Homework 2: [YOUR NAME HERE]

Introduction to Time Series, Fall 2024

Due Friday September 27

The total number of points possible for this homework is 39. The number of points for each question is written below, and questions marked as "bonus" are optional (points awarded for bonus problems can be used to earn back points that you may have lost on other parts of this homework but will not put you above full credit). Submit the **knitted pdf file** from this Rmd to Gradescope.

If you collaborated with anybody for this homework, put their names here:

Simple regression

1. (2 pts) Derive the population least squares coefficients, which solve

$$\min_{\beta_1, \beta_0} \mathbb{E}\big[(y - \beta_0 - \beta_1 x)^2\big],$$

by differentiating the criterion with respect to each β_j, setting equal to zero, and solving. Repeat the calculation but without intercept (without the β_0 coefficient in the model).

2. (2 pts) As in Q1, now derive the sample least squares coefficients, which solve

$$\min_{\beta_1, \beta_0} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Again, repeat the calculation but without intercept (no β_0 in the model).

3. (2 pts) Prove or disprove: in the model without intercept, the regression coefficient of x on y is the inverse of that from the regression of y on x. Answer the question for both the population version and the sample version.

4. (3 pts) Consider the following hypothetical. Let y be the height of a child and x be the height of their parent, and consider a regression of y on x, performed in a large population. Suppose that we estimate the regression coefficients separately for male and female parents (two separate regressions) and we find that the slope coefficient from the former regression $\hat\beta_1^{\text{dad}}$ is smaller than that from the latter $\hat\beta_1^{\text{mom}}$. Suppose however that we find (in this same population) the sample correlation between a father's height and their child's height is *larger* than that between a mother's height and their child's height. What is a plausible explanation for what is happening here?

Multiple regression

5. (2 pts) In class, we claimed that the multiple regression coefficients, with respect to responses y_i and feature vectors $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$, can be written in two ways: the first is

$$\hat\beta = \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \sum_{i=1}^n x_i y_i.$$

The second is

$$\hat\beta = (X^T X)^{-1} X^T y,$$

where $X \in \mathbb{R}^{n \times p}$ is a feature matrix, with i^{th} row x_i, and $y \in \mathbb{R}^n$ is a response vector, with i^{th} component y_i. Prove that these two expressions are equivalent.

6. (Bonus) Derive the population and sample multiple regression coefficients by solving the corresponding least squares problem (differentiating the criterion with respect to each β_j, setting equal to zero, and solving). For the sample least squares coefficient, deriving either representation in Q5 will be fine.

Covariance calculations

7. (3 pts) Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ be random vectors, and let $A \in \mathbb{R}^{k \times n}$ and $B \in \mathbb{R}^{\ell \times m}$ be fixed matrices. Prove that
$$\text{Cov}(Ax, By) = A\text{Cov}(x, y)B^T.$$
Prove as a consequence that $\text{Cov}(Ax) = A\text{Cov}(x)A^T$. Hint: you may use the rule for covariances of linear combinations (as reviewed in the lecture from week 2, "Measures of dependence and stationarity").

8. (2 pts) Suppose that $y = X\beta + \epsilon$, with X and β fixed, and where ϵ is a vector with white noise entries, with variance σ^2. Use the rule in Q7 to prove that for the sample least squares coefficients, namely, $\hat{\beta} = (X^T X)^{-1} X^T y$, it holds that
$$\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

9. (4 pts) An equivalent way to state the Gauss-Markov theorem is as follows. Under the model from Q8, if $\tilde{\beta}$ is any other unbiased linear estimator of β (where linearity means that $\tilde{\beta} = My$ for a fixed matrix M) then
$$\text{Cov}(\hat{\beta}) \lesssim \text{Cov}(\tilde{\beta})$$
where $\lesssim$ means less than or equal to in the *PSD (positive semidefinite) ordering*. Precisely, $A \lesssim B$ if and only if $B - A$ is a PSD matrix, which recall, means $z^T (B - A) z \geq 0$ for all vectors z. Prove that this is indeed equivalent to the statement of the Gauss-Markov theorem given in lecture.

Metrics matter

10. (4 pts) Below is some code to generate data `y` and predictions `yhat1` and `yhat2` from two hypothetical models. Plot the predictions from each model as a line overlaid on top of the original data. Give each line its own color, and use a legend to denote which model it refers to. Then compute and report the mean absolute error (MAE) and mean absolute percentage error (MAPE), as defined in lecture, for each model. Discuss and explain what you find.

```r
set.seed(0)
x = 1:50
y = rpois(n = 50, lambda = c(rep(5, 25), 5 + exp(0:24 * 0.2)))
yhat1 = c(rep(6.5, 25), 6.5 + exp(0:24 * 0.2))
yhat2 = c(rep(5, 25), 5 + exp(0:24 * 0.18))
```

11. (4 pts) Define a new vector of predictions `yhat3` from a third hypothetical model, by starting with those `yhat2` from the second model, and changing the multiplier in the exponent for the last 25 predictions from `0.18` to `0.22`. As in the last question, plot `yhat2` and `yhat3` as colored lines overlaid on top of the data and clearly mark them with a legend. Also, compute and compare the MAE and MAPE of the third model versus the second. It should be worse according to both metrics. Finally, for each of `yhat2` and `yhat3`, compute what is known as mean absolute relative error (MARE):
$$\text{MARE} = 100 \times \frac{1}{N} \sum_{t=1}^{N} \frac{|y_t - \hat{y}_t|}{|\hat{y}_t|}$$
Discuss and explain what you find.

Cross-validation

12. (3 pts) Recall the R code from lecture that performs time series cross-validation to evaluate the mean absolute error (MAE) of predictions from the linear regression of cardiovascular mortality on 4-week lagged particulate levels. Adapt this to evaluate the MAE of predictions from the regression of cardiovascular mortality on 4-week lagged particulate levels and 4-week lagged temperature (2 features). Fit each regression model using a trailing window of 200 time points (not all past). Plot the predictions, and print the MAE on the plot, following the code from lecture.

    Additionally (all drawn on the same figure), plot the fitted values on the training set. By the training set here, we mean what is also called the "burn-in set" in the lecture notes, and indexed by times 1 through `t0` in the code. The fitted values should come from the initial regression model that is fit to the burn-in set. These should be plotted in a different color from the predictions made in time series cross-validation pass. Print the MAE from the fitted values the training set somewhere on the plot (and label this as "Training MAE" to clearly differentiate it).

13. (2 pts) Repeat the same exercise as in Q12 but now with multiple lags per variable: use lags 4, 8, 12 for each of particulate level and temperature (thus 6 features in total). Did the training MAE go down? Did the cross-validated MAE go down? Discuss. Hint: you may find it useful to know that `lm()` can take a predictor matrix, as in `lm(y ~ x)` where `x` is a matrix; in this problem, you can form the predictor matrix by calling `cbind()` on the lagged feature vectors.

14. (2 pts) Repeat once more the same exercise as in the last question but now with many lags per variable: use lags 4, 5, . . . , through 50 for each of particulate level and temperature (thus 47 x 2 = 94 features in total). Did the training MAE go down? Did the cross-validated MAE go down? Are you surprised? Discuss.

More features, the merrier?

15. (2 pts) Let y_i be an arbitrary response, and $x_i \in \mathbb{R}^p$ be an arbitrary feature vector, for $i = 1, \dots, n$. Let

    $$\tilde{x}_i = (x_{i1}, \dots, x_{ip}, \tilde{x}_{i,p+1}), \quad i = 1, \dots, n$$

    be the result of appending one more feature. Let $\hat{y}_i$ denote the fitted values from the regression of y_i on x_i, and let $\tilde{y}_i$ denote the fitted values from the regression of y_i on $\tilde{x}_i$. Prove that

    $$\sum_{i=1}^n (y_i - \tilde{y}_i)^2 \leq \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

    In other words, *the training MSE will never get worse as we add features* to a given sample regression problem.

16. (2 pts) How many linearly independent features do we need (how large should p be) in order to achieve a perfect training accuracy, i.e., training MSE of zero? Why?

17. (Bonus) Implement an example in R in order to verify your answer to Q16 empirically. Extra bonus points if you do it on the cardiovascular mortality data, using enough lagged features. You should be able to plot the fitted values from the training set and see that they match the observations perfectly (and the CV predictions should look super wild).