
title: “Assign3” author: “Chen Ming” date: “May 20th” output: pdf_document: default html_document: default —

Part 1

1.1. Find descriptive statistics of the data and summarize them into a table

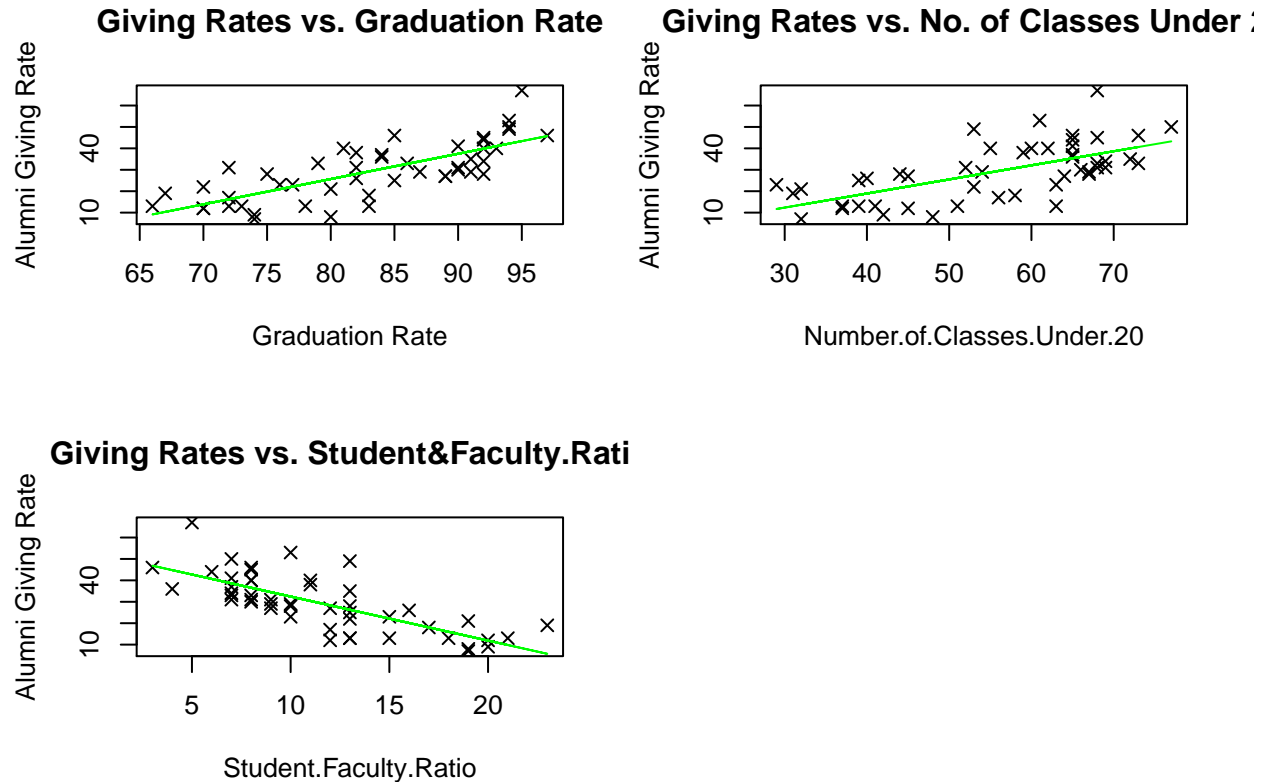
The table is constructed as follows.

| ## | Graduation.Rate | Number.of.Classes.Under.20 | Student.Faculty.Ratio |
|-------|-----------------|----------------------------|-----------------------|
| ## 1 | 85 | 39 | 13 |
| ## 2 | 79 | 68 | 8 |
| ## 3 | 93 | 60 | 8 |
| ## 4 | 85 | 65 | 3 |
| ## 5 | 75 | 67 | 10 |
| ## 6 | 72 | 52 | 8 |
| ## 7 | 89 | 45 | 12 |
| ## 8 | 90 | 69 | 7 |
| ## 9 | 91 | 72 | 13 |
| ## 10 | 94 | 61 | 10 |
| ## 11 | 92 | 68 | 8 |
| ## 12 | 84 | 65 | 7 |
| ## 13 | 91 | 54 | 10 |
| ## 14 | 97 | 73 | 8 |
| ## 15 | 89 | 64 | 9 |
| ## 16 | 81 | 55 | 11 |
| ## 17 | 92 | 65 | 6 |
| ## 18 | 72 | 63 | 13 |
| ## 19 | 90 | 66 | 8 |
| ## 20 | 80 | 32 | 19 |
| ## 21 | 95 | 68 | 5 |
| ## 22 | 92 | 62 | 8 |
| ## 23 | 92 | 69 | 7 |
| ## 24 | 87 | 67 | 9 |
| ## 25 | 72 | 56 | 12 |
| ## 26 | 83 | 58 | 17 |
| ## 27 | 74 | 32 | 19 |
| ## 28 | 74 | 42 | 20 |
| ## 29 | 78 | 41 | 18 |
| ## 30 | 80 | 48 | 19 |
| ## 31 | 70 | 45 | 20 |
| ## 32 | 84 | 65 | 4 |
| ## 33 | 67 | 31 | 23 |
| ## 34 | 77 | 29 | 15 |
| ## 35 | 83 | 51 | 15 |
| ## 36 | 82 | 40 | 16 |
| ## 37 | 94 | 53 | 13 |
| ## 38 | 90 | 65 | 7 |
| ## 39 | 76 | 63 | 10 |
| ## 40 | 70 | 53 | 13 |
| ## 41 | 66 | 39 | 21 |
| ## 42 | 92 | 44 | 13 |
| ## 43 | 70 | 37 | 12 |

| | | | |
|-------|--------------------|----|----|
| ## 44 | 73 | 37 | 13 |
| ## 45 | 82 | 68 | 9 |
| ## 46 | 82 | 59 | 11 |
| ## 47 | 86 | 73 | 7 |
| ## 48 | 94 | 77 | 7 |
| ## | Alumni.Giving.Rate | | |
| ## 1 | 25 | | |
| ## 2 | 33 | | |
| ## 3 | 40 | | |
| ## 4 | 46 | | |
| ## 5 | 28 | | |
| ## 6 | 31 | | |
| ## 7 | 27 | | |
| ## 8 | 31 | | |
| ## 9 | 35 | | |
| ## 10 | 53 | | |
| ## 11 | 45 | | |
| ## 12 | 37 | | |
| ## 13 | 29 | | |
| ## 14 | 46 | | |
| ## 15 | 27 | | |
| ## 16 | 40 | | |
| ## 17 | 44 | | |
| ## 18 | 13 | | |
| ## 19 | 30 | | |
| ## 20 | 21 | | |
| ## 21 | 67 | | |
| ## 22 | 40 | | |
| ## 23 | 34 | | |
| ## 24 | 29 | | |
| ## 25 | 17 | | |
| ## 26 | 18 | | |
| ## 27 | 7 | | |
| ## 28 | 9 | | |
| ## 29 | 13 | | |
| ## 30 | 8 | | |
| ## 31 | 12 | | |
| ## 32 | 36 | | |
| ## 33 | 19 | | |
| ## 34 | 23 | | |
| ## 35 | 13 | | |
| ## 36 | 26 | | |
| ## 37 | 49 | | |
| ## 38 | 41 | | |
| ## 39 | 23 | | |
| ## 40 | 22 | | |
| ## 41 | 13 | | |
| ## 42 | 28 | | |
| ## 43 | 12 | | |
| ## 44 | 13 | | |
| ## 45 | 31 | | |
| ## 46 | 38 | | |
| ## 47 | 33 | | |
| ## 48 | 50 | | |

1.2. Use graphical analysis to investigate the relationship between Alumni Giving Rate and each of the other variables

Scatter plot is demonstrated as follows. In the first and the third graphs, we can find that points cluster closely around fitted lines. While in the second graph, points seem to drift away from fitted line. From the graphical analysis, we can reasonably assume that alumni giving rate is more closely related to both graduation rate and faculty rate than number of classes under 20.



1.3. Develop a multiple linear regression model that could be used to predict the Alumni Giving Rate using the data provided

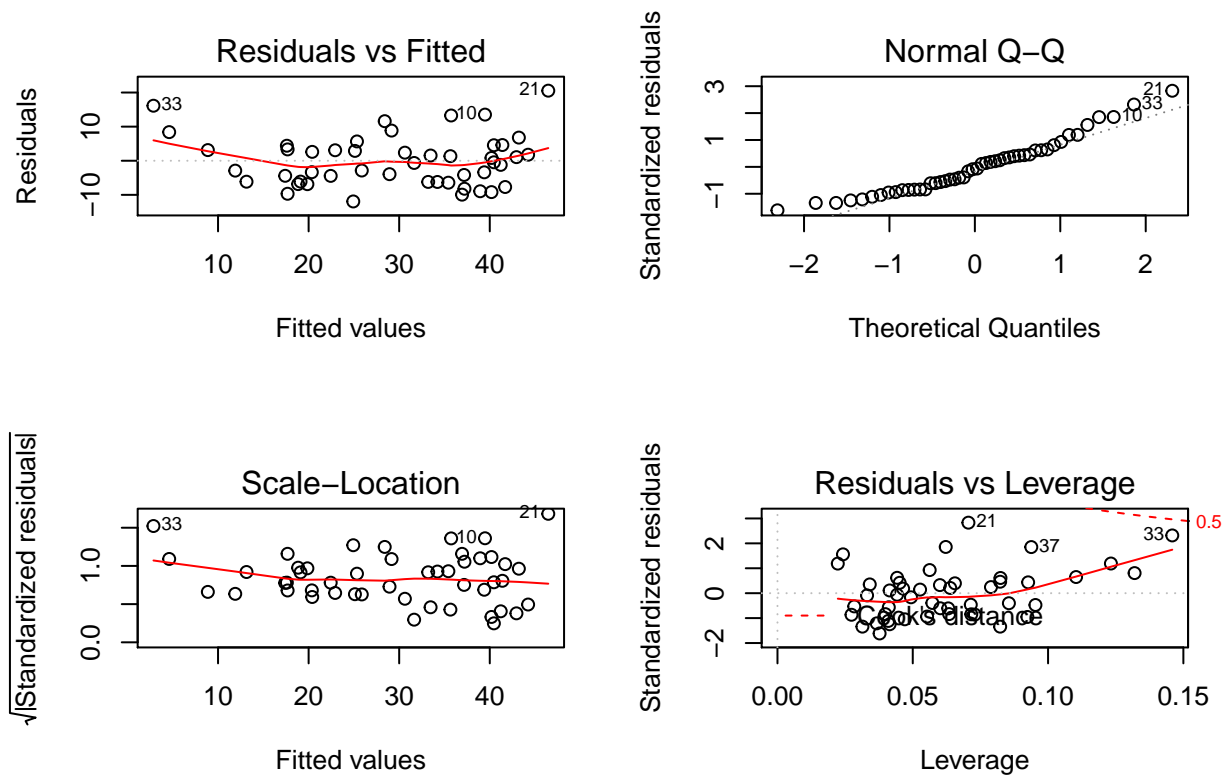
We can use function **stepAIC** to construct the best multi-linear regression model from a set of candidate variables. The experiment result indicates that, AIC value becomes smaller when the variable “Number of classes under 20” is deleted from the regression model. Since smaller AIC value means better fitting effect, we should construct a regression model with “Graduation Rates” and “Student & Faculty Ratios” as independent variables.

```
## Start: AIC=198.65
## Alumni.Giving.Rate ~ Graduation.Rate + Number.of.Classes.Under.20 +
## Student.Faculty.Ratio
##
##               Df Sum of Sq  RSS   AIC
## - Number.of.Classes.Under.20  1      2.52 2550.5 196.70
## <none>                                2547.9 198.65
## - Student.Faculty.Ratio         1     550.17 3098.1 206.03
## - Graduation.Rate               1     1176.92 3724.9 214.88
##
## Step: AIC=196.7
```

```
## Alumni.Giving.Rate ~ Graduation.Rate + Student.Faculty.Ratio
##
##              Df Sum of Sq    RSS    AIC
## <none>                2550.5 196.70
## - Student.Faculty.Ratio  1    1088.5 3639.0 211.76
## - Graduation.Rate        1    1260.9 3811.4 213.98
##
## Call:
## lm(formula = Alumni.Giving.Rate ~ Graduation.Rate + Student.Faculty.Ratio,
##     data = newdataframe)
##
## Coefficients:
##             (Intercept)      Graduation.Rate  Student.Faculty.Ratio
##                -19.1063                0.7557                 -1.2460
##
## Call:
## lm(formula = Alumni.Giving.Rate ~ Graduation.Rate + Student.Faculty.Ratio,
##     data = newdataframe)
##
## Coefficients:
##             (Intercept)      Graduation.Rate  Student.Faculty.Ratio
##                -19.1063                0.7557                 -1.2460
##
## Call:
## lm(formula = Alumni.Giving.Rate ~ Graduation.Rate + Student.Faculty.Ratio,
##     data = newdataframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9304  -6.1594  -0.5521   3.5910  20.5412
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -19.1063    15.5501  -1.229   0.226
## Graduation.Rate    0.7557     0.1602   4.717 2.35e-05 ***
## Student.Faculty.Ratio -1.2460     0.2843  -4.382 6.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.528 on 45 degrees of freedom
## Multiple R-squared:  0.6996, Adjusted R-squared:  0.6863
## F-statistic: 52.41 on 2 and 45 DF,  p-value: 1.765e-12
```

1.4. Check the model assumptions

In Q-Q graph, points closely cluster around the line, which proves that the assumption of **normality** is satisfied; there is no reason to assume that graduation ratio and faculty to student ratio is related. Therefore, the assumption of **independence** is satisfied; from graph one, we can observe that residuals is no systematic relationship between residuals and the predicted values. The model well captures systematic variance in the data, thereby proves that the assumption of **Linearity** is satisfied; the Scale-Location graph shows that the points form a random band around the horizontal line. Hence the assumption of **Homoscedasticity** is satisfied.



Part 2

2.1. Calculate the mean of Fertility and partition the provinces into two groups

The mean of fertility in stated provinces is 70.14255. Group1 and group2 is illustrated as follows.

```
## [1] 70.14255
```

| | Fertility | Agriculture | Examination | Education | Catholic |
|-----------------|-----------|-------------|-------------|-----------|----------|
| ## Courtelary | 80.2 | 17.0 | 15 | 12 | 9.96 |
| ## Delemont | 83.1 | 45.1 | 6 | 9 | 84.84 |
| ## Franches-Mnt | 92.5 | 39.7 | 5 | 5 | 93.40 |
| ## Moutier | 85.8 | 36.5 | 12 | 7 | 33.77 |
| ## Neuveville | 76.9 | 43.5 | 17 | 15 | 5.16 |
| ## Porrentruy | 76.1 | 35.3 | 9 | 7 | 90.57 |
| ## Broye | 83.8 | 70.2 | 16 | 7 | 92.85 |
| ## Glane | 92.4 | 67.8 | 14 | 8 | 97.16 |
| ## Gruyere | 82.4 | 53.3 | 12 | 7 | 97.67 |
| ## Sarine | 82.9 | 45.2 | 16 | 13 | 91.38 |
| ## Veveyse | 87.1 | 64.5 | 14 | 6 | 98.61 |
| ## Grandson | 71.7 | 34.0 | 17 | 8 | 3.30 |
| ## Oron | 72.5 | 71.2 | 12 | 1 | 2.40 |
| ## Payerne | 74.2 | 58.1 | 14 | 8 | 5.23 |
| ## Paysd'enhaut | 72.0 | 63.5 | 6 | 3 | 2.56 |
| ## Conthey | 75.5 | 85.9 | 3 | 2 | 99.71 |
| ## Herens | 77.3 | 89.7 | 5 | 2 | 100.00 |
| ## Martigwy | 70.5 | 78.2 | 12 | 6 | 98.96 |

| | | | | | |
|-----------------|--|------|----|----|-------|
| ## Monthey | 79.4 | 64.9 | 7 | 3 | 98.22 |
| ## Sierre | 92.2 | 84.6 | 3 | 3 | 99.46 |
| ## Sion | 79.3 | 63.1 | 13 | 13 | 96.83 |
| ## Boudry | 70.4 | 38.4 | 26 | 12 | 5.62 |
| ## Le Locle | 72.7 | 16.7 | 22 | 13 | 11.22 |
| ## Val de Ruz | 77.6 | 37.6 | 15 | 7 | 4.97 |
| ## | Infant.Mortality ynprovinces | | | | |
| ## Courtelary | 22.2 | | 1 | | |
| ## Delemont | 22.2 | | 1 | | |
| ## Franches-Mnt | 20.2 | | 1 | | |
| ## Moutier | 20.3 | | 1 | | |
| ## Neuveville | 20.6 | | 1 | | |
| ## Porrentruy | 26.6 | | 1 | | |
| ## Broye | 23.6 | | 1 | | |
| ## Glane | 24.9 | | 1 | | |
| ## Gruyere | 21.0 | | 1 | | |
| ## Sarine | 24.4 | | 1 | | |
| ## Veveyse | 24.5 | | 1 | | |
| ## Grandson | 20.0 | | 1 | | |
| ## Oron | 21.0 | | 1 | | |
| ## Payerne | 23.8 | | 1 | | |
| ## Paysd'enhaut | 18.0 | | 1 | | |
| ## Conthey | 15.1 | | 1 | | |
| ## Herens | 18.3 | | 1 | | |
| ## Martigwy | 19.4 | | 1 | | |
| ## Monthey | 20.2 | | 1 | | |
| ## Sierre | 16.3 | | 1 | | |
| ## Sion | 18.1 | | 1 | | |
| ## Boudry | 20.3 | | 1 | | |
| ## Le Locle | 18.9 | | 1 | | |
| ## Val de Ruz | 20.0 | | 1 | | |
| ## | Fertility Agriculture Examination Education Catholic | | | | |
| ## Aigle | 64.1 | 62.0 | 21 | 12 | 8.52 |
| ## Aubonne | 66.9 | 67.5 | 14 | 7 | 2.27 |
| ## Avenches | 68.9 | 60.7 | 19 | 12 | 4.43 |
| ## Cossonay | 61.7 | 69.3 | 22 | 5 | 2.82 |
| ## Echallens | 68.3 | 72.6 | 18 | 2 | 24.20 |
| ## Lausanne | 55.7 | 19.4 | 26 | 28 | 12.11 |
| ## La Vallee | 54.3 | 15.2 | 31 | 20 | 2.15 |
| ## Lavaux | 65.1 | 73.0 | 19 | 9 | 2.84 |
| ## Morges | 65.5 | 59.8 | 22 | 10 | 5.23 |
| ## Moudon | 65.0 | 55.1 | 14 | 3 | 4.52 |
| ## Nyone | 56.6 | 50.9 | 22 | 12 | 15.14 |
| ## Orbe | 57.4 | 54.1 | 20 | 6 | 4.20 |
| ## Rolle | 60.5 | 60.8 | 16 | 10 | 7.72 |
| ## Vevey | 58.3 | 26.8 | 25 | 19 | 18.46 |
| ## Yverdon | 65.4 | 49.5 | 15 | 8 | 6.10 |
| ## Entremont | 69.3 | 84.9 | 7 | 6 | 99.68 |
| ## St Maurice | 65.0 | 75.9 | 9 | 9 | 99.06 |
| ## La Chauxdfnd | 65.7 | 7.7 | 29 | 11 | 13.79 |
| ## Neuchatel | 64.4 | 17.6 | 35 | 32 | 16.92 |
| ## ValdeTravers | 67.6 | 18.7 | 25 | 7 | 8.65 |
| ## V. De Geneve | 35.0 | 1.2 | 37 | 53 | 42.34 |

| | | | | | |
|-----------------|------------------|-------------|----|----|-------|
| ## Rive Droite | 44.7 | 46.6 | 16 | 29 | 50.43 |
| ## Rive Gauche | 42.8 | 27.7 | 22 | 29 | 58.33 |
| ## | Infant.Mortality | ynprovinces | | | |
| ## Aigle | 16.5 | 0 | | | |
| ## Aubonne | 19.1 | 0 | | | |
| ## Avenches | 22.7 | 0 | | | |
| ## Cossonay | 18.7 | 0 | | | |
| ## Echallens | 21.2 | 0 | | | |
| ## Lausanne | 20.2 | 0 | | | |
| ## La Vallee | 10.8 | 0 | | | |
| ## Lavaux | 20.0 | 0 | | | |
| ## Morges | 18.0 | 0 | | | |
| ## Moudon | 22.4 | 0 | | | |
| ## Nyone | 16.7 | 0 | | | |
| ## Orbe | 15.3 | 0 | | | |
| ## Rolle | 16.3 | 0 | | | |
| ## Vevey | 20.9 | 0 | | | |
| ## Yverdon | 22.5 | 0 | | | |
| ## Entremont | 19.8 | 0 | | | |
| ## St Maurice | 17.8 | 0 | | | |
| ## La Chauxdfnd | 20.5 | 0 | | | |
| ## Neuchatel | 23.0 | 0 | | | |
| ## ValdeTravers | 19.5 | 0 | | | |
| ## V. De Geneve | 18.0 | 0 | | | |
| ## Rive Droite | 18.2 | 0 | | | |
| ## Rive Gauche | 19.3 | 0 | | | |
| ## | [,1] | | | | |
| ## [1,] | 1 | | | | |
| ## [2,] | 1 | | | | |
| ## [3,] | 1 | | | | |
| ## [4,] | 1 | | | | |
| ## [5,] | 1 | | | | |
| ## [6,] | 1 | | | | |
| ## [7,] | 1 | | | | |
| ## [8,] | 1 | | | | |
| ## [9,] | 1 | | | | |
| ## [10,] | 1 | | | | |
| ## [11,] | 1 | | | | |
| ## [12,] | 0 | | | | |
| ## [13,] | 0 | | | | |
| ## [14,] | 0 | | | | |
| ## [15,] | 0 | | | | |
| ## [16,] | 0 | | | | |
| ## [17,] | 1 | | | | |
| ## [18,] | 0 | | | | |
| ## [19,] | 0 | | | | |
| ## [20,] | 0 | | | | |
| ## [21,] | 0 | | | | |
| ## [22,] | 0 | | | | |
| ## [23,] | 0 | | | | |
| ## [24,] | 0 | | | | |
| ## [25,] | 1 | | | | |
| ## [26,] | 1 | | | | |

```
## [27,] 1
## [28,] 0
## [29,] 0
## [30,] 0
## [31,] 1
## [32,] 0
## [33,] 1
## [34,] 1
## [35,] 1
## [36,] 0
## [37,] 1
## [38,] 1
## [39,] 1
## [40,] 0
## [41,] 1
## [42,] 0
## [43,] 1
## [44,] 0
## [45,] 0
## [46,] 0
## [47,] 0
```

2.2. Use logistic regression to show the relationship between y and the other variables and then interpret the regression results

The experiment result indicates that fertility in selected provinces is significantly related to **Agriculture** and **Examination** under significance level of 0.05. It implies that fertility is closely related to agriculture situation and examination circumstance in the provinces.

```
##
## Call:
## glm(formula = y ~ Agriculture + Examination + Education + Catholic +
##      Infant.Mortality, family = binomial(), data = swiss)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85403  -0.45960   0.03648   0.55548   2.32911
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.82826    5.25607   0.919  0.3583
## Agriculture    -0.09615    0.04011  -2.397  0.0165 *
## Examination    -0.32116    0.13844  -2.320  0.0203 *
## Education      -0.12078    0.08610  -1.403  0.1607
## Catholic        0.02078    0.01376   1.509  0.1312
## Infant.Mortality 0.29078    0.21051   1.381  0.1672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 65.135  on 46  degrees of freedom
## Residual deviance: 32.887  on 41  degrees of freedom
## AIC: 44.887
##
```



```
## Number of Fisher Scoring iterations: 6
```

2.3. Choose a model selection criterion, for instances, AIC, BIC, adjusted R square or Cp, and use it to select a reasonable model

We can construct the best multi-linear regression model from a set of candidate variables. Under AIC criterion, regression model with smaller AIC value is considered better. The experiment result indicates that AIC value gets smaller when the variables “Catholic” and “Education” are deleted from the regression model. Hence we should construct a regression model with “**Infant.Mortality**”, “**Agriculture**” and “**Examination**” as independent variables.

```
## Start:  AIC=-84.75
## y ~ Agriculture + Examination + Education + Catholic + Infant.Mortality
##
##           Df Sum of Sq    RSS    AIC
## - Education      1   0.25537 6.2545 -84.792
## - Infant.Mortality 1   0.25579 6.2549 -84.789
## <none>                        5.9991 -84.751
## - Catholic       1   0.40468 6.4038 -83.683
## - Examination    1   0.93170 6.9308 -79.966
## - Agriculture    1   1.06427 7.0634 -79.075
##
## Step:  AIC=-84.79
## y ~ Agriculture + Examination + Catholic + Infant.Mortality
##
##           Df Sum of Sq    RSS    AIC
## - Catholic      1   0.20534 6.4598 -85.274
## <none>                        6.2545 -84.792
## - Infant.Mortality 1   0.36465 6.6191 -84.129
## - Agriculture    1   0.82599 7.0805 -80.962
## - Examination    1   2.43417 8.6887 -71.342
##
## Step:  AIC=-85.27
## y ~ Agriculture + Examination + Infant.Mortality
##
##           Df Sum of Sq    RSS    AIC
## <none>                        6.4598 -85.274
## - Infant.Mortality 1   0.4546  6.9144 -84.077
## - Agriculture    1   0.7943  7.2542 -81.823
## - Examination    1   3.7161 10.1760 -65.916
##
## Call:
## lm(formula = y ~ Agriculture + Examination + Infant.Mortality,
##     data = swiss)
##
## Coefficients:
##      (Intercept)      Agriculture      Examination  Infant.Mortality
##           1.05056          -0.00811           -0.05017            0.03501
##
## Call:
## lm(formula = y ~ Agriculture + Examination + Infant.Mortality,
##     data = swiss)
##
## Coefficients:
```

| | | | | |
|----|-------------|-------------|-------------|------------------|
| ## | (Intercept) | Agriculture | Examination | Infant.Mortality |
| ## | 1.05056 | -0.00811 | -0.05017 | 0.03501 |