TAKE-HOME CLASS QUIZ SOLUTIONS: DUE WEDNESDAY DECEMBER 4: ORDINARY LEAST SQUARES REGRESSION

MATH 196, SECTION 57 (VIPUL NAIK)

1. Performance review

25 people took this 8-question quiz. The score distribution was as follows:

- Score of 3: 1 person
- Score of 4: 2 people
- Score of 5: 1 person
- score of 6: 8 people
- Score of 7: 8 people
- Score of 8: 5 people

The mean score was 6.4.

The question-wise answers and performance review were as follows:

(1) Option (C): 23 people
(2) Option (C): 24 people
(3) Option (C): 22 people
(4) Option (D): 20 people
(5) Option (A): 13 people
(6) Option (D): 20 people
(7) Option (A): 18 people
(8) Option (A): 20 people

2. Solutions

PLEASE FEEL FREE TO DISCUSS *ALL* QUESTIONS.

(1) Assume no measurement error. Consider the situation where we have a function f of the form $f(x) = a_0 + a_1 x$ with unknown values of the parameters a_0 and a_1. We collect n distinct input-output pairs, i.e., we collect n distonct inputs and compute the outputs for them. The coefficient matrix for the system is a $n \times 2$ matrix (the rows correspond to the input values, and the columns correspond to the unknown parameters). What is the rank of this matrix?

(A) It is always 2
(B) It is always n
(C) It is always $\min\{2, n\}$
(D) It is always $\max\{2, n\}$

Answer: Option (C)

Explanation: Note that the rank is at most $\min\{2, n\}$ because the rank is at most the minimum of the number of rows and the number of columns. To see that the rank is exactly that value, consider the cases $n = 1$ and $n = 2$. In the case $n = 1$, we have a single nonzero row $\begin{bmatrix} 1 & x_1 \end{bmatrix}$ so the rank is exactly one. In the case $n = 2$, we have a coefficient matrix as follows:

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}$$

Note that, since $x_1 \neq x_2$, the second row is *not* a scalar multiple of the first. Thus, the matrix has rank two. The case $n \geq 2$ follows, because we have already achieved the maximum possible rank of two using the first two rows.

Performance review: 23 out of 25 people go this. 2 chose (A).

(2) Assume no measurement error. Consider the situation where we have a function f of the form $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$ with unknown values of the parameters $a_0, a_1, \ldots, a_m$. We collect n distinct input-output pairs, i.e., we collect n distinct inputs and compute the outputs for them. The coefficient matrix for the system is a $n \times (m+1)$ matrix (the rows correspond to the input values, and the columns correspond to the unknown parameters). What is the rank of this matrix?

(A) It is always $m + 1$
(B) It is always n
(C) It is always $\min\{m + 1, n\}$
(D) It is always $\max\{m + 1, n\}$

Answer: Option (C)

Explanation: The rank can be at *most* $\min\{m + 1, n\}$ because it is at most the minimum of the number of rows and the number of columns. See Section 3.1 of the lecture notes on "hypothesis testing, rank, and overdetermination" but note that the roles of m and n are interchanged there relative to this question.

Performance review: 24 out of 25 people got this. 1 chose (D).

(3) Assume no measurement error. Consider the situation where we have a function f of the form $f(x, y) = a_0 + a_1 x + a_2 y$ with unknown values of the parameters a_0, a_1, and a_2. We collect n distinct input-output pairs, i.e., we collect n distinct inputs (here an input specification involves specifying both the x-value and the y-value) and compute the outputs for them. The coefficient matrix for the system is a $n \times 3$ matrix (the rows correspond to the input values, and the columns correspond to the unknown parameters). What is the rank of this matrix?

(A) It is always $\min\{3, n\}$
(B) It is always $\max\{3, n\}$
(C) For $n = 1$, it is 1. For $n \geq 2$, it is 2 if the input points are all collinear in the xy-plane. Otherwise, it is 3.
(D) For $n = 1$, it is 1. For $n \geq 2$, it is 3 if the input points are all collinear in the xy-plane. Otherwise, it is 2.

Answer: Option (C)

Explanation: For $n \geq 2$, the first two rows are not scalar multiples of each other. Explicitly, the first two rows look like:

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{bmatrix}$$

In the case that all the points are collinear, later rows can be obtained as linear combinations of the first two rows. Equivalently, knowing the value of f at (x_1, y_1) and at (x_2, y_2) allows us to predict the value of f at all other points on the line joining these in $\mathbb{R}^2$, and therefore other points on the line do not reveal new information. Thus, the rank of the coefficient matrix is 2.

See the discussion in Section 2.1 of the lecture notes on hypothesis testing, rank and overdetermination. Also see the answer to Question 1 of the Friday October 11 take-home quiz on linear systems, and the answer to Question 5 of the Monday October 7 take-home class quiz on linear functions and equation-solving (part 2).

Performance review: 22 out of 25 people got this. 2 chose (A), 1 chose (D).

(4) Which of the following is closest to correct in the setting where we use a linear system to find the parameters using input-output pairs given a functional form that is linear in the parameters? Assume for simplicity that we are dealing with a functional form $y = f(x)$ with one input and one output, but possibly multiple parameters in the general description.

(A) The solutions to the linear system that we set up correspond to possibilities for the inputs to the function, and geometrically correspond to choices of points x for the graph $y = f(x)$.
(B) The solutions to the linear system that we set up correspond to possibilities for the inputs to the function, and geometrically correspond to different possible choices for the line or curve that is the graph $y = f(x)$.

(C) The solutions to the linear system that we set up correspond to possibilities for the parameters, and geometrically correspond to choices of points x for the graph $y = f(x)$.
(D) The solutions to the linear system that we set up correspond to possibilities for the parameters, and geometrically correspond to different possible choices for the line or curve that is the graph $y = f(x)$.

Answer: Option (D)

Explanation: We are trying to find the line or curve. That's the whole goal of regression. And the way we do this is by choosing a general functional form and then using regression to estimate the parameters in that functional form.

Performance review: 20 out of 25 people got this. 5 chose (C).

(5) Continuing with the notation and setup of the preceding question, consider the coefficient matrix of the linear system. This matrix defines a linear transformation from the vector space of possible parameter values to the vector space of the outputs of the function. What is the image of this linear transformation?
(A) The image is the set of possible output values for which the linear system is consistent, i.e., we can find *at least one* function f of the required functional form that fits all the input-output pairs with *no measurement error*.
(B) The image is the set of possible output values for which the linear system has *at most one solution*, i.e., the set of output valus for which we can find *at most one* function f of the required functional form that fits all the input-output pairs with *no measurement error*.

Answer: Option (A)

Explanation: This follows from the definition of image.

Performance review: 13 out of 25 people got this. 12 chose (B).

(6) Consider the case of polynomial regression for a polynomial function of one variable, allowing for measurement error. We believe that a function has the form of a polynomial. We can tentatively choose a degree m for the polynomial we are trying to fit, and a value n for the number of distinct inputs for which we compute the corresponding outputs to obtain input-output pairs (i.e., data points). We will get a $n \times (m + 1)$ coefficient matrix. Which of the following correctly describes what we should try for?
(A) We should choose n and $m + 1$ to be exactly equal, so that we get a unique polynomial.
(B) We should choose n to be greater than $m+1$, so that the system is guaranteed to be consistent and we can find the polynomial.
(C) We should choose n to be less than $m + 1$, so that the system is guaranteed to be consistent and we can find the polynomial.
(D) We should choose n to be greater than $m + 1$, so that the system is *not* guaranteed to be consistent, but we do have a unique solution after we project the output vector to a vector for which the system is consistent.
(E) We should choose n to be less than $m+1$, so that the system is *not* guaranteed to be consistent, but we do have a unique solution after we project the output vector to a vector for which the system is consistent.

Answer: Option (D)

Explanation: If we chose $n > m + 1$, the coefficient matrix has full column rank $m + 1$, and therefore the linear transformation is injective, i.e., we have at most one solution. However, it does not have full row rank, so the linear transformation is not surjective, i.e., we do not necessarily have a solution for every output vector.

Both aspects are features for us. The existence of at most one solution means that we can find the parameters uniquely after we project to the closest vector in the image. The fact that the system does not have full row rank means that there is a potential for inconsistency, and in the presence of measurement error, the system will probably be inconsistent with the measured output vector. This is good because the potential for inconsistency allows us to test the validity of the model better. Also, in the case of measurement error, the more the number of inputs that we use, the better our estimate of the function is likely to be.

Performance review: 20 out of 25 people got this. 3 chose (E), 1 each chose (B) and (C).

(7) Consider the general situation of linear regression. Denote by X the coefficient matrix for the linear system (also called the design matrix). Denote by $\vec{\beta}$ the parameter vector that we are trying to solve for. Denote by $\vec{y}$ an observed output vector. The idea in ordinary least squares regression is to choose a suitable vector $\vec{\varepsilon}$ such that the linear system $X\vec{\beta} = \vec{y} - \vec{\varepsilon}$ can be solved for $\vec{\beta}$. Among the many possibilities that we can choose for $\vec{\varepsilon}$, what criterion do we use to select the appropriate choice? Recall that the *length* of a vector is the square root of the sum of squares of its coordinates.

(A) We choose $\vec{\varepsilon}$ to have the minimum length possible subject to the constraint that $X\vec{\beta} = \vec{y} - \vec{\varepsilon}$ has a solution.

(B) We choose $\vec{\varepsilon}$ such that the system $X\vec{\beta} = \vec{y} - \vec{\varepsilon}$ can be solved and such that the solution vector $\vec{\beta}$ has the minimum possible length (among all such choices of $\vec{\varepsilon}$).

(C) We choose $\vec{\varepsilon}$ such that the system $X\vec{\beta} = \vec{y} - \vec{\varepsilon}$ can be solved and such that the difference vector $\vec{y} - \vec{\varepsilon}$ has the minimum possible length (among all such choices of $\vec{\varepsilon}$).

Answer: Option (A)

Explanation: We want to deviate as little as possible from the measured output. This is the idea behind using the orthogonal projection.

Performance review: 18 out of 25 people got this. 6 chose (C), 1 chose (B).

(8) We have data for the logarithm of annual per capita GDP for a country for the last 100 years. We want to see if this fits a polynomial model. The idea is to try to first fit a polynomial of degree 0 (i.e., per capita GDP remains constant), then fit a polynomial of degree ≤ 1 (i.e., per capita GDP grows or decays exponentially), then fit a polynomial of degree ≤ 2 (i.e., per capita GDP grows or decays as the exponential of a quadratic function), and so on.

What happens to the length of the error vector $\vec{\varepsilon}$ as we increase the degree of the polynomial that we are trying to fit?

(A) The error vector $\vec{\varepsilon}$ keeps getting smaller and smaller in length, with a near-certainty that it keeps *strictly* decreasing in length at each step, until the error vector becomes $\vec{0}$ (which we expect will happen when we get to the stage of trying to fit the function using a polynomial of degree 99).

(B) The error vector $\vec{\varepsilon}$ keeps getting larger and larger in length, with a near-certainty that it keeps *strictly* increasing in length at each step, until the error vector becomes $\vec{y}$ (which we expect will happen when we get to the stage of trying to fit the function using a polynomial of degree 99).

Answer: Option (A)

Explanation: The space that we are projecting on keeps getting bigger and bigger, so the distance from the vector $\vec{y}$ to the space keeps getting smaller and smaller. Note that it's highly unlikely that there would be *no* improvement possible by introducing a new parameter, so it's likely that the improvement would be strict at every stage until the time that we reach a polynomial of degree 99, where we can obtain a perfect fit because we have only 100 data points.

Performance review: 20 out of 25 people got this. 5 chose (B).