

Name: _____

Circle the letter of the best answer to each question. The notation below is the same as from the lectures: $x_i \in \mathbb{R}^p$ is a feature vector, $y_i \in \mathbb{R}$ is the response, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and the mean squared error of a parameter vector $\theta \in \mathbb{R}^p$ is

$$\text{MSE}(\theta) = \frac{1}{n} \|y - X\theta\|^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \theta^\top x_i)^2.$$

1. Least squares and projection.

[2 points]

Let $\hat{\theta}$ minimize $\text{MSE}(\theta)$, and let $\hat{y} = X\hat{\theta}$. Which statement is always true?

- (A) $\hat{y} = XX^\top y$.
- (B) $\hat{y}$ is the orthogonal projection of y onto the column space of X . ✓**
- (C) The residual $y - \hat{y}$ lies in the column space of X .
- (D) $\hat{y} = y$ whenever $n > p$.

Explanation. As θ ranges over $\mathbb{R}^p$, $X\theta$ ranges over the column space $\mathcal{C}(X)$. So minimizing $\|y - X\theta\|$ over θ is the same as finding the point of $\mathcal{C}(X)$ closest to y , and that point is the orthogonal projection of y onto $\mathcal{C}(X)$. The residual $y - \hat{y}$ is therefore *orthogonal* to $\mathcal{C}(X)$, not in it: $X^\top(y - \hat{y}) = 0$, which is the normal equation. The formula $XX^\top y$ is this projection only when the columns of X are orthonormal ($X^\top X = I$); in general the projection is $X(X^\top X)^{-1}X^\top y$. And $\hat{y} = y$ only when y happens to lie in $\mathcal{C}(X)$, which typically fails when $n > p$.

2. Increasing the penalty.

[2 points]

Let $\hat{\theta}_\lambda$ be the ridge solution, the minimizer of $\text{MSE}(\theta) + \lambda\|\theta\|^2$. As $\lambda > 0$ increases, which statement is always true?

- (A) The test MSE of $\hat{\theta}_\lambda$ never increases.
- (B) The training MSE of $\hat{\theta}_\lambda$ never decreases. ✓**
- (C) More and more entries of $\hat{\theta}_\lambda$ become exactly 0.
- (D) Eventually $X^\top X + n\lambda I$ becomes singular.

Explanation. Increasing λ makes the $\lambda\|\theta\|^2$ part of the ridge objective increasingly important relative to $\text{MSE}(\theta)$. So, as λ increases, the training MSE can only stay the same or increase (i.e., never decrease). Choice (A) fails on the advertising data, where the test MSE was 0.65 at $\lambda = 10^{-2}$ and 1.04 at $\lambda = 10^{-1}$. Ridge shrinks every coefficient toward 0 but does not typically set any of them exactly to 0, so (C) is wrong. One way to see that (D) is wrong is to consider the case when λ is extremely large, in which case $X^\top X + n\lambda I$ is essentially just the $n\lambda I$ part, which is invertible.

3. Two temperature columns.

[2 points]

A data set records each day's temperature twice, in degrees Celsius c_i and in degrees Fahrenheit $f_i = 32 + 1.8c_i$. We fit a linear regression with the design matrix $X = [\mathbf{1} \quad c \quad f]$ (a column of ones, the Celsius column, and the Fahrenheit column). Which statement is true?

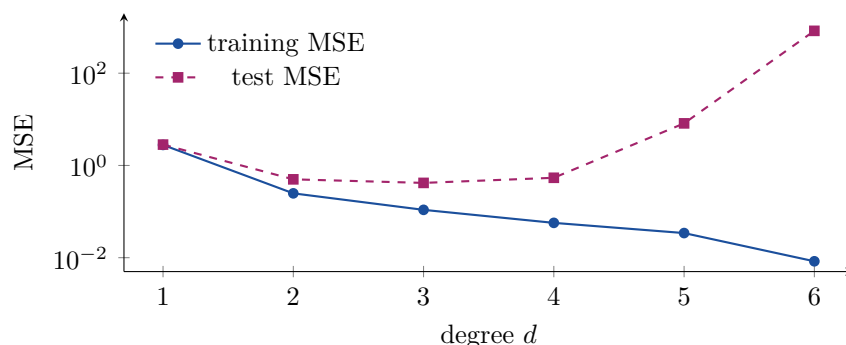
- (A) $X^\top X$ is singular. Infinitely many θ minimize the MSE, and they all give the same fitted values $\hat{y} = X\theta$. ✓
- (B) $X^\top X$ is singular, so no θ minimizes the MSE.
- (C) $X^\top X$ is invertible, because neither temperature column is a multiple of the other.
- (D) The minimizing θ is unique, but the fitted values $\hat{y}$ are not.

Explanation. The Fahrenheit column is a linear combination of the other two: $f = 32 \cdot \mathbf{1} + 1.8c$, which means there are infinitely many minimizers of the MSE and $X^\top X$ singular. Choice (C) misses that the combination uses the column of ones. A minimizer still exists, because the closest point of the column space to y always exists; this rules out (B). If $\hat{\theta}$ is one minimizer, then so is $\hat{\theta} + sv$ for every $s \in \mathbb{R}$ where $v = [32, 1.8, -1]^\top$, since $X(\hat{\theta} + sv) = X\hat{\theta}$. So the coefficients are not unique, but the fitted values are, which rules out (D).

4. Training and test error.

[2 points]

We predict sales from the TV, radio and newspaper budgets of 200 markets, using 100 markets for training and the other 100 for testing. For each degree $d = 1, \dots, 6$, we fit least squares using all products of the budgets of degree at most d (3 features at $d = 1$, up to 83 features at $d = 6$). The plot shows the training and test MSE of each fit.



Which conclusion is best supported by the plot?

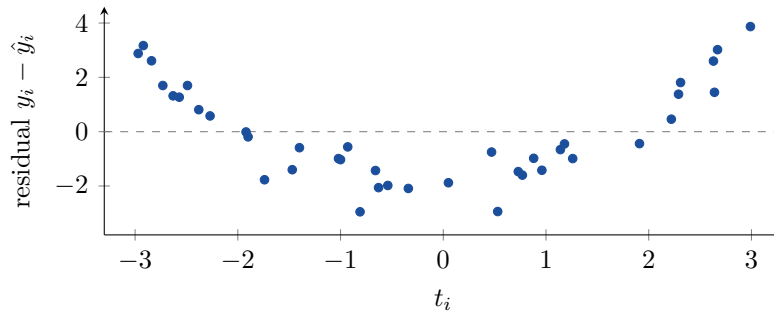
- (A) Degree 6 gives the best model, because its training MSE is the smallest.
- (B) The test MSE is large at degree 6 only because these 100 test markets are unusual; a different random split would not show this.
- (C) At degrees 5 and 6 the fits overfit: they match the training markets very closely but predict new markets poorly. ✓
- (D) At degree 6 the model is too simple for the data, and more features would lower the test MSE.

Explanation. The features of degree at most d include those of lower degree, so the training MSE can only go down as d grows, from 2.78 at $d = 1$ to 0.008 at $d = 6$. The training MSE therefore cannot be used to choose d , which rules out (A). What we care about is the error on new markets, which the test MSE estimates: it is smallest near $d = 3$ (0.42) and then rises to 8.2 at $d = 5$ and 828 at $d = 6$. Small training error with large test error is overfitting. It is not an unlucky split: over 100 random splits, the 83-feature fit did worse than the 3 budgets on the test markets every time.

5. Reading a residual plot.

[2 points]

Each of 40 examples is a single measurement $t_i \in \mathbb{R}$ with a response y_i . We fit the line $\hat{y}_i = \hat{\theta}_1 + \hat{\theta}_2 t_i$ by least squares and plot the residuals $y_i - \hat{y}_i$ against t_i .



What is the best conclusion?

- (A) The model overfits the data, so we should use fewer features.
- (B) The residuals are large, so y is too noisy to be predicted from t .
- (C) The relationship between t and y is curved, so linear regression cannot describe it and a different method is needed.
- (D) The line misses a curved trend. We should add the feature t_i^2 and fit $\theta_1 + \theta_2 t_i + \theta_3 t_i^2$ by least squares. ✓

Explanation. The residuals are not scattered evenly around 0: they are positive at both ends and negative in the middle. So the data follow a curve that no line can follow, and the pattern tells us what is missing: a term that is large for large $|t|$. Least squares only needs the model to be linear in θ , not in t , so the feature vector $x_i = [1, t_i, t_i^2]^\top$ is still linear regression. This rules out (C). On these data, adding t_i^2 lowers the training MSE from 3.18 to 0.31. Choice (A) has it backward: a line has too little flexibility here, not too much.