Model Selection for the Seatpos Dataset

Ridhwan Choudahri

2025-09-24

Contents

1. Analysis Overview

This report details a model selection exercise for the `seatpos` dataset from the `faraway` library. The goal is to find the best subset of predictor variables to create a simple yet effective model for the response variable, `hipcenter`.

The process involves: * Fitting a full model with all predictors. * Using multiple model selection techniques (all possible subsets, forward, backward, and stepwise regression) to identify the best predictors. * Fitting and interpreting the final, optimized model.

---

2. R Code and Initial Model

First, we load the required libraries and the `seatpos` dataset. We then fit an initial linear model that includes all available predictors to serve as our baseline.

```r
library(faraway)
library(olsrr)
data("seatpos")
model <- lm(hipcenter ~ ., data = seatpos)
summary(model)
```

Initial Interpretation: The summary of the full model shows an **Adjusted R-squared** of 0.6001, meaning about 60% of the variability in `hipcenter` is explained by all predictors combined. However, several predictors have high p-values ($p > 0.05$), suggesting they are not statistically significant and could be removed to create a simpler, more effective model.

---

3. Model Selection Procedures

We will use several automated methods from the `olsrr` package to find the best-fitting, most parsimonious model.

3.1 All Possible Subsets Regression

This method evaluates every possible combination of predictors. We will use **Mallows' Cp** and the **Akaike Information Criterion (AIC)** to identify the best model. In both cases, a lower value indicates a better model.

```
fit <- ols_step_all_possible(model)
result <- fit[["result"]]

# Best model by Mallows' Cp
c_1 <- result$cp - result$n
result$predictors[which(c_1 == min(c_1))]

# Best model by AIC
a_1 <- result$aic
result$predictors[which(a_1 == min(a_1))]
```

Result Interpretation: Based on the R output, both the Mallows' Cp and AIC criteria identify the model with the predictors **`Age`**, **`Ht`**, and **`Leg`** as the best choice among all possible combinations.

3.2 Automated Stepwise Methods

Next, we use three common automated procedures to confirm the results.

Forward Selection (based on Adjusted R^2) This method starts with no predictors and adds the most significant variable at each step.

```
fit1 <- ols_step_forward_adj_r2(model)
fit1$model
```

Result: Forward selection also chooses the model `hipcenter ~ Ht + Leg + Age`.

Backward Elimination (based on AIC) This method starts with the full model and removes the least significant variable at each step.

```
fit2 <- ols_step_backward_aic(model)
fit2$model
```

Result: Backward elimination results in a model with `Age`, `HtShoes`, and `Leg`. This differs slightly from the other methods, suggesting `Ht` and `HtShoes` may be highly correlated.

Stepwise Regression (based on AIC) This hybrid method adds or removes variables at each step to find the model with the lowest AIC.

```r
fit3 <- ols_step_both_aic(model)
fit3$model
```

Result: Stepwise regression also selects the model `hipcenter ~ Age + Ht + Leg`, which is consistent with the all-subsets and forward selection methods.

---

4. Final Model and Interpretation

Given that three of the four methods consistently identified the same best model, we will proceed with the model using `Age`, `Ht`, and `Leg` to predict `hipcenter`.

```r
fit4 <- lm(hipcenter ~ Age + Ht + Leg, data = seatpos)
summary(fit4)
```

Interpretation of the Final Model

Output:

```
Call:
lm(formula = hipcenter ~ Age + Ht + Leg, data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max 
-79.715 -22.758  -4.102  21.394  60.576 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 452.1976   100.9482   4.480 8.04e-05 ***
Age           0.5807     0.3790   1.532   0.1347    
Ht           -2.3254     1.2545  -1.854   0.0725 .  
Leg          -6.7390     4.1050  -1.642   0.1099    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.12 on 34 degrees of freedom
Multiple R-squared:  0.6814,	Adjusted R-squared:  0.6533 
F-statistic: 24.24 on 3 and 34 DF,  p-value: 1.426e-08
```

1. Overall Model Significance: The **F-statistic** is 24.24 with a very small **p-value** (1.426e-08), which is highly significant. This indicates that the model as a whole is useful for predicting `hipcenter`.

2. Model Fit (Adjusted R-squared): The **Adjusted R-squared** is **0.6533**. This means that approximately **65.3%** of the variance in the `hipcenter` measurement is explained by the predictors `Age`, `Ht`, and `Leg`. This is an improvement over the full model's adjusted R-squared (0.6001), and our new model is much simpler.

3. Coefficients: In this final model, none of the individual predictors are statistically significant at the traditional $p < 0.05$ level. However, `Ht` is significant at $p < 0.1$.

- `Age` **($p = 0.1347$):** The coefficient is **0.5807**. Holding other variables constant, for each additional year of age, the `hipcenter` is predicted to increase by 0.58 mm. This effect is not statistically significant.
- `Ht` **($p = 0.0725$):** The coefficient is **-2.3254**. For each one-unit (mm) increase in height, the `hipcenter` is predicted to *decrease* by 2.33 mm, holding other variables constant. This is significant at the $p < 0.1$ level.
- `Leg` **($p = 0.1099$):** The coefficient is **-6.7390**. For each one-unit (mm) increase in leg length, the `hipcenter` is predicted to **decrease** by 6.74 mm, holding other variables constant. This effect is not statistically significant.

Even though the individual predictors are not all significant, the model as a whole is strong (as shown by the F-statistic), likely due to correlations between the predictors.