Assignment 08

Cross-Validation

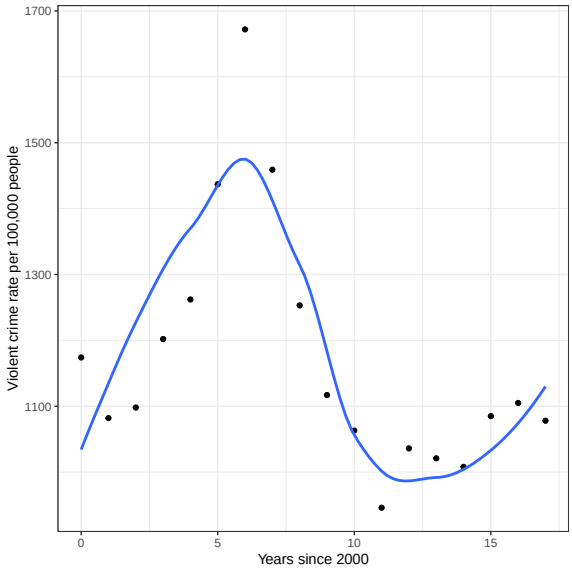
Answer Key

This assignment is worth 20 points.

Part I: Minneapolis Violent Crime

1. Create a scatterplot showing the violent crime rate as a function of time.

Figure 1. *Plot of violent crime rate versus years since 2000. The loess smoother is also displayed.*



2. Based on the plot, describe the trend in violent crime rate over time.

The plot of the data suggests that crime rate is non-linear and both increases and decreases at various points since 2000.

3. If you were going to fit a polynomial model to these data, what degree polynomial would you fit to? Explain.

The data suggest a cubic polynomial may fit the data. The loess line suggests two changes in the direction of the trend.

4. Fit a series of polynomial models starting with a linear model, and then models that also include higher order polynomials that allow you to evaluate your response to Question #3. Be sure to fit models up to degree + 1, where is the degree you hypothesized in Question #3. Analyze each of the polynomial terms (including the linear term) by using a series of nested F-tests. Report these results in an ANOVA table. (Note: If you need a refresher on fitting polynomial models and carrying out a nested F-test, see the Polynomial Regression notes from EPsy 8252.)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
16	492484.7	NA	NA	NA	NA
15	437112.6	1	55372.09	2.8775748	0.1136202
14	266051.2	1	171061.37	8.8897121	0.0106084
13	250154.1	1	15897.13	0.8261416	0.3799330

5. Based on these results, which polynomial model would you adopt? Explain.

Based on these results we would adopt the cubic polynomial model. Adding the cubic term in Model 3 results in a statistically significant increase in variance explained from the quadratic model, $F(1, 14) = 8.89$, $p = 0.011$. The subsequent increase in variance when we add the 4th-degree polynomial term is not statistically significant, $F(1, 13) = 0.83$, $p = 0.380$.

6. Write and include syntax that will carry out the LOOCV.

```
# Set up empty vector to store results
mse_1 = rep(NA, 18)
mse_2 = rep(NA, 18)
mse_3 = rep(NA, 18)
mse_4 = rep(NA, 18)


# Loop through the cross-validation
for(i in 1:18){
  train = mpls %>% filter(row_number() != i)
  validate = mpls %>% filter(row_number() == i)

  lm.1 = lm(crime_rate ~ 1 + year2, data = train)
  lm.2 = lm(crime_rate ~ 1 + year2 + I(year2^2), data = train)
  lm.3 = lm(crime_rate ~ 1 + year2 + I(year2^2) + I(year2^3), data = train)
  lm.4 = lm(crime_rate ~ 1 + year2 + I(year2^2) + I(year2^3) + I(year2^4), data = train)

  yhat_1 = predict(lm.1, newdata = validate)
  yhat_2 = predict(lm.2, newdata = validate)
  yhat_3 = predict(lm.3, newdata = validate)
  yhat_4 = predict(lm.4, newdata = validate)

  mse_1[i] = (validate$crime_rate - yhat_1) ^ 2
  mse_2[i] = (validate$crime_rate - yhat_2) ^ 2
  mse_3[i] = (validate$crime_rate - yhat_3) ^ 2
  mse_4[i] = (validate$crime_rate - yhat_4) ^ 2
}
```

7. Report the cross-validated MSE for each of the models in your set of polynomial models.

Model	MSE
Linear	32950.63
Quadratic	33150.82
Cubic	30172.30
Quartic	45402.68

8. Based on these results, which degree polynomial model should be adopted? Explain.

Based on these results, we will adopt the cubic model. It has the lowest cross-validated MSE of the four candidate models.

Part II: Course Evaluations

9. Using average course evaluation scores (y), compute the total sum of squares (SST). Show your work.

```
sum((evals$avg_eval - mean(evals$avg_eval)) ^ 2)
```

```
## [1] 19.7745
```

10. Using average course evaluation scores (y) and the predicted values from the model ($\hat{y}$), compute the sum of squared errors (SSE). Show your work.

```
sum((evals$avg_eval - fitted(lm.full)) ^ 2)
```

```
## [1] 13.80276
```

11. Compute the model R^2 value using the formula: $1 - \frac{\text{SSE}}{\text{SST}}$.

```
## [1] 0.301992
```

12. Write and include syntax that will carry out the 5-fold cross-validation. In this syntax use `set.seed(1000)` so that you and the answer key will get the same results.

```
# Create 5 folds
set.seed(1000)
my_cv = evals %>% crossv_kfold(k = 5)

# Fit model to training set and get augmented data based on validation set
cv_1 = my_cv %>%
  mutate(
    model = map(train, ~lm(avg_eval ~ 1 + beauty + female + num_courses + native_english, data = .)),
    out = map2(model, test, ~ augment(.x, newdata = .y))
    ) %>%
  tidyr::unnest(out)
```

13. Report the five R^2 values from your analysis and the cross-validated R^2 value.

```
# Group by validation set; compute SST, SSE, and R2
all_r2s = cv_1 %>%
  group_by(.id) %>%
  summarize(
    SSE = sum((avg_eval - .fitted) ^ 2),
    SST = sum((avg_eval - mean(avg_eval))^2)
  ) %>%
  mutate(
    R2 = 1 - SSE/SST
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
all_r2s
```

```
## # A tibble: 5 x 4
##   .id     SSE   SST      R2
##   <chr> <dbl> <dbl>   <dbl>
## 1 1      5.54  6.41  0.135
## 2 2      2.71  2.56 -0.0584
## 3 3      3.77  4.81  0.216
## 4 4      1.69  2.03  0.169
## 5 5      1.55  3.26  0.524
```

```
# Compute cross-validated R2
all_r2s %>%
  summarize(
    Mean_R2 = mean(R2)
  )
```

```
## # A tibble: 1 x 1
##   Mean_R2
##     <dbl>
## 1   0.197
```

14. How does this value compare to the R^2 value you computed in Question #11, based on the data.

The R^2 value based on the cross-validation (0.219) is much lower than the data-based R^2 value (0.302).

15. Explain why the cross-validated estimate of R^2 is a better estimate than the data-based R^2.

We expect the cross-validated R^2 value to be a better estimate since the errors (and thus the SSE) are minimized based on the data. Thus the data-based R^2 value is optimistic. The errors from the cross-validated data will be larger and thus the R^2 value will be smaller, and a better estimate of both how the model will perform on new data sets.

Part III: Credit Balance

16. Use the `lm.ridge()` function to fit the same sequence of d values you used in the FOR loop from Assignment 6, Question 7. Running `MASS::select()` on this output, provides d values based on different criteria. Report the d value associated with the generalized cross-validation (GCV) metric.

```
## modified HKB estimator is 0.4054568
## modified L-W estimator is 0.5629809
## smallest value of GCV  at 1.849
```

The modified d value that has the lowest CV-MSE based on fitting a ridge regression is 1.849.

17. Re-run the FOR loop from Assignment 6, Question 7. Except this time compute the AICc and select the d value based on using the AICc. How does this compare to the d value you found using the GCV metric from the previous question? (Show your syntax.)

```
##       d     aicc
## 1 1.867 1564.432
```

The d value associated with the lowest AICc was $d = 1.87$. This is quite comparable to the d value chosen with the GCV metric ($d = 1.85$).

18. Use 10-fold cross-validation to find the modified d value that has the lowest CV-MSE based on fitting an elastic net. (Prior to carrying out this analysis, set the seed for the random number generation to 100.) Then use this modified d value to re-fit the elastic net. Report the modified d value and the fitted equation from the elastic net.

```
## [1] 0.0025585
```

```
## # A tibble: 6 x 5
##   term       step estimate  lambda dev.ratio
##   <chr>     <dbl>    <dbl>   <dbl>     <dbl>
## 1 income        1  -0.570  0.00256     0.878
## 2 limit         1   0.678  0.00256     0.878
## 3 rating        1   0.638  0.00256     0.878
## 4 cards         1   0.0358 0.00256     0.878
## 5 age           1  -0.0328 0.00256     0.878
## 6 education     1   0.0120 0.00256     0.878
```

The modified d value chosen by the 10-fold cross-validation was 0.003. The fitted equation is:

$$\widehat{\text{Balance}}_i = -0.57(\text{Income}_i) + 0.68(\text{Credit Limit}_i) + 0.64(\text{Credit Rating}_i) + 0.04(\text{Credit Cards}_i) - 0.03(\text{Age}_i) + 0.01(\text{Education}_i)$$

19. Compare the coefficients from the elastic net (from Question 18) to those from the ridge regression analyses fitted using the d value you found in Question 16. How are they different? Explain based on the penalty terms in the two models.

Table 1. *Coefficients for the ridge regression (d=1.833) and elastic net (modified d = 0.0026).*

Predictor	Ridge Model	Elastic Net
Income	-0.566	-0.570
Credit Limit	0.647	0.678
Credit Rating	0.665	0.638
Credit Cards	0.036	0.036
Age	-0.034	-0.033
Education	0.013	0.012

The coefficients from the two models are quite similar. The elastic net has shrunk them slightly more than the ridge model, but the difference is negligible. This is expected as the elastic net produces a slightly larger penalty than the ridge model.

20. Use the elastic net to compute a multiple R^2 value. Remember that multiple R^2 is the squared correlation between the observed and predicted values. Report this value. (Show your work.)

```
##           [,1]
## [1,] 0.8782096
```

The model explains 87.8% of the variation in credit balance.