Thursday Exam 3 - Solutions - STAT 324

## Question 7:

1. Only, write out the full regression equation for Model 4 using the estimated coefficients.
2. Using Model 4, what is the predicted price for a home where square footage is 2 units above the mean and number of bedrooms is also 2 units above the mean? Show all work and explain the interpretation of your result.

Solution:

### **Solution**

#### **a) Full Regression Equation**

Let:

* = predicted price of home
* = sqft\_centered (i.e., square footage minus the mean sqft)
* = beds\_centered (i.e., number of bedrooms minus the mean number of bedrooms)

[I did not take off points if you do not include the above but for the final exam I will expect you to do this, if you did no points will be taken away ]

#### **b) Prediction for 2 Units Above the Mean on Both Variables**

Given:

Substitute into the equation:

### **Interpretation**

A house that is 2 sqft units above the average and has 2 more bedrooms than average is predicted to cost **approximately $198,466**.

* The negative interaction term () indicates that the *marginal effect* of square footage on price **decreases** as the number of bedrooms increases (and vice versa).
* In this case, having more of both sqft and bedrooms *together* reduces the predicted price more than if we considered them independently.

PTS: 22

## Question 8:

What does the 154963.526 value in Model 7 represent? Provide as much detail as possible.

### **Solution**

The value 154,963.526 in Model 7 is the intercept shift for the “RegionGroup 4 – Auburn & Foothills” relative to the baseline group (RegionGroup 1 – Sacramento Core), in a regression model with an interaction between sqft and RegionGroup.

PTS: 16

## Question 9:

Using Model 7, predict the price for a house in RegionGroup 5 - South & Rural with 2000 sqft. Show your calculation and interpret the result.

### **Solution**

#### **Calculation:**

#### **Interpretation:**

A home located in **RegionGroup 5 – South & Rural** with **2000 sqft** is predicted to cost approximately **$295,204.85**.

PTS: 22

## Question 10:

1. Which model (from 1 to 7) would you choose to predict price and why? b) What additional criteria or steps would you use to compare the models more rigorously?

### **Solution**

#### **a) Model Choice**

I would choose **Model 7** to predict housing prices.

**Why:** **Highest R-squared**: Model 7 has the highest , indicating that it explains the largest proportion of variation in price.

#### **b) Additional Criteria for Model Comparison**

To compare the models more rigorously, I would consider the following:

1. **Cross-Validation** 
   * Evaluate how well each model generalizes to new data by estimating prediction error using k-fold cross-validation or the Prediction Residual Sum of Squares (PRESS).
2. **AIC / BIC (Information Criteria)**
   * Compare model complexity and fit. AIC and BIC penalize more complex models, helping us assess whether added complexity improves performance.
3. **Residual Assumptions**
   * Check for model assumptions: linearity, homoscedasticity (constant variance), normality of residuals, and absence of outliers or influential points.
4. **Multicollinearity**
   * Use Variance Inflation Factor (VIF) to ensure predictors are not overly correlated.
5. **Practicality and Interpretability**
   * Although Model 7 fits best statistically, a simpler model (e.g., Model 4 or 5) might be chosen if interpretability or data availability is a priority.
6. **Test Set Performance**
   * Split data into training and test sets to evaluate each model’s true predictive performance on unseen data.

[You do not need to mention all of these just some deeper thought would suffice]

PTS: 12