**Questions week 1**

**Please download from CANVAS the file “HKRE.csv”. The file contains information about apartments from Baywater bay in Hongkong.**

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| price | Price an apartment was sold for in million USD (outcome variable) in 1995 |
| beds | Number of bedrooms |
| baths | Number of bathrooms |
| area | square footage of apartment |
| age | how old is the property in years (since it was built) |
| floor | level above ground |
| view | Coded as 0,1,2 = none, partial view of water, good view of water |

1. **Let’s start with some exploratory analysis and check all correlations (include numeric and binary coded dummy variables, if there are any, but not other categorical factors).**
   1. **Which input variable has the greatest correlation with price?**
   2. **Which one has the weakest correlation with price?**
2. **Decide on how to model the outcome variable. Should we log-transform price?**

**Imagine you local real-estate agent to build a linear regression model to predict the sales prices of apartments were hired by the out of sample – your model will help the real-estate agent set the expected sales price for the auction right (as a benchmark for sellers). The goal is to obtain an as accurate forecast of the likely sales price as possible.**

1. **Let’s estimate our prediction model for prices of new apartments.** 
   1. **Which input variables would you include in your model?**
   2. **Report all coefficients and p-values for these variables. How many are statistically significant at a level of p<0.05?**
   3. **What is the AIC/ adjusted R^2 of this model?**
2. **Now find the best prediction model based on in-sample AIC.** 
   1. **Run a stepwise AIC optimisation (both directions) for the model of C. Which variables are included in this model?**
   2. **Run further stepwise AIC optimised models:**

**Run one model with log-transformation of all (numeric) input variables, one model with squared terms for all numeric variables, two more models with different polynomials (more than squared terms), and at least one model specification with a two-way interaction.**

***Note: This means you need to run in total 5 additional model specifications (in addition the basic one from a). That means, you will have 6 models in total.***

* 1. **Report all coefficients and the respective p-values for the variables included in your final model that is the one with the best AIC from the six candidates.**
  2. **What is the AIC of the best-performing model from c?**

**Now imagine you were hired by a property developer who has access to the same data set. The developer wants you to build a linear model to help him understand which elements drive prices so that he can build / configure his units with the goal of achieving a high sales price.**

1. **Let’s first run simple linear regressions for each of the input variables that is numeric or already dummy coded and you think should be in the explanatory model (without transformations or interactions).** 
   1. **Which input variables would you include in this model?**
   2. **How many variables in this model are statistically significant at a level of p<0.05?**
   3. **Let’s plot a bar chart comparing the coefficients for all input variables from the simple regression model (from a above) and a multiple regression model (keep categorical variables in the model but only compare the numeric/dummy coded coefficients). (0.1)**
   4. **What are the most and least influential input variables affecting price according to the simple and multiple regression model? (0.2)**
2. **Now let’s standardise the coefficients of all input variables and rerun the simple and multiple linear regressions.** 
   1. **Plot a bar chart that compares the standardised coefficients from both types of regressions. What do you observe in terms of differences now in comparison to the previous bar chart from Eb?**
   2. **What are the most and least influential input variables affecting price according to the simple and multiple regression models now?**
   3. **Check whether the residuals of the multiple regression model appear normally distributed using a qq-plot. Are they normally distributed?**
   4. **Let’s carry out a Breusch-Pagan test. Can we assume homoscedasticity?**
3. **Would you remove the non-significant input variables from your model for the developer?** **Please explain briefly why or why not.**
4. **Do you think there are any possible confounder variables missing in your analysis? Draw a causal diagram.**