Q1.

1. What makes a model linear? Linear in what?
   When the independent and dependent terms have a linear relationship(makes a line rather than a curve) in the expression.

2. How do you interpret the coefficient for a dummy/one-hot-encoded variable? (This is a trick question, and the trick involves how you handle the intercept of the model.)

   The coefficient for a dummy variable represents the average difference in the dependent variable between the group represented by the dummy variable and the reference group.

3. Can linear regression be used for classification? Explain why, or why not.

   Linear regression predicts continuous values, but classification requires discrete values.

4. What are signs that your linear model is over-fitting?

   High variance is one sign that the model is over-fitting meaning the model is sensitive to small changes in the training data. Also, if there is a big difference in performance (accuracy) of the training and testing set. For example, if the training set is highly accurate, but then the model does not perform well on new data, this indicates that the model is likely over-fitting.

5. Clearly explain multi-colinearity using the two-stage least squares technique.

   Multicollinearity is when the independent variables are highly correlated. The two stage least squares technique can minimize the effects of this problem. In the first stage, you regress the endogenous variable on the exogenous variable. The goal here is to predict an endogenous variable based on the exogenous one to predict a “clean” value. In the second stage, the dependent variable is regressed on the predicted values (from stage one). The goal here is to estimate the relationship between the dependent and predicted variable to lessen the impact of multicollinearity.

6. How can you incorporate nonlinear relationships between your target/response/dependent/outcome variable y and your features/control/response/independent variables x into your analysis?

   One way to incorporate nonlinear relationships between x and y variables is to use tree models like random forests, which can capture complex nonlinear relationships.

7. What is the interpretation of the intercept? A slope coefficient for a variable? The coefficient for a dummy/one-hot-encoded variable?

   The intercept is the predicted value of y when the x is zero. The slope is the value of a one unit increase in x. The coefficient of a dummy variable is the difference of the mean of the dependent variable between the category and reference category.

① 6 $\hat{y}_i = b_0 + b_1 z_{i1} + b_2 z_{i2}$

$\frac{1}{N} \sum_{i=1}^{N} z_{ij} = 0$

1. SSE for the model

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

2. partial derivative w/ respect to b_0, b_1, b_2

$\frac{\partial \hat{y}_i}{\partial b_0} = 1$ $\frac{\partial \hat{y}_i}{\partial b_1} = z_{i1}$ $\frac{\partial \hat{y}_i}{\partial b_2} = z_{i2}$

3. average error is zero $e \cdot z = 0$

• error term $e_i = y_i - \hat{y}_i$

$\sum_{i=1}^{N} e_i = 0$ $\sum_{i=1}^{N} e_i z_{ij} = 0$ for $j = 1, 2$

$\sum_{i=1}^{N} (y_i - b_0 - b_1 z_{i1} - b_2 z_{i2}) = 0$

average error $= \frac{1}{N} \sum_{i=1}^{N} e_i = 0$

$\sum_{i=1}^{N} z_{i1} (y_i - b_0 - b_1 z_{i1} - b_2 z_{i2}) = 0$

$\sum_{i=1}^{N} z_{i2} (y_i - b_0 - b_1 z_{i1} - b_2 z_{i2}) = 0$

$\sum_{i=1}^{N} z_{i1} e_i = 0$ $\sum_{i=1}^{N} z_{i2} e_i = 0$

4. optimal intercept $b_0^* = \bar{y}$

$$\frac{\partial SSE}{\partial b_0} = \sum_{i=1}^{N} 2(y_i - b_0 - b_1 z_{i1} - b_2 z_{i2})(-1)$$

$$= -2 \sum_{i=1}^{N} (y_i - b_0 - b_1 z_{i1} - b_2 z_{i2})$$

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} (b_0 + b_1 z_{i1} + b_2 z_{i2})$$

$$\frac{1}{N} \sum_{i=1}^{N} y_i = b_0 + b_1 \frac{1}{N} \sum_{i=1}^{N} z_{i1} + b_2 \frac{1}{N} \sum_{i=1}^{N} z_{i2}$$

$$\frac{1}{N} \sum_{i=1}^{N} z_{i1} = 0, \quad \frac{1}{N} \sum_{i=1}^{N} z_{i2} = 0$$

$$b_0^* = \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{y} \qquad (b_0^* = \bar{y})$$

5. results as matrix equation $Ab = C$

$$\begin{bmatrix} \sum z_{i1}^2 & \sum z_{i1} z_{i2} \\ \sum z_{i1} z_{i2} & \sum z_{i2}^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum z_{i1} y_i \\ \sum z_{i2} y_i \end{bmatrix}$$

$$A \qquad\qquad B = C$$

6. What is matrix A? What is vector c? Explain intuition.

$$\frac{1}{N} \begin{bmatrix} \sum z_{i1}^2 & \sum z_{i1} z_{i2} \\ \sum z_{i1} z_{i2} & \sum z_{i2}^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{1}{N} \begin{bmatrix} \sum z_{i1} y_i \\ \sum z_{i2} y_i \end{bmatrix}$$

$$A = \frac{1}{N} \begin{bmatrix} \sum_{i=1}^{N} z_{i1}^2 - \frac{1}{N} \left(\sum_{i=1}^{N} z_{i1} \right)^2 & \\ \sum_{i=1}^{N} z_{i1} z_{i2} - \frac{1}{N} \sum_{i=1}^{N} \sum_{i=1}^{N} z_{i2} & \end{bmatrix}$$

$\Sigma \rightarrow$

$$\begin{bmatrix} \sum_{i=1}^{N} z_{i1} z_{i2} - \frac{1}{N} \sum_{i=1}^{N} z_{i1} \sum_{i=1}^{N} z_{i2} \\ \sum_{i=1}^{N} z_{i2}^2 - \frac{1}{N} \left(\sum_{i=1}^{N} z_{i2} \right)^2 \end{bmatrix}$$

$$c = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^{N} z_{i1} \bar{y}_i \\ \frac{1}{N} \sum_{i=1}^{N} z_{i2} \bar{y}_i \end{bmatrix}$$

The vector c represents the covariances between the centered predictor & centered response variable.

Q7 $a^* = \bar{y}$

$$b^* = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

1. when will b^* be large/small w/ relationship between X, y & variance of X?

-the numerator is the covariance between y & X. Since they are both in the numerator, this means if y increases, then X does & vice versa. The denominator is the variance of X (amount x varies around the mean($\bar{x}$)). b^* will be large when X & y have a strong relationship (covariance) or X has a small variance. b^* will be small if X & y have a weak relationship or X has a large variance.

2. affect intercept of regression? What happen to b^* coefficient? How does it affect ability to predict?

- The intercept is independent, so the intercept will not be affected by measurement error in x_i.
- rather than x_i measure error, b^* will tend to be biased toward zero,
- The ability to predict accurately will lessen. There will be greater uncertainty

3. noise n_i is independent

- Basically, the noise does not introduce bias (b/c of zero mean independent), but it does reduce precision of our estimate of the relationship of X & Y. The added noise of X inflates the variance. This leads to attenuation of b^*.

4. Attenuation is important in the cost-benefit analysis of data quality. When measurement error is large, cleaner data/ accurate measures are largely beneficial for predictions. Alternately, if the measurement error is small, more effort to clean data is unlikely not worth the cost.