Home Assignment 1 Solutions

STAT 151A Linear Modelling: Theory and Applications

Ajay Sharma — Spring 2025

Problem 1: Matrix Basics

(a) *Computing Trace and Determinant*
Let $A = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix}$. The trace of A, $\text{tr}(A)$, is the sum of its diagonal entries

$$\text{tr}(A) = 2 + 4 = 6.$$

The determinant of A is given by

$$\det(A) = (2 \times 4) - (1 \times 3) = 8 - 3 = 5.$$

(b) *Computing Transpose and Inverse*
The transpose of A, $A^\top$, is

$$A^\top = \begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}.$$

The inverse of A, A^{-1}, using the formula $A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ is

$$A^{-1} = \frac{1}{5} \begin{pmatrix} 4 & -1 \\ -3 & 2 \end{pmatrix} = \begin{pmatrix} 4/5 & -1/5 \\ -3/5 & 2/5 \end{pmatrix}.$$

(c) *Computing Eigenvalue-Eigenvector Pairs*
Solving for eigenvalues, we find the characteristic polynomial

$$p(\lambda) = \det(A - \lambda I) = \lambda^2 - 6\lambda + 5.$$

Solving $p(\lambda) = 0$ gives eigenvalues $\lambda_1 = 5, \lambda_2 = 1$. Corresponding eigenvectors can be found by finding $\ker(A - \lambda_j I_2) := \{v \in \mathbb{R}^2 : (A - \lambda_j I_2)v = 0\}$, for $\lambda_{j \in [2]}$.
For $\lambda_1 = 5$, solve $(A - 5I)v = 0$

$$\begin{pmatrix} -3 & 1 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which simplifies to $-3x_1 + x_2 = 0 \Rightarrow x_1 = x_2/3$, where $x_2 \in \mathbb{R}$ is a free variable. Therefore, the eigenvector corresponding to $\lambda_1 = 5$ is given by

$$v_1 = \alpha \cdot \begin{pmatrix} 1/3 \\ 1 \end{pmatrix} \text{ for any } \alpha \in \mathbb{R}.$$

For $\lambda_2 = 1$, solve $(A - I)v = 0$:

$$\begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which simplifies to $x_1 + x_2 = 0 \Rightarrow x_1 = -x_2$, where $x_2 \in \mathbb{R}$ is a free variable. Therefore, the eigenvector corresponding to $\lambda_2 = 1$ is $v_2 = \beta \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ for any $\beta \in \mathbb{R}$.

(d) *Making a Conjecture*

Claim: The trace of any square matrix $A \in \mathbb{R}^{dxd}$ is computed as $\mathrm{tr}(A) = \sum_{j=1}^{d} \lambda_j$, where $\lambda_{j \in [d]}$ are eigenvalues of A.

Proof: We can write $\mathrm{tr}(A) = \sum_{j=1}^{d} e_j^{\mathrm{T}} A e_j$, where $e_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$. Additionally, by the Spectral Theorem we have $A = V \Lambda V^{\mathrm{T}}$, where V is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_d)$. Putting this together, we have

$$\begin{aligned} \mathrm{tr}(A) &= \sum_{j=1}^{d} e_j^{\mathrm{T}} (\lambda_j v_j v_j^{\mathrm{T}}) e_j \\ &= \sum_{j=1}^{d} \lambda_j e_j^{\mathrm{T}} v_j v_j^{\mathrm{T}} e_j \\ &= \sum_{j=1}^{d} \lambda_j e_j^{\mathrm{T}} v_j e_j^{\mathrm{T}} v_j \\ &= \sum_{j=1}^{d} \lambda_j \langle e_j, v_j \rangle^2 \\ &= \sum_{j=1}^{d} \lambda_j. \end{aligned}$$

Note that the last equality comes from the fact that e_j and v_j are orthogonal vectors and therefore $\langle e_j, v_j \rangle = 1$. This proves our claim. Now we'll verify by direct computation from parts (a) and (c) that $\mathrm{tr}(A) = 6 = \lambda_1 + \lambda_2$. Indeed this agrees!

(e) Let $B = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$. Now observe that $AB = \begin{pmatrix} 1 & 3 \\ -1 & 7 \end{pmatrix} \neq BA = \begin{pmatrix} 5 & 5 \\ 1 & 3 \end{pmatrix}$.

(f) Using the result from part (e), we have $\mathrm{tr}(AB) = \mathrm{tr}(BA) = 8$.

(g) From part (e) we have $\det(B) = 2, \det(AB) = \det(BA) = \det(A) \times \det(B) = 10$.

Problem 2: About $X^{\mathrm{T}}X$

(a) *Positive Semi-Definite Matrices*
Let $X \in \mathbb{R}^{n\times d}$. Then by Singular Value Decomposition we can write $X = U\Sigma V^{\mathrm{T}}$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices, $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_r, 0_{r+1}, ..., 0_{\min(n,d)})$. So, we have $X^{\mathrm{T}}X = V\Sigma^{\mathrm{T}}U^{\mathrm{T}}U\Sigma V^{\mathrm{T}} = V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}$. Then note $\Sigma^{\mathrm{T}}\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$. In order to show $X^{\mathrm{T}}X$ is PSD: for any $v \in \mathbb{R}^d$ we have $v^{\mathrm{T}}(X^{\mathrm{T}}X)v \geq 0$.

Consider $v^{\mathrm{T}}(X^{\mathrm{T}}X)v = v^{\mathrm{T}}(V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}})v$. Now let $w = V^{\mathrm{T}}v$. Then we obtain

$$v^{\mathrm{T}}(X^{\mathrm{T}}X)v = w^{\mathrm{T}}(\Sigma^{\mathrm{T}}\Sigma)w = ||\Sigma w||_2^2 \geq 0, \text{ as claimed.}$$

(b) *Showing* $\ker(X) \subseteq \ker(X^{\mathrm{T}}X)$
Take $v \in \ker(X) \iff \{v \in \mathbb{R}^d : Xv = 0\}$, by definition of kernel (null-space). In particular, we consider the equation $Xv = 0$ and left-multiply by X^{T}, we obtain $X^{\mathrm{T}}Xv = 0$. Therefore, we have shown that $v \in \ker(X^{\mathrm{T}}X)$ and indeed we have $\ker(X) \subseteq \ker(X^{\mathrm{T}}X)$, as claimed.

(c) *Showing* $\ker(X) \supseteq \ker(X^{\mathrm{T}}X)$
Take $v \in \ker(X^{\mathrm{T}}X) \iff \{v \in \mathbb{R}^d : X^{\mathrm{T}}Xv = 0\}$, by definition of the kernel. In particular, consider the equation $X^{\mathrm{T}}Xv = 0$ and left-multiply by v^{T} to obtain $v^{\mathrm{T}}X^{\mathrm{T}}Xv = 0$. Then set $w = Xv$. So we have $w^{\mathrm{T}}w = ||w||_2^2 = 0 \iff w = 0$. In other words, v satisfies $Xv = 0$ so $v \in \ker(X)$. Hence $\ker(X) \supseteq \ker(X^{\mathrm{T}}X)$.

(d) *Rank of Matrices*
Having shown from part (b) that $\ker(X) \subseteq \ker(X^{\mathrm{T}}X)$ and from part (c) that $\ker(X) \supseteq \ker(X^{\mathrm{T}}X)$, we conclude $\ker(X) = \ker(X^{\mathrm{T}}X)$.

Now suppose X has full-column rank. Then this means all solutions to $Xv = 0$ are linearly independent for any $v \in \mathbb{R}^d$. In particular $\ker(X) = \{0\} = \ker(X^{\mathrm{T}}X)$, using the previous result. Therefore $X^{\mathrm{T}}X \in \mathbb{R}^{d\times d}$ also has full-column rank. In other words, $\mathrm{rank}(X^{\mathrm{T}}X) = d$.

(e) *Positive Definite Matrices*
Ignore the assumption we made about X earlier. We would like to show that $M := X^{\mathrm{T}}X + \lambda I_d$ is always positive definite for $\lambda > 0$. In other words for any $v \in \mathbb{R}^d$, show that $v^{\mathrm{T}}Mv > 0$.

Consider $v^{\mathrm{T}}Mv = v^{\mathrm{T}}(X^{\mathrm{T}}X + \lambda I_d)v = v^{\mathrm{T}}X^{\mathrm{T}}Xv + \lambda v^{\mathrm{T}}v$. Recall that from part (a) we have shown that X^TX is PSD. Then observe $\lambda v^{\mathrm{T}}v = \lambda||v||_2^2 > 0$, assuming that $v \neq 0$. So adding a strictly positive quantity to a non-negative quantity will also be strictly positive and therefore M is positive definite as claimed.

(f) *Conclusion*
Claim: M as defined above is invertible because it is a positive-definite matrix.
Proof: To show that M is invertible, we'll make use of the Fundamental Theorem of Linear Maps (FTL), which states for any square matrix A

$$\mathrm{rank}(A) + \ker(A) = n.$$

Given $M \in \mathbb{R}^{d\times d}$, we need to show that $\ker(M) = 0$, which means the only solution to $Mv = 0$ is $v = 0$. In other words, consider

$$(X^T X + \lambda I)v = 0 \Rightarrow X^T Xv + \lambda Iv = 0 \Rightarrow X^T Xv = -\lambda Iv.$$

Left multiplication of both sides by v^T gives

$$v^T X^T Xv = -\lambda v^T Iv \Rightarrow v^T X^T Xv = -\lambda \|v\|_2^2.$$

Since $X^T X$ is PSD, $v^T X^T Xv \geq 0$. Given $\|v\|_2^2 \geq 0$ and $\lambda > 0$, we have

$$-\lambda \|v\|_2^2 \leq 0 \Rightarrow \|v\|_2^2 = 0 \Rightarrow v = 0.$$

Therefore $\ker(M) = 0$. By FTL we have $\text{rank}(M) + \ker(M) = d \Rightarrow \text{rank}(M) = d$, which means M has full-column rank and is invertible.

Before we proceed with the remaining exercises, consider the following fact.

Fact: For a matrix $X \in \mathbb{R}^{n\times d}$ representing the predictor variables, and a vector $y \in \mathbb{R}^n$ representing the response variable, the least-squares regression problem can be formulated

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^d}{\arg\min}\ \|y - X\beta\|_2^2 = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

Proof: Define a function $\mathcal{L}(\beta) = \|y - X\beta\|_2^2 = (y - X\beta)^{\mathrm{T}}(y - X\beta)$. Expanding,

$$\mathcal{L}(\beta) = y^{\mathrm{T}}y - 2\beta^{\mathrm{T}}X^{\mathrm{T}}y + \beta^{\mathrm{T}}X^{\mathrm{T}}X\beta.$$

To find the minimum, we take the gradient of $\mathcal{L}(\beta)$ and set it to zero

$$\nabla_\beta \mathcal{L} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\beta = 0.$$

Solving for β, we obtain $X^{\mathrm{T}}X\beta = X^{\mathrm{T}}y$. If $X^{\mathrm{T}}X$ is invertible, the closed form solution is

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

Problem 3: Prediction $\hat{\beta}_0 + \hat{\beta}_1 a$

(a) *A Simple Linear Regression Model*
Using the fact above, we can express

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

where X has columns of ones and the independent variables $x_1, \ldots, x_n$. The matrix $X^T X$ and vector $X^T y$ are given by

$$X^T X = \begin{pmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{pmatrix}, \quad X^T y = \begin{pmatrix} \sum y_j \\ \sum x_j y_j \end{pmatrix}.$$

Then the inverse of $X^T X$ is

$$(X^T X)^{-1} = \frac{1}{n \sum x_j^2 - (\sum x_j)^2} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix}.$$

Multiplying by $X^T y$, we get

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^T X)^{-1} X^T y = \frac{1}{n \sum x_j^2 - (\sum x_j)^2} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix} \begin{pmatrix} \sum y_j \\ \sum x_j y_j \end{pmatrix}.$$

Simplifying this expression yields the standard formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$, where

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum (x_j - \bar{x})(y_j - \bar{y})}{\sum (x_j - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}.$$

We also know the relation $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Therefore we can write

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 a &= \bar{y} + \hat{\beta}_1 (a - \bar{x}) \\ &= \frac{1}{n} \sum y_j + (a - \bar{x}) \left(\frac{\sum (x_j - \bar{x})(y_j - \frac{1}{n} \sum y_i)}{\sum (x_j - \bar{x})^2} \right) \\ &= \frac{1}{n} \sum y_j + (a - \bar{x}) \left(\frac{\sum (x_j - \bar{x}) y_j - \frac{1}{n} \sum (x_j - \bar{x}) \sum y_i)}{\sum (x_j - \bar{x})^2} \right) \\ &= \sum_{j=1}^{n} y_j \left\{ \frac{1}{n} + \frac{(a - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right\}. \end{aligned}$$

(b) *Computing the Variance*

We start with the model $y_j = \beta_0 + \beta_1 x_j + \epsilon_j$, where ϵ_j are independent errors with $\mathbb{E}[\epsilon_j] = 0$ and $\mathrm{Var}(\epsilon_j) = \sigma^2$. We also know that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} + \bar{\epsilon}$ and from part (a) the estimator $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{\sum(x_j - \bar{x})(y_j - \bar{y})}{\sum(x_j - \bar{x})^2}.$$

As a side calculation: let's expand the numerator $y_j - \bar{y}$ and observe

$$y_j - \bar{y} = (\beta_0 + \beta_1 x_j + \epsilon_j) - (\beta_0 + \beta_1 \bar{x} + \bar{\epsilon}) = (x_j - \bar{x})\beta_1 + (\epsilon_j - \bar{\epsilon}).$$

Thus, the expression becomes

$$\sum(x_j - \bar{x})(y_j - \bar{y}) = \sum(x_j - \bar{x})((x_j - \bar{x})\beta_1 + (\epsilon_j - \bar{\epsilon})) = \sum(x_j - \bar{x})^2\beta_1 + (x_j - \bar{x}_j)\epsilon_j.$$

Hence the only term that contributes to the variance of $\hat{\beta}_1$ involves ϵ_j, which is random (or has a noise). Furthermore, since ϵ_j are independent of x_j and have zero mean, we can now compute the variance of $\hat{\beta}_1$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\mathrm{Var}(\sum(x_j - \bar{x})\epsilon_j)}{\left(\sum(x_j - \bar{x})^2\right)^2} = \frac{\sigma^2 \sum(x_j - \bar{x})^2}{\left(\sum(x_j - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum(x_j - \bar{x})^2}.$$

Thus, the variance of the prediction $\hat{\beta}_0 + \hat{\beta}_1 a$ is

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 a | x_1, \ldots, x_n) &= \mathrm{Var}(\bar{y} + (a - \bar{x})\hat{\beta}_1 | x_1, \ldots, x_n) \\
&= \frac{\sigma^2}{n} + (a - \bar{x})^2 \mathrm{Var}(\hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2 (a - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

(c) *Finding the Local Minimizer*

We want to show $\bar{x} = \arg\min_{a \in \mathbb{R}} \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 a | x_1, \ldots, x_n)$. Then define a function $\mathcal{L}(a) = \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 a | x_1, \ldots, x_n)$. Using the expression from part (b), we find that

$$\frac{\partial \mathcal{L}}{\partial a} = \frac{2\sigma^2 (a - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$$

gives $a = \bar{x}$, as claimed. Note that $\sigma^2 > 0$, so we don't run into any issues. Now we'll verify using the second derivative test that our choice of a is indeed a minimum. So

$$\frac{\partial^2 \mathcal{L}}{\partial a^2} = \frac{2\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} > 0.$$

Indeed $\mathcal{L}(a)$ attains a minimum at $a = \bar{x}$, so we're done.

Problem 4: Linear Regression without Intercept
For the purposes of these exercises consider $\{(x_j, y_j)_{j\in[n]}\}$ and suppose $\bar{y} = 0$ when $x = 0$. Finally define a model by $\mathbb{E}(y|x) = \beta_1 x$.

(a) *Least Squares Estimate*
We'll proceed by writing $X \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ in matrix notation.

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Then using the fact above, the least squares estimate is given by the formula

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

For the simple regression case (without intercept) $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}y$ are defined as

$$X^{\mathrm{T}}X = \sum_{i=1}^{n} x_i^2, \quad X^{\mathrm{T}}y = \sum_{i=1}^{n} x_i y_i.$$

Therefore we have,

$$\hat{\beta}_1 = \frac{X^{\mathrm{T}}y}{X^{\mathrm{T}}X} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(b) *Unbiased Estimator*
From part (a) we know the least squares estimator is given by

$$\hat{\beta}_1 = \frac{X^{\mathrm{T}}y}{X^{\mathrm{T}}X}.$$

Then recall this estimator is unbiased if $\mathbb{E}(\hat{\beta}_1) - \hat{\beta} = 0$. Computing this quantity and substituting $y = X\beta_1 + \epsilon$ gives

$$\mathbb{E}(\hat{\beta}_1) - \beta_1 = \mathbb{E}\left(\frac{X^{\mathrm{T}}(X\beta_1 + \epsilon)}{X^{\mathrm{T}}X}\right) - \beta_1 = \frac{X^{\mathrm{T}}X\beta_1 + X^{\mathrm{T}}\mathbb{E}(\epsilon)}{X^{\mathrm{T}}X} - \beta_1.$$

Since $\mathbb{E}(\epsilon) = 0$, we obtain

$$\mathbb{E}(\hat{\beta}_1) - \beta_1 = \frac{X^{\mathrm{T}}X\beta_1}{X^{\mathrm{T}}X} - \beta_1 = 0.$$

Thus, $\hat{\beta}_1$ is an unbiased estimator of β_1.

(c) *Computing the Variance*

Using the expression obtained in part (a) and substituting $y = X\beta_1 + \epsilon$ gives

$$\hat{\beta}_1 = \frac{X^{\mathrm{T}}y}{X^{\mathrm{T}}X} = \frac{X^{\mathrm{T}}(X\beta_1 + \epsilon)}{X^{\mathrm{T}}X}.$$

Then rewriting the expression and taking the variance

$$\hat{\beta}_1 = \beta_1 + \frac{X^{\mathrm{T}}\epsilon}{X^{\mathrm{T}}X} \Rightarrow \mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{X^{\mathrm{T}}\epsilon}{X^{\mathrm{T}}X}\right).$$

Since $\mathrm{Var}(\epsilon) = \sigma^2 I$, we have

$$\mathrm{Var}(X^{\mathrm{T}}\epsilon) = X^{\mathrm{T}}\mathrm{Var}(\epsilon)X = \sigma^2 X^{\mathrm{T}}X.$$

Therefore

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2 X^{\mathrm{T}}X}{(X^{\mathrm{T}}X)^2} = \frac{\sigma^2}{X^{\mathrm{T}}X}.$$