Classwork 1

Rento Saijo

Department of Mathematics, Connecticut College

STA336: Statistical Machine Learning

Yan Zhuang, Ph.D.

February 2, 2026

I. Exploration

Let us explore the 2nd income CSV.

A. Basics

In the set-up chunk, tikZ was set to produce LaTeX figures, all necessary libraries were loaded, and the seed (123) was set. Let us load the income data and check the basics: data types, missingness, and summary.

```
# Read from CSV.
INCOME <- readr::read_csv('data/Income2.csv', show_col_types = FALSE) %>%
  select(education = Education, seniority = Seniority, income = Income)


# Check data types and # of observations.
tibble::glimpse(INCOME)
```

```
## Rows: 30
## Columns: 3
## $ education <dbl> 21.58621, 18.27586, 12.06897, 17.03448, 19.93103, 18.27586, ~
## $ seniority <dbl> 113.10345, 119.31034, 100.68966, 187.58621, 20.00000, 26.206~
## $ income    <dbl> 99.91717, 92.57913, 34.67873, 78.70281, 68.00992, 71.50449, ~
```

The data contains $n = 30$ observations and $p = 3$ numeric variables: `education`, `seniority`, and `income`.

```
# Check NAs.
colSums(is.na(INCOME))
```

```
## education seniority    income
##         0         0         0
```

There are 0 NAs in every column, so no imputation or row removal is needed before modeling.

```
# Check summary.
summary(INCOME)
```

```
##    education       seniority          income
##  Min.   :10.00   Min.   : 20.00   Min.   :17.61
##  1st Qu.:12.48   1st Qu.: 44.83   1st Qu.:36.39
##  Median :17.03   Median : 94.48   Median :70.80
```

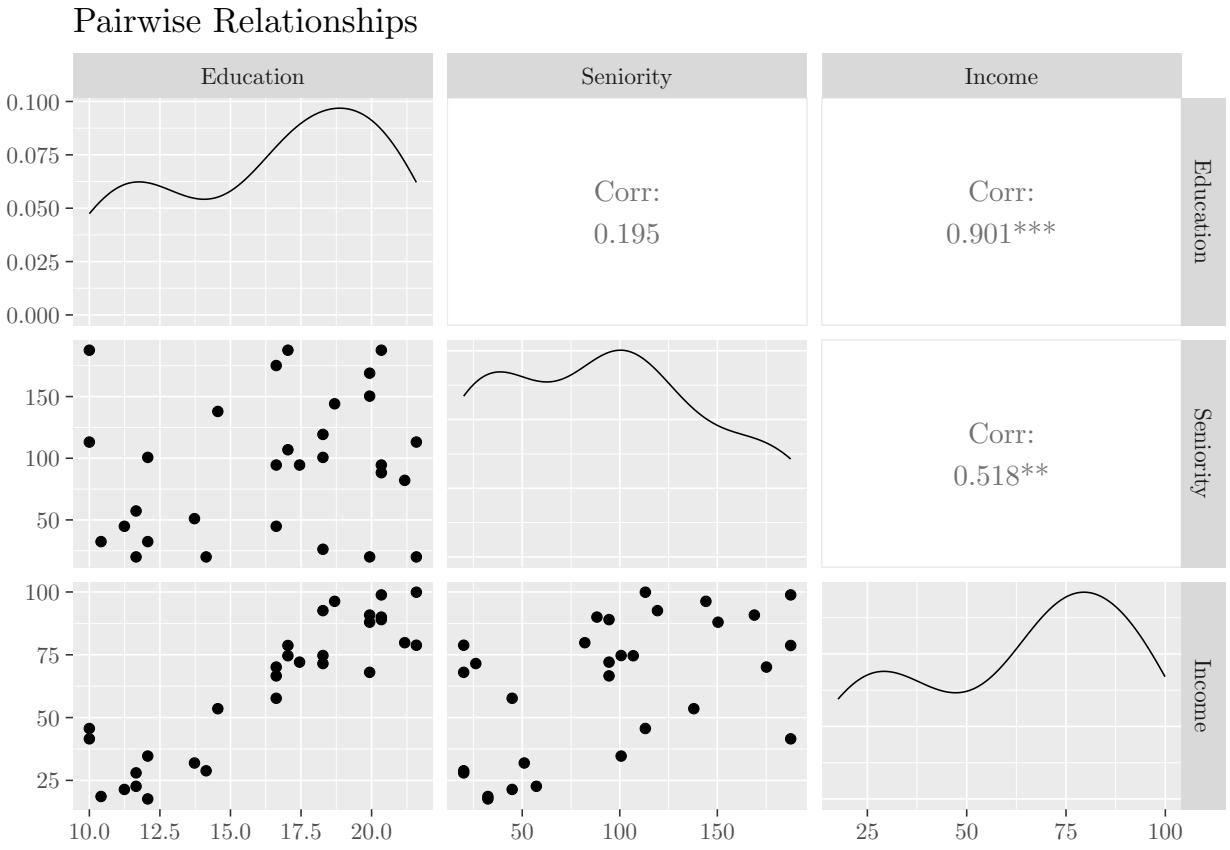
```
##  Mean   :16.39   Mean   : 93.86   Mean   :62.74
##  3rd Qu.:19.93   3rd Qu.:133.28   3rd Qu.:85.93
##  Max.   :21.59   Max.   :187.59   Max.   :99.92
```

The summary statistics indicate that `education`ranges from 10.00 to 21.59 with median 17.03 and mean 16.39, suggesting a roughly centered distribution over a moderate spread. `seniority` has a much wider range (20.00 to 187.59) with median 94.48 and mean 93.86, indicating substantial variability but a fairly balanced center. Finally, `income` ranges from 17.61 to 99.92 with median 70.80 exceeding the mean 62.74, which suggests a left-skew (a few relatively low income values pull the mean downward).

B. Pair-wise Plot

Let us build a pair-wise plot to view any kind of relationships among the 3 variables.

```
# Build pair-wise plot.
GGally::ggpairs(
  INCOME,
  columnLabels = c('$\\mathrm{Education}$', '$\\mathrm{Seniority}$', '$\\mathrm{Income}$'),
  title        = 'Pairwise Relationships'
)
```



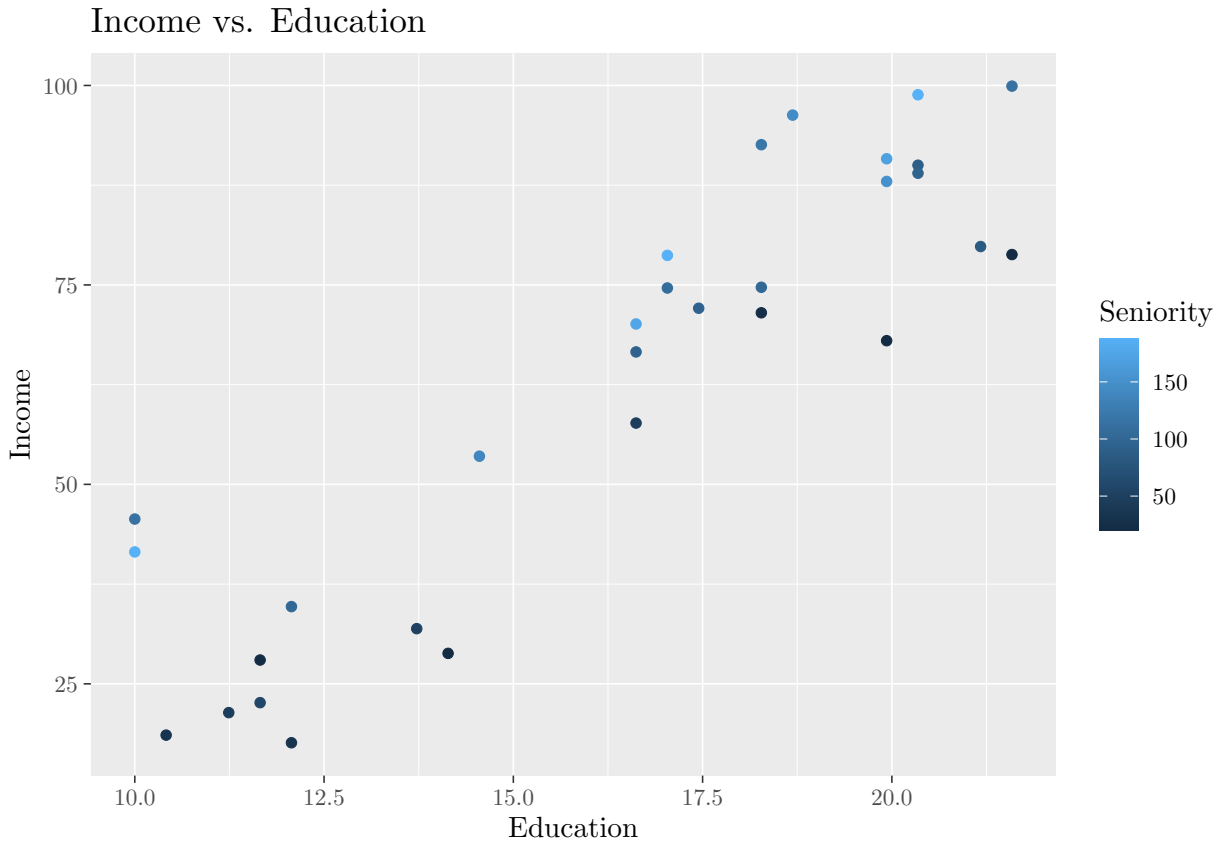
`income` appears to be most strongly related to `education`. The upper-triangle correlations confirm this: Corr(Education, Income) ≈ 0.901 (high and statistically significant), suggesting a strong positive linear association where higher education tends to correspond to higher income. There is also a moderate positive relationship between `seniority` and `income` with Corr(Seniority, Income) ≈ 0.518, indicating that income tends to increase with seniority as well, though with noticeably more scatter than the education-income relationship. In contrast, `education` and `seniority` have only a weak association (Corr ≈ 0.195), which suggests these two predictors are not strongly linearly related to each other. This is useful for modeling because it reduces concerns about multicollinearity when including both variables as predictors of income. Overall, the plot supports a regression model where `education` is likely the strongest single predictor of `income` with `seniority` potentially providing additional explanatory power.

C. Income vs. Education by Seniority

Let us take a look at if there's any distinction in the income-education relationship by seniority.

```
ggplot2::ggplot(INCOME, ggplot2::aes(x = education, y = income)) +
  ggplot2::geom_point(ggplot2::aes(color = seniority)) +
  ggplot2::labs(
```

```
    title  = 'Income vs. Education',
    x      = '$\\mathrm{Education}$',
    y      = '$\\mathrm{Income}$',
    color  = '$\\mathrm{Seniority}$'
)
```



The effect of seniority is not perfectly uniform across the plot, though, there is still noticeable vertical spread in income for similar education values, meaning education does not fully explain income by itself. Overall, this visualization suggests an additive relationship where both `education` and `seniority` contribute to predicting `income`, and it also motivates considering an interaction term if we want to test whether the education–income slope changes with seniority.

II. Modeling

Let us build several models to predict `income`.

A. Train-Test Split

Let us split the data into training and testing splits.

```
split <- caret::createDataPartition(INCOME$income, p = 0.5, list = FALSE)
train <- INCOME[split, ]
test  <- INCOME[-split, ]
```

B. Simple Linear Regression

Let us build a simple linear regression to predict `income` based solely on `education`.

i. Summary

```
# Fit SLR.
slr <- lm(income ~ education, data = train)


# Compute MSE for test set.
pred_test <- predict(slr, newdata = test)
mse_test  <- mean((test$income - pred_test)^2)
extra_row <- data.frame(
  term        = 'Test MSE',
  SLR         = sprintf('%.3f', mse_test),
  check.names = FALSE
)
```

```
# Create table 1.
modelsummary::modelsummary(
  list('SLR' = slr),
  add_rows = extra_row,
  output = 'latex',
  title = 'SLR Summary'
)
```

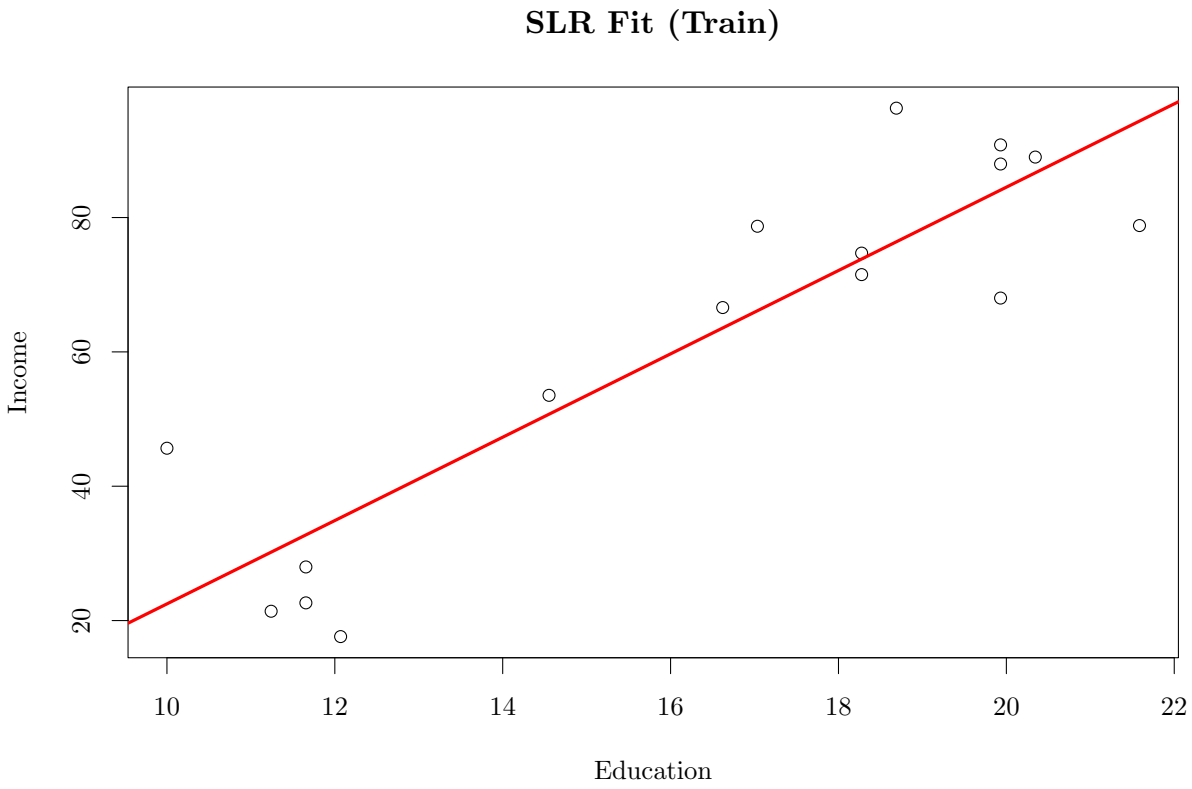
Table 1: SLR Summary

	SLR
(Intercept)	−39.547
	(13.948)
education	6.203
	(0.831)
Num.Obs.	16
R2	0.799
R2 Adj.	0.785
AIC	130.2
BIC	132.5
Log.Lik.	−62.079
F	55.760
RMSE	11.72
Test MSE	129.487

Table 1 summarizes the simple linear regression model predicting `income` from `education`. The estimated slope for `education` is positive ($\hat{\beta}_1 \approx 6.203$), meaning that, on average, an additional year (or unit) of education is associated with about a 6.2 unit increase in income. The model explains a substantial portion of the variation in income on the training set ($R^2 \approx 0.799$), indicating a strong linear relationship between education and income in this sample. However, the test-set error is noticeably larger: the reported test MSE is about 129.487, which corresponds to a test RMSE of roughly $\sqrt{129.487} \approx 11.38$. This gap between training fit (high R^2) and test performance suggests that, while education is an important predictor, the single-predictor model may miss additional structure in the data and may not generalize perfectly to unseen observations.

ii. Plot

```
# See training scatterplot.
plot(
  train$education, train$income,
  xlab = '$\\mathrm{Education}$',
  ylab = '$\\mathrm{Income}$',
  main = 'SLR Fit (Train)'
)
abline(slr, col = 'red', lwd = 3)
```



The training scatterplot shows a clear positive relationship between `education` and `income`, and the fitted regression line captures the overall upward trend. Most points lie reasonably close to the line, which is consistent with the relatively strong training fit reported earlier. That said, there is still noticeable vertical spread around the line (especially at higher education levels), suggesting that `education` alone does not fully explain variation in `income`.

C. Smoothing Splines

Let us manually fit smoothing splines with df = 2 and df = 3. These choices are motivated by the training scatterplot, where a mild parabola (df = 2) or a slightly more flexible 'S' shape (df = 3) could plausibly capture curvature in the education-income relationship.

i. MSE Comparison

```r
# Fit smoothing splines with df = 2 and df = 3.
ss2 <- smooth.spline(train$education, train$income, df = 2)
ss3 <- smooth.spline(train$education, train$income, df = 3)

# Compute test MSE for df = 2 and df = 3.
```

```r
pred2 <- predict(ss2, x = test$education)$y
pred3 <- predict(ss3, x = test$education)$y
mse2 <- mean((test$income - pred2)^2)
mse3 <- mean((test$income - pred3)^2)
data.frame(
  df = c(2, 3),
  test_mse = c(mse2, mse3)
)
```

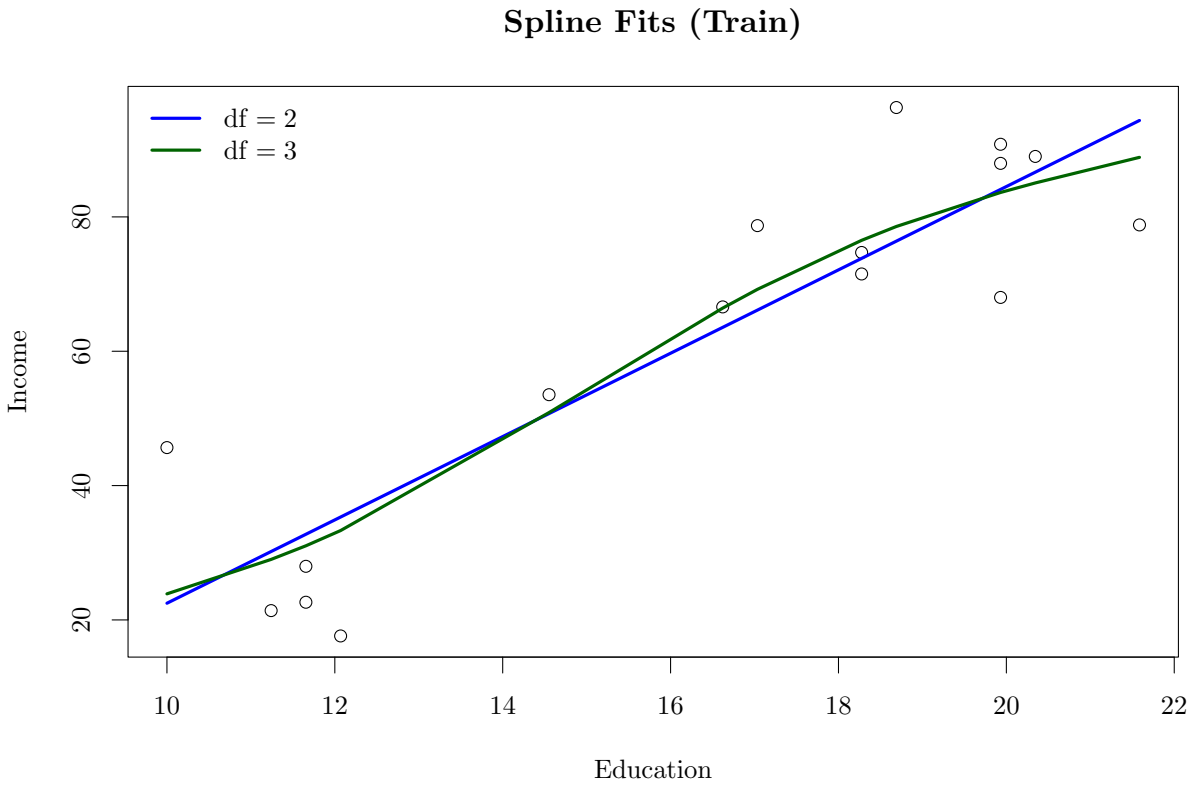
```
##   df test_mse
## 1  2 129.4866
## 2  3 117.9341
```

With the manual smoothing spline fits, df = 3 achieves a lower test-set MSE than df = 2 (about 117.93 vs. 129.49). This suggests that allowing slightly more flexibility improves predictive performance on the held-out data, which is consistent with the idea that the education-income relationship may have mild curvature that a very smooth (nearly quadratic) fit cannot fully capture. At the same time, both MSE values are in the same general range, so the improvement from df = 3 is moderate rather than dramatic; this motivates using cross-validation over a wider range of degrees of freedom to choose the level of smoothness more systematically.

ii. Plot

```r
# Plot training scatter with fitted spline curves.
plot(
  train$education, train$income,
  xlab = '$\\mathrm{Education}$',
  ylab = '$\\mathrm{Income}$',
  main = 'Spline Fits (Train)'
)
lines(ss2, col = 'blue', lwd = 3)
lines(ss3, col = 'darkgreen', lwd = 3)
legend(
  'topleft',
```

```
  legend = c('$\\mathrm{df}=2$', '$\\mathrm{df}=3$'),
  col = c('blue', 'darkgreen'),
  lwd = 3,
  bty = 'n'
)
```



The fitted spline curves show that both df = 2 and df = 3 capture the overall increasing trend between `education` and `income`, but the df = 3 fit introduces a small amount of curvature. In particular, the df = 3 curve bends slightly upward in the mid-range of education (roughly 15-19) and then flattens somewhat at the highest education values, which is consistent with a mild 'S'-shaped relationship. By comparison, the df = 2 curve is closer to a nearly linear/quadratic trend and does not adapt as much to local changes in the pattern. Visually, the additional flexibility of df = 3 appears to better follow the central tendency of the points without becoming overly wiggly, which aligns with its lower test MSE in the previous comparison.

C. Cross-validated Smoothing Splines

Let us find the degrees of freedom that minimizes the test MSE via cross-validation.

i. Determining DF

```
# Determine df that minimizes test MSE.
dfs  <- seq(1.01, 10, by = 0.01)
mses <- c()
for (df in dfs) {
  ss        <- smooth.spline(train$education, train$income, df = df)
  predicted <- predict(ss, x = test$education)$y
  actual    <- test$income
  res <- caret::postResample(pred = predicted, obs = actual)
  mse <- as.numeric(res['RMSE'])^2
  mses <- c(mses, mse)
}
dfs_mses <- data.frame(df = dfs, mse = mses)
min_mse <- dfs_mses %>% dplyr::filter(mse == min(mse))
min_mse
```

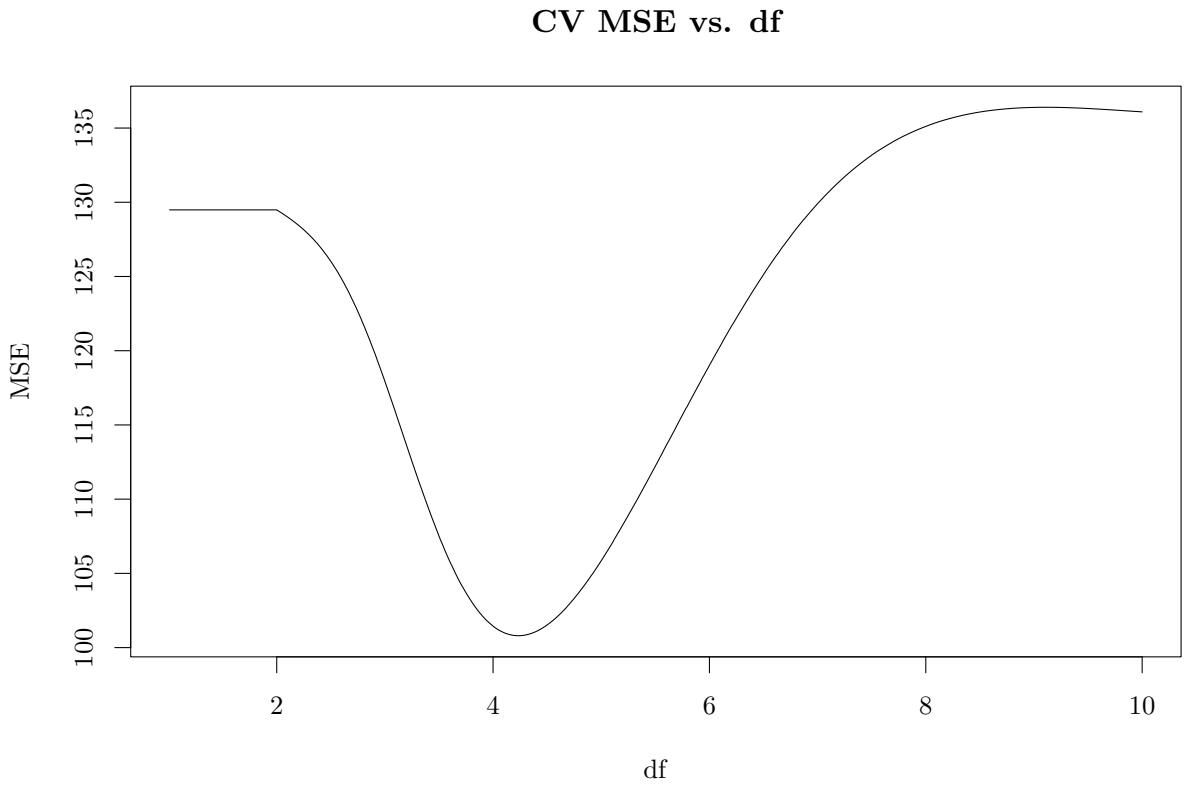
```
##     df      mse
## 1 4.23 100.8009
```

Sweeping over degrees of freedom from 1.01 to 10 shows that the lowest test-set MSE occurs around df ≈ 4.23, with a minimum MSE of about 100.80. Compared to the manual fits (df = 2 and df = 3), this suggests that a slightly more flexible spline can better capture the curvature in the education-income relationship and improve generalization to the test set. At the same time, the optimal df is still relatively small, indicating that the underlying pattern is smooth (mild curvature) rather than highly wiggly, so we do not need a very large df to achieve good predictive performance.

ii. Cross-validation Plot

```
# Create cross-validation plot.
plot(
  dfs, mses, type = 'l',
  xlab = '$\\mathrm{df}$',
  ylab = '$\\mathrm{MSE}$',
  main = 'CV MSE vs. df'
```

```
)
```



The cross-validation curve shows a clear U-shaped pattern in test MSE as the spline degrees of freedom increase. Starting from very low df (an overly smooth fit), the test MSE decreases as the spline becomes flexible enough to capture curvature in the relationship. The minimum occurs around df ≈ 4.23, where the test MSE is lowest (about 100.8). After this point, the test MSE rises again as df increases, suggesting that higher flexibility begins to overfit noise in the training data and hurts generalization.