Untitled

Nathan Krieger

2026-02-18

K folds with varying degrees - REAL DATA

```
# Predicting medv based on lstat

set.seed(123)


library(ISLR2)
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(ggplot2)
#install.packages('caret')

k = 5 # Number of folds

max_degree <- 9
errors <- numeric(max_degree)

# Create folds
flds <- createFolds(Boston$medv, k, list = TRUE)
flds[[1]]
```

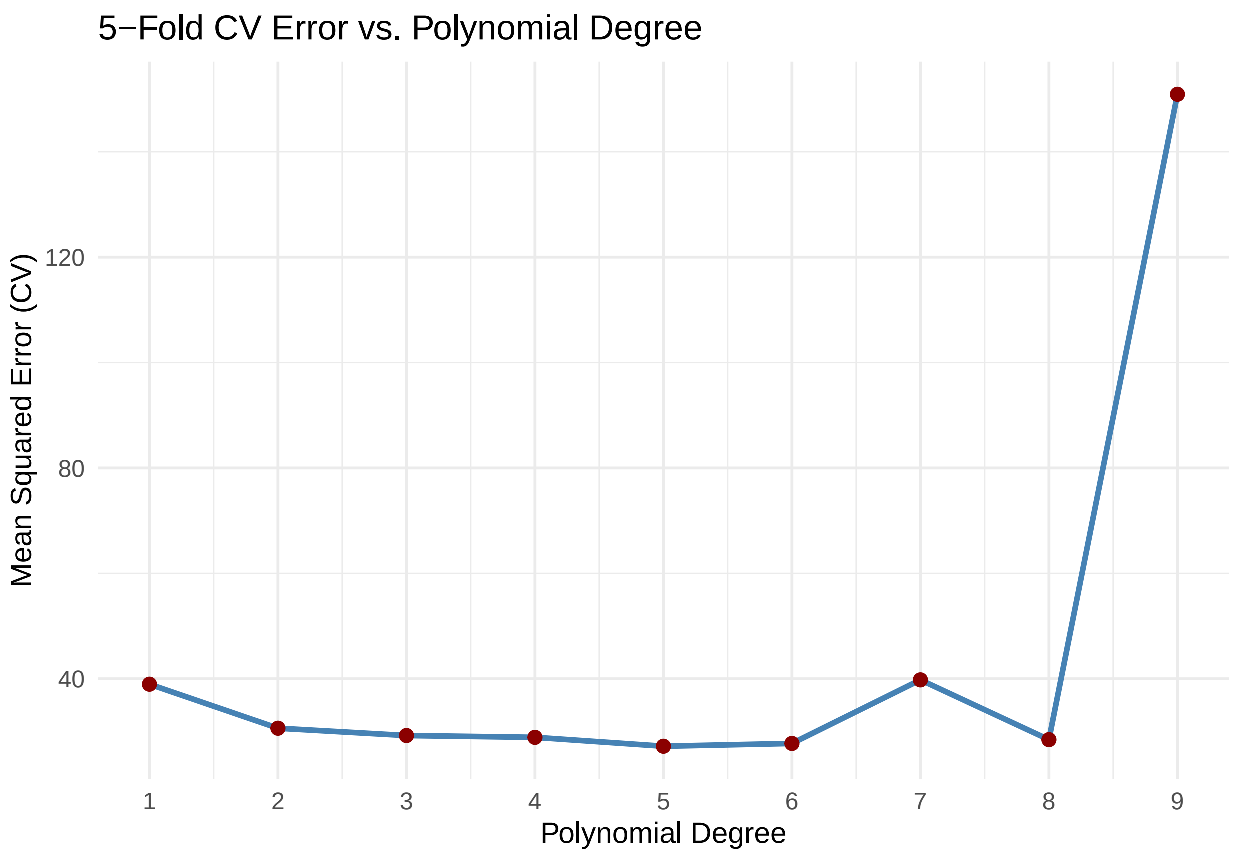
```
##  [1]   2  11  15  16  26  29  41  42  44  49  52  53  54  63  69  72  77  79
## [19]  87  96  97 105 119 121 127 128 129 130 131 133 135 136 137 141 142 149
## [37] 154 156 159 163 169 176 183 184 189 195 200 203 213 218 220 228 238 240
## [55] 248 253 259 262 272 283 288 289 292 297 299 302 306 308 312 327 331 332
## [73] 349 350 363 385 388 391 392 400 402 410 413 421 423 436 446 449 455 457
## [91] 458 460 461 465 486 490 495 499 500 501
```

```
for (d in 1:max_degree) {

  folds <- numeric(k)

  for(i in 1:k){
    # Get test indices for this fold
    test_index = flds[[i]]
    test = Boston[test_index, ]
    train = Boston[-test_index, ]

    m <- lm(medv ~ poly(lstat, d), data = train)
    predict <- predict(m, newdata = test)

    folds[i] = mean((test$medv - predict)^2)

  }

  errors[d] = mean(folds)
}

cv_results <- data.frame(Degree = 1:max_degree, CV_Error = errors)
print(cv_results)
```

```
##   Degree  CV_Error
## 1      1  38.97214
## 2      2  30.61871
## 3      3  29.23357
## 4      4  28.88822
## 5      5  27.20094
## 6      6  27.72832
## 7      7  39.79336
## 8      8  28.45044
## 9      9 150.90530
```

```
ggplot(cv_results, aes(x = Degree, y = CV_Error)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color = "darkred", size = 2) +
  scale_x_continuous(breaks = 1:9) +
  labs(title = "5-Fold CV Error vs. Polynomial Degree",
       x = "Polynomial Degree",
       y = "Mean Squared Error (CV)") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once per session.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



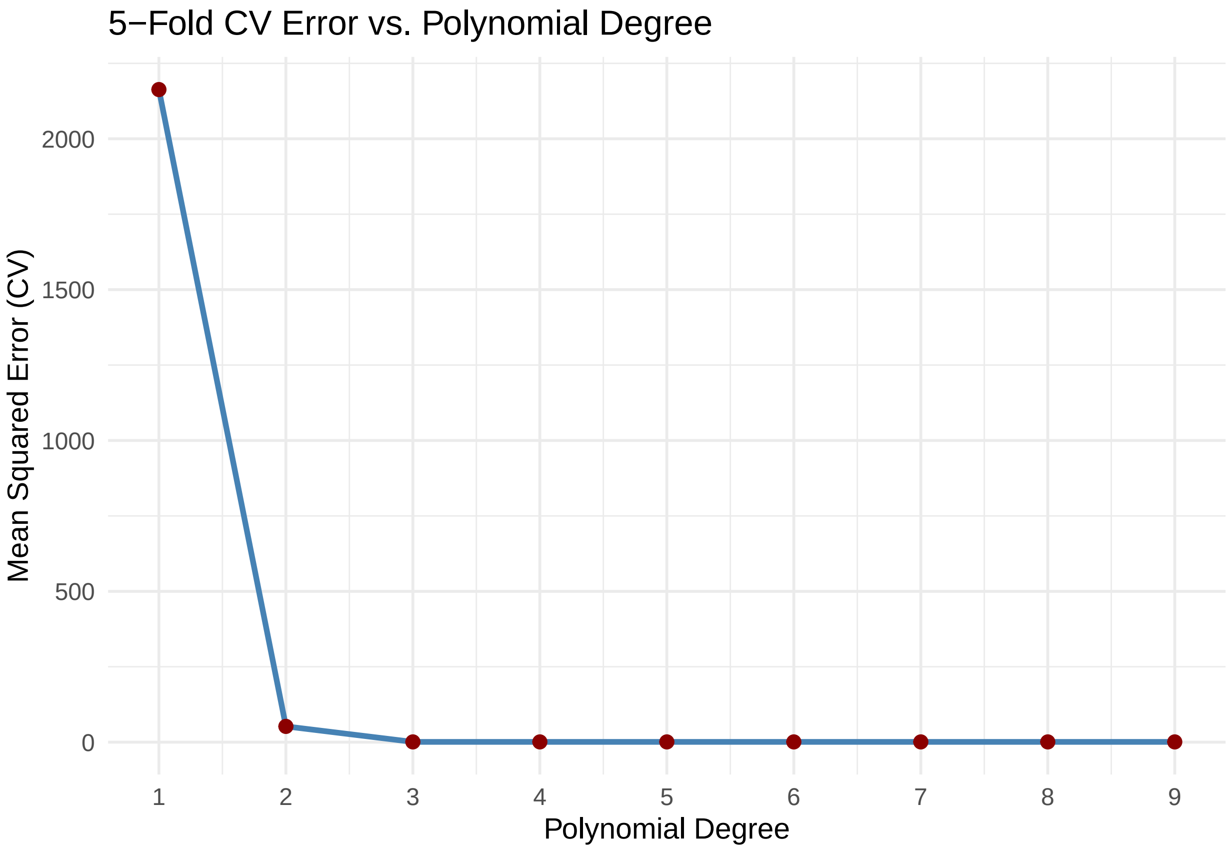
K folds with varying degrees - FAKE DATA

```
set.seed(123)

n <- 500

X1 <- seq(0, 5, length.out = n)

beta_0 <- 1
beta_1 <- 1
beta_2 <- 2
beta_3 <- 3

error = rnorm(n,0,1)

y <- beta_0 + beta_1 * X1 + beta_2 * X1^2 + beta_3 * X1^3 + error

df <- data.frame(X1, y)

?data.frame

library(ISLR2)
library(caret)
library(ggplot2)
#install.packages('caret')
```

```r
k = 5 # Number of folds

max_degree <- 9
errors <- numeric(max_degree)

# Create folds
flds <- createFolds(df$y, k, list = TRUE)
#flds[[1]]

for (d in 1:max_degree) {

  folds <- numeric(k)

  for(i in 1:k){
    # Get test indices for this fold
    test_index = flds[[i]]
    test = df[test_index, ]
    train = df[-test_index, ]

    m <- lm(y ~ poly(X1, d), data = train)

    predict <- predict(m, newdata = test)

    folds[i] = mean((test$y - predict)^2)

  }

  errors[d] = mean(folds)
}

cv_results <- data.frame(Degree = 1:max_degree, CV_Error = errors)
print(cv_results)
```

```
##   Degree     CV_Error
## 1      1 2163.1389959
## 2      2   52.3397377
## 3      3    0.9503528
## 4      4    0.9528262
## 5      5    0.9569451
## 6      6    0.9575593
## 7      7    0.9555550
## 8      8    0.9566102
## 9      9    0.9564446
```

```r
ggplot(cv_results, aes(x = Degree, y = CV_Error)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color = "darkred", size = 2) +
  scale_x_continuous(breaks = 1:9) +
  labs(title = "5-Fold CV Error vs. Polynomial Degree",
       x = "Polynomial Degree",
       y = "Mean Squared Error (CV)") +
  theme_minimal()
```

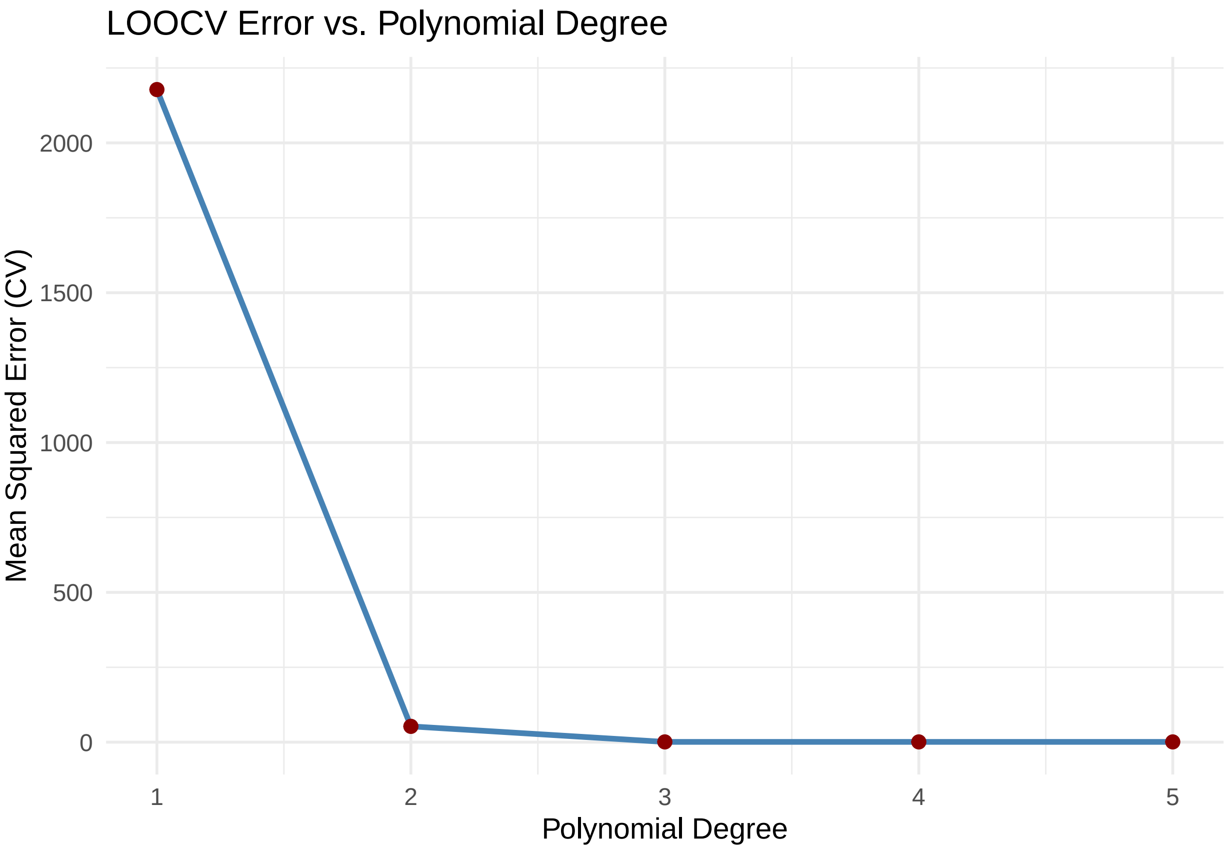


LOOCV with varying degrees - FAKE DATA

```
set.seed(123)

n <- 500

X1 <- seq(0, 5, length.out = n)

beta_0 <- 1
beta_1 <- 1
beta_2 <- 2
beta_3 <- 3

error = rnorm(n,0,1)

y <- beta_0 + beta_1 * X1 + beta_2 * X1^2 + beta_3 * X1^3 + error

df <- data.frame(X1, y)

max_degree <- 5
errors <- numeric(max_degree)

for (d in 1:max_degree) {

  abc <- numeric(n)
  for(i in 1:n){
    test = df[i, ]
    train = df[-i, ]

    m <- lm(y ~ poly(X1, d), data = train)

    predict <- predict(m, newdata = test)

    abc[i] = mean((test$y - predict)^2)

  }

  errors[d] = mean(abc)
}

cv_results <- data.frame(Degree = 1:max_degree, CV_Error = errors)
print(cv_results)
```

```
##   Degree     CV_Error
## 1      1 2177.9593624
## 2      2   52.8534003
## 3      3    0.9573797
## 4      4    0.9579840
## 5      5    0.9611512
```

```
ggplot(cv_results, aes(x = Degree, y = CV_Error)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color = "darkred", size = 2) +
  scale_x_continuous(breaks = 1:9) +
  labs(title = "LOOCV Error vs. Polynomial Degree",
       x = "Polynomial Degree",
       y = "Mean Squared Error (CV)") +
  theme_minimal()
```


LOOCV Error vs. Polynomial Degree
Mean Squared Error (CV)
2000
1500
1000
500
0
1
2
3
4
5
Polynomial Degree