Mathematical Statistics Recitation 12

TA: Sonia Reilly

April 24, 2026

Lecture Review

Lecture 22

Recall:

$$y = X\beta + e, \qquad \mathbb{E}[e] = 0, \quad \Sigma_{ee} = \sigma^2 I$$
$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y$$

Expectation and covariance of $\hat{\beta}$:

$$\mathbb{E}[\hat{\beta}] = \beta, \quad \Sigma_{\hat{\beta}\hat{\beta}} = \sigma^2 (X^T X)^{-1}$$

Unbiased estimate of σ^2:

$$s^2 = \frac{\|y - X\hat{\beta}\|^2}{n - p}$$

Estimated standard error:

$$\hat{s}_{\beta_i} = s\sqrt{(X^T X)^{-1}_{ii}}$$

For CI's and hypothesis tests, use

$$\frac{\hat{\beta}_i - \beta_i}{\hat{s}_{\beta_i}} \xrightarrow[n\to\infty]{} N(0,1) \quad \text{or if } e \text{ normal,} \quad \frac{\hat{\beta}_i - \beta_i}{\hat{s}_{\beta_i}} \sim t_{n-p}$$

If x_i random, $\hat{\beta}$ is still unbiased but $\Sigma_{\hat{\beta}\hat{\beta}} = \sigma^2 \mathbb{E}[(X^T X)^{-1}]$.

Lecture 23

Heteroskedasticity "Different variance", $\mathrm{Cov}(e) = \Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.
$\hat{\beta} = (X^T X)^{-1} X^T y$ still unbiased, but

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

Standard error is still $\sqrt{\mathrm{Cov}(\hat{\beta})_{ii}}$, but now instead of estimating σ^2 overall by s^2, estimate each σ_i^2 by, e.g., $(y - X\hat{\beta})_i^2$ to construct approx $\hat{\Omega}$ and plug in.

If Ω known (rare in practice): Weighted Least Squares, minimizes variance by weighting less noisy data points more highly.

$$\hat{\beta}^{WLS} = \arg\min_{\beta \in \mathbb{R}^p} \sum_i \frac{1}{\sigma_i^2} (y - X\beta)_i^2 = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

Regularization If $n < p$ (fewer equations than unknowns), ordinary least squares is not unique.

ℓ_2 regularization / ridge regression: encourages small norm of β:

$$\hat{\beta} = \arg\min_{\beta} \ \|y - X\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2 = (X^T X + \lambda I)^{-1} X^T y.$$

$X^T X + \lambda I$ is always invertible. $\hat{\beta}$ is biased (increases with λ), but variance is lower (decreases with λ). May have lower mean squared error

$$\mathbb{E}\|\hat{\beta} - \beta\|^2 = \|\mathbb{E}\hat{\beta} - \beta\|^2 + \mathbb{E}\|\hat{\beta} - \mathbb{E}\hat{\beta}]\|^2 = \text{bias}^2 + \text{variance}$$

than we could achieve with an unbiased estimator, so also used in some $n > p$ problems.

ℓ_1 regularization / LASSO:

$$\hat{\beta} = \arg\min_{\beta} \ \|y - X\beta\|_2^2 + \lambda \sum_{i=0}^{p-1} |\beta_i|$$

No closed expression for $\hat{\beta}$, but typically sparse (few non-zero entries). Used to choose which features are important and which aren't.

Cross-validation: to choose best model, e.g., which features to include, or best λ,

- split data randomly into test/validation and training data
- compute $\hat{\beta}$ for each model using training data
- pick model with lowest error on test data

Problems

1. The time it takes to apply a certain algorithm to a problem of size n is quadratic in n, but the coefficients of the formula for the time are unknown. Four tests were run with different n's and the results were as follows.

n	time
1	3
2	6
3	15
4	22

The time measurements have some measurement error which you can assume to have uniform variance. If you build your X matrix correctly, you may take the following as given:

$$(X^T X)^{-1} = \begin{bmatrix} \frac{31}{4} & -\frac{27}{4} & \frac{5}{4} \\ -\frac{27}{4} & \frac{129}{20} & -\frac{5}{4} \\ \frac{5}{4} & -\frac{5}{4} & \frac{1}{4} \end{bmatrix} \quad \text{and} \quad (X^T X)^{-1} X^T = \begin{bmatrix} \frac{9}{4} & -\frac{3}{4} & -\frac{5}{4} & \frac{3}{4} \\ -\frac{31}{20} & \frac{23}{20} & \frac{27}{20} & -\frac{19}{20} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

(a) Write down an equation for the time T in terms of n. Set up a least squares problem in matrix form to find the coefficients of the equation for the time.

(b) Compute the coefficients (you may want to use a calculator/phone).

(c) What is the estimated standard error of each coefficient?

(d) Think about which coefficients have a smaller standard error. Is the order you observe intuitive?

(e) If we had the same data, but tried to fit it to a cubic model, how big would our residuals be? Does this mean that even if we know our data is quadratic, it's better to fit it using a cubic model?

(f) If we were given more data points, but not told whether the data was quadratic or cubic, how would we determine which model to use?

Solution:

(a) Model:

$$T_i = a_0 + a_1 n_i + a_2 n_i^2 + e_i, \qquad e \sim \mathcal{N}(0, \sigma^2 I),$$

so $T = Xa + e$, with

$$T = \begin{bmatrix} 3 \\ 6 \\ 15 \\ 22 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \qquad a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

$$\hat{a} = \arg\min_{a \in \mathbb{R}^3} \|T - Xa\|_2^2$$

(b)

$$\hat{a} = (X^\top X)^{-1} X^\top T = \begin{bmatrix} \frac{9}{4} & -\frac{3}{4} & -\frac{5}{4} & \frac{3}{4} \\ -\frac{31}{20} & \frac{23}{20} & \frac{27}{20} & -\frac{19}{20} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 3 \\ 6 \\ 15 \\ 22 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{8}{5} \\ 1 \end{bmatrix}$$

(c)

$$s_{\hat{a}_i} = s\sqrt{(X^\top X)^{-1}_{ii}} \qquad \text{where} \qquad s^2 = \frac{\|T - X\hat{a}\|^2}{n - p}$$

$$e = T - X\hat{a} = \begin{bmatrix} 3 \\ 6 \\ 15 \\ 22 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{8}{5} \\ 1 \end{bmatrix} = \begin{bmatrix} 3 - \frac{8}{5} - 1 \\ 6 - \frac{16}{5} - 4 \\ 15 - \frac{24}{5} - 9 \\ 22 - \frac{32}{5} - 16 \end{bmatrix} = \begin{bmatrix} 2/5 \\ -6/5 \\ 6/5 \\ -2/5 \end{bmatrix}$$

$$s^2 = \frac{\|T - X\hat{a}\|^2}{n - p} = \left(\frac{2}{5}\right)^2 + \left(\frac{6}{5}\right)^2 + \left(\frac{6}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{80}{25} = 3.2$$

Reading off the diagonal elements of $(X^\top X)^{-1}$,

$$s_{\hat{a}_0} = \sqrt{3.2}\sqrt{7.75} \approx 5$$

$$s_{\hat{a}_1} = \sqrt{3.2}\sqrt{6.45} \approx 4.5$$

$$s_{\hat{a}_2} = \sqrt{3.2}\sqrt{0.25} \approx 0.9$$

(d) Yes – the quadratic coefficient has lowest standard error because the quadratic term has the strongest effect on the data – changes in that coefficient have big effects on the data, so it is the easiest to recover accurately.

(e) If we use a cubic model, we add a linearly independent column to X, so given 4 data points, the system

$$T = Xa$$

has a unique exact solution, which is the same as the least squares solution. Therefore the residuals $e = T - Xa$ are 0. Even though it fits our data better, we should not use this model if we know our data is quadratic – this is called overfitting the noise and gives you incorrect estimates for your true coefficients.

(f) Cross-validation or LASSO regularization.

2. Consider the standard linear model $y = X\beta + e$, where $e \sim N(0, \sigma^2 I)$. Suppose we assign a Gaussian prior to the parameters, $\beta \sim N(0, \tau^2 I)$. Use the fact that a Gaussian $N(\mu, \Sigma)$ in higher dimensions has pdf

$$f(x) \propto \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

(a) Write down the likelihood function $f_{\text{like}}(y|\beta)$, i.e., the pdf of the data y, up to a proportionality constant. Do the same for the prior.

(b) Use Bayes' rule to show that the posterior mode (argmax of the posterior $f_{\text{post}}(\beta|y)$) is the same as the least squares estimator with ridge regression, for a particular λ.

(c) Explain the expression you get for λ. When is it high? When is it low? What does it say about our confidence in the data vs. the prior?

(d) Reverse the calculation you did above to find which prior is equivalent to introducing LASSO regression (up to a proportionality constant). Roughly sketch the prior in 1 dimension.

Solution:

(a) Since e is random but $X\beta$ is not, the likelihood $f_{\text{like}}(y|\beta)$ is the density of a Gaussian $N(X\beta, \sigma^2 I)$. Up to a proportionality constant:

$$f_{\text{like}}(y|\beta) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

For the parameters $\beta \sim N(0, \tau^2 I)$, the prior density is:

$$f_{\text{prior}}(\beta) \propto \exp\left(-\frac{1}{2\tau^2}\beta^T\beta\right)$$

(b) Using Bayes' rule, the posterior density is $f_{\text{post}}(\beta|y) \propto f_{\text{like}}(y|\beta) f_{\text{prior}}(\beta)$. To find the posterior mode we maximize the product or, equivalently, minimize the negative log-posterior:

$$-\log f_{\text{post}}(\beta|y) = \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) + \frac{1}{2\tau^2}\beta^T\beta$$

Multiplying by the constant $2\sigma^2$, we see that this is equivalent to solving:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \frac{\sigma^2}{\tau^2}\|\beta\|_2^2 \right\}.$$

(c) The regularization parameter is given by the ratio $\lambda = \frac{\sigma^2}{\tau^2}$.

- High regularization corresponds to noisy data (high σ^2) and a narrow prior peaked at 0 (low τ^2), so this is the case where we're confident in our prior guess but not very confident in our data. The estimator will be close to 0.
- Conversely, low regularization corresponds to the case where we're confident in our data but not in our prior guess. The estimator will be closer to the unregularized least squares solution.

(d) The LASSO estimator minimizes the objective function

$$\|y - X\beta\|_2^2 + \lambda \sum_{j=0}^{p-1} |\beta_j|.$$

To reverse the process from before, set this equal to $-\log f_{\text{post}}(\beta|y)$, so with the same likelihood as before,

$$-\log f_{\text{prior}}(\beta) \propto \frac{\lambda}{2\sigma^2} \sum_{j=0}^{p-1} |\beta_j| \quad \Longrightarrow \quad f_{\text{prior}}(\beta) \propto \prod_{j=0}^{p-1} \exp\left(-|\beta_j|\right).$$

This is called a Laplace distribution. It has a sharp peak at 0, which is one way to see why LASSO tends to give sparse solutions. The LASSO estimator is the Bayesian estimator if your prior guess weights 0 highly.