**ISOM 2600 Introduction to Business Analytics**

**Weekly Exercise 6**

Q1:

To deal with collinearity, which of the following is **NOT** a correct solution?

1. Drop one of the paired predictors from the model.
2. Combine the correlated predictor with other predictors.
3. Live with it.
4. Include more predictors to the model.

[Q2 – Q6] A real estate valuation analyst would like to build a model for the sales price of a house. He regressed the house price on a set of predictors. The result of data analysis is given below:

Table

Description automatically generated

Q2.

Which of the following statement about the model is correct?

1. The model is significant overall as the R-squared is higher than 0.5.
2. The model is insignificant overall as the adjusted R-squared is lower than 0.5.
3. The model is significant overall because the p-value for F-statistic is less than 0.05.
4. The model is insignificant overall as there are insignificant predictors in the model.

Q3.

Which statement is false?

1. The regression can explain 50.6% of the variation in house selling prices.
2. If the number of observations in the regression is larger, then r-squared and adj r-squared will be closer to each other.
3. If drop Lot size from the regression, r-squared will not increase.
4. If drop Lot size from the regression, adj r-squared will not increase.

Solution: If we drop an irrelevant x-variable from a regression, adj r-squared may increase.

Q4.

You are given the sample standard deviation of house price, Sy = 64992. Then the approximated RMSE should be:

* 1. 32106
  2. 179
  3. 2.09\*109
  4. 45680

Solution: Estimate of residual std = RMSE = 45679.6665

Q5.

Now we consider models with different combination of predictors. A list of models and its performance is given below:

| **Model** | **features** | **SSE** | **R\_squared** | **adj\_R2** |
| --- | --- | --- | --- | --- |
| **1** | **(Property tax,)** | **317286593808** | **0.105761** | **0.094987** |
| **2** | **(Living area,)** | **250457310875** | **0.294112** | **0.285608** |
| **3** | **(Living area, Year built)** | **216615445413** | **0.389492** | **0.374602** |
| **4** | **(Bathrooms, Lot size, Year built)** | **224028658719** | **0.368599** | **0.345214** |

Which model has the best prediction performance according to adjusted R2?

1. Model 1
2. Model 2
3. Model 3
4. Model 4

Solution: Model 3 has the highest adjusted R2.

Q6

Compute the RMSE for model 4.

1. 51953
2. 5568
3. 5843
4. 52591

Solution: RMSE = sqrt(SSE/(n – k – 1)). We can find n = 85 from the regression table, and k =3, then RMSE = sqrt(224028658719/ (85 – 3 – 1)) = 52590.7