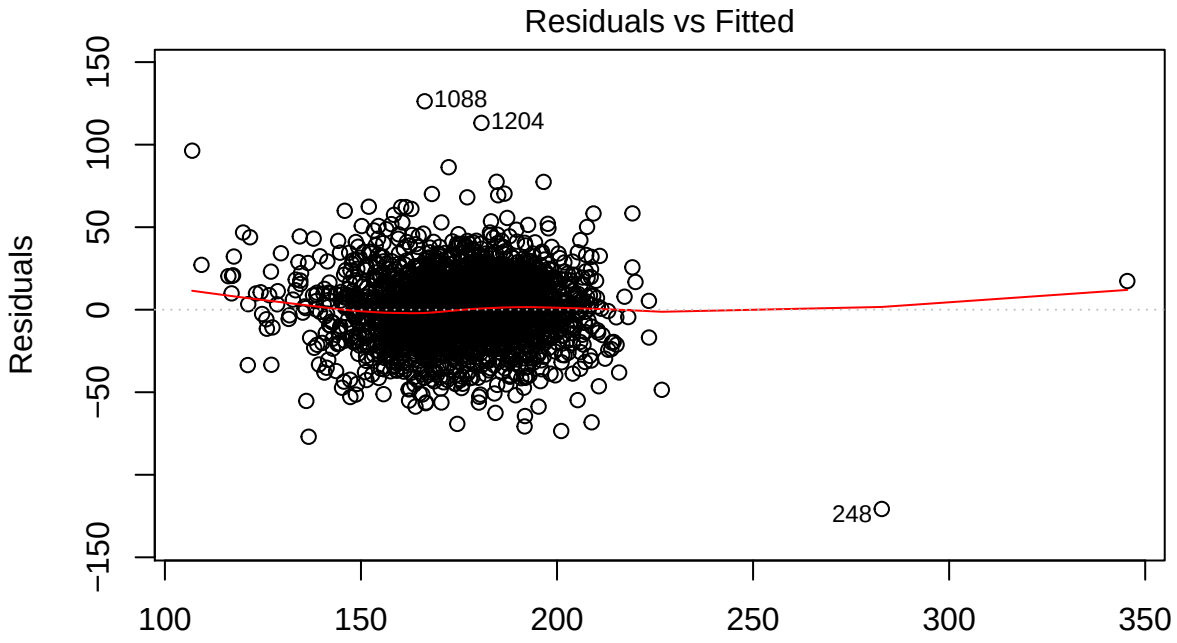

Residuals vs Fitted
Residuals
Fitted values
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + medi ..
30
332
101

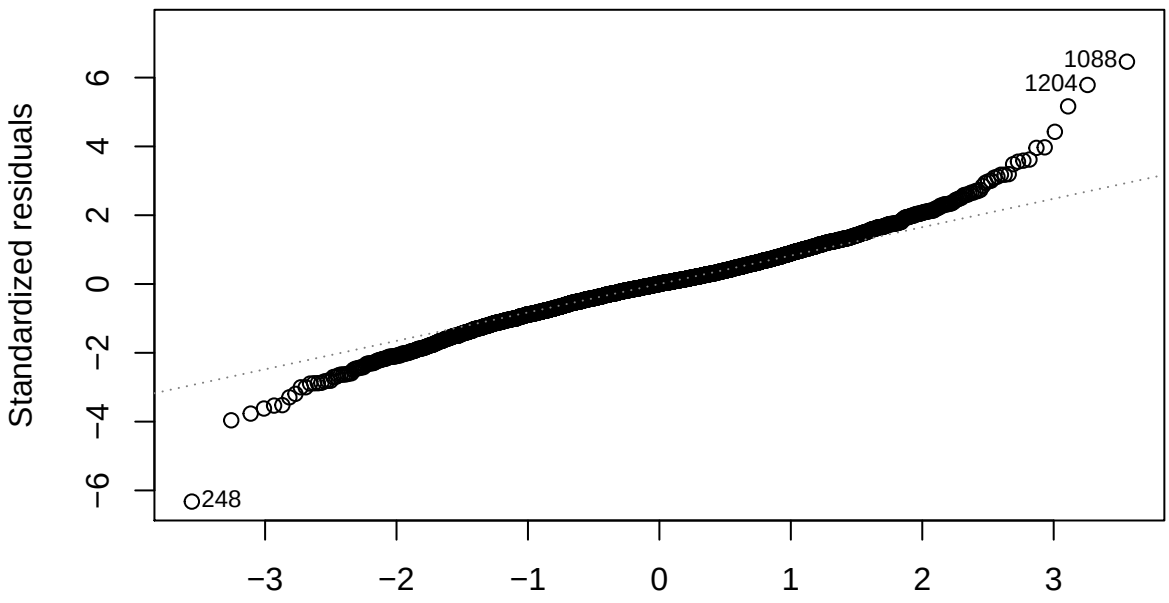

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + medi .
30
332
101

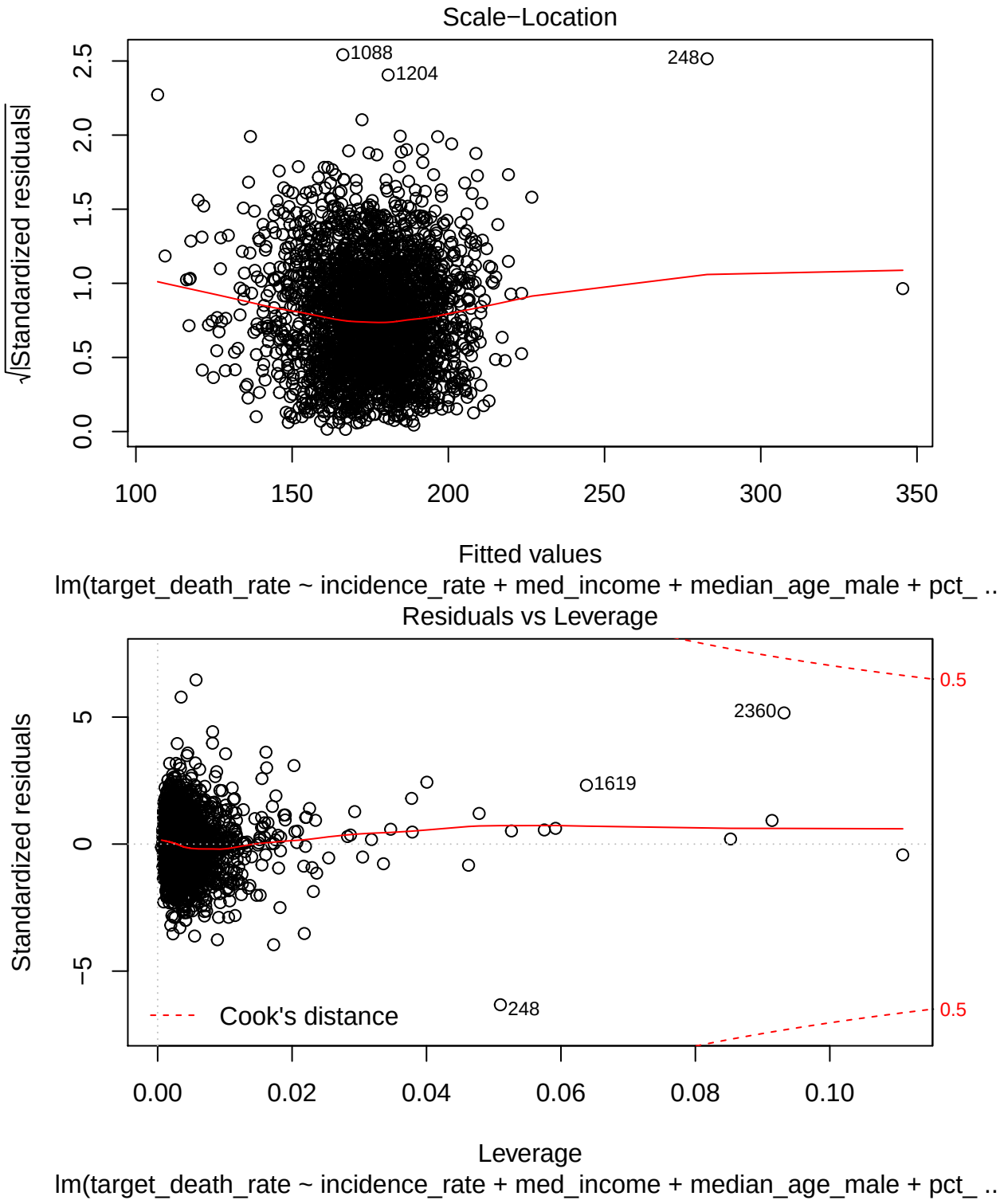





```
plot(backward_model_high)
```


Residuals vs Fitted
Residuals
Fitted values
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + pct_ ..
1088
1204
248


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + pct_ ..
1088
1204
248



citerion based approach

low income

```
library(leaps)
criterion_stats_low =
```

```
  regsubsets(target_death_rate ~ ., nvmax = 12, data = income_low_data) %>%
  summary()

criterion_stats_low
```

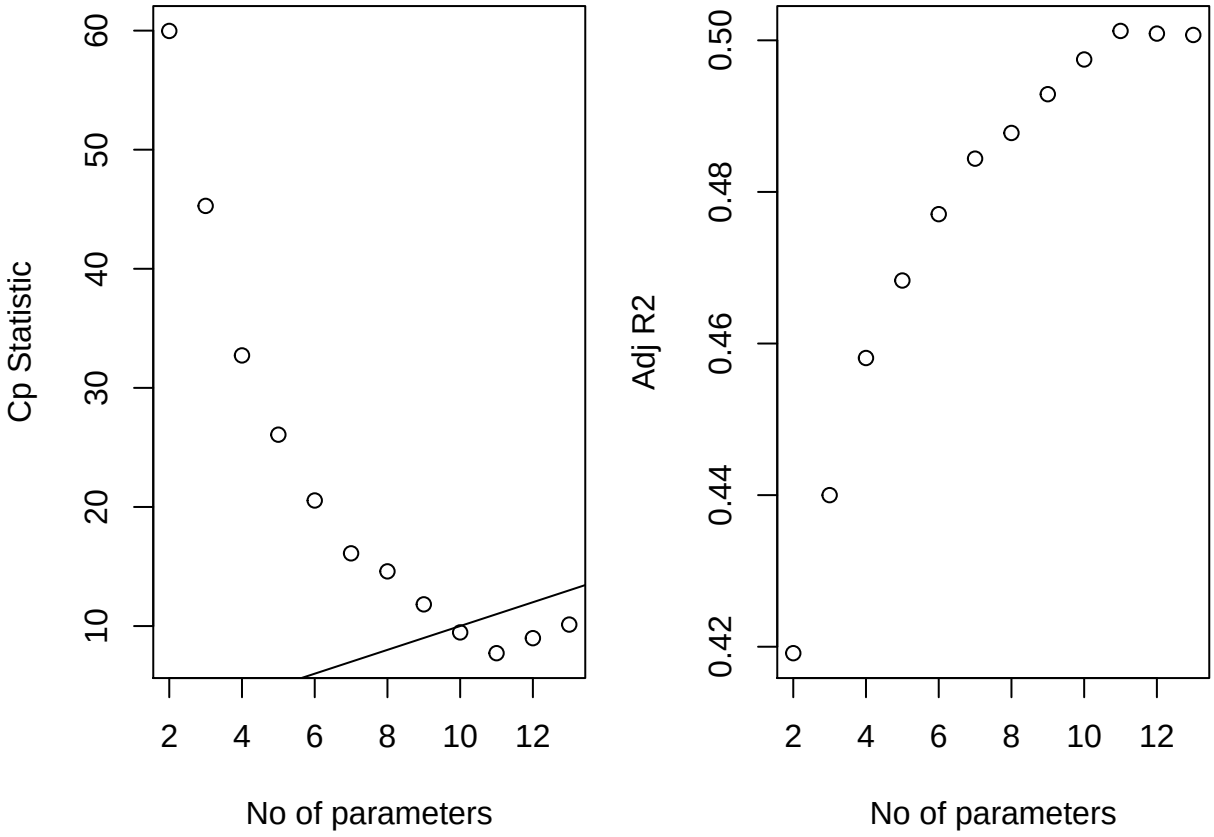
```
## Subset selection object
## Call: regsubsets.formula(target_death_rate ~ ., nvmax = 12, data = income_low_data)
## 15 Variables  (and intercept)
##                          Forced in Forced out
## incidence_rate               FALSE      FALSE
## med_income                   FALSE      FALSE
## poverty_percent              FALSE      FALSE
## median_age                   FALSE      FALSE
## median_age_male              FALSE      FALSE
## median_age_female            FALSE      FALSE
## avg_household_size           FALSE      FALSE
## pct_unemployed16_over        FALSE      FALSE
## pct_white                    FALSE      FALSE
## pct_black                    FALSE      FALSE
## pct_asian                    FALSE      FALSE
## pct_other_race               FALSE      FALSE
## pct_married_households       FALSE      FALSE
## pct_upto_hs18_24             FALSE      FALSE
## pct_with_coverage            FALSE      FALSE
## 1 subsets of each size up to 12
## Selection Algorithm: exhaustive
##           incidence_rate med_income poverty_percent median_age
## 1  ( 1 )  "*"            " "        " "             " "
## 2  ( 1 )  "*"            "*"        " "             " "
## 3  ( 1 )  "*"            "*"        " "             " "
## 4  ( 1 )  "*"            "*"        " "             " "
## 5  ( 1 )  "*"            "*"        " "             " "
## 6  ( 1 )  "*"            "*"        " "             " "
## 7  ( 1 )  "*"            "*"        " "             " "
## 8  ( 1 )  "*"            "*"        " "             " "
## 9  ( 1 )  "*"            "*"        " "             " "
## 10  ( 1 ) "*"            "*"        " "             " "
## 11  ( 1 ) "*"            "*"        " "             " "
## 12  ( 1 ) "*"            "*"        "*"             " "
##           median_age_male median_age_female avg_household_size
## 1  ( 1 )  " "             " "               " "
## 2  ( 1 )  " "             " "               " "
## 3  ( 1 )  " "             " "               " "
## 4  ( 1 )  " "             " "               " "
## 5  ( 1 )  " "             " "               " "
## 6  ( 1 )  " "             " "               " "
## 7  ( 1 )  " "             " "               " "
## 8  ( 1 )  "*"             "*"               " "
## 9  ( 1 )  "*"             "*"               " "
## 10  ( 1 ) "*"             "*"               " "
## 11  ( 1 ) "*"             "*"               " "
## 12  ( 1 ) "*"             "*"               " "
##           pct_unemployed16_over pct_white pct_black pct_asian
## 1  ( 1 )  " "                   " "       " "       " "
```

```
## 2  ( 1 )  " "                   " "        " "        " "
## 3  ( 1 )  " "                   " "        " "        " "
## 4  ( 1 )  " "                   " "        " "        " "
## 5  ( 1 )  " "                   " "        " "        " "
## 6  ( 1 )  " "                   " "        "*"        " "
## 7  ( 1 )  " "                   "*"        "*"        " "
## 8  ( 1 )  " "                   " "        "*"        " "
## 9  ( 1 )  " "                   "*"        "*"        "*"
## 10  ( 1 ) " "                   "*"        "*"        "*"
## 11  ( 1 ) " "                   "*"        "*"        "*"
## 12  ( 1 ) " "                   "*"        "*"        "*"
##           pct_other_race pct_married_households pct_upto_hs18_24
## 1  ( 1 )  " "            " "                    " "
## 2  ( 1 )  " "            " "                    " "
## 3  ( 1 )  "*"            " "                    " "
## 4  ( 1 )  "*"            " "                    " "
## 5  ( 1 )  "*"            " "                    "*"
## 6  ( 1 )  "*"            " "                    "*"
## 7  ( 1 )  "*"            " "                    "*"
## 8  ( 1 )  "*"            " "                    "*"
## 9  ( 1 )  "*"            " "                    " "
## 10  ( 1 ) "*"            " "                    "*"
## 11  ( 1 ) "*"            "*"                    "*"
## 12  ( 1 ) "*"            "*"                    "*"
##           pct_with_coverage
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  "*"
## 5  ( 1 )  "*"
## 6  ( 1 )  "*"
## 7  ( 1 )  "*"
## 8  ( 1 )  "*"
## 9  ( 1 )  "*"
## 10  ( 1 ) "*"
## 11  ( 1 ) "*"
## 12  ( 1 ) "*"
```

```r
par(mar = c(4,4,1,1))
par(mfrow = c(1,2))

plot(2:13, criterion_stats_low$cp, xlab = "No of parameters", ylab = "Cp Statistic")
abline(0,1)

plot(2:13, criterion_stats_low$adjr2, xlab = "No of parameters", ylab = "Adj R2")
```




```
## 9 predictors are sufficient
criterion_model_low <- lm(formula = target_death_rate ~ incidence_rate + med_income +
    median_age_male + median_age_female + pct_white + pct_black + pct_asian +
    pct_other_race +  pct_with_coverage, data = income_low_data)
```

high income

```
criterion_stats_high =
  regsubsets(target_death_rate ~ ., nvmax = 12, data = income_high_data) %>%
  summary()

criterion_stats_high
```

```
## Subset selection object
## Call: regsubsets.formula(target_death_rate ~ ., nvmax = 12, data = income_high_data)
## 15 Variables  (and intercept)
##                         Forced in Forced out
## incidence_rate              FALSE      FALSE
## med_income                  FALSE      FALSE
## poverty_percent             FALSE      FALSE
## median_age                  FALSE      FALSE
## median_age_male             FALSE      FALSE
## median_age_female           FALSE      FALSE
## avg_household_size          FALSE      FALSE
## pct_unemployed16_over       FALSE      FALSE
## pct_white                   FALSE      FALSE
## pct_black                   FALSE      FALSE
## pct_asian                   FALSE      FALSE
```

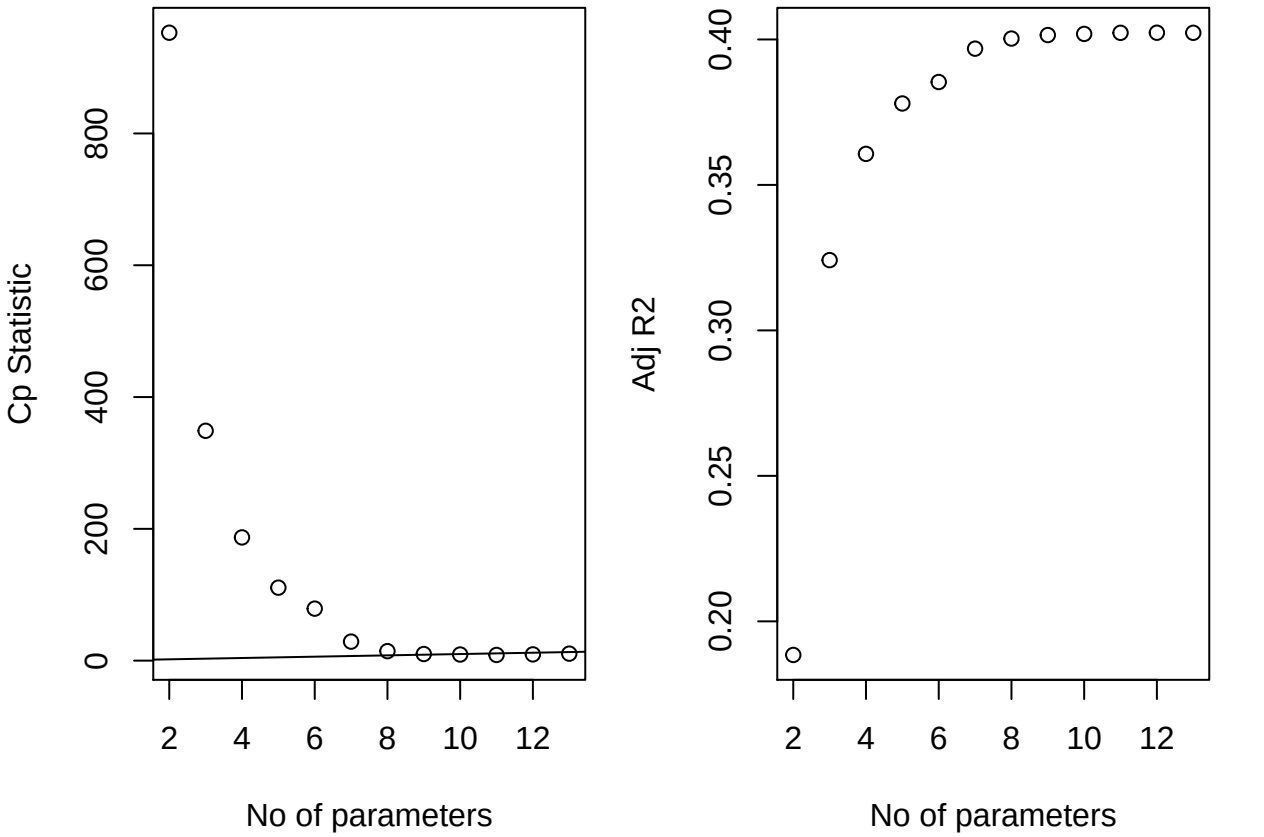
```
## pct_other_race             FALSE      FALSE
## pct_married_households     FALSE      FALSE
## pct_upto_hs18_24           FALSE      FALSE
## pct_with_coverage          FALSE      FALSE
## 1 subsets of each size up to 12
## Selection Algorithm: exhaustive
##           incidence_rate med_income poverty_percent median_age
## 1  ( 1 )  "*"            " "        " "             " "
## 2  ( 1 )  "*"            "*"        " "             " "
## 3  ( 1 )  "*"            " "        "*"             " "
## 4  ( 1 )  "*"            "*"        " "             " "
## 5  ( 1 )  "*"            "*"        " "             " "
## 6  ( 1 )  "*"            "*"        " "             " "
## 7  ( 1 )  "*"            "*"        " "             " "
## 8  ( 1 )  "*"            "*"        " "             " "
## 9  ( 1 )  "*"            "*"        " "             " "
## 10  ( 1 ) "*"            "*"        " "             " "
## 11  ( 1 ) "*"            "*"        " "             " "
## 12  ( 1 ) "*"            "*"        "*"             " "
##           median_age_male median_age_female avg_household_size
## 1  ( 1 )  " "             " "               " "
## 2  ( 1 )  " "             " "               " "
## 3  ( 1 )  " "             " "               " "
## 4  ( 1 )  " "             " "               " "
## 5  ( 1 )  " "             " "               " "
## 6  ( 1 )  "*"             " "               " "
## 7  ( 1 )  " "             "*"               " "
## 8  ( 1 )  "*"             " "               " "
## 9  ( 1 )  " "             "*"               " "
## 10  ( 1 ) "*"             " "               " "
## 11  ( 1 ) "*"             "*"               " "
## 12  ( 1 ) "*"             "*"               " "
##           pct_unemployed16_over pct_white pct_black pct_asian
## 1  ( 1 )  " "                   " "       " "       " "
## 2  ( 1 )  " "                   " "       " "       " "
## 3  ( 1 )  " "                   " "       " "       " "
## 4  ( 1 )  "*"                   " "       " "       " "
## 5  ( 1 )  "*"                   " "       " "       " "
## 6  ( 1 )  "*"                   " "       " "       " "
## 7  ( 1 )  "*"                   " "       "*"       " "
## 8  ( 1 )  "*"                   " "       "*"       " "
## 9  ( 1 )  "*"                   " "       "*"       " "
## 10  ( 1 ) "*"                   " "       "*"       "*"
## 11  ( 1 ) "*"                   " "       "*"       "*"
## 12  ( 1 ) "*"                   " "       "*"       "*"
##           pct_other_race pct_married_households pct_upto_hs18_24
## 1  ( 1 )  " "            " "                   " "
## 2  ( 1 )  " "            " "                   " "
## 3  ( 1 )  " "            " "                   "*"
## 4  ( 1 )  " "            " "                   "*"
## 5  ( 1 )  "*"            " "                   "*"
## 6  ( 1 )  "*"            " "                   "*"
## 7  ( 1 )  "*"            " "                   "*"
## 8  ( 1 )  "*"            " "                   "*"
```

```
## 9  ( 1 )  "*"            "*"                    "*"
## 10  ( 1 ) "*"            "*"                    "*"
## 11  ( 1 ) "*"            "*"                    "*"
## 12  ( 1 ) "*"            "*"                    "*"
##           pct_with_coverage
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  "*"
## 9  ( 1 )  "*"
## 10  ( 1 ) "*"
## 11  ( 1 ) "*"
## 12  ( 1 ) "*"
```

```r
par(mar = c(4,4,1,1))
par(mfrow = c(1,2))

plot(2:13, criterion_stats_high$cp, xlab = "No of parameters", ylab = "Cp Statistic")
abline(0,1)

plot(2:13, criterion_stats_high$adjr2, xlab = "No of parameters", ylab = "Adj R2")
```

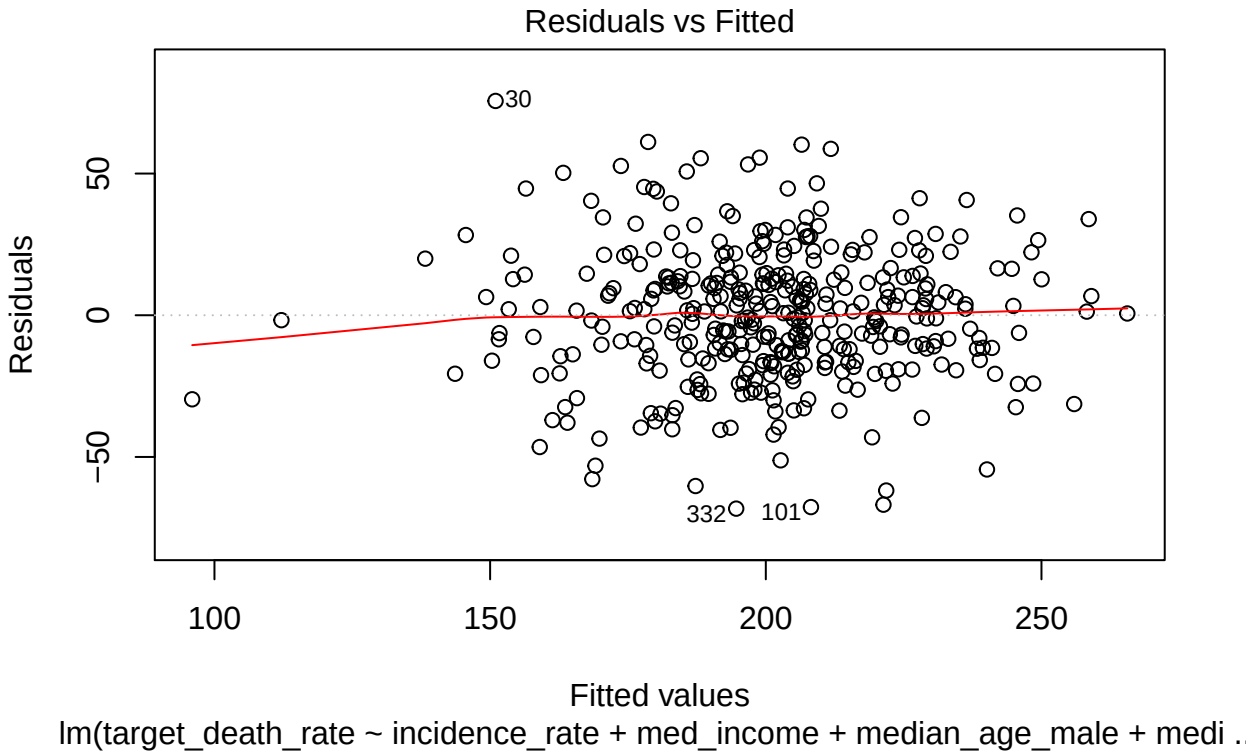


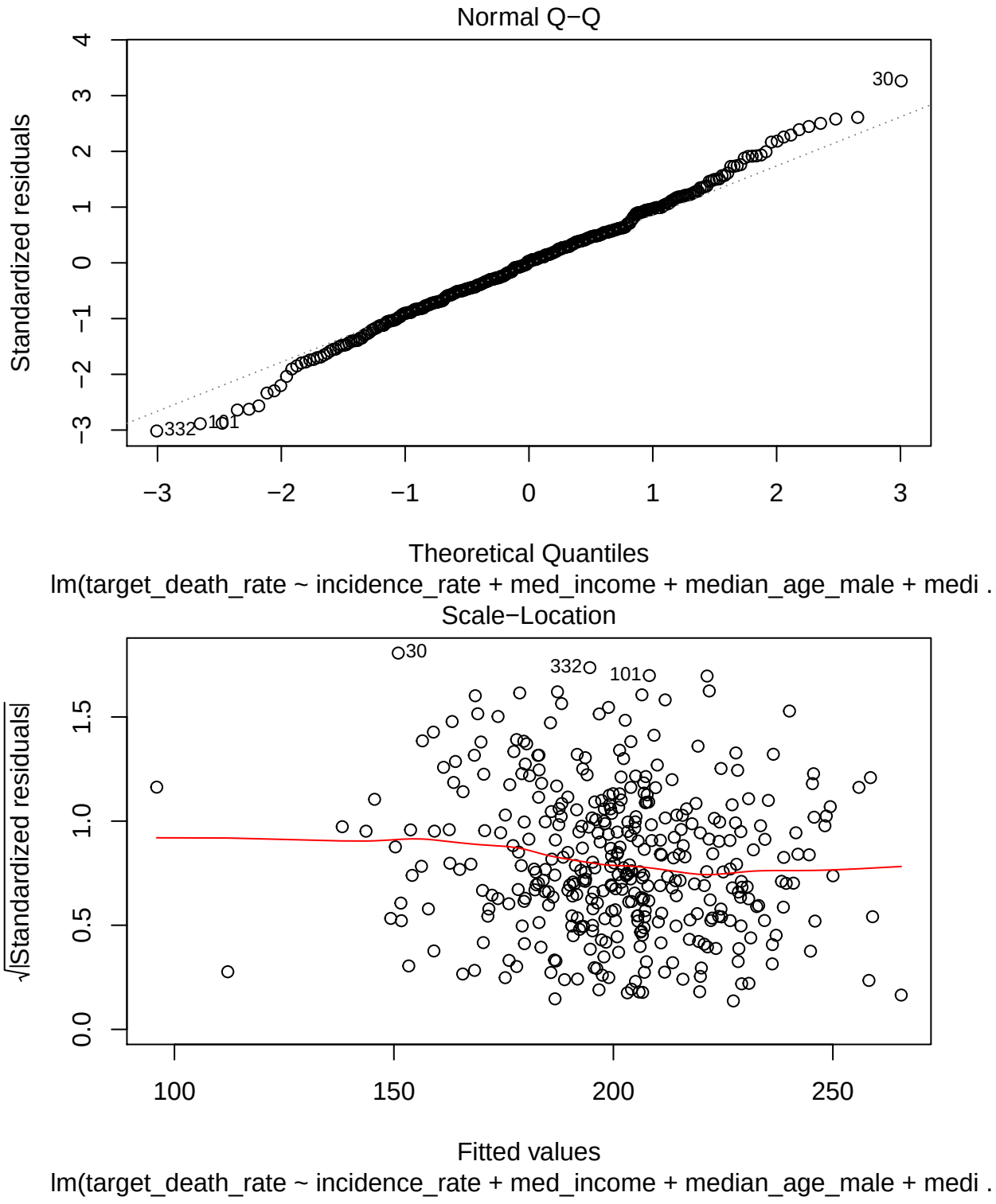

```r
# 7 predictors are sufficients
criterion_model_high <- lm(formula = target_death_rate ~ incidence_rate + med_income +
```

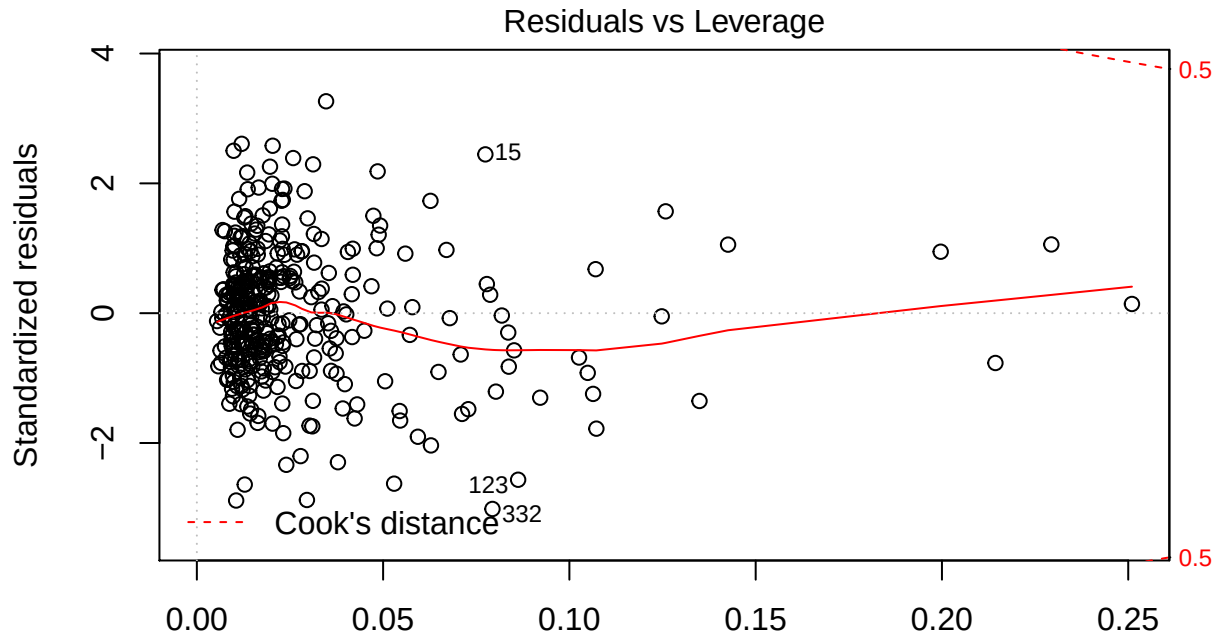
```
                            median_age_female + pct_unemployed16_over + pct_black + pct_other_race +
                            pct_upto_hs18_24,
    data = income_high_data)
```

check assumption and influential points

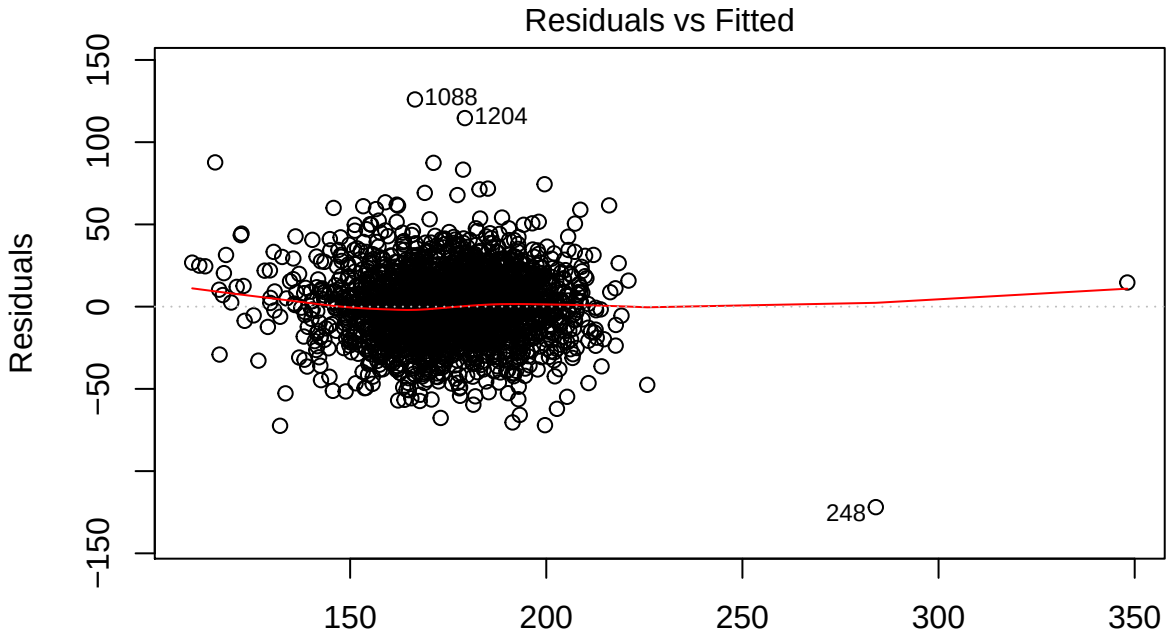
```
plot(criterion_model_low)
```

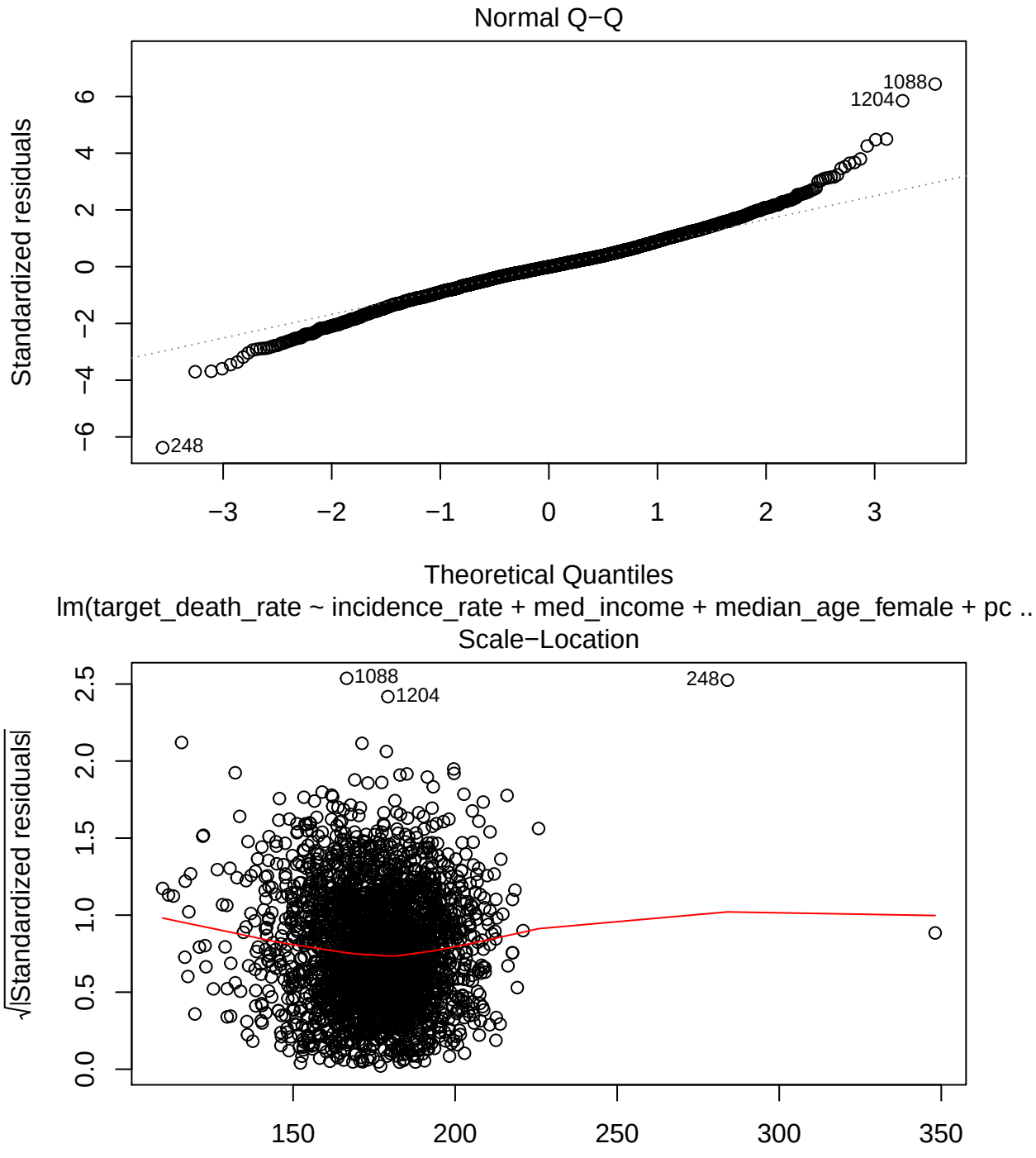


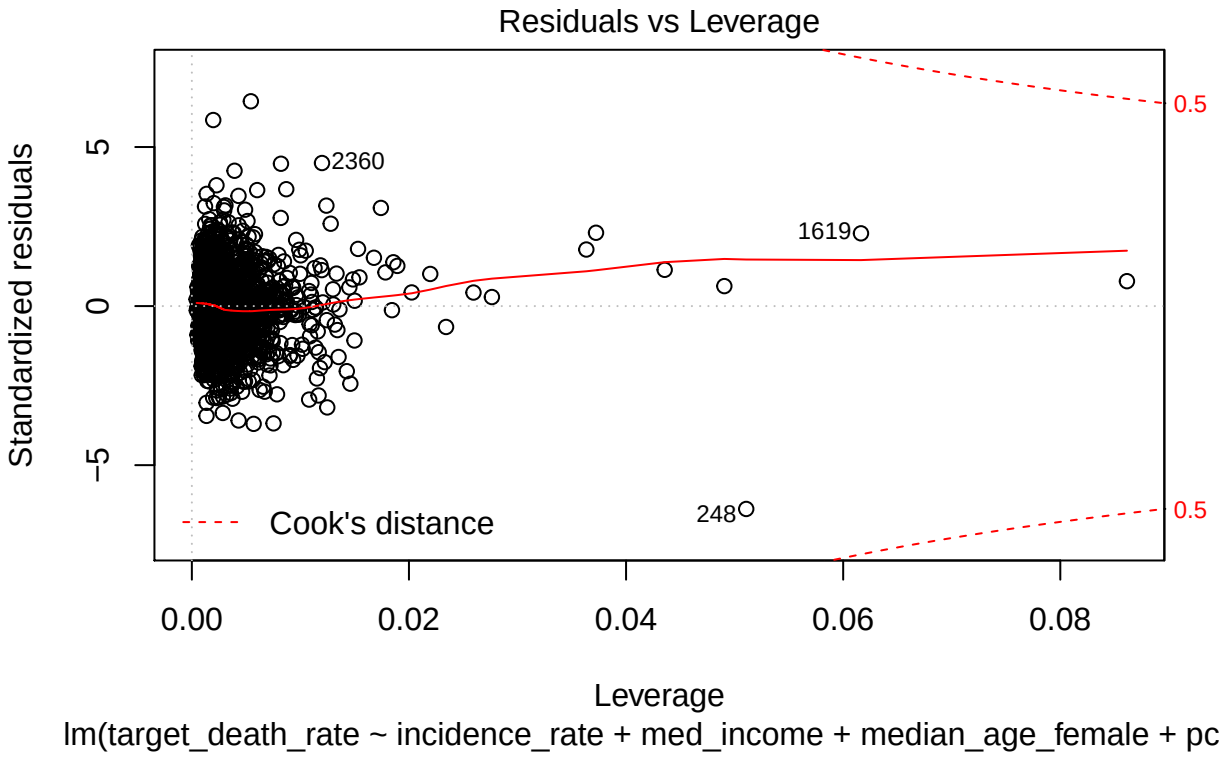

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + medi .
Scale−Location
√|Standardized residuals|
Fitted values
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + medi .




```
plot(criterion_model_high)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_death_rate ~ incidence_rate + med_income + median_age_female + pc ..
Scale−Location
√|Standardized residuals|
Fitted values
lm(target_death_rate ~ incidence_rate + med_income + median_age_female + pc ..
1088
1204
248



Model comparison

backward_model_low <- lm(target_death_rate ~ incidence_rate + med_income + median_age_male + median_age_female + pct_white + pct_black + pct_asian + pct_other_race + pct_upto_hs18_24 + pct_with_coverage, data = income_low_data) 10 predictor

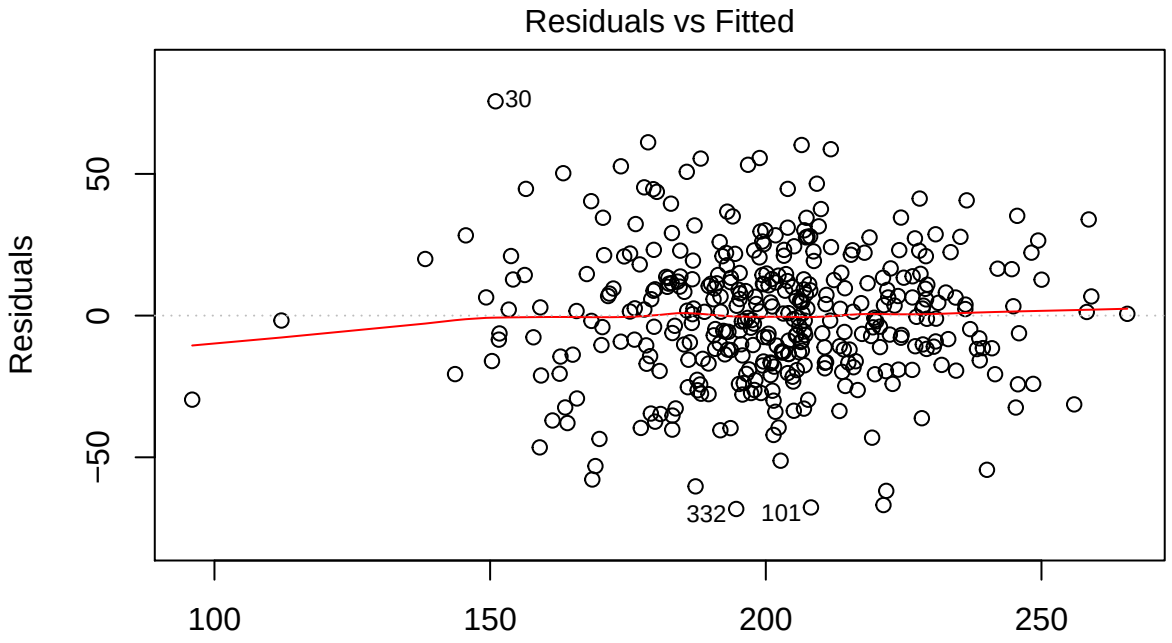
criterion_model_low <- lm(target_death_rate ~ incidence_rate + med_income + median_age_male + median_age_female + pct_white + pct_black + pct_asian + pct_other_race + pct_with_coverage, data = income_low_data) 9 predictor

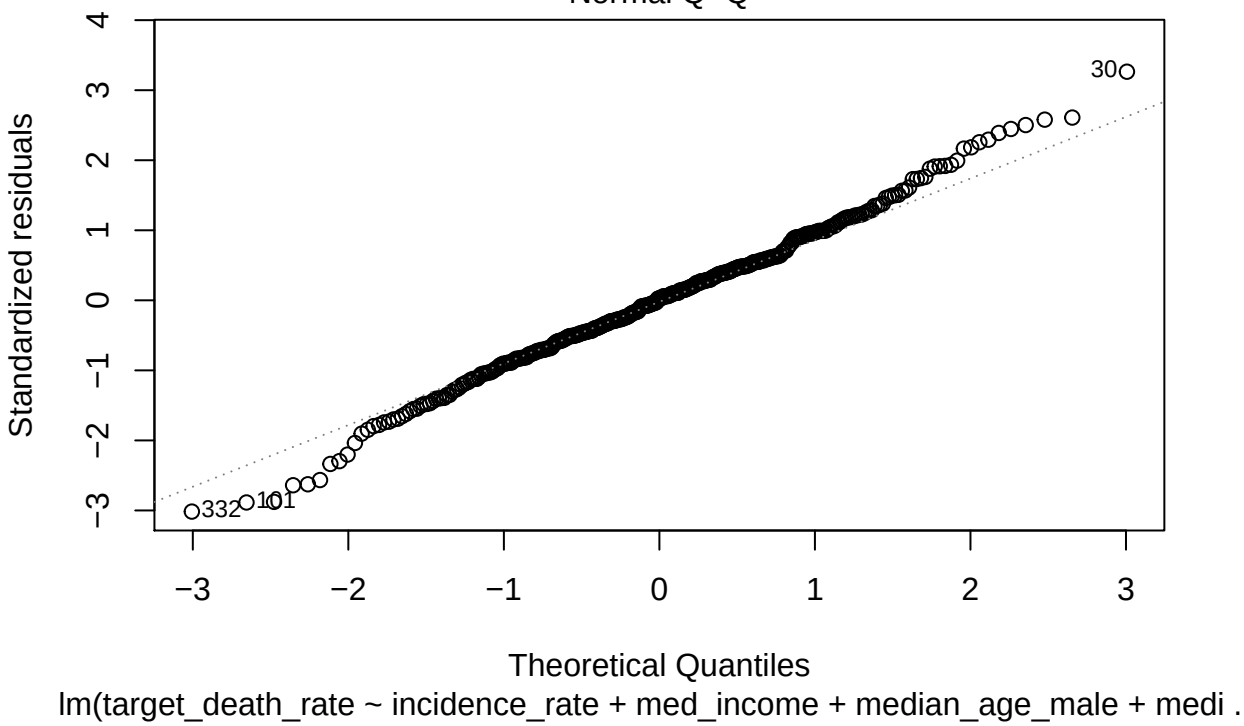
backward_model_high <- lm(target_death_rate ~ incidence_rate + med_income + median_age_male + pct_unemployed16_over + pct_black + pct_asian + pct_other_race + pct_married_households + pct_upto_hs18_24 + pct_with_coverage, data = income_high_data) 10 predictor

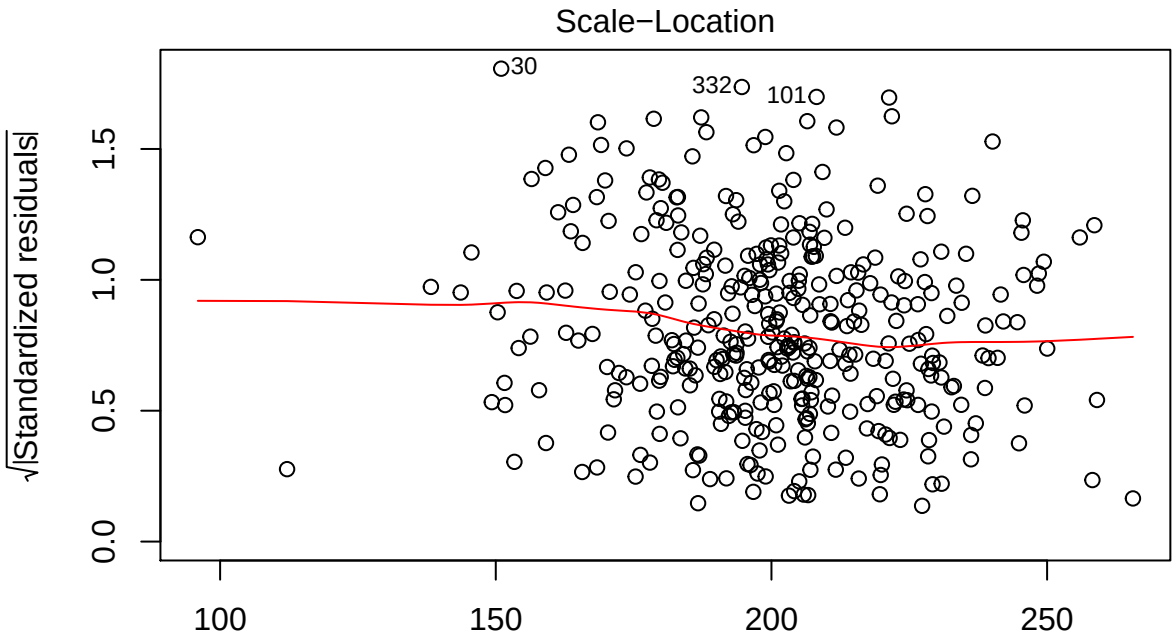
criterion_model_high <- lm(target_death_rate ~ incidence_rate + med_income + median_age_female + pct_unemployed16_over + pct_black + pct_other_race + pct_upto_hs18_24, data = income_high_data) 7 predictor

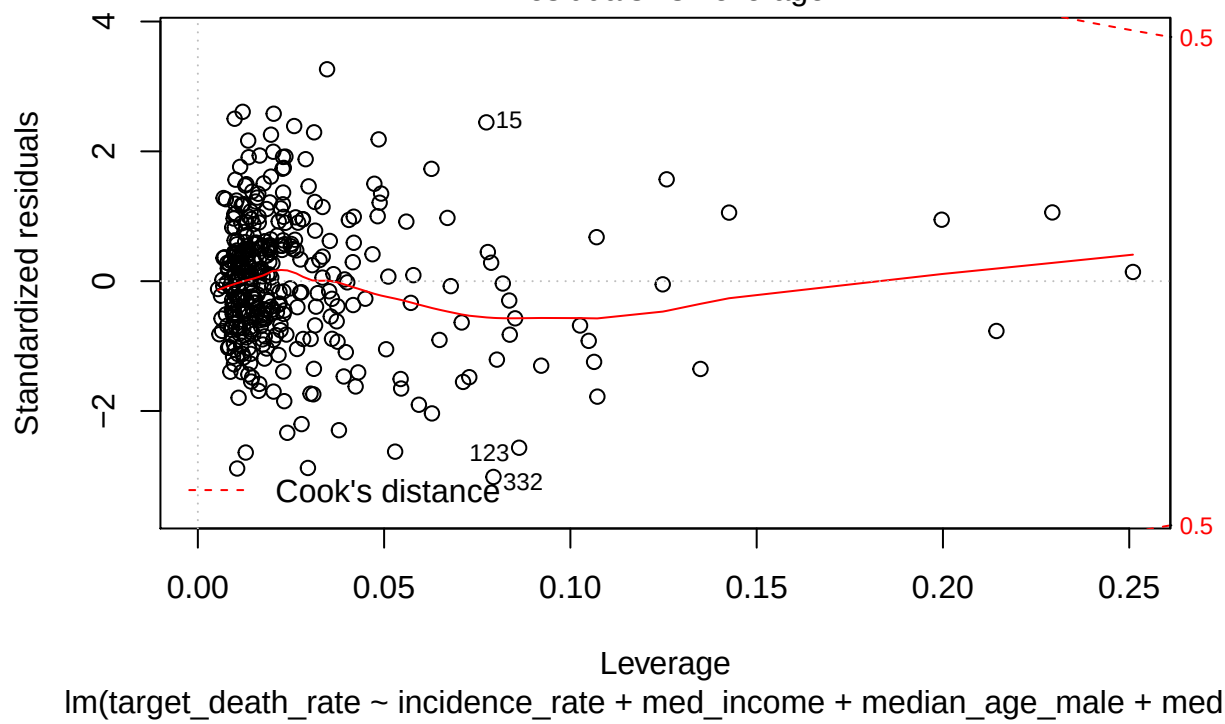
select model with least predictors

```
low_model <- lm(formula = target_death_rate ~ incidence_rate + med_income +
    median_age_male + median_age_female + pct_white + pct_black + pct_asian +
    pct_other_race +  pct_with_coverage, data = income_low_data)
high_model <- lm(formula = target_death_rate ~ incidence_rate + med_income +
                             median_age_female + pct_unemployed16_over + pct_black + pct_other_race +
                             pct_upto_hs18_24,
    data = income_high_data)
plot(low_model)
```

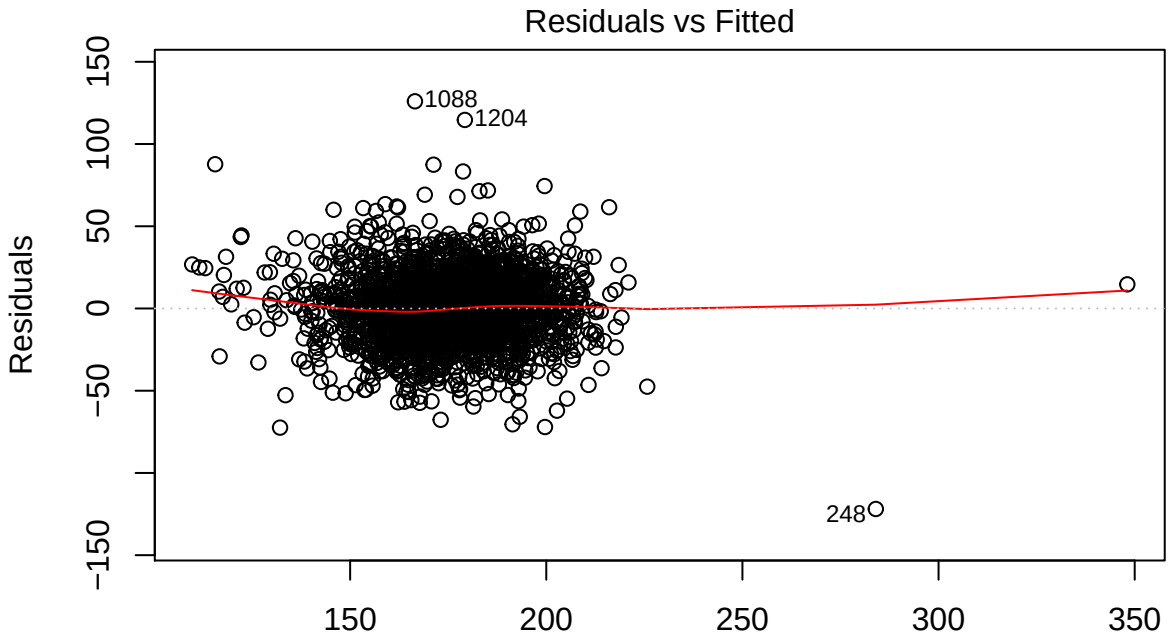

Residuals vs Fitted
Residuals
Fitted values
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + medi ..
30
332
101
50
0
−50
100
150
200
250

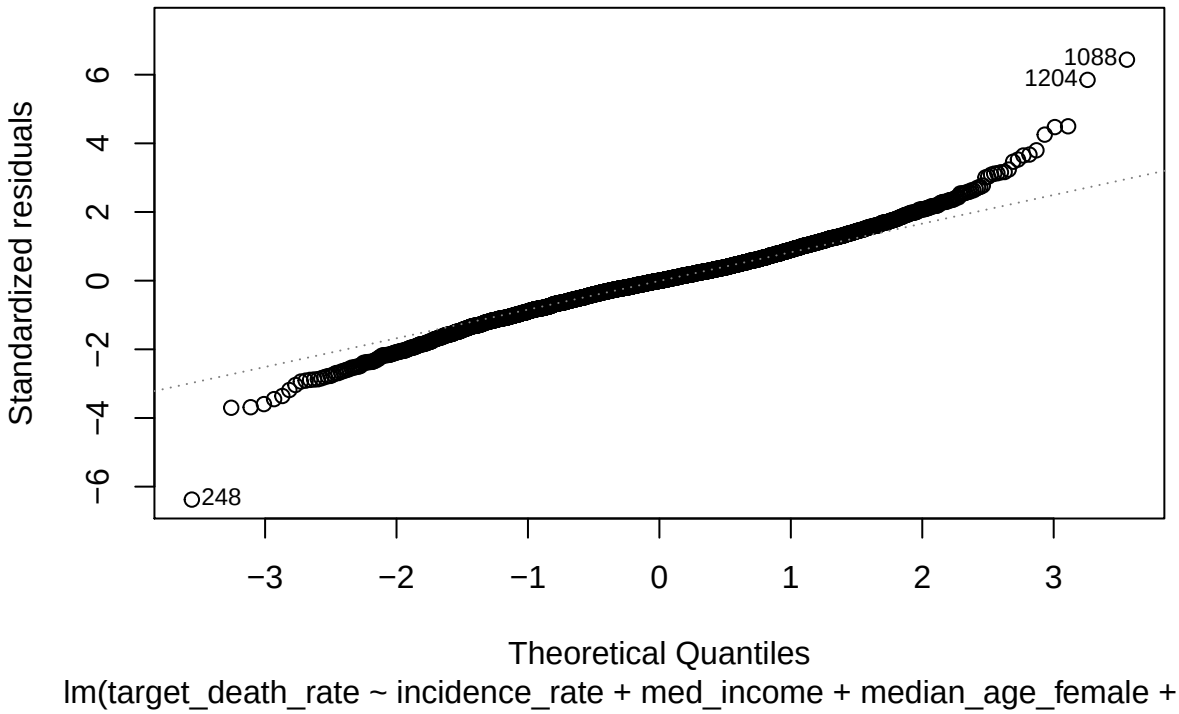

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_death_rate ~ incidence_rate + med_income + median_age_male + medi .
30
332
101
4
3
2
1
0
−1
−2
−3

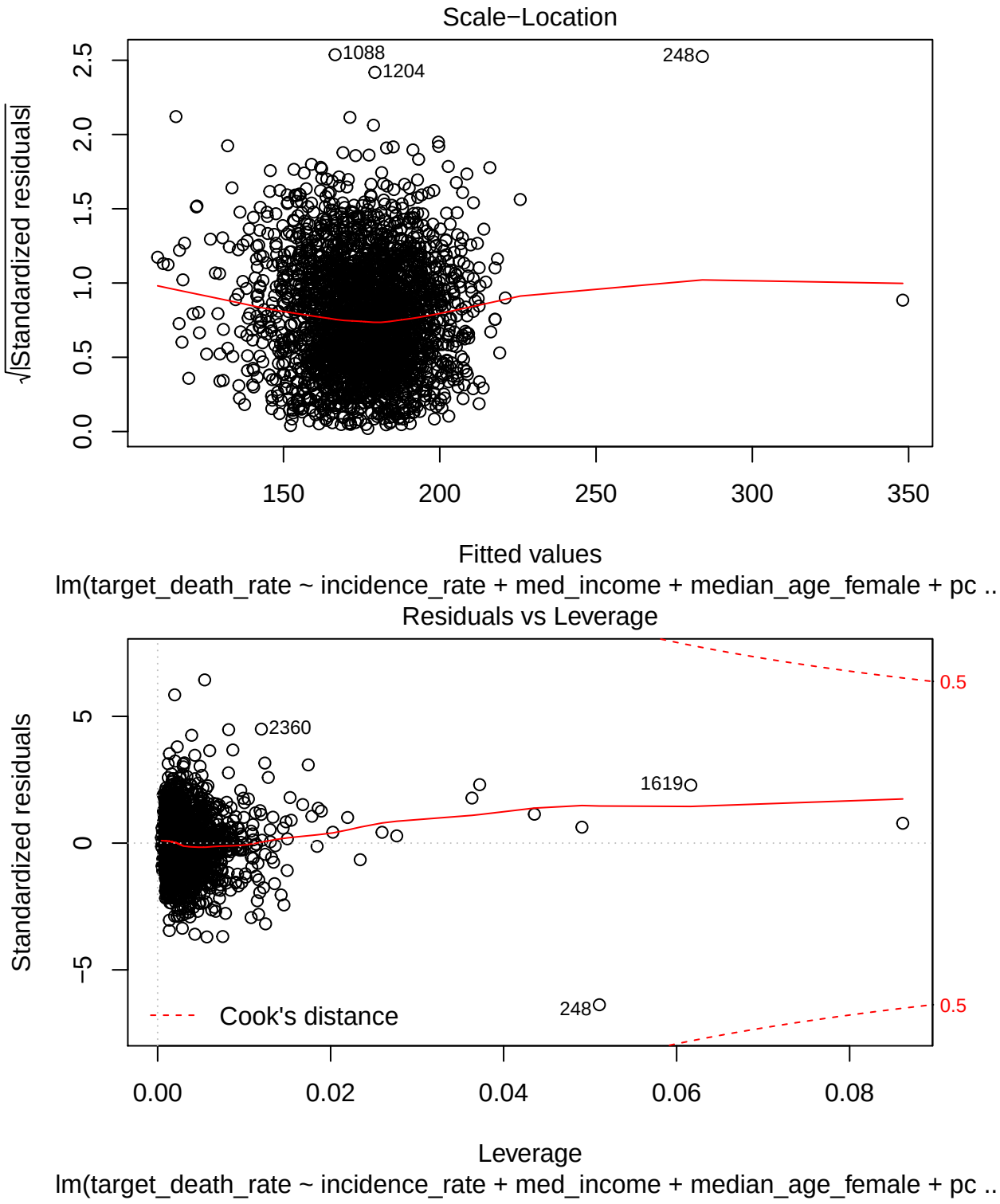





```
plot(high_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(target_death_rate ~ incidence_rate + med_income + median_age_female + pc ..
1088
1204
248
150
100
50
0
−50
−150
150
200
250
300
350


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_death_rate ~ incidence_rate + med_income + median_age_female + pc ..
1088
1204
248
6
4
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3



Influential points

remove influential points in low income

```r
income_low_rm <- income_low_data[-c(30,101,332),]

low_model_rm<- lm(formula = target_death_rate ~ incidence_rate + med_income +
    median_age_male + median_age_female + pct_white + pct_black + pct_asian +
    pct_other_race +  pct_with_coverage, data = income_low_rm)
summary(low_model_rm)
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_male + median_age_female + pct_white + pct_black +
##     pct_asian + pct_other_race + pct_with_coverage, data = income_low_rm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.422 -13.429   0.047  13.349  60.908
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       98.8552983 23.8724038   4.141 4.30e-05 ***
## incidence_rate     0.3077773  0.0196854  15.635  < 2e-16 ***
## med_income        -0.0022657  0.0004811  -4.710 3.53e-06 ***
## median_age_male   -1.4354244  0.5698558  -2.519 0.012198 *
## median_age_female  1.3398542  0.5629921   2.380 0.017831 *
## pct_white         -0.2142685  0.1117099  -1.918 0.055881 .
## pct_black         -0.3772086  0.1113062  -3.389 0.000778 ***
## pct_asian         -5.1084653  1.8393838  -2.777 0.005765 **
## pct_other_race    -0.9765686  0.3077001  -3.174 0.001632 **
## pct_with_coverage  0.8360271  0.2766087   3.022 0.002685 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.72 on 365 degrees of freedom
## Multiple R-squared:  0.5371, Adjusted R-squared:  0.5257
## F-statistic: 47.05 on 9 and 365 DF,  p-value: < 2.2e-16
```

```r
summary(low_model)
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_male + median_age_female + pct_white + pct_black +
##     pct_asian + pct_other_race + pct_with_coverage, data = income_low_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.234 -14.251   0.675  13.363  75.616
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       93.5355565 24.7371687   3.781 0.000182 ***
## incidence_rate     0.2869252  0.0199783  14.362  < 2e-16 ***
## med_income        -0.0024180  0.0004974  -4.861 1.73e-06 ***
## median_age_male   -1.7544224  0.5841892  -3.003 0.002854 **
```

```
## median_age_female  1.6709740  0.5763207   2.899 0.003963 **
## pct_white         -0.2520327  0.1154077  -2.184 0.029604 *
## pct_black         -0.4205350  0.1150302  -3.656 0.000294 ***
## pct_asian         -4.5604501  1.9050976  -2.394 0.017174 *
## pct_other_race    -1.1299233  0.3164051  -3.571 0.000403 ***
## pct_with_coverage  1.1143671  0.2812032   3.963 8.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.58 on 368 degrees of freedom
## Multiple R-squared:  0.5095, Adjusted R-squared:  0.4975
## F-statistic: 42.47 on 9 and 368 DF,  p-value: < 2.2e-16
```

remove influential points in high income

```
income_high_rm <- income_high_data[-c(1088, 1204, 248),]

high_model_rm <- lm(formula = target_death_rate ~ incidence_rate + med_income +
                            median_age_female + pct_unemployed16_over + pct_black + pct_other_race +
                            pct_upto_hs18_24,
    data = income_high_rm)
summary(high_model_rm)
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_female + pct_unemployed16_over + pct_black + pct_other_race +
##     pct_upto_hs18_24, data = income_high_rm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.530 -11.236  -0.015  10.874  90.088
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.062e+02  5.615e+00  18.919  < 2e-16 ***
## incidence_rate         2.056e-01  7.533e-03  27.289  < 2e-16 ***
## med_income            -6.011e-04  3.695e-05 -16.267  < 2e-16 ***
## median_age_female     -5.518e-01  7.994e-02  -6.902 6.39e-12 ***
## pct_unemployed16_over  1.224e+00  1.458e-01   8.394  < 2e-16 ***
## pct_black              1.402e-01  3.708e-02   3.782 0.000159 ***
## pct_other_race        -8.934e-01  1.190e-01  -7.506 8.29e-14 ***
## pct_upto_hs18_24       4.159e-01  3.285e-02  12.657  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.2 on 2658 degrees of freedom
## Multiple R-squared:  0.4185, Adjusted R-squared:  0.417
## F-statistic: 273.3 on 7 and 2658 DF,  p-value: < 2.2e-16
```

```
summary(high_model)
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     median_age_female + pct_unemployed16_over + pct_black + pct_other_race +
```

```
##     pct_upto_hs18_24, data = income_high_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.907  -11.203   -0.072   10.883  125.976
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.086e+02  5.726e+00  18.974  < 2e-16 ***
## incidence_rate          1.955e-01  7.531e-03  25.953  < 2e-16 ***
## med_income             -5.932e-04  3.772e-05 -15.727  < 2e-16 ***
## median_age_female      -5.125e-01  8.155e-02  -6.285 3.83e-10 ***
## pct_unemployed16_over   1.203e+00  1.489e-01   8.076 1.00e-15 ***
## pct_black               1.600e-01  3.781e-02   4.231 2.40e-05 ***
## pct_other_race         -9.101e-01  1.216e-01  -7.484 9.76e-14 ***
## pct_upto_hs18_24        4.188e-01  3.356e-02  12.479  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.62 on 2661 degrees of freedom
## Multiple R-squared:  0.4019, Adjusted R-squared:  0.4003
## F-statistic: 255.4 on 7 and 2661 DF,  p-value: < 2.2e-16
```

cross validation

CV for low and high income model

```
cross_df_low = crossv_mc(income_low_data, n = 100, test = 0.2)
cross_df_high = crossv_mc(income_high_data, n = 100, test = 0.2)

cross_result_low =
  cross_df_low %>%
  mutate(
    step_mod = map(train, ~lm(target_death_rate ~ incidence_rate + med_income +
    median_age_male + median_age_female + pct_white + pct_black + pct_asian +
    pct_other_race +  pct_with_coverage, data = .x)),
    rmse_train = map2_dbl(step_mod, train, ~rmse(model = .x, data = .y)),
    rmse_test = map2_dbl(step_mod, test, ~rmse(model = .x, data = .y))
    )


mse_results_low = cross_result_low %>%
  dplyr::select(rmse_train, rmse_test) %>%
  summarize(mse_train_low = (mean(rmse_train))^2,
            mse_test_low = (mean(rmse_test))^2) #mse results

cross_result_high =
  cross_df_high %>%
  mutate(
    step_mod = map(train, ~lm(target_death_rate ~ incidence_rate + med_income +
                              median_age_female + pct_unemployed16_over + pct_black + pct_other_race +
                              pct_upto_hs18_24, data = .x)),
    rmse_train = map2_dbl(step_mod, train, ~rmse(model = .x, data = .y)),
```

```
    rmse_test = map2_dbl(step_mod, test, ~rmse(model = .x, data = .y))
    )


mse_results_high = cross_result_high %>%
  dplyr::select(rmse_train, rmse_test) %>%
  summarize(mse_train_high = (mean(rmse_train))^2,
            mse_test_high = (mean(rmse_test))^2)

#LOOCV
glm.fit_low = glm(target_death_rate ~ incidence_rate + med_income +
    median_age_male + median_age_female + pct_white + pct_black + pct_asian +
    pct_other_race +  pct_with_coverage, data = income_low_data)

cv.err_low = cv.glm(income_low_data, glm.fit_low)

glm.fit_high = glm(target_death_rate ~ incidence_rate + med_income +
                             median_age_female + pct_unemployed16_over + pct_black + pct_other_race +
                             pct_upto_hs18_24, data = income_high_data)

cv.err_high = cv.glm(income_high_data, glm.fit_high)

# The two delta values should be similar: we use the first one
# The second value is bias corrected
cv.err_low$delta
```

```
## [1] 579.0604 579.0086
```

```
anova(low_model)
```

```
## Analysis of Variance Table
##
## Response: target_death_rate
##                    Df Sum Sq Mean Sq  F value    Pr(>F)
## incidence_rate      1 175440  175440 315.6066 < 2.2e-16 ***
## med_income          1   9292    9292  16.7156 5.336e-05 ***
## median_age_male     1   1048    1048   1.8856   0.17053
## median_age_female   1   1873    1873   3.3699   0.06720 .
## pct_white           1   3267    3267   5.8774   0.01582 *
## pct_black           1   1888    1888   3.3955   0.06618 .
## pct_asian           1   3277    3277   5.8958   0.01566 *
## pct_other_race      1   7648    7648  13.7578   0.00024 ***
## pct_with_coverage   1   8730    8730  15.7042 8.896e-05 ***
## Residuals         368 204565     556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(low_model)[10, 3] #MSE: 431
```

```
## [1] 555.8834
```

```
anova(high_model)[8, 3]
```

```
## [1] 385.1276
```

```r
mse_low =
  tibble(
  mse_model = anova(low_model)[10, 3],
  mse_LOOCV = cv.err_low$delta[1],
  mse_CV_train = mse_results_low$mse_train_low,
  mse_CV_test = mse_results_low$mse_test_low
)

mse_high =
  tibble(
  mse_model = anova(high_model)[8, 3],
  mse_LOOCV = cv.err_high$delta[1],
  mse_CV_train = mse_results_high$mse_train_high,
  mse_CV_test = mse_results_high$mse_test_high
)

rbind(mse_low, mse_high) %>% mutate(dataset = c("low income", "high income")) %>%
  dplyr::select(dataset, everything()) %>% knitr::kable(digits = 3)
```

dataset	mse_model	mse_LOOCV	mse_CV_train	mse_CV_test
low income	555.883	579.060	537.696	571.741
high income	385.128	387.517	382.965	389.582