HW1 Part 1

Week 1

**Q1**. Which of the following indicate that the result from a simple linear regression model could be potentially misleading?

* 1. The dependent and the independent variable show a linear pattern.
  2. The error terms exhibit homoscedasticity
  3. The error terms follow a normal distribution
  4. The nth error term could be predicted with e\_n = 0.91\*e\_{n-1}

Answer: d. Module 1 Slide 37,38,39

**Q2**. Consider a multiple linear regression model: Y = 0.55 + 0.93\*x1 + 1.88\*x2. Which of the following interpretation of the coefficients is correct?

* 1. A unit increase in x1 is associated with an 0.93 increase in Y.
  2. Y is predicted to be equal to 0.55 when both x1 and x2 take the value of 1
  3. A unit increase in x2 is associated with a 1.88 increase in Y, keeping all else constant.
  4. A 0.93 increase in x1 is associated with a 1.88 increase in x2

Answer: c. Module 1 Slide 22

**Q3**. When testing our predictive variables for Multicollinearity we create a model in R of lm( pred1 ~ pred2 + pred3, data = dataset) and we get an R Squared of 0.85. What is the VIF for pred1?

* 1. 0.85
  2. 0.15
  3. 6.667
  4. 0.5405

Answer C from Wk1 page 47 slide 1

Week 2

**Q4**. Consider a linear regression model estimating the fuel efficiency of a car in terms miles per gallon of gas (mpg) based on its origin (region A, B or C) and number of cylinders with the following formula:

mpg = b0 + b1\*RegionB + b2\*RegionC + b3\*Cylinders

The estimated values of the regression coefficients are provided below:

b0 = 40.7

b1 = -0.91

b2 = 2.66

b3 = -3.15

Based on this model, if X is the mpg of a car with 4 cylinders originated from region B, and Z is the mpg of a car with 3 cylinders originated from region A, what is the value of X - Z:

-0.91

-3.15

-4.06

-6.72

Answer: c (page 17, slide 1)

**Q5**. True or False, using factor variables directly in R is an alternative to coding dummy variables:

* 1. True
  2. False

Answer: A (Week2.IndicatorVariables.R line 59 and 60)

**Q6**. From the following regression model: Gold\_Price\_Per\_oz = B0 + B1\*M2+B2\*VIX+B3\*War:

Where M2 is a continuous variable of the M2 money supply, VIX is a continuous variable of the VIX index, and War is a categorical variable (0 is Time period at peace, 1 is Time period at war), the base case is?

* 1. Time period at war
  2. Time period at peace
  3. A high VIX index
  4. Period of inflation

Answer b page 7 slide 2

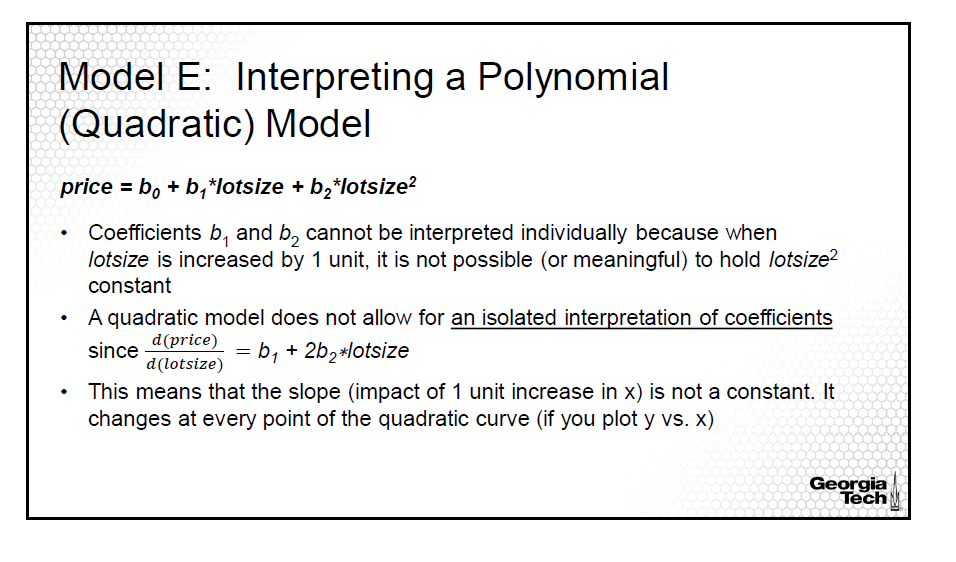
Week 3

**Q7**. Given the following model: price=b0+ b1∗lotsize+b2∗lotsize2 ; how can one interpret the coefficients? Select the best answer

* 1. A quadratic model **does** allow for an isolated interpretation of coefficients
  2. dfl
  3. A quadratic model **does not** allow for an isolated interpretation of coefficients
  4. Coefficients [Equation] **can** be interpreted individually because when lotsize is increased by 1 unit, it is **possible** (or meaningful) to hold [Equation] constant

Answer: B

Pg 19

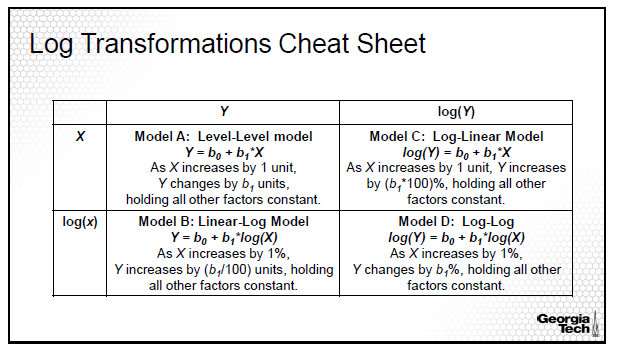


**Q8**. Select the model approximation that best matches the following statement: ‘As X increases by 1%, Y increases by (b\_1 /100) units, holding all other factors constant.’

* 1. Y=b0+b1∗X
  2. Y=b0+b1∗log(X)
  3. log(Y)=b0+b1∗X
  4. log(Y)=b0+b1∗log(X)

Answer: B; Linear-Log

Pg 17



**Q9**. Assume that you have concluded to use a log transformation on your data to model a relationship. However, on investigating the dataset, you found negative or zero values. How will you proceed?

1. Throw out the data points which are negative or zero
2. Use Log(x+1), where x is the variable you want to transform
3. Use log(x + c +1), where c is the absolute value of the most negative number
4. Use log(10 \*x)

Answer: **C,** Module 3 page 6

While log(x +1) will work for zero values, we need to offset the transformation to accommodate negative values, hence adding c . Here, c is the value of the most negative number.

Option D is not correct because it models the relationship inefficiently. If x is negative, multiplying by 10 still results in the value within log to be negative. I.e It doesn’t exactly handle the negative aspect.

Week 4

**Q10**. The logit function is the log of the ratio of the probability of success (belonging to a group) to the probability of failure (not belonging to a group). It is also known as the log odds function.

1. True
2. False

Answer: a (Slide 10)

**Q11**. Using the following confusion matrix, what is the sensitivity and specificity of the model?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Predicted Value | | | |
| Actual  Value |  | 1 | 0 | Total |
| 1 | 107 | 72 | 179 |
| 0 | 23 | 798 | 821 |
| Total | 120 | 870 | 1000 |

1. Specificity: 0.917, Sensitivity: 0.823
2. Specificity: 0.972, Sensitivity: 0.598
3. Specificity: 0.598, Sensitivity: 0.972
4. Specificity: 0.591, Sensitivity: 0.778

Answer: b (Slide 28). Flipped 1/0 so Confusion Matrix is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Predicted Value | | | |
| Actual  Value |  | 1 | 0 | Total |
| 1 | TP | FN |  |
| 0 | FP | TN |  |
| Total |  |  |  |

**Q12**. Which of the following case is referred to Type II error?

1. Null is false and we reject it.
2. Null is True, but we fail to reject it.
3. Null is True but we mistakenly reject it.
4. Null is false but we fail to reject it.

Answer: d (Slide 27)