Stoke Data

Miles Tweed

5/11/2021

```
Stroke <- read_csv('../../../Data/healthcare-dataset-stroke-data.csv')
Stroke$bmi <- Stroke$bmi %>% as.numeric()
```

```
## Warning in Stroke$bmi %>% as.numeric(): NAs introduced by coercion
```

```
Stroke <- Stroke %>% mutate(bmi2 = ifelse(is.na(bmi), median(bmi, na.rm = TRUE), bmi)) %>% select(-bmi)
```

Expectation 2

```
# If you got Linux
Stroke$ever_married <- factor(Stroke$ever_married)
Stroke$smoking_status <- factor(Stroke$smoking_status)


con.table <- table(Stroke$ever_married, Stroke$smoking_status)

chisq.test(con.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  con.table
## X-squared = 599.05, df = 3, p-value < 2.2e-16
```

```
pchisq(599.05, df=3, lower.tail=F)
```

```
## [1] 1.619348e-129
```

Expectation 3

```
lm.obj <- lm(bmi2~., Stroke)

summary(lm.obj)
```

```
##
## Call:
## lm(formula = bmi2 ~ ., data = Stroke)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.960  -4.382  -1.160   3.222  67.585
##
```

```
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.931e+01  5.013e-01  38.523  < 2e-16 ***
## id                           -8.254e-07  4.476e-06  -0.184 0.853707
## genderMale                    6.780e-02  1.948e-01   0.348 0.727859
## genderOther                  -7.514e+00  6.775e+00  -1.109 0.267500
## age                          -1.474e-02  7.228e-03  -2.040 0.041413 *
## hypertension                  2.207e+00  3.365e-01   6.560 5.93e-11 ***
## heart_disease                -8.912e-01  4.425e-01  -2.014 0.044060 *
## ever_marriedYes               2.049e+00  2.800e-01   7.320 2.86e-13 ***
## work_typeGovt_job             8.372e+00  4.860e-01  17.227  < 2e-16 ***
## work_typeNever_worked         5.206e+00  1.474e+00   3.532 0.000417 ***
## work_typePrivate              8.375e+00  4.037e-01  20.747  < 2e-16 ***
## work_typeSelf-employed        7.911e+00  4.975e-01  15.901  < 2e-16 ***
## Residence_typeUrban           1.023e-02  1.896e-01   0.054 0.956986
## avg_glucose_level             1.855e-02  2.189e-03   8.471  < 2e-16 ***
## smoking_statusnever smoked -3.692e-01  2.799e-01  -1.319 0.187340
## smoking_statussmokes       -2.587e-01  3.350e-01  -0.772 0.440019
## smoking_statusUnknown      -7.917e-01  3.164e-01  -2.502 0.012378 *
## stroke                     -7.709e-01  4.593e-01  -1.678 0.093329 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.765 on 5092 degrees of freedom
## Multiple R-squared:  0.2305, Adjusted R-squared:  0.2279
## F-statistic: 89.73 on 17 and 5092 DF,  p-value: < 2.2e-16
```

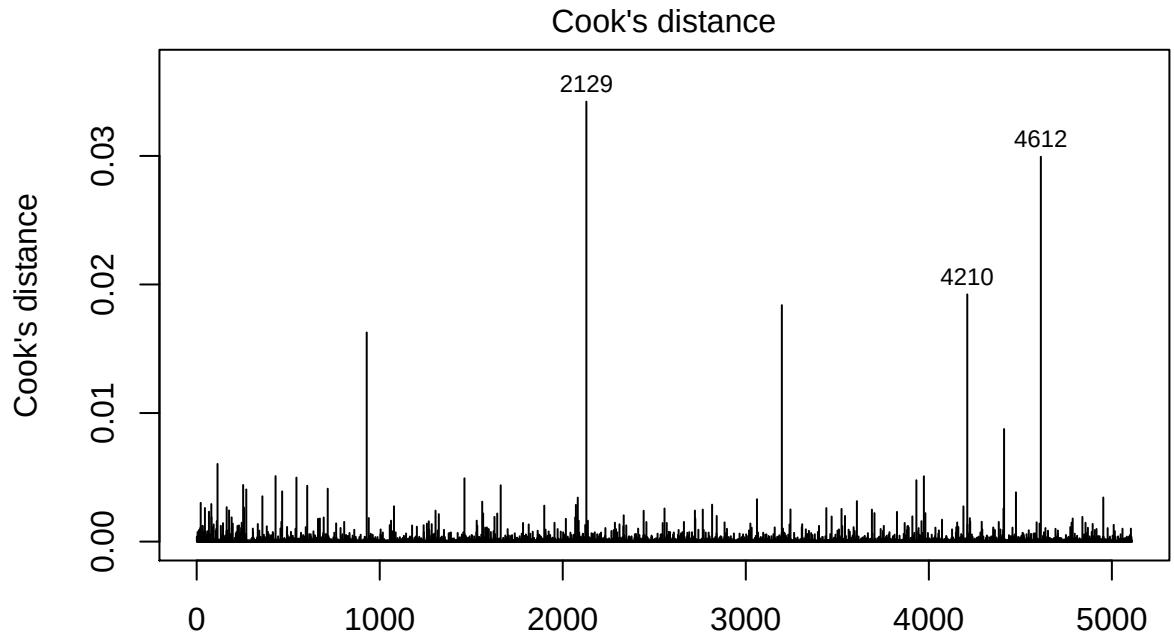
```
step(lm.obj, trace = 0)
```

```
##
## Call:
## lm(formula = bmi2 ~ age + hypertension + heart_disease + ever_married +
##     work_type + avg_glucose_level + smoking_status + stroke,
##     data = Stroke)
##
## Coefficients:
##                (Intercept)                        age
##                   19.31341                   -0.01466
##               hypertension              heart_disease
##                    2.21029                   -0.87789
##           ever_marriedYes          work_typeGovt_job
##                    2.05603                    8.35619
##      work_typeNever_worked           work_typePrivate
##                    5.20647                    8.35698
##     work_typeSelf-employed          avg_glucose_level
##                    7.89336                    0.01854
## smoking_statusnever smoked       smoking_statussmokes
##                   -0.36663                   -0.25045
##      smoking_statusUnknown                     stroke
##                   -0.78351                   -0.77102
```

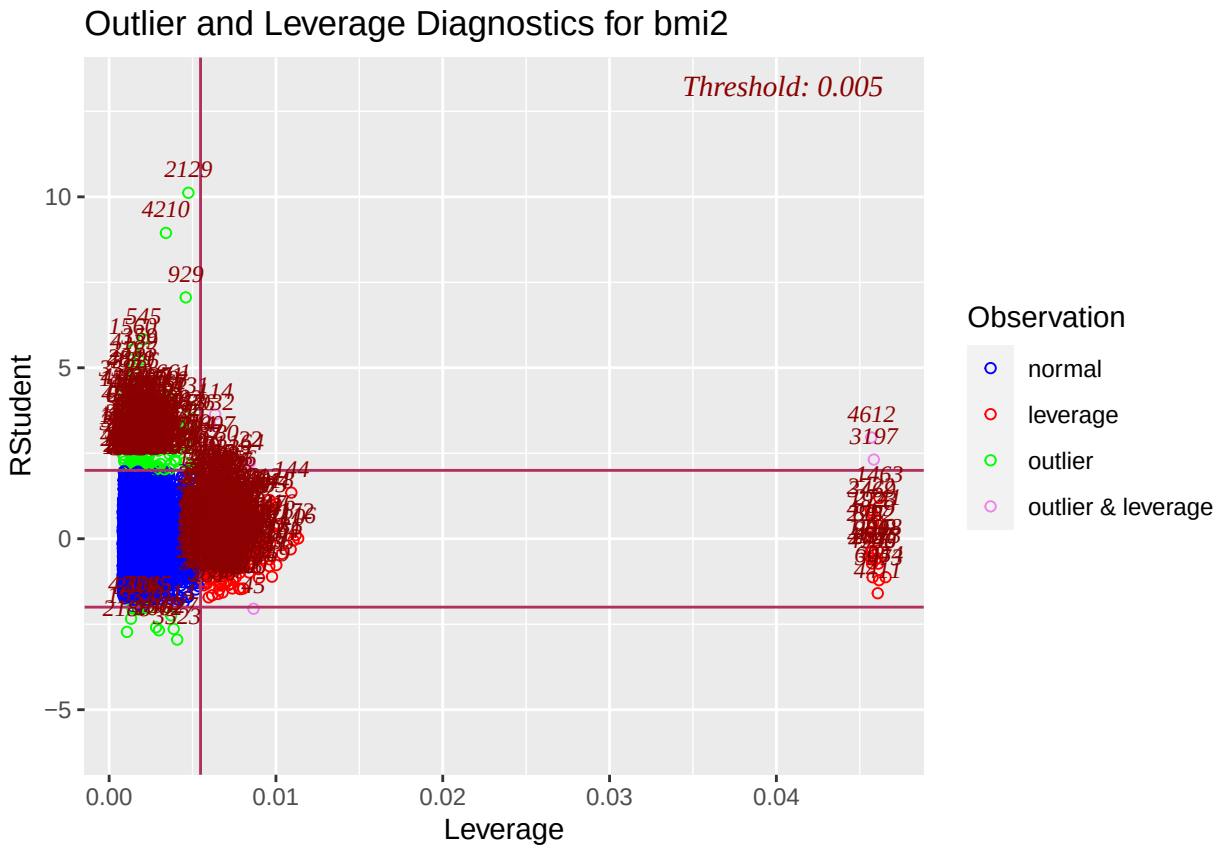
```
lm.reduced <- lm(bmi2 ~ age + hypertension + heart_disease + ever_married +
    work_type + avg_glucose_level + smoking_status + stroke, data = Stroke)

summary(lm.reduced)
```

```
##
## Call:
## lm(formula = bmi2 ~ age + hypertension + heart_disease + ever_married +
##     work_type + avg_glucose_level + smoking_status + stroke,
##     data = Stroke)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.893  -4.401  -1.165   3.215  67.608
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  19.313413   0.451889  42.739  < 2e-16 ***
## age                          -0.014657   0.007225  -2.029 0.042551 *
## hypertension                  2.210288   0.336323   6.572 5.46e-11 ***
## heart_disease                -0.877893   0.440972  -1.991 0.046555 *
## ever_marriedYes               2.056033   0.279794   7.348 2.32e-13 ***
## work_typeGovt_job             8.356190   0.485056  17.227  < 2e-16 ***
## work_typeNever_worked         5.206471   1.473367   3.534 0.000413 ***
## work_typePrivate              8.356983   0.402613  20.757  < 2e-16 ***
## work_typeSelf-employed        7.893360   0.496253  15.906  < 2e-16 ***
## avg_glucose_level             0.018543   0.002186   8.483  < 2e-16 ***
## smoking_statusnever smoked -0.366628   0.278968  -1.314 0.188829
## smoking_statussmokes       -0.250454   0.334741  -0.748 0.454373
## smoking_statusUnknown      -0.783508   0.316104  -2.479 0.013221 *
## stroke                     -0.771019   0.459162  -1.679 0.093177 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.764 on 5096 degrees of freedom
## Multiple R-squared:  0.2303, Adjusted R-squared:  0.2283
## F-statistic: 117.3 on 13 and 5096 DF,  p-value: < 2.2e-16
```

```
plot(lm.reduced, which=4)
```




```
ols_plot_resid_lev(lm.reduced)
```



Expectation 6

```
glm.obj <- glm(stroke~., Stroke, family='binomial')

summary(glm.obj)
```

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = Stroke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1457  -0.3206  -0.1640  -0.0869   3.5512
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.843e+00  7.959e-01  -8.598  < 2e-16 ***
## id                            1.572e-06  3.220e-06   0.488 0.625378
## genderMale                    1.212e-02  1.419e-01   0.085 0.931949
## genderOther                  -1.056e+01  1.455e+03  -0.007 0.994212
## age                           7.475e-02  5.831e-03  12.819  < 2e-16 ***
## hypertension                  4.017e-01  1.651e-01   2.433 0.014973 *
## heart_disease                 2.783e-01  1.913e-01   1.455 0.145630
## ever_marriedYes              -1.852e-01  2.255e-01  -0.821 0.411407
## work_typeGovt_job            -9.446e-01  8.359e-01  -1.130 0.258465
## work_typeNever_worked        -1.033e+01  3.092e+02  -0.033 0.973343
## work_typePrivate             -8.009e-01  8.198e-01  -0.977 0.328621
## work_typeSelf-employed       -1.175e+00  8.406e-01  -1.397 0.162351
## Residence_typeUrban           8.519e-02  1.384e-01   0.616 0.538197
## avg_glucose_level             4.044e-03  1.199e-03   3.374 0.000741 ***
## smoking_statusnever smoked -2.073e-01  1.760e-01  -1.178 0.238668
## smoking_statussmokes          1.112e-01  2.154e-01   0.516 0.605588
## smoking_statusUnknown        -7.013e-02  2.084e-01  -0.336 0.736524
## bmi2                          1.119e-03  1.133e-02   0.099 0.921328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1990.4  on 5109  degrees of freedom
## Residual deviance: 1580.9  on 5092  degrees of freedom
## AIC: 1616.9
##
## Number of Fisher Scoring iterations: 14
```

```
step(glm.obj, trace = 0)
```

```
##
## Call:  glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
##     family = "binomial", data = Stroke)
##
## Coefficients:
##       (Intercept)                age       hypertension      heart_disease
##         -7.489396           0.068926           0.381410           0.329965
## avg_glucose_level
```

```
##          0.004121
##
## Degrees of Freedom: 5109 Total (i.e. Null);  5105 Residual
## Null Deviance:       1990
## Residual Deviance: 1591  AIC: 1601
```

```
glm.reduced <- glm(stroke ~ age + hypertension + heart_disease + avg_glucose_level, data = Stroke, famil

summary(glm.reduced)
```

```
##
## Call:
## glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
##     family = "binomial", data = Stroke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0587  -0.3215  -0.1731  -0.0828   3.7707
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.489396   0.357879 -20.927  < 2e-16 ***
## age                0.068926   0.005140  13.410  < 2e-16 ***
## hypertension       0.381410   0.162599   2.346  0.01899 *
## heart_disease      0.329965   0.187724   1.758  0.07880 .
## avg_glucose_level  0.004121   0.001162   3.546  0.00039 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1990.4  on 5109  degrees of freedom
## Residual deviance: 1591.5  on 5105  degrees of freedom
## AIC: 1601.5
##
## Number of Fisher Scoring iterations: 7
```