Penalized Regression

University of St. Gallen
School of Management, Economics, Law,
Social Sciences, International Affairs
and Computer Science

Assignment 2

Data Analytics I: Predictive Econometrics
Prof. Jana Mareckova

submitted by

Cyril Janak, 16-611-287
Jonas Husmann, 16-610-917
Niklas Kampe, 16-611-618
Robin Scherrer, 18-617-969

08.12.2021

Contents

Requirements

To solve the following tasks, the required libraries and the data sets are loaded first.

```
library(glmnet)
library(corrplot)
library(ggplot2)
library(dplyr)

load("GHA/student-mat-train.RData")
load("GHA/student-mat-test.RData")
```

Exercise 1

There are 214 observations in the training data set and 143 observations in the test data set.

```
(n_obs_train <- nrow(train))
```

```
## [1] 214
```

```
(n_obs_test <- nrow(test))
```

```
## [1] 143
```

Exercise 2

The average grade is ~11.64, the minimum grade is 4 and the maximum grade is 19. All numbers were calculated using the training data.

```
(avg_grade <- mean(train$G3))
```

```
## [1] 11.64019
```

```
(min_grade <- min(train$G3))
```

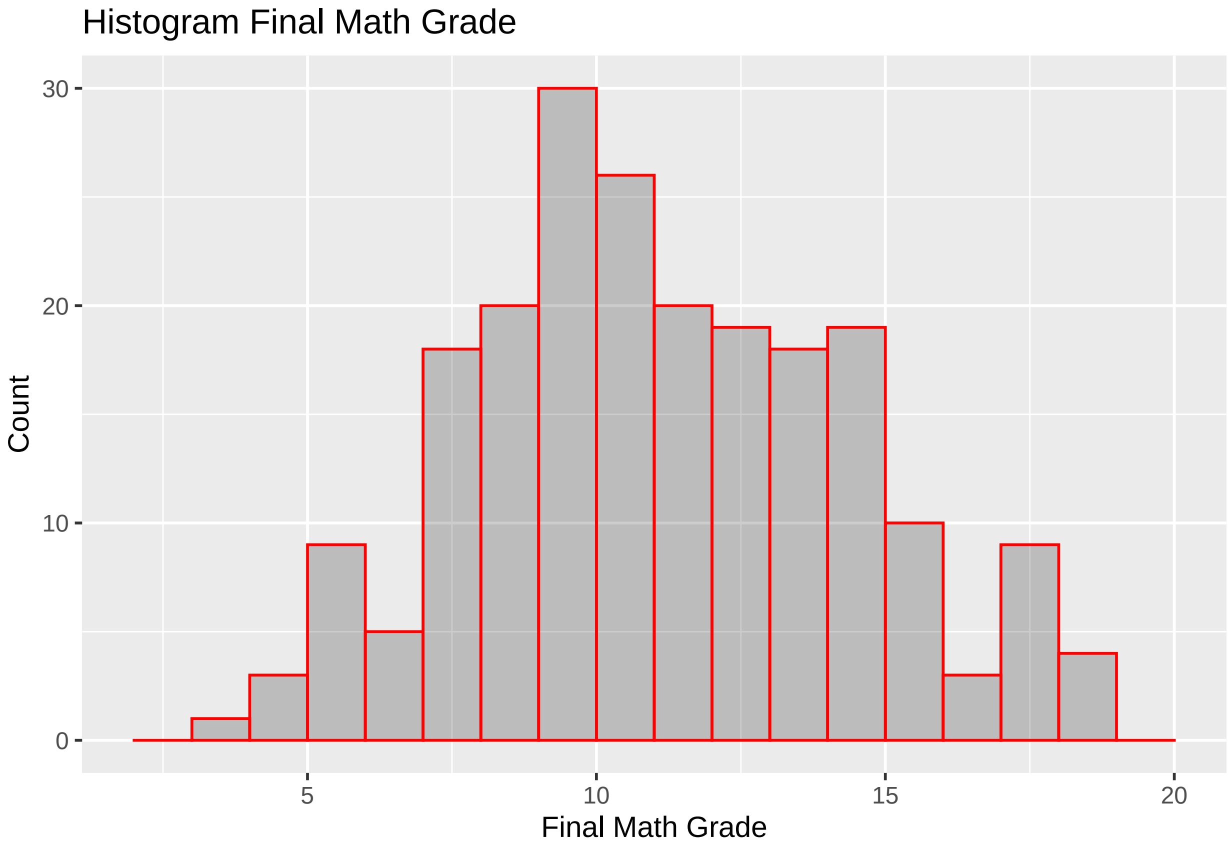
```
## [1] 4
```

```
(max_grade <- max(train$G3))
```

```
## [1] 19
```

Exercise 3

```r
(final_grade_hist <- ggplot(data=train, aes(G3)) +
    geom_histogram(breaks=seq(2,20, by=1),
                   col="red",
                   fill="black",
                   alpha = 0.2)+
    labs(title="Histogram Final Math Grade", x="Final Math Grade", y="Count"))
```



Exercise 4

Predictive modeling is used to predict an object of interest (e.g. forecasting or nowcasting) by using predictors (covariates). The goal is to get as good out-of-sample predictions as possible (e.g. predicting unemployment). The goal of causal modeling, in contrast, is to establish a causal relationship between the explanatory variables (covariates) and the object of interest (e.g. causal effect of inflation, GDP per capita, . . . on unemployment). While for predictive modeling only the full rank condition is mandatory, for causal modeling both the full rank condition and the exclusion restriction must be fulfilled.

Exercise 5

```r
OLS1 <- lm(G3 ~ . ,
            data=select(train, G3, Medu, Fedu, studytime, schoolsup, higher))
```

```
OLS2 <- lm(G3 ~ . + .^2,
            data=select(train, G3, Medu, Fedu, studytime, schoolsup, higher))

MSE_IS_OLS1 <- round(mean((train$G3 - OLS1$fitted.values)^2), digits = 4)
MSE_IS_OLS2 <- round(mean((train$G3 - OLS2$fitted.values)^2), digits = 4)
R2_IS_OLS1 <- round(summary(OLS1)$r.squared, digits = 4)
R2_IS_OLS2 <- round(summary(OLS2)$r.squared, digits = 4)
(MSE_R2_IS <- data.frame(model = c("OLS1_IS", "OLS2_IS"),
                      MSE = c(MSE_IS_OLS1, MSE_IS_OLS2),
                      R2 = c(R2_IS_OLS1, R2_IS_OLS2)))
```

```
##     model    MSE     R2
## 1 OLS1_IS 8.9888 0.1501
## 2 OLS2_IS 8.3804 0.2076
```

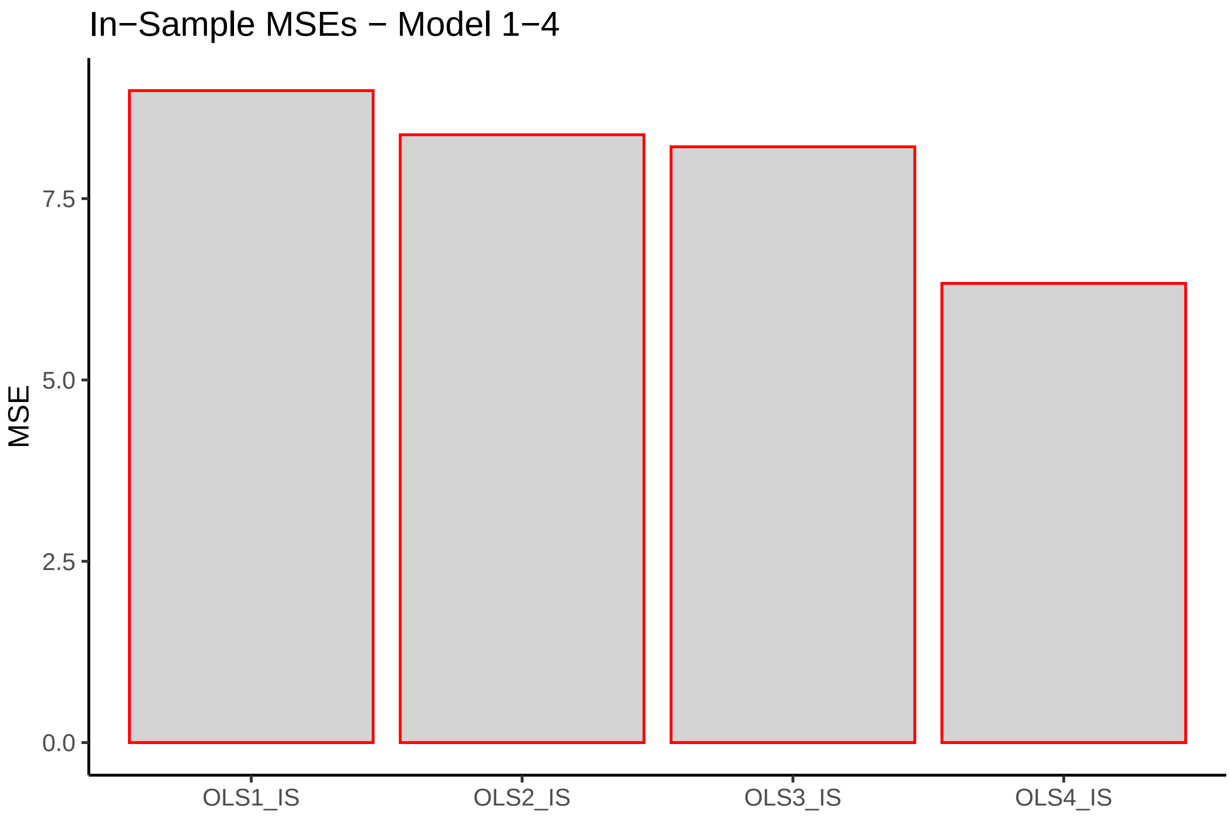
In order to elaborate the in-sample fit of the two models, we define the coefficient of determination R^2 as well as the in-sample MSE as the key fit determinants. From the results, we can observe that the first linear model with five covariates has a R^2 of 0.15 and an in-sample MSE of 8.99, whereas the second linear model including the first order interactions has an R^2 of 0.21 and an in-sample MSE of 8.38. From these results, we can conclude that the second model performns better in both fit coefficients, which is in accordance with the general result that an increased number of covariates often leads to better in-sample fits (or delivers the same model fit). Nevertheless, the both R^2s aand MSEs are relatively low/high, latter compared to the level of the dependent variable, which concludes an overall weak model fit.

Exercise 6

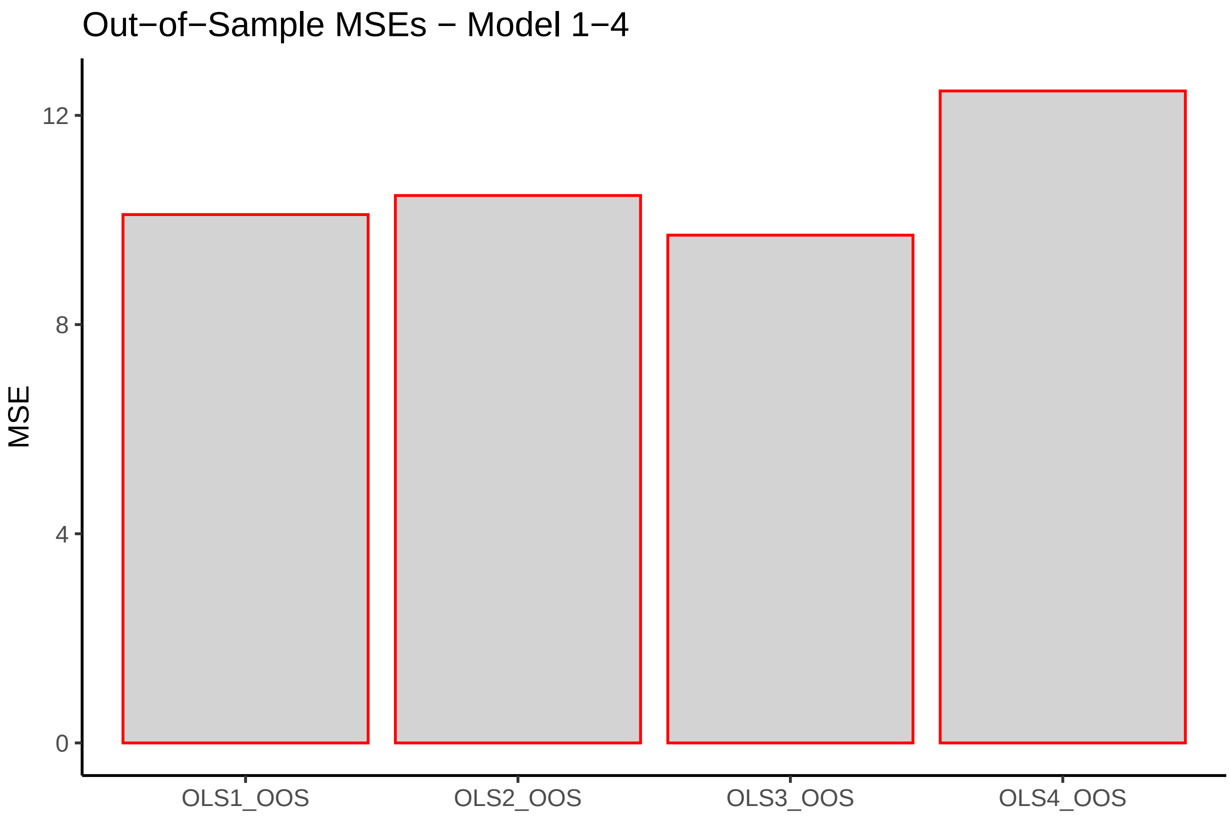
```
OLS3 <- lm(G3 ~ . ,
            data=select(train, G3, Medu, Fedu, studytime, schoolsup, higher, Pstatus,
                        famrel, failures, famsup,internet))

OLS4 <- lm(G3 ~ . + .^2,
            data=select(train, G3, Medu, Fedu, studytime, schoolsup, higher, Pstatus,
                        famrel, failures, famsup,internet))

fit_OLS1 <- predict(OLS1, newdata = test)
fit_OLS2 <- predict(OLS2, newdata = test)
fit_OLS3 <- predict(OLS3, newdata = test)
fit_OLS4 <- predict(OLS4, newdata = test)

MSE_IS_OLS3<- round(mean((train$G3 - OLS3$fitted.values)^2), digits = 4)
MSE_IS_OLS4 <- round(mean((train$G3 - OLS4$fitted.values)^2), digits = 4)
MSE_OOS_OLS1<- round(mean((test$G3 - fit_OLS1)^2), digits = 4)
MSE_OOS_OLS2 <- round(mean((test$G3 - fit_OLS2)^2), digits = 4)
MSE_OOS_OLS3<- round(mean((test$G3 - fit_OLS3)^2), digits = 4)
MSE_OOS_OLS4 <- round(mean((test$G3 - fit_OLS4)^2), digits = 4)
R2_IS_OLS3 <- round(summary(OLS3)$r.squared, digits = 4)
R2_IS_OLS4 <- round(summary(OLS4)$r.squared, digits = 4)
R2_OOS_OLS1 <- "-"
R2_OOS_OLS2 <- "-"
R2_OOS_OLS3 <- "-"
R2_OOS_OLS4 <- "-"

(MSE_R2_IS_OOS <- data.frame(model =
                    c("OLS1_IS", "OLS2_IS", "OLS3_IS", "OLS4_IS",
                      "OLS1_OOS", "OLS2_OOS", "OLS3_OOS", "OLS4_OOS"),
                    MSE = c(MSE_IS_OLS1, MSE_IS_OLS2, MSE_IS_OLS3, MSE_IS_OLS4,
                            MSE_OOS_OLS1, MSE_OOS_OLS2, MSE_OOS_OLS3, MSE_OOS_OLS4),
                    R2 = c(R2_IS_OLS1, R2_IS_OLS2, R2_IS_OLS3, R2_IS_OLS4,
                           R2_OOS_OLS1, R2_OOS_OLS2, R2_OOS_OLS3, R2_OOS_OLS4)))
```

```
##      model     MSE     R2
## 1  OLS1_IS  8.9888 0.1501
## 2  OLS2_IS  8.3804 0.2076
## 3  OLS3_IS  8.2146 0.2233
## 4  OLS4_IS  6.3310 0.4014
## 5 OLS1_OOS 10.1030      -
## 6 OLS2_OOS 10.4676      -
## 7 OLS3_OOS  9.7090      -
## 8 OLS4_OOS 12.4666      -
```

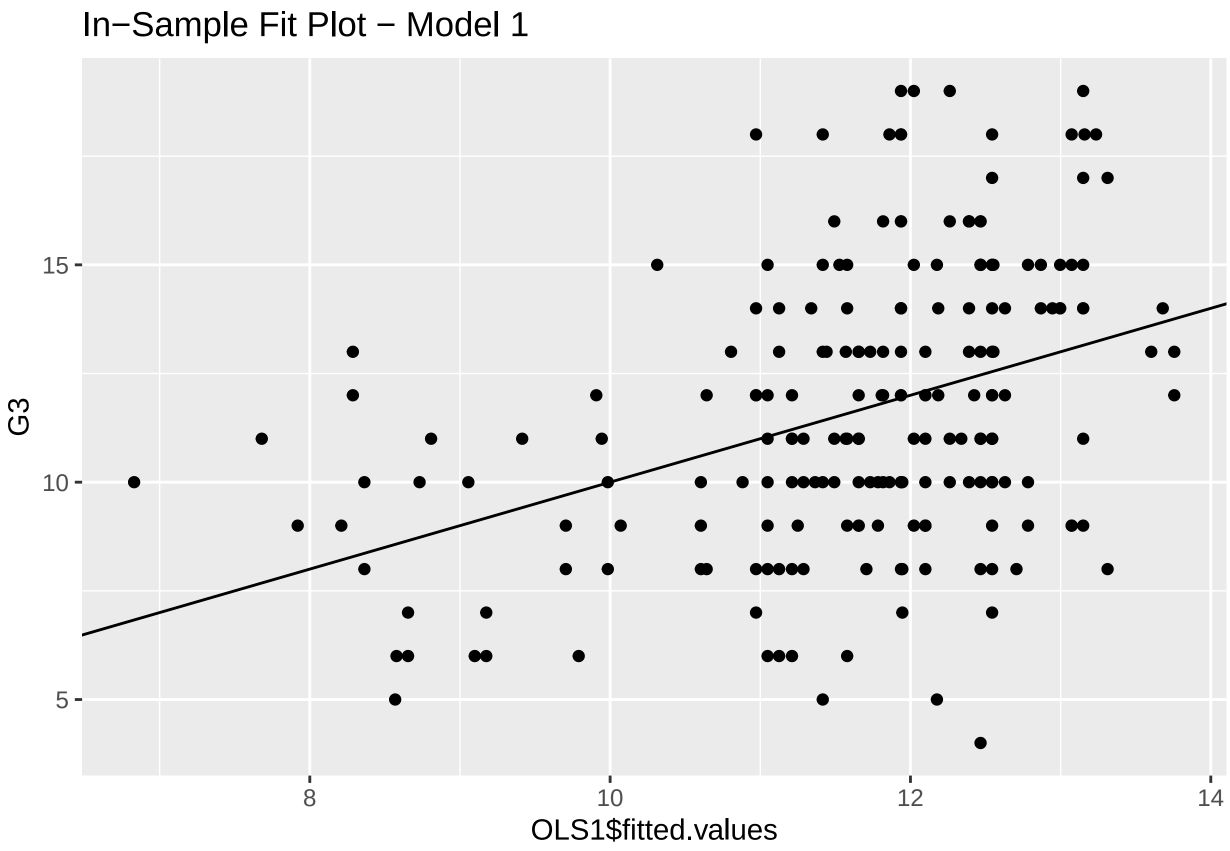
```
(ggplot(slice(MSE_R2_IS_OOS, 1:4), aes(model, MSE)) +
    geom_col(col="red", fill="lightgrey",) +
    ggtitle("In-Sample MSEs - Model 1-4") +
    xlab("") +
    theme_classic())
```



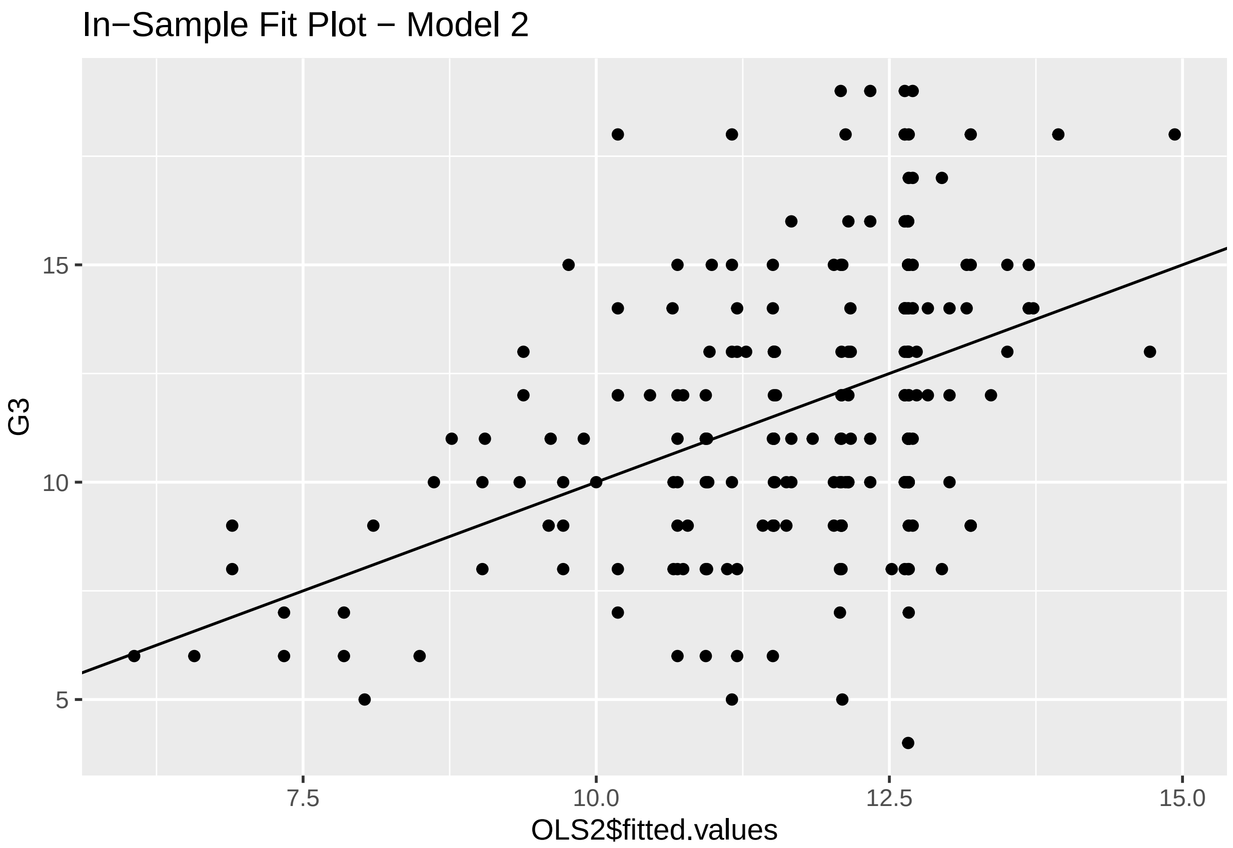
```
(ggplot(slice(MSE_R2_IS_OOS, 5:8), aes(model, MSE)) +
  geom_col(col="red", fill="lightgrey",) +
  ggtitle("Out-of-Sample MSEs - Model 1-4") +
  xlab("") +
  theme_classic())
```



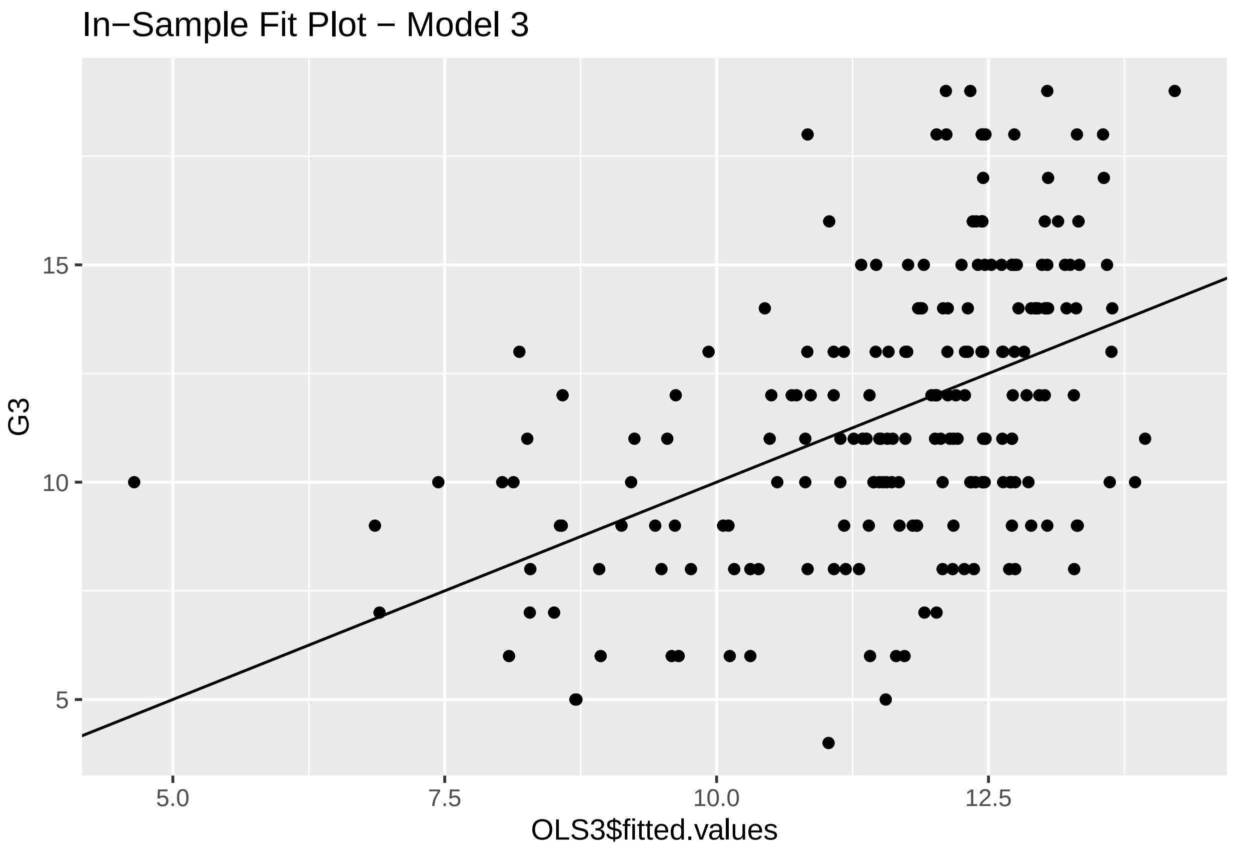
```
(ggplot(train, aes(x=OLS1$fitted.values, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("In-Sample Fit Plot - Model 1"))
```



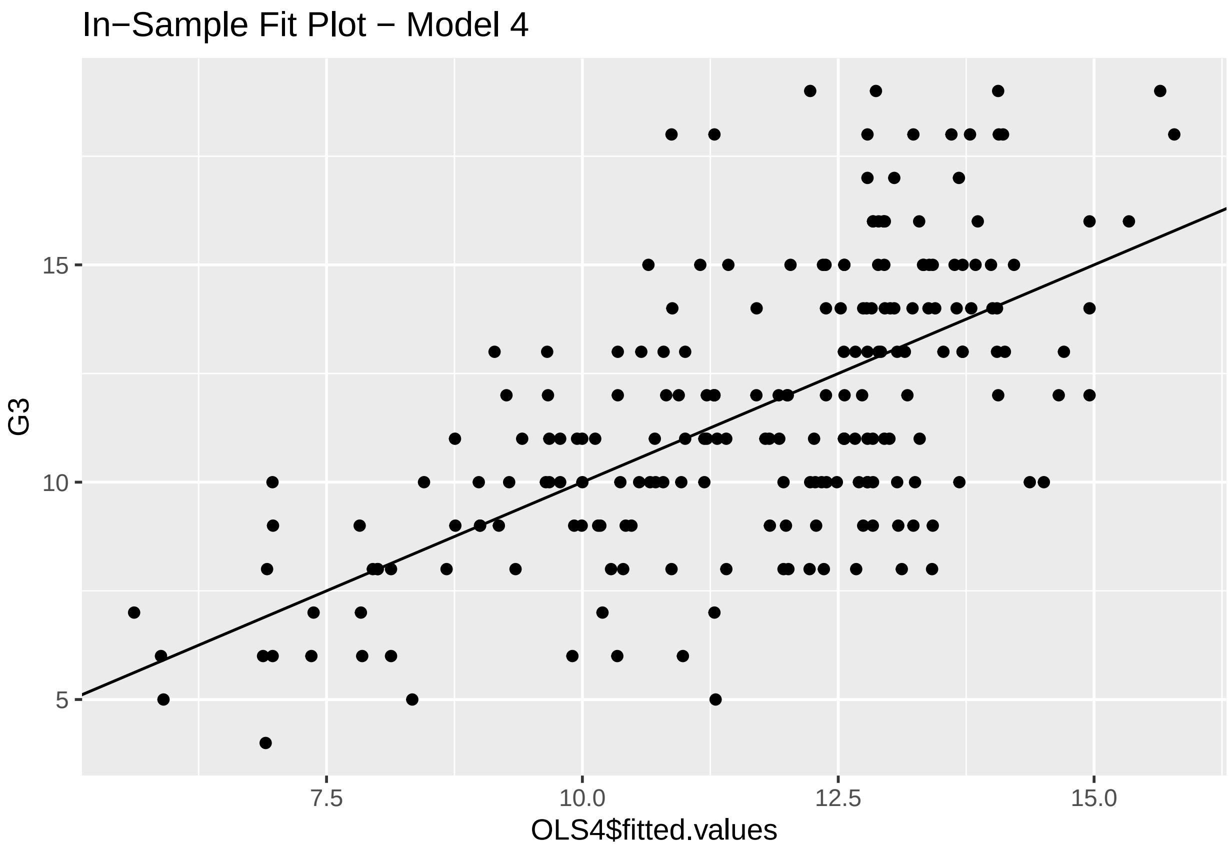
```
(ggplot(train, aes(x=OLS2$fitted.values, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("In-Sample Fit Plot - Model 2"))
```



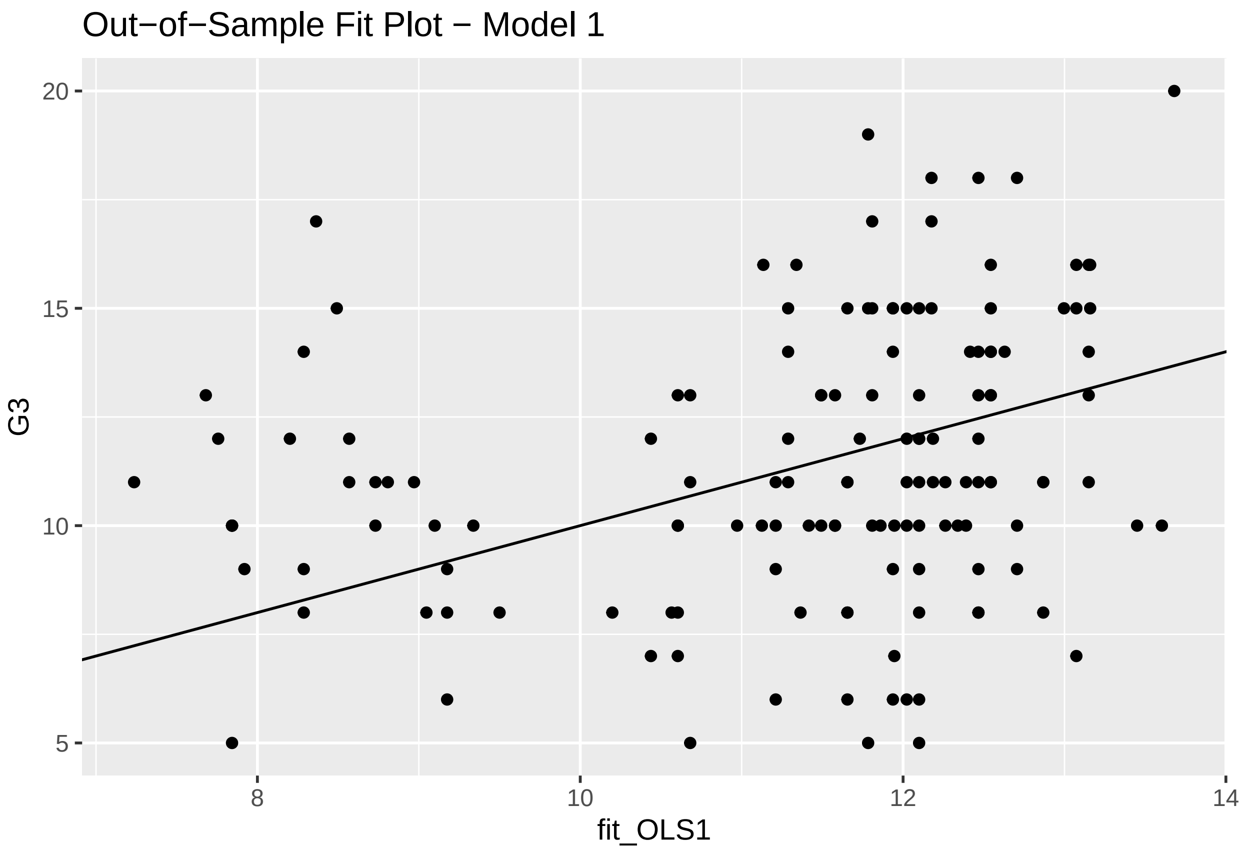
```
(ggplot(train, aes(x=OLS3$fitted.values, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("In-Sample Fit Plot - Model 3"))
```



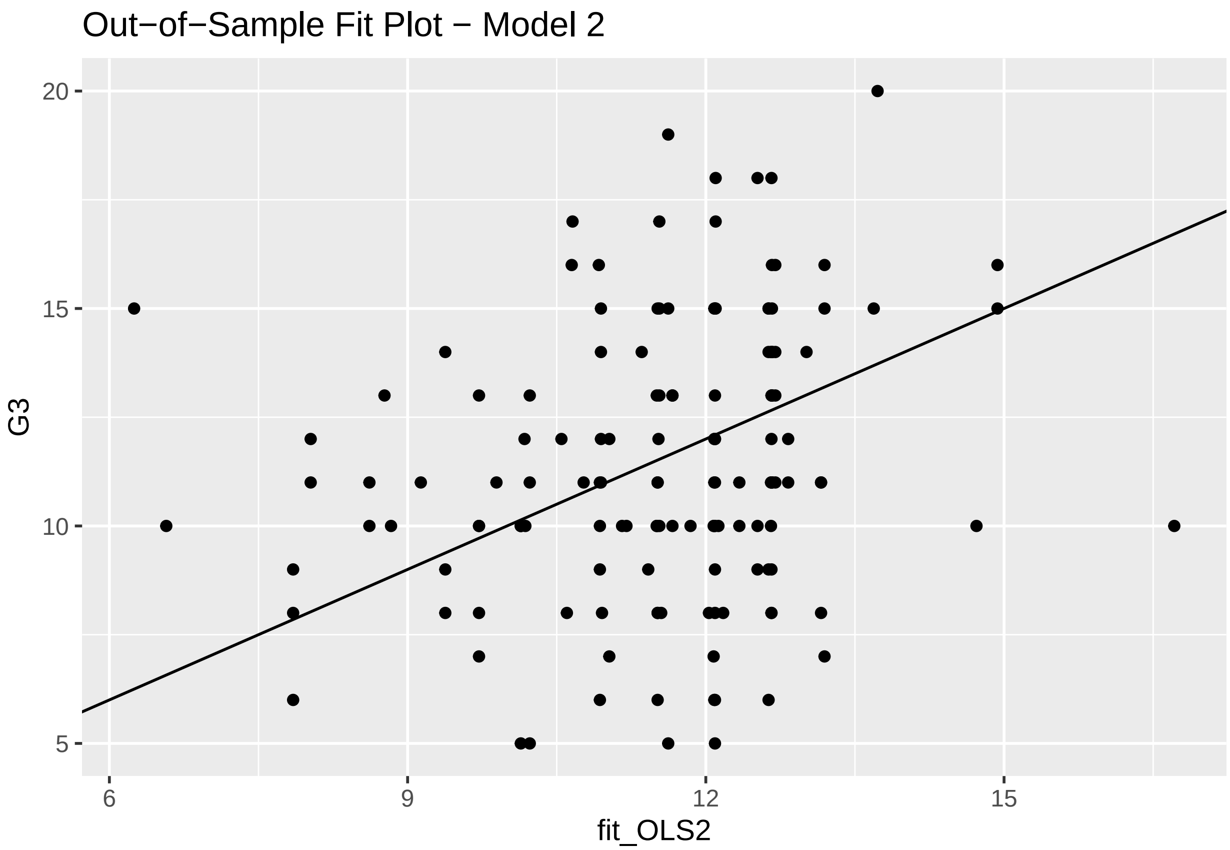
```
(ggplot(train, aes(x=OLS4$fitted.values, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("In-Sample Fit Plot - Model 4"))
```



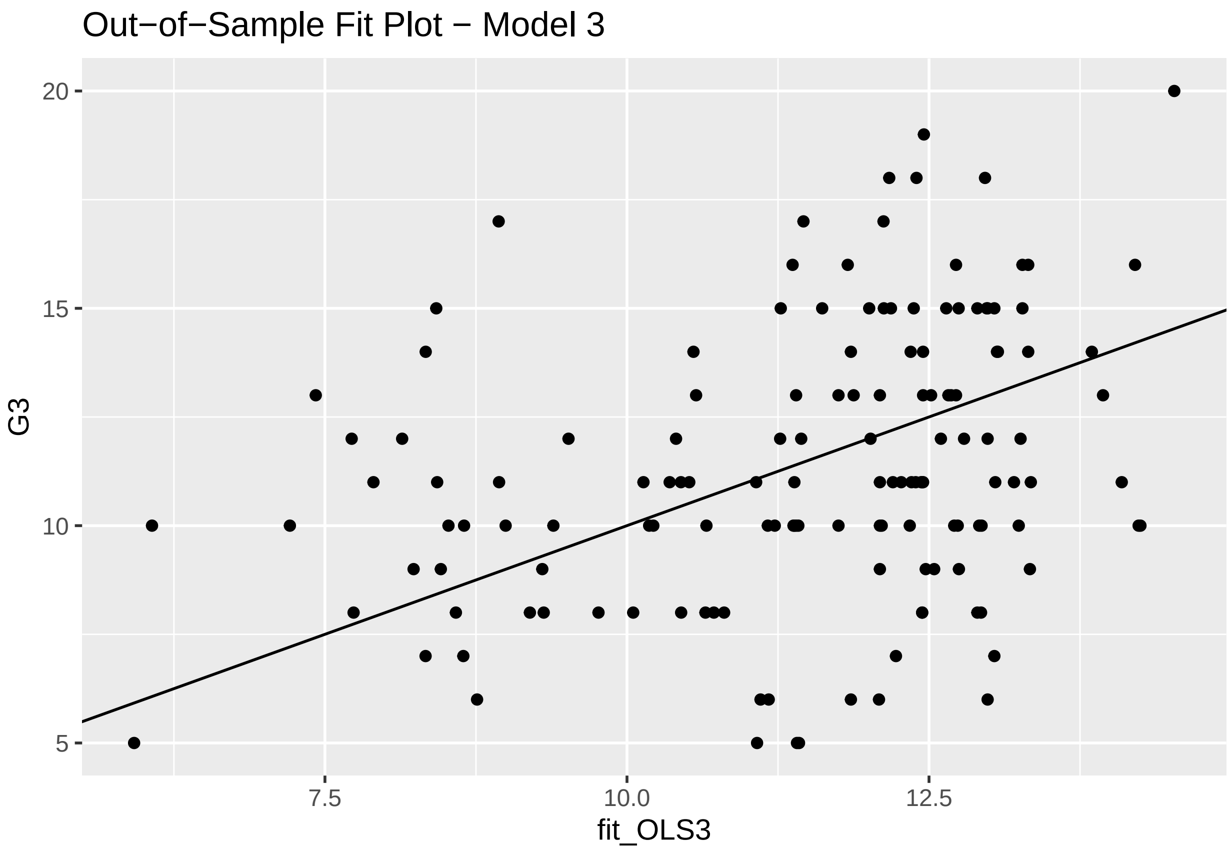
```
(ggplot(test, aes(x=fit_OLS1, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("Out-of-Sample Fit Plot - Model 1"))
```



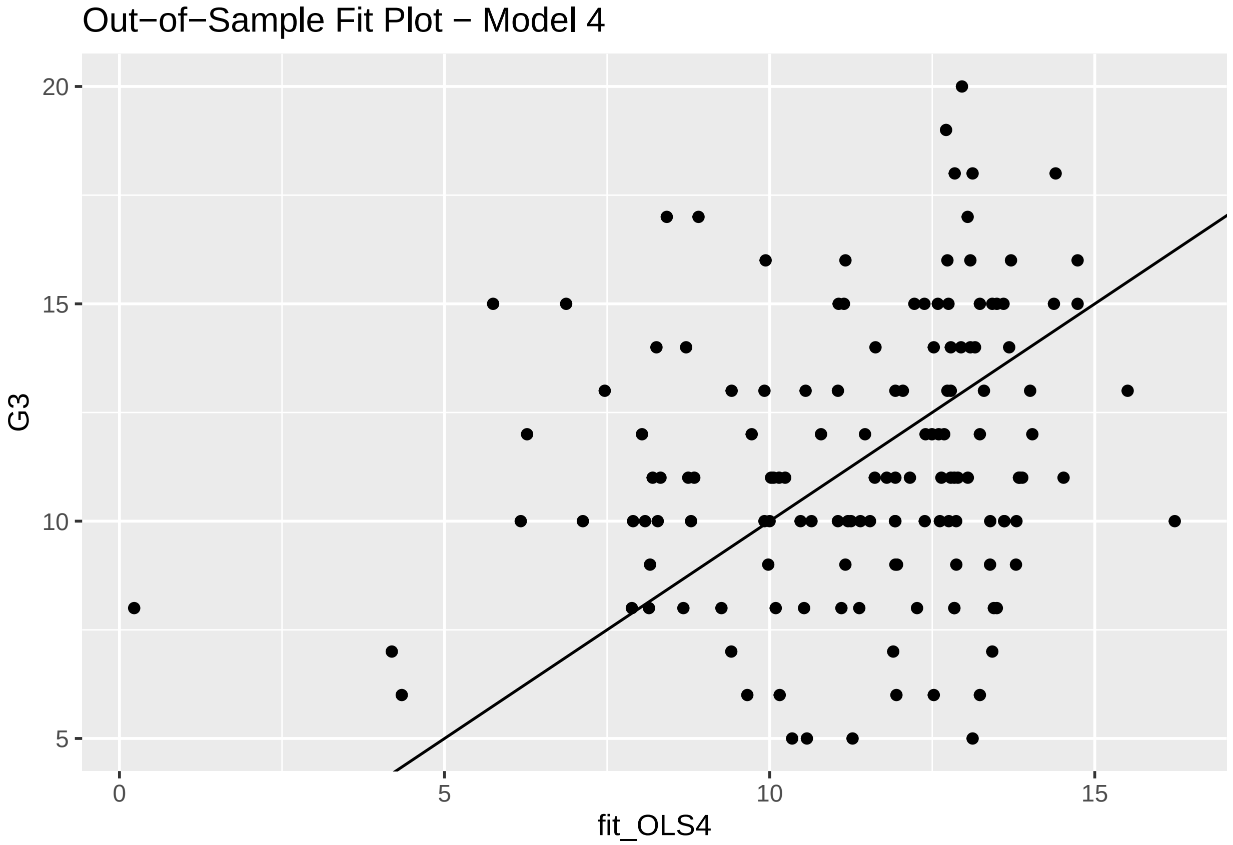
```
(ggplot(test, aes(x=fit_OLS2, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("Out-of-Sample Fit Plot - Model 2"))
```



```
(ggplot(test, aes(x=fit_OLS3, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("Out-of-Sample Fit Plot - Model 3"))
```



```
(ggplot(test, aes(x=fit_OLS4, y = G3)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
    ggtitle("Out-of-Sample Fit Plot - Model 4"))
```



In order to determine the best-performing model, we define the out-of-sample fit plot as well as the out-of-sample MSE as the main determinants, based on the fact that the purpose of a prediction model is to perform best on the test sample. From the results, we can observe that the four OLS models, as described in the formulas above, have out-of-sample MSEs of 10.10, 10.47, 9.71 and 12.47, respectively. This result is further underlined in the fit plots in which the prediction of the third model shows the least deviation from the true values of the dependent variable in the test sample. Hence, we can conclude that the third model, namely the linear regression based on OLS with ten different covariates, performs best in the out-of-sample/test data. Thus, a better prediction performance on new data is expected compared to the three other models, even if the fourth model shows the best in-sample performance, which leads to the conclusion of overfitting in the third model.