Homework Assignment 2

DSO 530 Applied Modern Statistical Learning Methods

Spring 2025

Deadline: The assignment is due **Sunday, Feb 2, end of day (Los Angeles time)**.

Submission Instructions: Submit on Gradescope

- The link to Gradescope is provided on Brightspace under the "Homework" section.
- Log in to Gradescope and upload your completed assignment as a PDF.
- Ensure that all pages are clear and legible before submission. Scanned handwritten solutions are acceptable, as long as they are **clear and legible**.
- **Mark Parts Appropriately:** After uploading your submission, you must mark each question on Gradescope by associating the relevant pages with the corresponding problems. If you fail to mark the questions correctly, the grader will not grade your work. For detailed guidance on submitting assignments and marking questions in Gradescope, refer to this help link.

Collaboration Policy: You are allowed to discuss this assignment with your study group to clarify concepts and brainstorm ideas. However, everyone must submit their own individual writeup in their own words. Include the names of your collaborators at the top of your submission.

Grading Policy: The grading criteria are outlined in the syllabus. Each part of each problem will be graded as follows:

- **0 points**: No attempt was made.
- **2 points**: Attempted, but there are substantial methodological errors or major conceptual misunderstandings.
- **3 points**: Attempted with no substantial errors or misunderstandings.

Late Day Policy: To accommodate unforeseen challenges that may arise during the quarter, you have three late days for the problem sets. Each late day allows you to turn in an assignment up to 24 hours late. (Any fraction of a late day counts as one late day.) You may use multiple late days on the same problem set. Work submitted beyond the allowed late days will not receive credit.

Professional Writeup: Present your work professionally. Clearly label each problem and ensure that your answers are easy to read. Provide full explanations for your solutions where required, and write your solutions in a logical, structured, and concise manner. If your solutions include code, ensure that it is properly formatted and that the output is clearly presented. For mathematical problems, show all steps to fully demonstrate your understanding.

Understanding the Simple Linear Regression Model

Assume the following model: $Y_i = 10.0 + 0.5X_i + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$.

(a) $\mathbb{E}[Y|X=0] = ?, \quad \mathbb{E}[Y|X=-1] = ?, \quad \text{var}[Y|X] = ?$

(b) What is the probability of $Y > 10$, given $X = 2$?

(c) If X has a mean of zero and variance of 20, what are $\mathbb{E}[Y]$ and $\text{Var}(Y)$?

(d) What is $\text{Cov}(X, Y)$?

Simulation from the SLR Model

Generate $n = 100$ samples of $X \sim N(0, \sigma_X^2)$, with $\sigma_X^2 = 2$. For each draw, simulate Y_i from the simple linear regression model $Y_i = 2.5 - 1.0X_i + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$, with $\sigma_\epsilon^2 = 3$.

(a) Show the scatter plot of Y versus X along with the true regression line.

(b) Split the sample into 2 subsets of size 25 and 75. For each subset, run the regression of Y on X. Add each fitted regression line (use color) to your plot from **(a)**. Why are they not the same?

(c) What is the marginal sample mean for Y? What is the true marginal mean?

(d) Start a fresh scatter plot of Y versus X and add the true regression line and the estimated version (using the full sample).

(i) Add the bounds of the 90% prediction interval to your plot.

(ii) What percentage of your observations are outside of this interval?

(iii) Add the bounds of the *true* 90% prediction interval to your plot. This is the interval that assumes you know the true β_0, β_1, σ_X^2, and σ_ϵ^2 and don't have to use estimates. Thus, estimation of these won't factor into the uncertainty of $\hat{Y}$.

(iv) What percentage of your observations are outside of this *true* interval?

(e) Repeat part **(d)** for different values of n, σ_X^2, and σ_ϵ^2. What do you learn? What effect do these values have?

Maintenance Costs

The cost of the maintenance of a certain type of tractor seems to increase with age. The file `tractor.csv` contains ages (years) and 6-monthly maintenance costs for $n = 17$ such tractors.

(a) Create a plot of tractor maintenance `cost` versus `age`.

(b) Find the least squares fit to the model

$$\texttt{cost}_i = b_0 + b_1 \texttt{age}_i + e_i$$

in two ways: first using the 'statsmodels' package and second by calculating a correlation and standard deviations [verify that the answers are identical]. Add the fitted line to the scatterplot.

(c) Suppose you were considering buying a tractor that is three years old, what would you expect your six-monthly maintenance costs to be? What is the 95% predictive (cost) interval for the six-monthly maintenance of your tractor? Compare the endpoints of the interval to the observed values of `cost`. What do you conclude about your prediction from this? Why or why not is this conclusion surprising?

Broadway Box Office

Let X and Y denote the weekly reports on the box office ticket sales for plays on Broadway in New York for two consecutive weeks, respectively, in October 2017. (You can actually download similar data from `www.playbill.com`. The regression output for this data set in shown in the table below:

Variable	Coefficient	s.e.	t-value	p-value
Intercept	6805	9929	0.685	0.503
X	0.9821	0.01443	68.071	$< 2 \times 10^{-16}$
	$n = 18$	$R^2 = 0.9966$	$s_\varepsilon = 18007.56$	

Suppose that the model satisfies the usual SLR model assumptions, and that the SST for Y is 1.507773×10^{12}.

(a) What were the degrees of freedom used in calculating s_ε? What are the SSE and SSR?

(b) Compute the sample variance for Y (s_Y^2) and sample correlation between X and Y (r_{XY}).

(c) Suppose that the ticket sales in the first week for a particular play was \$822,000. What is the expected sales for the same play in the following week?

(d) Suppose further that $\bar{X} = 822186.6$ and $s_X = 302724.5$. Construct the 95% forecast interval for the estimate in **(c)**.

(e) Construct the 95% confidence interval for the slope of the true regression line β_1.

(f) Some Broadway plays use the rule of thumb that next week's gross box office results will be the same as this week's. Is this reasonable? (Justify/Refute using an appropriate hypothesis test.)

(g) If Y and X were reversed in the above regression, what would you expect R^2 to be? In a simple linear regression, R^2 is equal to the square of the correlation coefficient between X and Y. R^2 is discussed in class next week.