Assignment 2 Group 26

Vanshita Sharma Kumar Lucian Truşcă Xander Poortvliet

2025-10-10

Excercise 2.1

2.1a

We first will start with the full multi-regression model

```
sat = read.delim("sat.txt", header=TRUE, sep = "", stringsAsFactors = FALSE)
model_full = lm(total ~ expend + ratio + salary + takers, data=sat)
```

```
## The AIC score for the full model = 497.3694
```

Step Up method: With the forward selection we will first start with no predictors and add variables one by one based on the lowest AIC

```
model_StepUp = lm(total ~ expend + takers, data=sat)
```

```
## The AIC score for the step-up method  = 494.7994
```

Step-down Method: we start from the full model and iteratively remove variables that worsen AIC the least.

```
model_StepDown <- lm(total ~ expend + takers, data=sat)
```

```
## The AIC score for the step-down method  = 494.7994
```

Model interpretation: SAT performance is best explained by school spending and participation rate. Other variables (ratio, salary) don't significantly improve model fit.

2.1b

```
sat$takers2 = sat$takers^2
## 2) Stepwise model selection (AIC)
# forward (start from intercept)
m0    = lm(total ~ 1, data = sat)
scope = ~ expend + ratio + salary + takers + takers2
m_fwd = step(m0, scope = list(lower = ~1, upper = scope),
             direction = "forward", trace = 0)
```

Where the result for the AIC is 473.9 (rounded up from 473.85).

```
## [1] 473.8576
```

```
## The AICS without takers2 is:  494.7994
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + takers
## Model 2: total ~ takers + takers2 + expend
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     47 49520
## 2     46 31298  1     18222 26.783 4.872e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In a nested-model ANOVA comparing M_1 : total $\sim$ expend+takers to M_2 : total $\sim$ expend+takers+takers^2, adding the quadratic term reduces the residual sum of squares from 49,520 to 31,298, a drop of 18,222 with one additional parameter (df $= 1$), yielding $F(1, 46) = 26.783$ and $p = 4.872 \times 10^{-6}$. This highly significant improvement leads us to reject $H_0 : \beta_{\text{takers}^2} = 0$ and conclude that **takers**^{2} is a useful predictor: it captures curvature in the relationship between SAT scores and participation that the linear-only specification misses.

2.1c

Comparing the reduced model M_1 to the expanded model M_2, the ANOVA shows a large and statistically significant drop in residual sum of squares as seen previously, where this drop implies the rejection of $H_0 : \beta_{\text{takers}^2} = 0$ and confirming that the quadratic term is informative; this statistical improvement is mirrored by information criteria, with AIC falling from ≈ 492.8 for M_1 to ≈ 471.9 for M_2, indicating that the model including takers^2 provides a substantially better fit despite its extra parameter.