DATA 100: Vitamin 10 Solutions

November 1, 2019

1 Simple Linear Probability Model

In lecture, we defined a simple linear probability model as

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where Y is a random variable, β_0, β_1 are the true parameter values, and ϵ is random noise.
Which of the following statements/assumptions about this model are true?

- ☑ $\mathbb{E}(\epsilon) = 0$
- ☐ The size of the errors depend on x, β_0, and β_1
- ☑ If the linear model holds, the least squares regression parameters are unbiased

Explanation: In a simple linear probability model, the expectation of the random noise term term is assumed to be zero, and its variance is assumed to be constant. This implies that the values and the size of these error terms do not depend on x, β_0, and β_1. Assuming the linear model holds, then the least squares regression parameters are unbiased estimates of the true intercept and slope.

2 Variance of Y

What is the variance of random variable Y, defined in the previous question? Select all that apply. Assume that $Var(\epsilon) = \sigma^2$.

- ☑ $Var(\beta_0 + \beta_1 x + \epsilon)$
- ☐ $\beta_0 + Var(\beta_1 x + \epsilon)$
- ☐ ${\beta_1}^2 Var(x) + Var(\epsilon)$
- ☑ σ^2

Explanation:

$Var(Y) = Var(\beta_0 + \beta_1 x + \epsilon)$

Since this model is implicitely conditional on X, e.g. $X = x$,

$\beta_0 + \beta_1 x$ are constant terms.

By the properties of variance, that is $Var(c + dZ) = d^2 Var(Z)$:

$$\begin{aligned} Var(Y) &= Var(\beta_0 + \beta_1 x + \epsilon) \\ &= Var(\epsilon) \\ &= \sigma^2 \end{aligned}$$

Therefore, the first and last options are correct.

3 Model Assessment

Fill in the blanks:
As model complexity increases, so too does the risk of ____ the model to the data. Therefore, we should determine how well our model generalizes using ____.

- ☐ underfitting, test data
- ☐ overfitting, training data
- ☑ overfitting, test data
- ☐ underfitting, training data

Explanation: As model complexity increases (e.g. increasing the number of features in a linear model), the risk of overfitting to the data increases. A good approach to identify when a model is overfitting is to evaluate it's error over a test dataset. A test data set is a set of data that is similar to the data on which the model was trained, but completely disjoint.

4 Bias-Variance Tradeoff

The bias-variance tradeoff relates to the relationship between a model's complexity, generalizability, and similarity to the true model. Which of the following statements are true?

- ☑ As model complexity increases, so does model variance
- ☑ Very biased models tend underfit to the data
- ☐ Models with large model variance are said to generalize well

Explanation: The bias-variance tradeoff defines a balancing act between a model's ability to represent the true data generating process, and its ability to generate good predictions. As model complexity increases, its variance increases, and so too does its sensitivity to variation in the training data. This often leads to good predictive accuracy in the training data, but the model does not generalize well (e.g. poor predictive accuracy on the test data).

5 Regularization and the Bias-Variance Tradeoff

Fill in the blanks:

Ridge and LASSO linear regression allow us to control model complexity. As their regularization parameters increase, the estimated coefficients shrink towards 0. This increases the _____ of each model, but decreases the _____.

- ☐ error, predictive accuracy
- ☑ bias, variance
- ☐ variance, bias
- ☐ predictive accuracy, generalizability

Explanation: As the estimated coefficients shrink towards 0, the models become increasingly biased, but have less variation.