AAE 722 Machine Learning in Applied Economic Analysis
Homework 1

Due: Oct 2, 2024, 11:59 Central Time

Homework Instructions

Please submit your completed Jupyter Notebook, which should include the following components:

- **Markdown cells**: Use these for explanations, hypotheses, and interpretations.
- **Code cells**: Include all related code following the homework questions.
- **Output**: Ensure that all code cells have been executed, and the outputs are included in the notebook.

Once complete, commit your Jupyter Notebook to your GitHub repository.

1 Question 1 (23 points)

	Coefficient	Std. error	*t*-statistic	*p*-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	−0.001	0.0059	−0.18	0.8599

1.1 Replicating the Regression Table (9 points)

Using the **Advertising** dataset from the ISLP library and the linear regression to model the relationship between **sales** and the three predictors: **TV**, **radio**, and **newspaper**, replicate the regression table similar to the one provided above.

1.2 Hypotheses for the p-values (7 points)

- **Question**: For each predictor (**TV, radio, and newspaper**), describe the null hypothesis that corresponds to the p-values given in the regression table.
- Explain what the null hypothesis means in the context of the data, not just in terms of the coefficients.

1.3 Interpreting the Results (7 points)

- **Question**: Based on the p-values from the regression output, explain what conclusions you can draw about the relationship between sales and the predictors (TV, radio, newspaper).

2 Question 2 (10 point)

Explain the concepts of the K-Nearest Neighbors (KNN) Classifier and K-Nearest Neighbors (KNN) Regression methods. Discuss how each method works and their differences.

3 Question 3 (20 points)

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is **linear**, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer. (5 points)

(b) Answer (a) using test rather than training RSS (Note: The term 'test' refers to the test data). (5 points)

(c) Suppose that the true relationship between X and Y is **not linear, but we don't know how far it is from linear**. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer. (5 points)

(d) Answer (c) using test rather than training RSS. (5 points)

4 Question 4 (22 points)

In this problem, we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor $\mathbf{x}$ and a response $\mathbf{y}$ as follows.

```
rng = np.random.default_rng(1)
x = rng.normal(size=100)
y = 2 * x + rng.normal(size=100)
```

(a) Perform a simple linear regression of $\mathbf{y}$ onto $\mathbf{x}$, **without** an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the keywords argument `intercept=False` to `ModelSpec()`). (4 points)

(b) Now perform a simple linear regression of $\mathbf{x}$ onto $\mathbf{y}$ without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (4 points)

(c) What is the relationship between the results obtained in (a) and (b)? (3 points)

(d) For the regression of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $\hat{\beta}/\mathrm{SE}(\hat{\beta})$, where $\hat{\beta}$ is given by (3.38), and where

$$\mathrm{SE}(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i'=1}^{n} x_{i'}^2}}.$$

(These formulas are slightly different from those given in lecture notes, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in Python, that the t-statistic can be written as

$$\frac{(\sqrt{n-1})\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} x_i y_i\right)^2}}.$$

(4 points)

(e) Using the results from (d), argue that the t-statistic for the regression of $\mathbf{y}$ onto $\mathbf{x}$ is the same as the t-statistic for the regression of $\mathbf{x}$ onto $\mathbf{y}$. (3 points)

(f) In Python, show that when regression is performed **with an intercept**, the t-statistic for $H_0 : \beta_1 = 0$ is the same for the regression of $\mathbf{y}$ onto $\mathbf{x}$ as it is for the regression of $\mathbf{x}$ onto $\mathbf{y}$. (4 points)