In the Name of God

Foundations of Data Science
Fall 2026 (1405-1)
Instructors: Dr. Babak Hossein Khalaj and
Dr. Amir Hossein Saberi
Department of Electrical Engineering



Homework 1

Probability, Statistics, and Regression

Due date: Sunday, Aban 3rd, 23:59

Instructions

- If you have any questions about this homework, you may contact the homework designer, Mohammad Ali Naderi.
- Please remember to write your name on your submission.
- Show your reasoning clearly. Unsupported final answers may receive little or no credit.
- You may discuss the problems with classmates, but you must write up the solutions individually, in your own words, and acknowledge your collaborators.
- Submit a single PDF before the due date. Late submissions follow the course policy.

1. (15 points) **The Cocountable σ-Algebra and Probability Space**

   Let $\Omega = \mathbb{R}$, and define the collection of subsets

   $$\Sigma = \{A \subseteq \Omega : A \text{ is countable or } A^c \text{ is countable}\},$$

   where $A^c = \Omega \setminus A$. Define a set function $P : \Sigma \to [0, 1]$ by

   $$P(A) = \begin{cases} 0, & \text{if } A \text{ is countable,} \\ 1, & \text{if } A^c \text{ is countable.} \end{cases}$$

   Show that (Ω, Σ, P) is a well-defined probability space. Specifically:

   (a) Verify that P is well-defined on Σ (i.e., that no set $A \in \Sigma$ satisfies both conditions).

   (b) Prove that Σ is a σ-algebra on Ω.

   (c) Prove that P satisfies the Kolmogorov axioms (non-negativity, unit measure, and countable additivity).

2. (10 points) **Measurability and Indicator Decomposition of Random Variables**

Let (Ω, Σ, P) be a probability space, and let $X, Y : \Omega \to \mathbb{R}$ be real-valued random variables (i.e., Borel measurable functions). For a fixed event $A \in \Sigma$, define the function $Z : \Omega \to \mathbb{R}$ by

$$Z(\omega) = \begin{cases} X(\omega), & \text{if } \omega \in A, \\ Y(\omega), & \text{if } \omega \in A^c. \end{cases}$$

Equivalently, $Z = X\mathbf{1}_A + Y\mathbf{1}_{A^c}$, where $\mathbf{1}_A$ denotes the indicator function of A. Show that Z is a well-defined random variable on (Ω, Σ, P).

Hint: Recall that a function $Z : \Omega \to \mathbb{R}$ is a random variable if for every Borel set $B \in \mathcal{B}(\mathbb{R})$ (or equivalently, for every ray $(-\infty, c]$ with $c \in \mathbb{R}$), the pre-image satisfies $Z^{-1}(B) \in \Sigma$.

3. (12 points) **Bias, Variance, and Model Complexity**

Assume that

$$Y = f(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathrm{Var}(\varepsilon) = \sigma^2.$$

Let $\hat{f}(x_0)$ denote an estimate of $f(x_0)$ obtained from a random training set.

(a) Starting from

$$\mathbb{E}\Big[(Y_0 - \hat{f}(x_0))^2\Big],$$

derive the bias–variance decomposition

$$\mathbb{E}\Big[(Y_0 - \hat{f}(x_0))^2\Big] = \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Bias}(\hat{f}(x_0))^2 + \sigma^2.$$

Clearly state the assumptions used to eliminate the cross terms.

(b) On a single graph, sketch the typical behavior of the following quantities as model flexibility increases: squared bias, variance, training MSE, test MSE, and irreducible error. Briefly justify the shape of each curve.

(c) At a particular point x_0, two learning methods have the following properties:

	$\mathrm{Bias}(\hat{f}(x_0))$	$\mathrm{Var}(\hat{f}(x_0))$
A	0.9	1.6
B	0.3	2.2

and $\sigma^2 = 0.8$.

Compute the expected test error of each method at x_0 and determine which method is preferable. If σ^2 were doubled, would your choice between the two methods change? Explain.

4. (12 points) **Statistical Inference and Prediction**

A simple linear regression is fitted using $n = 60$ observations. The estimated slope and its standard error are

$$\hat{\beta}_1 = 1.65, \qquad \mathrm{SE}(\hat{\beta}_1) = 0.60.$$

Assume the usual linear-regression assumptions hold. You may use

$$t_{0.975,58} = 2.002.$$

(a) Test

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_A : \beta_1 \neq 0$$

at significance level 5%. Compute the test statistic and state your conclusion.

(b) Construct a 95% confidence interval for β_1.

(c) Explain why the statement "β_1 is statistically significantly different from zero" does *not* necessarily imply that "X can accurately predict Y."

(d) At $x = x_0$, suppose that $\hat{y}_0 = 26$. The estimated standard error for estimating the *mean response* is 1.10, whereas the estimated standard error for predicting a *new observation* is 4.20. Construct the corresponding approximate 95% intervals and explain why they have different widths.

5. (13 points) **Properties of Least-Squares Regression**

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

The least-squares estimates minimize

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

(a) By differentiating the RSS, derive the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(b) Use your result to prove that the fitted regression line necessarily passes through $(\bar{x}, \bar{y})$.

(c) Let $e_i = y_i - \hat{y}_i$ be the residuals. Prove that

$$\sum_{i=1}^{n} e_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_i e_i = 0.$$

Interpret these two identities geometrically or statistically.

(d) Suppose the predictor is transformed according to

$$Z = 3X - 5.$$

A regression of Y on Z is now fitted. If the original fitted coefficients are $\hat{\beta}_0$ and $\hat{\beta}_1$, derive the intercept and slope of the new regression in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$. Show that the fitted values $\hat{Y}_i$ do not change.

6. (13 points) **Multiple Regression and Collinearity**

A data set contains $n = 120$ observations and two strongly correlated predictors X_1 and X_2. The model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

is fitted and yields

$$\hat{\beta}_1 = 2.40, \quad \mathrm{SE}(\hat{\beta}_1) = 1.35, \qquad \hat{\beta}_2 = 1.70, \quad \mathrm{SE}(\hat{\beta}_2) = 0.95,$$

with $TSS = 1800$ and $RSS = 1500$. You may use

$$t_{0.975,117} \approx 1.98, \qquad F_{0.95;2,117} \approx 3.07.$$

(a) Test separately $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ at significance level 5%.

(b) Compute the global F-statistic for $H_0 : \beta_1 = \beta_2 = 0$. What conclusion do you obtain?

(c) Your answers to the previous two parts may appear contradictory. Explain carefully why they are not.

(d) Explain how strong collinearity can affect (i) coefficient estimates, (ii) their standard errors, (iii) individual t-statistics, and (iv) predictive performance. Can a regression model still make useful predictions even if individual coefficients cannot be estimated precisely? Explain.

7. (10 points) **Interaction Terms and Interpretation**

Suppose starting salary Y, measured in thousands of dollars, is modeled by

$$\begin{aligned} Y = 48 &+ 18\,GPA + 0.08\,IQ + 32\,Female \\ &+ 0.015(GPA \times IQ) - 9(GPA \times Female) + \varepsilon, \end{aligned}$$

where

$$Female = \begin{cases} 1, & \text{female}, \\ 0, & \text{male}. \end{cases}$$

(a) Write the fitted regression model separately for males and females.

(b) For fixed GPA and IQ, derive

$$\mathbb{E}[Y \mid Female = 1] - \mathbb{E}[Y \mid Female = 0].$$

At what GPA do the fitted salaries for males and females coincide?

(c) Predict the salary of a female student with $GPA = 3.8$ and $IQ = 115$.

(d) A student states: "The coefficient 0.015 of the GPA–IQ interaction is extremely small, so the interaction must be unimportant." Explain why this conclusion is not valid.

(e) State the joint null hypothesis that should be tested if we want to determine whether gender affects either the baseline salary or the relationship between GPA and salary.

8. (15 points) **Optimism of the Training Error in Empirical Risk Minimization**

Consider the standard linear regression setting. Let $\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$ be a training dataset drawn i.i.d. from an unknown joint distribution $\mathcal{P}$ on $\mathbb{R}^p \times \mathbb{R}$. We fit a parameter vector $\beta \in \mathbb{R}^p$ via Ordinary Least Squares (OLS) by minimizing the empirical training risk:

$$\mathcal{R}_{\text{tr}}(\beta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2.$$

Let $\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \mathcal{R}_{\text{tr}}(\beta)$ denote an empirical risk minimizer computed on $\mathcal{D}_{\text{tr}}$. Now suppose we have an independent test dataset $\mathcal{D}_{\text{te}} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{M}$ drawn i.i.d. from the identical distribution $\mathcal{P}$, with empirical test risk defined as:

$$\mathcal{R}_{\text{te}}(\beta) = \frac{1}{M} \sum_{i=1}^{M} (\tilde{y}_i - \beta^T \tilde{x}_i)^2.$$

Prove that the expected training risk underestimates the expected test risk, i.e.,

$$\mathbb{E}_{\mathcal{D}_{\text{tr}}}\big[\mathcal{R}_{\text{tr}}(\hat{\beta})\big] \leq \mathbb{E}_{\mathcal{D}_{\text{tr}},\mathcal{D}_{\text{te}}}\big[\mathcal{R}_{\text{te}}(\hat{\beta})\big],$$

where the expectations are taken over the random sampling of $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$.