Day10

Olivia Wu

2024-03-19

Problem 4.8

a) $\hat{TsqrtMDs} = -3.17 + 6.79 Hospitals$

```
##
## Call:
## lm(formula = TsqrtMDs ~ Hospitals, data = Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.582  -6.362  -2.918   8.277  23.170
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.1695     2.6915  -1.178    0.247
## Hospitals     6.7853     0.5284  12.841 2.19e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.627 on 33 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8282
## F-statistic: 164.9 on 1 and 33 DF,  p-value: 2.194e-14
```

b) The cross-validation correlation is 0.953.

```
Prediction <- predict(modelT, Holdout)
Prediction
```

```
##        1        2        3        4        5        6        7        8
## 64.68383 23.97183 17.18649 17.18649 17.18649 30.75716 10.40116 10.40116
##        9       10       11       12       13       14       15       16
## 71.46916 23.97183 10.40116 10.40116 44.32783 10.40116 10.40116 10.40116
##       17       18
## 10.40116 10.40116
```

```
cor(Prediction,sqrt(Holdout$MDs))
```

```
## [1] 0.9531439
```

c) The shrinkage is $0.8332 - 0.908 = -0.075$, which is close to zero, so our coefficient of determinations are similar. The model for our training sample seems to be effective.

Problem 4.9

a) $\hat{GPA} = 1.147 + 0.466HSGPA + 0.015HU + 0.199White$

All predictors are significant. The estimated standard deviation of the error term is 0.3773, and $R^2 = 0.2842$. This shows that a small percent of variability is explained by the model.

```
##
## Call:
## lm(formula = GPA ~ HSGPA + HU + I(White), data = Training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09844 -0.23079  0.03517  0.23600  0.82933
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.147478   0.311524   3.683 0.000323 ***
## HSGPA       0.466053   0.088393   5.273 4.75e-07 ***
## HU          0.015328   0.004091   3.747 0.000257 ***
## I(White)    0.199174   0.076152   2.615 0.009846 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3773 on 146 degrees of freedom
## Multiple R-squared:  0.2842, Adjusted R-squared:  0.2695
## F-statistic: 19.32 on 3 and 146 DF,  p-value: 1.319e-10
```

b)

```
##        1        2        3        4        5        6        7        8
## 3.152561 3.040603 2.960027 2.730315 2.944078 3.268243 3.604737 3.245045
##        9       10       11       12       13       14       15       16
## 3.410132 3.061938 2.901408 3.073537 3.014712 2.904827 3.065147 3.413655
##       17       18       19       20       21       22       23       24
## 3.061895 3.092490 2.990063 3.045884 3.201649 3.466889 3.354205 2.957334
##       25       26       27       28       29       30       31       32
## 2.718183 2.578471 3.191812 3.288337 3.151835 3.002475 2.926162 2.550298
##       33       34       35       36       37       38       39       40
## 3.306568 3.192433 2.954020 3.324895 3.017908 2.965929 2.780842 2.575170
##       41       42       43       44       45       46       47       48
## 2.759209 3.317013 2.960027 3.181870 3.187047 3.022686 2.640300 3.277669
##       49       50       51       52       53       54       55       56
## 3.394287 2.882145 3.195629 3.192223 2.963844 2.962720 2.894557 3.014607
##       57       58       59       60       61       62       63       64
## 2.437825 2.681637 3.257576 3.210957 3.252399 2.740256 3.245771 3.257060
##       65       66       67       68       69
## 3.089176 3.122512 2.945009 2.970695 2.810164
```

```
##             1             2             3             4             5
##  0.1774394377  0.7093971450  0.7199727824 -0.5703145067 -0.4140782175
##             6             7             8             9            10
## -0.0382432643 -0.2047372186 -0.6550454987  0.3798684276  0.4580618768
```

```
##           11           12           13           14           15
##  0.1985923190  0.2664629939 -0.2747119553  0.2351734504 -0.7751471876
##           16           17           18           19           20
## -0.2536553431 -0.0018953044 -0.3524895565 -0.3600627205  0.0141157790
##           21           22           23           24           25
##  0.7683507724 -0.7868886116  0.2257948302  0.6026659164 -0.0881825223
##           26           27           28           29           30
## -0.6084714179  0.4681877699  0.2216631324 -0.0818348277  0.1975249313
##           31           32           33           34           35
## -0.3461618179 -0.6202984124  0.2034320264 -0.4424330626  0.1059798830
##           36           37           38           39           40
## -0.4148954015 -0.4579081385 -0.3659294161 -0.3708417600 -0.0751703326
##           41           42           43           44           45
##  0.1907910433 -0.5770129223 -0.1300272176  0.0181296694  0.4429532055
##           46           47           48           49           50
## -0.5626864554  0.3596999810  0.2423307665 -0.2842874743 -0.1221447143
##           51           52           53           54           55
##  0.1743707542 -0.6322232584 -0.5838442333  0.7272796483 -0.7945570625
##           56           57           58           59           60
##  0.1053929468 -0.1978247748 -0.4516368696  0.4624243698  0.4690425865
##           61           62           63           64           65
##  0.4176008337 -0.2202564063  0.0842287668  0.0529403002  0.0408244100
##           66           67           68           69
## -0.2425121782 -0.2950094662 -0.0006948518 -0.1901644095
```

c) The mean (-0.059) is reasonably close to 0, and the standard deviation of the error term (0.407) is also close to the one provided by the output.

```
mean(Error)
```

```
## [1] -0.05947226
```

```
sd(Error)
```

```
## [1] 0.4065554
```

d) The cross-validation correlation is 0.596.

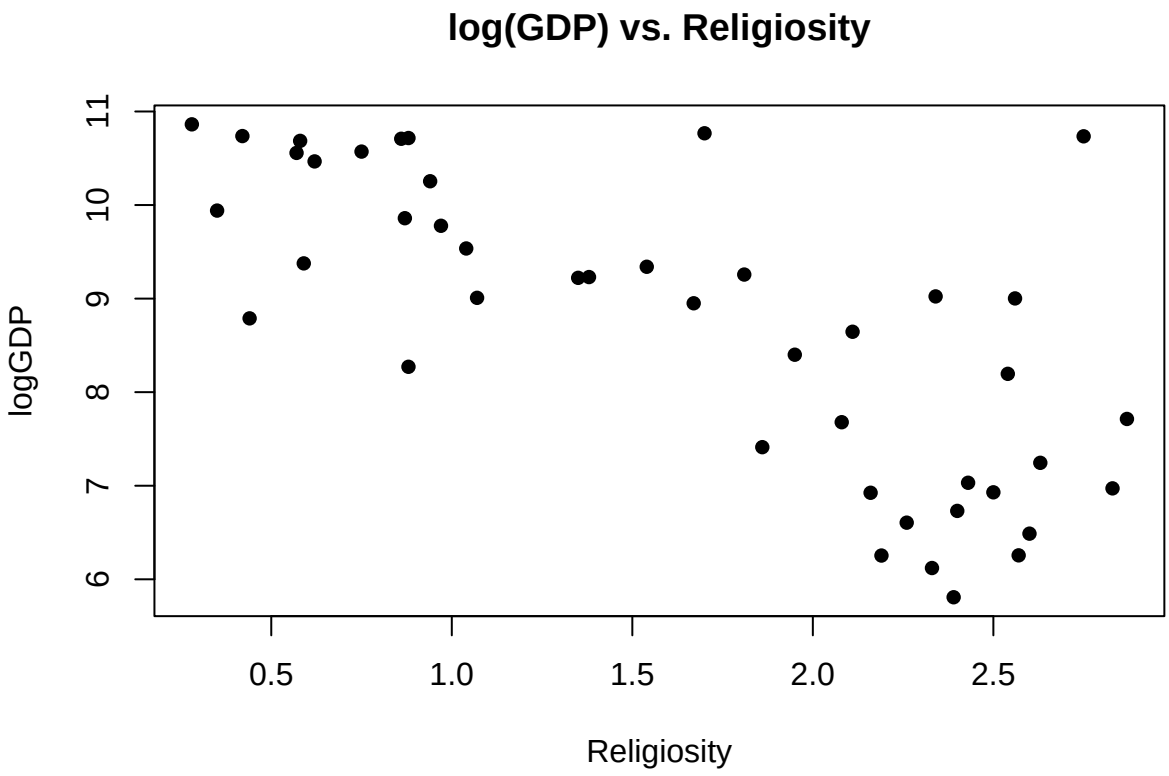
```
## [1] 0.5960115
```

e) $0.2842 - 0.596^2 = -0.071$. There is little change in the amount of variability explained.

```
## [1] -0.07102966
```

Problem 4.11

a)



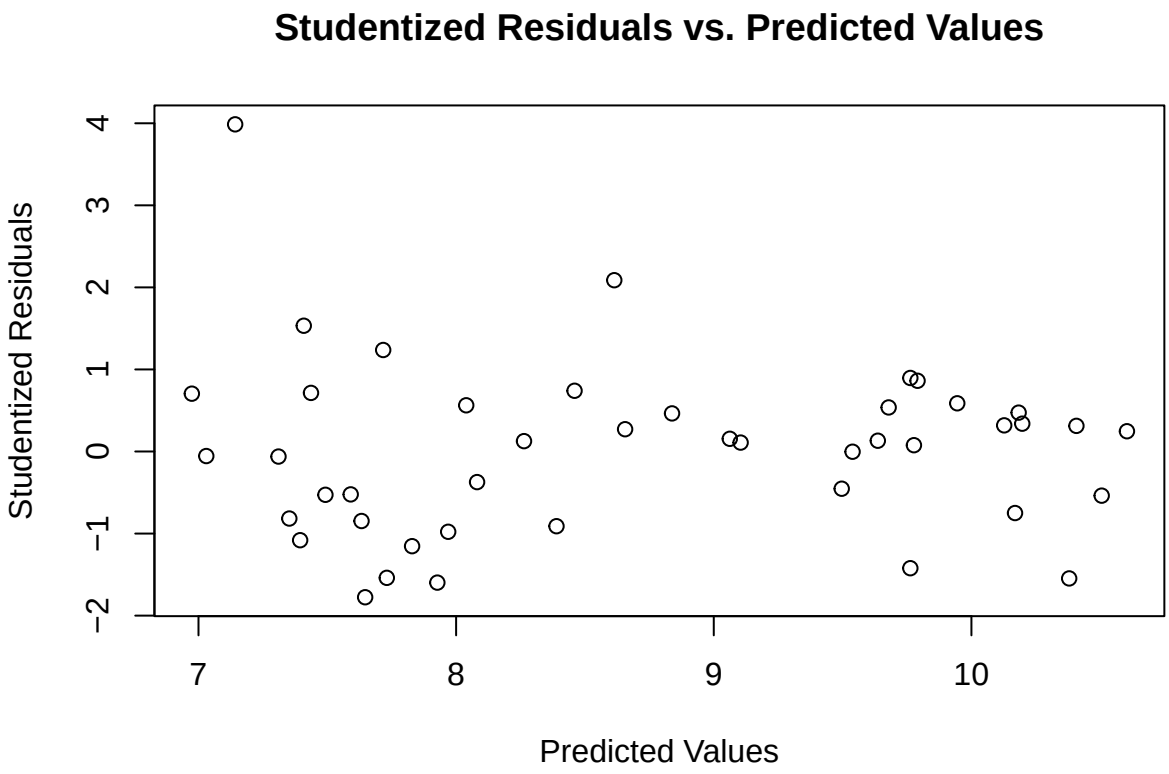
b) $log(\hat{GDP}) = 11 - 1.4 Religiosity$

53.88% of the variability in log(GDP) is explained by this model.

```
##
## Call:
## lm(formula = logGDP ~ Religiosity, data = GDP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8387 -0.8108  0.1272  0.5833  3.5923
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.9961     0.3656  30.079  < 2e-16 ***
## Religiosity  -1.4013     0.2001  -7.005 1.43e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.085 on 42 degrees of freedom
## Multiple R-squared:  0.5388, Adjusted R-squared:  0.5278
## F-statistic: 49.06 on 1 and 42 DF,  p-value: 1.432e-08
```

c) For every one percentage point of increase in Religiosity, the GDP of that country tends to decrease by 1.4.

d) The magnitude of the residual for Kuwait is 3.987.



e)$log(\hat{GDP}) = 10.8 - 0.998Religiosity - 1.59Africa - 0.608Asia + 0.344MiddleEast - 0.803EastEurope + 0.84WestEurope$

```
##
## Call:
## lm(formula = logGDP ~ Religiosity + I(Africa) + I(Asia) + I(MiddleEast) +
##     I(EastEurope) + I(WestEurope) + I(Americas), data = GDP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5274 -0.5720 -0.0760  0.5457  2.3395
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.7956     0.5365  20.124  < 2e-16 ***
## Religiosity    -0.9979     0.2852  -3.498  0.00124 **
## I(Africa)      -1.5937     0.4778  -3.336  0.00195 **
## I(Asia)        -0.6081     0.4689  -1.297  0.20265
## I(MiddleEast)   0.3437     0.4852   0.708  0.48314
## I(EastEurope)  -0.8035     0.5304  -1.515  0.13830
## I(WestEurope)   0.4601     0.5479   0.840  0.40646
## I(Americas)         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8947 on 37 degrees of freedom
```

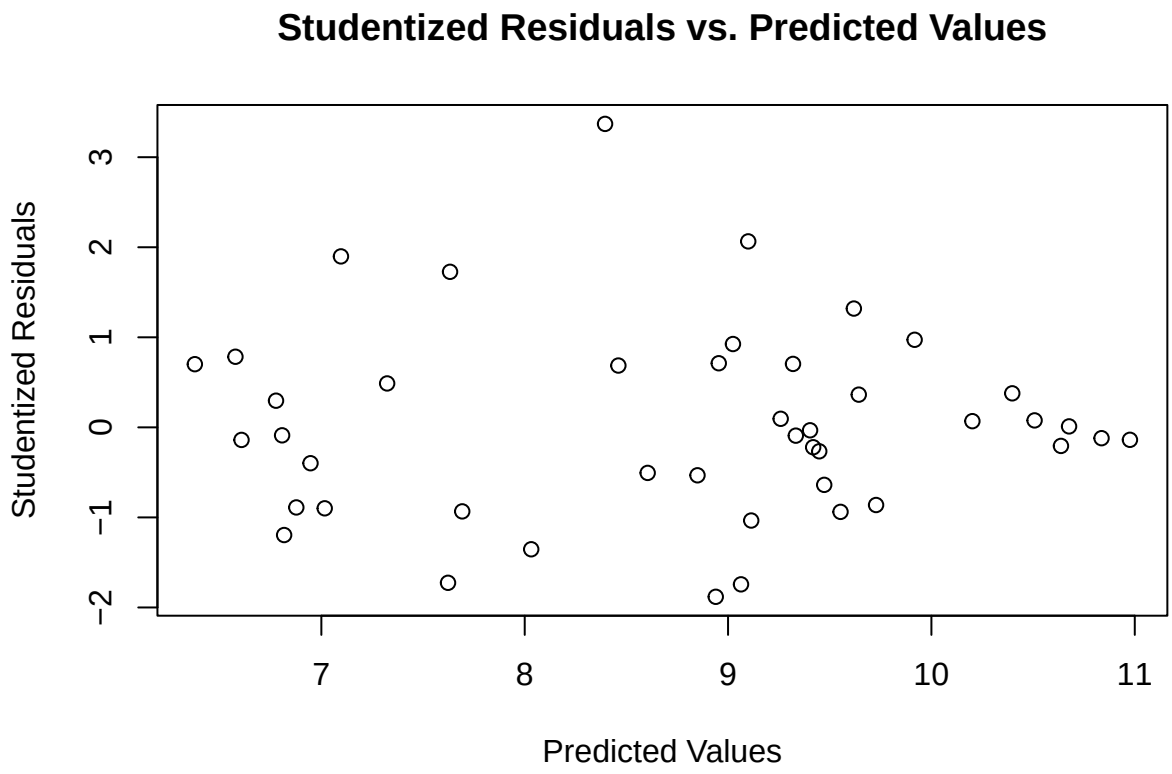
```
## Multiple R-squared:  0.7235, Adjusted R-squared:  0.6787
## F-statistic: 16.14 on 6 and 37 DF,  p-value: 5.095e-09
```

f) For every one percentage point of increase in Religiosity, the GDP of that country tends to decrease by 0.998.

g) We can use a nested F-test. The output shows signiicance at the 0.05 level.

```
## Analysis of Variance Table
##
## Model 1: logGDP ~ Religiosity
## Model 2: logGDP ~ Religiosity + I(Africa) + I(Asia) + I(MiddleEast) +
##     I(EastEurope) + I(WestEurope) + I(Americas)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     42 49.405
## 2     37 29.615  5     19.79 4.9449 0.001448 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

h) The magnitude of the residual for Kuwait is 3.37, which is better than before.

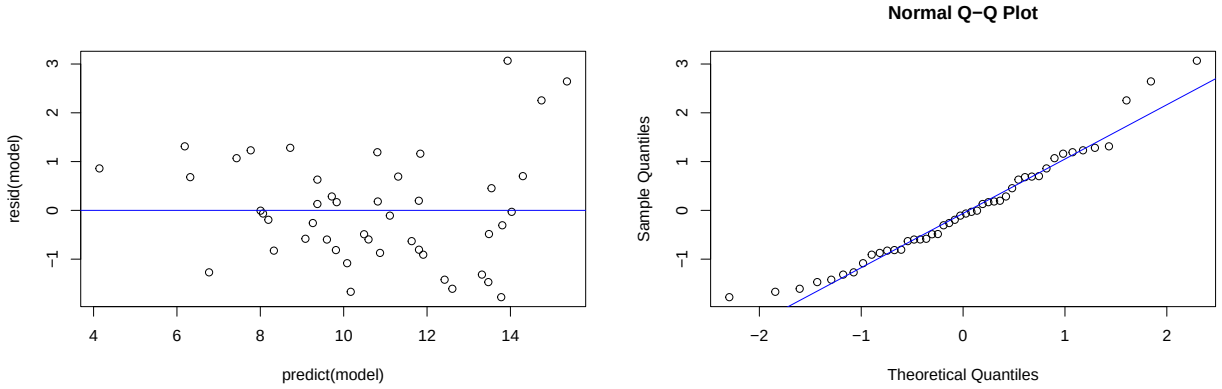


Problem 4.12

a) We can use backward selection, but removing the predictor with the highest p-value (Ascent) reduces the R^2 value. We can keep all four predictors.

```
##
## Call:
## lm(formula = Time ~ Difficulty + Ascent + Elevation + Length,
##     data = HP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77942 -0.81216 -0.08647  0.68962  3.06736
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9567864  2.2307630   2.670  0.01082 *
## Difficulty   0.8654527  0.2285275   3.787  0.00049 ***
## Ascent       0.0006011  0.0003310   1.816  0.07669 .
## Elevation   -0.0016703  0.0005183  -3.223  0.00249 **
## Length       0.4440084  0.0812523   5.465 2.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.171 on 41 degrees of freedom
## Multiple R-squared:  0.8401, Adjusted R-squared:  0.8245
## F-statistic: 53.84 on 4 and 41 DF,  p-value: 8.738e-16
```

b) The residuals look to be randomly scattered and have constant variance, and the normal quantile plot is roughly linear, suggesting that the residuals are normally distributed. There are 3 observations that have unusually high residuals.



c) There are three mountains that have residuals greater than 2, and none that are less than -2. These mountains are listed below.

```
##       24       40       33       12       39        2
## 2.964556 2.562853 2.103006 1.214045 1.166073 1.149155
```

```
##        46        20        44         9         8         3
## -1.194827 -1.347085 -1.438332 -1.445827 -1.522155 -1.650153
```

```
## [1] "Seward Mtn."   "Mt. Emmons"    "Mt. Donaldson"
```

d)

```
##         45          1         44         36         31         20         43
## 0.27592666 0.22312686 0.21784754 0.21773847 0.18153489 0.17180037 0.16953496
##          2         10         12          4          5         38         26
## 0.15684071 0.14783751 0.13746680 0.12972137 0.12479187 0.12375495 0.12361326
##          7         40          3         39         11         46         16
## 0.12253277 0.11873558 0.11613924 0.11018597 0.10954176 0.10526535 0.10192553
##         41          8         13         33         21          6         35
## 0.09962786 0.09462390 0.09280489 0.09191579 0.09036746 0.08883646 0.08859683
##         18         34         15         19         37          9         32
## 0.08768826 0.08689236 0.08448517 0.08251557 0.07822965 0.07488939 0.07367008
##         24         42         29         25         30         22         28
## 0.07072971 0.06695087 0.06630653 0.06032788 0.05940688 0.05885257 0.05747866
##         27         17         14         23
## 0.05243718 0.03895260 0.03770663 0.02984441
```

```
## [1] "hi2: 0.217391304347826 hi3: 0.326086956521739"
```

There are four mountains that have a moderately high leverage. They are

```
## [1] "Nye Mtn."   "Mt. Marcy"  "Cliff Mtn." "Sawteeth"
```

The highest Cook's D is only 0.1558, which is less than 0.5. There are no mountains that are an influential case.

```
##         40         24         44         33         20          3
## 0.15582827 0.11242804 0.11231391 0.08263319 0.07381836 0.06867450
```