Home Assignment 2 Solutions

STAT 151A Linear Modelling: Theory and Applications

Ajay Sharma — Spring 2025

The following fact will prove useful in the mathematical exercises.

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 1: Fitting y on x versus fitting x on y

(a) *Least Squares Estimates for y on x*
The least squares estimates $(\hat{\beta}_0, \hat{\beta}_1)$ in the model $\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$ are given by

$$\begin{cases} \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}.$$

(b) *Least Squares Estimates for x on y*
The least squares estimates $(\hat{\alpha}_0, \hat{\alpha}_1)$ in the model $\mathbb{E}(x_i|y_i) = \alpha_0 + \alpha_1 y_i$ are given by

$$\begin{cases} \hat{\alpha_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} \\ \hat{\alpha_0} = \bar{x} - \hat{\alpha}_1 \bar{y} \end{cases}.$$

(c) *Bound on $\hat{\alpha}_1\hat{\beta}_1$*
Intuition should suggest that $\hat{\alpha_1} = 1/\hat{\beta}_1 \Leftrightarrow \hat{\alpha}_1\hat{\beta}_1 = 1$. Instead, we'll consider the product

$$\hat{\alpha}_1\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} \cdot \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\mathrm{Cov}^2(x, y)}{\mathrm{Var}(x)\mathrm{Var}(y)} = r^2.$$

In particular, this simplifies to the correlation of determination r^2, where r is the Pearson correlation coefficient. It can be shown using the Schwarz Inequality that $-1 \leq r \leq 1$ and so $0 \leq r^2 \leq 1$. Hence $\hat{\alpha}_1\hat{\beta}_1 \leq 1$ and is only equal to 1 when there is perfect correlation i.e., $r = \pm 1$. So in general, our intuition is wrong.

Problem 2: Weighing Strategies with Least Squares

(a) *Least Squares Estimation*
Suppose $n = 8$. Then the design matrix X is given by

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

The least squares estimate for $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$. Also that in this case $X^{\mathrm{T}}X \in \mathbb{R}^{4\times 4}$ and $X^{\mathrm{T}}y \in \mathbb{R}^4$. By computation (plugging into WolframAlpha), we get

$$X^T X = \begin{pmatrix} 8 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix} = 8I_4.$$

Then its inverse is given by

$$(X^T X)^{-1} = \frac{1}{8} I_4.$$

The computation of $X^{\mathrm{T}}y$ is simple. In particular, we get

$$X^{\mathrm{T}}Y = \begin{pmatrix} \sum y_i \\ \sum (-1)^{j+1} y_i \\ \sum (-1)^{\lfloor \frac{j-1}{2} \rfloor} y_i \\ \sum (-1)^{[(j \bmod 2)+(j \bmod 4)]} y_i \end{pmatrix}$$

Thus, the least squares estimate is $\hat{\boldsymbol{\beta}} = \frac{1}{8} X^{\mathrm{T}}Y$.

(b) *Covariance Matrix of* $\hat{\beta}$
We'll use the property that $\mathrm{Var}(Az) = A\mathrm{Var}(z)A^{\mathrm{T}}$ as we proceed with the problem.

First substitute $y = X\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Then using our expression from part (a), we obtain

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}\left(\frac{1}{8} X^{\mathrm{T}}(X\beta + \epsilon)\right) = \frac{1}{64}\mathrm{Var}(X^{\mathrm{T}}X\beta + X^{\mathrm{T}}\epsilon).$$

Note that since $X^{\mathrm{T}}X\beta$ is deterministic, then $\mathrm{Var}(X^{\mathrm{T}}X\beta) = 0$. And using the property mentioned above we obtain $\mathrm{Var}(X^{\mathrm{T}}\epsilon) = X^{\mathrm{T}}\mathrm{Var}(\epsilon)X = \sigma^2 X^{\mathrm{T}}X = \frac{\sigma^2}{8} I_4$. Therefore, our final expression is given by

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{8} I_4.$$

This result is interesting compared to the naive weighing schemed with 32 weighing because we're only using 8 weighings instead of 32 and yet achieve the same variance in this estimation.

Problem 3: Weighted Least Squares with Heteroscedastic Gaussian Noise

(a) *Deriving the WLS Estimator*
Write $y_j = x_j^{\mathrm{T}}\beta + \epsilon_j$, where $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ and $j \in [n]$. Then consider the function

$$\mathcal{L}(\beta) := \sum_{j=1}^{n} \frac{(y_j - x_j^T\beta)^2}{\sigma_j^2}.$$

We want to show that $\hat{\beta_{\mathrm{WLS}}} = \underset{\beta\in\mathbb{R}^n}{\arg\min}\, \mathcal{L}(\beta) = \underset{\beta\in\mathbb{R}^n}{\arg\min}\, \|y - X\beta\|_W^2 = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy$,

$$W = \mathrm{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \ldots, \frac{1}{\sigma_n^2}\right).$$

We can expand the objective function as $\mathcal{L}(\beta) = (y - X\beta)^T W(y - X\beta)$ and take the gradient with respect to β. Then set it to zero to find the minimizer i.e.

$$\nabla_\beta \mathcal{L}(\beta) = \nabla_\beta(y^{\mathrm{T}}Wy - 2y^{\mathrm{T}}WX\beta + (X\beta)^T W(X\beta)) = 2X^{\mathrm{T}}WX\beta - 2y^{\mathrm{T}}WX = 0.$$

Solving for β gives the unique minimizer

$$\hat{\beta}_{\mathrm{WLS}} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy.$$

(b) *Showing $\hat{\beta_{\mathrm{WLS}}}$ is an Unbiased Estimator*
We say $\hat{\beta_{\mathrm{WLS}}}$ is an unbiased estimator of β iff $\mathbb{E}(\hat{\beta_{\mathrm{WLS}}}) - \beta = 0$. In other words, we want to compute this difference by substituting $y = X\beta + \epsilon$. This gives

$$\mathbb{E}[\hat{\beta}_{\mathrm{WLS}}] - \beta = \mathbb{E}[(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W(X\beta + \epsilon)] - \beta.$$

Simplifying this expression

$$\mathbb{E}[\hat{\beta}_{\mathrm{WLS}}] - \beta = \mathbb{E}((X^T WX)^{-1}X^T WX\beta) - \mathbb{E}((X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\epsilon) - \beta = I_n\beta - \beta = 0.$$

Since β is deterministic, $\mathbb{E}(\beta) = \beta$, which yields the above expression for the WLS estimator. And so we conclude $\hat{\beta}_{\mathrm{WLS}}$ is an unbiased estimator of β.

(c) *Variance-Covariance Matrix of $\hat{\beta}_{\mathrm{WLS}}$*
Using our expression from part (b) and substituting $y = X\beta + \epsilon$ we obtain

$$\mathrm{Var}(\hat{\beta}_{\mathrm{WLS}}) = \mathrm{Var}(\beta + (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\epsilon).$$

Using the property from 2(b) we have

$$\mathrm{Var}(\hat{\beta}_{\mathrm{WLS}}) = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\mathrm{Var}(\epsilon)WX(X^{\mathrm{T}}WX)^{-1}.$$

As a side calculation note that since $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $\mathrm{Var}(\epsilon_j) = \mathbb{E}(\epsilon_j^2) - [\mathbb{E}(\epsilon_j)]^2$. Therefore $\mathbb{E}(\epsilon_j^2) = \sigma_j^2$ and so $\mathrm{Var}(\epsilon) = \mathbb{E}(\epsilon\epsilon^{\mathrm{T}}) = \mathrm{diag}(\sigma_1^2, ..., \sigma_n^2) = \Lambda$. Going back to our original expression for $\mathrm{Var}(\hat{\beta}_{\mathrm{WLS}})$ we have that

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_{\mathrm{WLS}}) &= (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\Lambda WX(X^{\mathrm{T}}WX)^{-1} \\
&= (X^{\mathrm{T}}WX)^{-1}(X^{\mathrm{T}}WX)(X^{\mathrm{T}}WX)^{-1} \\
&= (X^{\mathrm{T}}WX)^{-1}.
\end{aligned}$$

(d) *Comparison to OLS Estimator*
Using the result we established about OLS estimators we know that

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

Then substituting $y = X\beta + \epsilon$ and computing the variance we can write

$$\begin{aligned}\text{Var}(\hat{\beta}_{\text{OLS}}) &= \text{Var}(\beta + (X^T X)^{-1}X^{\text{T}}\epsilon) \\ &= (X^{\text{T}}X)^{-1}X^{\text{T}}\text{Var}(\epsilon)X(X^{\text{T}}X)^{-1} \\ &= \sigma^2(X^{\text{T}}X)^{-1}.\end{aligned}$$

Since $\text{Var}(\epsilon) = \sigma^2 I_n = \Sigma$. Note if the errors are homoscedastic ($\sigma_i^2 = \sigma_j^2, \forall\, i, j \in [n]$), then the OLS is the best linear unbiased estimator BLUE. However, if the errors are heteroscedastic, then the WLS estimator is better (by the Gauss-Markov theorem) with $\text{Var}(\hat{\beta}_{\text{WLS}}) \leq \text{Var}(\hat{\beta}_{\text{OLS}})$.

Now consider the observation with $W = \text{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \ldots, \frac{1}{\sigma_n^2}\right)$. If σ_j^2 is large, then $w_j = 1/\sigma_j^2$ is small and if σ_j^2 is small, then w_j is large. In other words, WLS gives more weight for observations with low variance and less weight to observations with high variance, which has a balancing effect.

Problem 4: Frisch-Waugh-Lovell (FWL) Theorem

(a) *Verifying the Identity*
We know the inverse of the block matrix is given by

$$(X^{\text{T}}X)^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

with the components

$$\begin{cases} S_{11} = (X_1^{\text{T}}X_1)^{-1} + (X_1^{\text{T}}X_1)^{-1}X_1^{\text{T}}X_2(\tilde{X}_2^{\text{T}}\tilde{X}_2)^{-1}X_2^{\text{T}}X_1(X_1^{\text{T}}X_1)^{-1} \\ S_{21} = -(\tilde{X}_2^{\text{T}}\tilde{X}_2)^{-1}X_2^{\text{T}}X_1(X_1^{\text{T}}X_1)^{-1} \end{cases}.$$

Then the first block in the matrix multiplication of

$$(X^{\text{T}}X) \cdot \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} X_1^{\text{T}}X_1 & X_1^{\text{T}}X_2 \\ X_2^{\text{T}}X_1 & X_2^{\text{T}}X_2 \end{pmatrix} \cdot \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

In particular, we have

$$X_1^{\text{T}}X_1S_{11} + X_1^{\text{T}}X_2S_{21} = I_{p_1} \quad \text{as claimed.}$$

This is because

$$\begin{aligned}X_1^{\text{T}}X_1S_{11} &= X_1^{\text{T}}X_1((X_1^{\text{T}}X)^{-1} + ((X_1^{\text{T}}X)^{-1})X_1^{\text{T}}X_2(\tilde{X}_2^{\text{T}}\tilde{X}_2)^{-1}X_2^{\text{T}}X_1(X_1^{\text{T}}X_1)^{-1}) \\ &= I_{p_1} + X_1^{\text{T}}X_2(\tilde{X}_2^{\text{T}}\tilde{X}_2)^{-1}X_2^{\text{T}}X_1(X_1^{\text{T}}X_1)^{-1}\end{aligned},$$

and

$$\begin{aligned}X_1^{\text{T}}X_2S_{22} &= X_1^{\text{T}}X_2(-(\tilde{X_2^{\text{T}}}\tilde{X}_2)^{-1}X_2^{\text{T}}X_1(X_1^{\text{T}}X_1)^{-1}) \\ &= -X_1^{\text{T}}X_2(\tilde{X_2^{\text{T}}}\tilde{X}_2)^{-1}X_2^{\text{T}}X_1(X_1^{\text{T}}X_1)^{-1}\end{aligned}.$$

Combining these expressions the second term of $X_1^{\text{T}}X_1S_{11}$ cancels with $X_1^{\text{T}}X_2S_{22}$ and we are left with I_{p_1}.

(b) *Writing $\hat{\beta}_2$ in Terms of $S_{21}, S_{22}, X_1, X_2, y$*
From the OLS solution we can write $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$. Then we have

$$\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}S_{11} & S_{12} \\ S_{21} & S_{22}\end{pmatrix} \cdot \begin{pmatrix}{X_1}^{\mathrm{T}}y \\ {X_2}^{\mathrm{T}}y\end{pmatrix}$$

So we obtain $\hat{\beta}_2 = S_{21}X_1^{\mathrm{T}}y + S_{22}X_2^{\mathrm{T}}y$.

(c) *Writing $\hat{\beta}_2$ Using Projection Matrix H_1*
Let $H_1 := X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}$ and let

$$\begin{cases} S_{21} = -(\tilde{X_2^{\mathrm{T}}}\tilde{X_2})^{-1}X_2^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1} \\ S_{22} = (\tilde{X_2^{\mathrm{T}}}\tilde{X_2})^{-1} \end{cases}$$

Then we can substitute the known quantities S_{21} and S_{22} in $\hat{\beta}_2$. In other words,

$$\begin{aligned} \hat{\beta}_2 &= (-(\tilde{X_2^{\mathrm{T}}}\tilde{X_2})^{-1}X_2^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1})X_1^{\mathrm{T}}y + (\tilde{X_2^{\mathrm{T}}}\tilde{X_2})X_2^{\mathrm{T}}y \\ &= (\tilde{X_2^{\mathrm{T}}}\tilde{X_2})X_2^{\mathrm{T}}(y - X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}y) \\ &= (\tilde{X_2^{\mathrm{T}}}\tilde{X_2})X_2^{\mathrm{T}}(I - H_1)y. \end{aligned}$$

(d) *Writing $\hat{\beta}_2$ in Terms of $\tilde{X_2}$ and $\tilde{\epsilon_1}$*
We should note that $1 - H_1$ is a projection matrix that projects vectors onto the orthogonal complement of $\mathrm{col}(X_1)$. Then define $\tilde{X_2} = (1 - H_1)X_2$, which is the part of X_2 that is uncorrelated with X_1 and similarly define $\tilde{\epsilon_1} = (1 - H_1)y$ as the residual that is the part of y unexplained by y.

Note that $1 - H_1$ is a projection matrix it is idempotent and symmetric. So we have $(1 - H_1)^2 = 1 - H_1$ and $(1 - H_1)^{\mathrm{T}} = 1 - H_1$. Now consider the quantity $\tilde{X_2^{\mathrm{T}}} = X_2^{\mathrm{T}}(1 - H_1)$. Let's right multiply it by $\tilde{\epsilon_1}$ i.e.,

$$\tilde{X_2^{\mathrm{T}}}\tilde{\epsilon_1} = X_2^{\mathrm{T}}(1 - H_1)(1 - H_1)y = X_2^{\mathrm{T}}(1 - H_1)y.$$

In other words we can write $\tilde{\beta}_2 = (\tilde{X_2^{\mathrm{T}}}\tilde{X_2})^{-1}\tilde{X_2^{\mathrm{T}}}\tilde{\epsilon}_1$, as claimed.

(e) *Showing that $\tilde{\beta}_2 = \hat{\beta}_2$*
This is basically proved in part (d) but we'll restate our results anyway. In particular using our expression from part (c) we can write

$$\hat{\beta}_2 = (\tilde{X_2^{\mathrm{T}}}\tilde{X_2})^{-1}X_2^{\mathrm{T}}(1 - H_1)y = (\tilde{X_2^{\mathrm{T}}}\tilde{X_2})^{-1}\tilde{X_2^{\mathrm{T}}}\tilde{\epsilon_1} = \tilde{\beta}_2.$$

This concludes the proof of the Frisch-Waugh-Lovell (FWL) Theorem.