# M13 Activity

Courtney Hodge & Hannah Valenty

2024-08-01

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
df <- read.csv('data/expectancy.csv')
df2 <- select(df, Life.expectancy, Status, Adult.Mortality,
        infant.deaths,HIV.AIDS,BMI, GDP,Schooling)%>%
        na.omit()
```

# Task 1

```
mod0 <- lm(Life.expectancy~., data = df2)
summary(mod0)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9677 -2.0529  0.3311  2.0589 10.3389
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.820e+01  2.350e+00  24.771  < 2e-16 ***
## StatusDeveloping -1.319e+00  8.666e-01  -1.522 0.130260
## Adult.Mortality  -2.858e-02  3.934e-03  -7.265 2.21e-11 ***
## infant.deaths    -1.835e-03  3.101e-03  -0.592 0.554963
## HIV.AIDS         -9.398e-01  2.490e-01  -3.774 0.000234 ***
## BMI               2.264e-03  1.621e-02   0.140 0.889108
## GDP               3.243e-05  2.626e-05   1.235 0.218912
## Schooling         1.473e+00  1.499e-01   9.826  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.293 on 143 degrees of freedom
## Multiple R-squared:  0.844,  Adjusted R-squared:  0.8363
## F-statistic: 110.5 on 7 and 143 DF,  p-value: < 2.2e-16
```

```
aic <- MASS::stepAIC(mod0, direction='both', Trace=FALSE)
```

```
## Start:  AIC=367.69
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     HIV.AIDS + BMI + GDP + Schooling
##
##                   Df Sum of Sq    RSS    AIC
## - BMI              1      0.21 1550.8 365.71
## - infant.deaths    1      3.80 1554.3 366.06
## - GDP              1     16.53 1567.1 367.29
## <none>                         1550.5 367.69
## - Status           1     25.11 1575.7 368.12
## - HIV.AIDS         1    154.47 1705.0 380.03
## - Adult.Mortality  1    572.31 2122.9 413.13
## - Schooling        1   1046.90 2597.5 443.60
##
## Step:  AIC=365.71
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     HIV.AIDS + GDP + Schooling
##
##                   Df Sum of Sq    RSS    AIC
## - infant.deaths    1      4.00 1554.8 364.10
## - GDP              1     17.17 1567.9 365.38
## <none>                         1550.8 365.71
## - Status           1     24.94 1575.7 366.12
## + BMI              1      0.21 1550.5 367.69
## - HIV.AIDS         1    154.42 1705.2 378.05
## - Adult.Mortality  1    577.89 2128.7 411.54
## - Schooling        1   1324.88 2875.6 456.96
##
## Step:  AIC=364.1
## Life.expectancy ~ Status + Adult.Mortality + HIV.AIDS + GDP +
##     Schooling
##
##                   Df Sum of Sq    RSS    AIC
## - GDP              1     17.32 1572.1 363.77
## <none>                         1554.8 364.10
## - Status           1     24.52 1579.3 364.46
## + infant.deaths    1      4.00 1550.8 365.71
## + BMI              1      0.42 1554.3 366.06
## - HIV.AIDS         1    152.02 1706.8 376.19
## - Adult.Mortality  1    591.69 2146.5 410.80
## - Schooling        1   1378.83 2933.6 457.97
##
## Step:  AIC=363.77
## Life.expectancy ~ Status + Adult.Mortality + HIV.AIDS + Schooling
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         1572.1 363.77
## + GDP              1     17.32 1554.8 364.10
## - Status           1     31.21 1603.3 364.74
## + infant.deaths    1      4.15 1567.9 365.38
## + BMI              1      1.23 1570.8 365.66
## - HIV.AIDS         1    146.69 1718.8 375.24
## - Adult.Mortality  1    630.36 2202.4 412.69
## - Schooling        1   1553.18 3125.3 465.53
```

```
summary(aic)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + HIV.AIDS +
##     Schooling, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9300 -2.0243  0.3127  2.1598 10.3146
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57.871590   2.304972  25.107  < 2e-16 ***
## StatusDeveloping -1.443373   0.847760  -1.703 0.090776 .
## Adult.Mortality  -0.029506   0.003856  -7.651 2.48e-12 ***
## HIV.AIDS         -0.912691   0.247281  -3.691 0.000315 ***
## Schooling         1.536868   0.127964  12.010  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.281 on 146 degrees of freedom
## Multiple R-squared:  0.8418, Adjusted R-squared:  0.8375
## F-statistic: 194.2 on 4 and 146 DF,  p-value: < 2.2e-16
```

Make a new model only using Adult Mortality, HIV/AIDS, and Schooling. Check mulitcolliniarity across this new model.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.1
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.1
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
mod1 <- lm(Life.expectancy~Adult.Mortality+HIV.AIDS+Schooling, data = df2)
summary(mod1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + HIV.AIDS + Schooling,
##     data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9228 -1.8432  0.1605  2.0037 10.2611
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     55.310506   1.757712  31.467  < 2e-16 ***
## Adult.Mortality -0.030206   0.003859  -7.827 9.03e-13 ***
## HIV.AIDS        -0.870038   0.247593  -3.514 0.000587 ***
## Schooling        1.647743   0.110863  14.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.303 on 147 degrees of freedom
## Multiple R-squared:  0.8387, Adjusted R-squared:  0.8354
## F-statistic: 254.7 on 3 and 147 DF,  p-value: < 2.2e-16
```

```
vif(mod1)
```

```
## Adult.Mortality        HIV.AIDS       Schooling
##        1.934791        1.722737        1.368752
```

Now cross validate using by 10 fold.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.1
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
control <- trainControl(method = 'cv', number = 10)

mod2 <- train(Life.expectancy~Adult.Mortality+HIV.AIDS+Schooling, method='lm', trControl=control, data=df2)
summary(mod2)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9228 -1.8432  0.1605  2.0037 10.2611
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     55.310506   1.757712  31.467  < 2e-16 ***
## Adult.Mortality -0.030206   0.003859  -7.827 9.03e-13 ***
## HIV.AIDS        -0.870038   0.247593  -3.514 0.000587 ***
## Schooling        1.647743   0.110863  14.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.303 on 147 degrees of freedom
## Multiple R-squared:  0.8387, Adjusted R-squared:  0.8354
## F-statistic: 254.7 on 3 and 147 DF,  p-value: < 2.2e-16
```

```
mod2$results$RMSE
```

```
## [1] 3.366739
```

The RMSE of the model for 10 folds is 3.386653.

# Task 2

```
# select row to predict on
prediction <- df2[1,]
pred_with_mod <- predict(mod2, prediction)

# find residual with difference of actual - predicted
residual <- prediction$Life.expectancy - pred_with_mod
residual
```

```
##        1
## 1.078414
```

The residual of our prediction using the first row of the original data is 1.078414.

# Task 3.A

```
library(broom)
# same parameters but with lm() function model

diag <- mod1 %>%
  augment(data=df2)
head(diag)
```

```
## # A tibble: 6 × 15
##   .rownames Life.expectancy Status  Adult.Mortality infant.deaths HIV.AIDS   BMI
##   <chr>               <dbl> <chr>             <int>         <int>    <dbl> <dbl>
## 1 1                    65   Develo…             263            62      0.1  19.1
## 2 2                    77.8 Develo…              74             0      0.1  58
## 3 3                    75.6 Develo…              19            21      0.1  59.5
## 4 4                    52.4 Develo…             335            66      1.9  23.3
## 5 5                    76.4 Develo…              13             0      0.2  47.7
## 6 6                    76.3 Develo…             116             8      0.1  62.8
## # ℹ 8 more variables: GDP <dbl>, Schooling <dbl>, .fitted <dbl>, .resid <dbl>,
## #   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

```
# -- influential
infl <- filter(diag, .cooksd > 4/nrow(df2))
infl
```

```
## # A tibble: 12 × 15
##    .rownames Life.expectancy Status Adult.Mortality infant.deaths HIV.AIDS   BMI
##    <chr>               <dbl> <chr>             <int>         <int>    <dbl> <dbl>
##  1 4                    52.4 Devel…             335            66      1.9  23.3
##  2 26                   59.9 Devel…              26            38      0.6  19.4
##  3 48                   63.5 Devel…             241             1      2.1  35
##  4 53                   58.2 Devel…              32             3      4.2  24.5
##  5 71                   63.5 Devel…              24            14      0.5  49.9
##  6 93                   53.7 Devel…             484             4      9.3  32.6
##  7 98                   65.5 Devel…              22            28      0.3   2.5
##  8 104                  63.1 Devel…              25             8      0.9   3.8
##  9 139                  67.5 Devel…              19             0      0.2   3.9
## 10 144                  51   Devel…             413            22      0.5  24.4
## 11 182                  61.8 Devel…              33            27      4.1  23.4
## 12 183                  67   Devel…             336            22      6.2  31.8
## # ℹ 8 more variables: GDP <dbl>, Schooling <dbl>, .fitted <dbl>, .resid <dbl>,
## #   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

The influential points are listed above within the ‘infl’ variable, there were 57 points identified.

Now remove these points and rebuild the model.

```
# new data
no_infl <- anti_join(df2, infl)
```

```
## Joining with `by = join_by(Life.expectancy, Status, Adult.Mortality,
## infant.deaths, HIV.AIDS, BMI, GDP, Schooling)`
```

```
control <- trainControl(method = 'cv', number = 10)

mod3 <- train(Life.expectancy~Adult.Mortality+HIV.AIDS+Schooling, method='lm', trControl=control, data=no_infl)
summary(mod3)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7668 -1.3013  0.1159  1.5778  6.6317
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     62.569044   1.708814  36.615   <2e-16 ***
## Adult.Mortality -0.045862   0.004122 -11.126   <2e-16 ***
## HIV.AIDS        -0.505904   0.279670  -1.809   0.0727 .
## Schooling        1.285759   0.100069  12.849   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.502 on 135 degrees of freedom
## Multiple R-squared:  0.8932, Adjusted R-squared:  0.8908
## F-statistic: 376.2 on 3 and 135 DF,  p-value: < 2.2e-16
```

# Task 3.B

```
mod3$results$RMSE
```

```
## [1] 2.525145
```

The RMSE of the model based on a 10-fold cross-validation is valued at 1.474635.

# Task 3.C

```
# same prediction value but with new model
prediction <- df2[1,]
pred_with_mod3 <- predict(mod3, prediction)

# find residual with difference of actual - predicted
residual_mod3 <- prediction$Life.expectancy - pred_with_mod3
residual_mod3
```

```
##        1
## 1.557049
```

The residual for the new model using the same point as in question 2 is valued at 1.530593. This compares to the original residual value which is 1.078414. The difference between the original model with influence points and the new model without the points is 0.452179.