Individual Coursework Submission Form

Specialist Masters Programme

Surname: Heijliger-Krogulski	**First Name: Boris**
MSc in: Actuarial Science	**Student ID number: 230063332**
Module Code: SMM634	
Module Title: Analytics Methods for Business	
Lecturer: Rosalba Radice	**Submission Date: 27/10/2023**

Declaration:
By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: | | **Final Mark:** | %

Choosing a Regression Model & Justification

To start, we must look at the data set given, which contains 6 response variables and 6 binary "yes" "no" variables. For RStudio to read this data set correctly, we must import it, making sure we let the first column be the indicator, and transforming the string variables into factors containing 2 levels. This consequently allows us to work on this data correctly.

```
HousePrices             546 obs. of 12 variables
   $ price     : int  42000 38500 49500 60500 61000 66000 66000 69000 83800 88500 ...
   $ lotsize   : int  5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...
   $ bedrooms  : int  3 2 3 3 2 3 3 3 3 3 ...
   $ bathrooms : int  1 1 1 1 1 1 2 1 1 2 ...
   $ stories   : int  2 1 1 2 1 1 2 3 1 4 ...
   $ driveway  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
   $ recreation: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 2 2 ...
   $ fullbase  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 2 1 2 1 ...
   $ gasheat   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
   $ aircon    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 2 ...
   $ garage    : int  1 0 0 0 0 0 2 0 0 1 ...
   $ prefer    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Fig1.1: initial look at HousePrices Dataset

To get a baseline model, we start by assuming all the quantitative variables (price, lotsize, bedrooms, bathrooms, stories, garage) are all related to the factor variables (driveway, recreation, fullbase, gasheat, aircon, prefer) so we start by constructing a linear model with this property.

lm.price <- lm(price ~ (lotsize + bedrooms + bathrooms + stories + garage) * (driveway + recreation +fullbase + gasheat + aircon + prefer))

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.162e+04  8.753e+03   2.470 0.013842 *
lotsize                 1.944e+00  1.509e+00   1.288 0.198171
bedrooms                2.824e+02  2.896e+03   0.098 0.922361
bathrooms               4.759e+03  4.044e+03   1.177 0.239804
stories                 6.067e+03  4.130e+03   1.469 0.142446
garage                 -2.152e+03  3.916e+03  -0.550 0.582847
drivewayyes            -1.257e+04  9.174e+03  -1.370 0.171188
recreationyes           1.935e+04  9.738e+03   1.987 0.047445 *
fullbaseyes             2.995e+03  7.709e+03   0.389 0.697770
gasheatyes             -1.552e+04  1.536e+04  -1.010 0.312826
airconyes              -7.844e+03  7.746e+03  -1.013 0.311690
preferyes              -2.591e+03  9.117e+03  -0.284 0.776379
lotsize:drivewayyes     1.201e+00  1.544e+00   0.778 0.436984
lotsize:recreationyes  -3.876e+00  1.165e+00  -3.329 0.000936 ***
lotsize:fullbaseyes     2.218e+00  7.858e-01   2.822 0.004955 **
lotsize:gasheatyes     -1.157e+00  1.491e+00  -0.776 0.438114
lotsize:airconyes       5.342e-03  7.675e-01   0.007 0.994449
lotsize:preferyes      -4.609e-03  7.920e-01  -0.006 0.995359
bedrooms:drivewayyes    1.311e+03  3.010e+03   0.436 0.663320
bedrooms:recreationyes  2.953e+03  3.302e+03   0.894 0.371647
bedrooms:fullbaseyes    9.687e+02  2.478e+03   0.391 0.696044
bedrooms:gasheatyes     1.265e+04  5.275e+03   2.397 0.016878 *
bedrooms:airconyes      1.909e+03  2.443e+03   0.782 0.434865
bedrooms:preferyes     -1.238e+03  3.077e+03  -0.402 0.687634
bathrooms:drivewayyes   8.809e+03  4.286e+03   2.055 0.040346 *
bathrooms:recreationyes -2.474e+03 3.529e+03  -0.701 0.483565
bathrooms:fullbaseyes  -5.101e+03  3.224e+03  -1.582 0.114296
bathrooms:gasheatyes   -4.500e+03  7.674e+03  -0.586 0.557869
bathrooms:airconyes     6.695e+03  3.207e+03   2.088 0.037342 *
bathrooms:preferyes     7.710e+03  3.633e+03   2.122 0.034311 *
stories:drivewayyes    -1.753e+02  4.175e+03  -0.042 0.966529
stories:recreationyes  -1.497e+03  2.158e+03  -0.694 0.488013
stories:fullbaseyes    -2.280e+03  2.375e+03  -0.960 0.337411
stories:gasheatyes     -5.230e+03  5.445e+03  -0.961 0.337177
stories:airconyes       1.276e+03  1.853e+03   0.689 0.491404
stories:preferyes       2.016e+03  1.957e+03   1.030 0.303559
garage:drivewayyes      3.675e+03  3.966e+03   0.927 0.354599
garage:recreationyes    5.407e+03  2.396e+03   2.257 0.024429 *
garage:fullbaseyes     -7.717e+02  1.896e+03  -0.407 0.684184
garage:gasheatyes       1.293e+04  4.141e+03   3.123 0.001891 **
garage:airconyes        3.566e+03  1.771e+03   2.013 0.044633 *
garage:preferyes        1.382e+03  2.011e+03   0.687 0.492268
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14720 on 504 degrees of freedom
Multiple R-squared:  0.719,     Adjusted R-squared:  0.6962
F-statistic: 31.46 on 41 and 504 DF,  p-value: < 2.2e-16
```

Fig1.2: Summary of lm.price with associated p-values

Our analysis of this model tells us that it can explain 69.62% of the variation within our model, which is strong. However, we can observe certain insignificant ($p > 0.05$) relations. For example, whether a driveway is present or not doesn't relate to the number of bedrooms, and whether the house is in a preferred location doesn't relate to the number of garages. Hence, this model may not be a good fit. The second model produced is one where certain assumptions have been made:

1. The lot size is related to the presence of a driveway and (maybe) a recreational room since,
    a. A larger driveway would require a bigger plot of land, and
    b. The recreational room may make the house size bigger, requiring more land.
2. The number of bedrooms/bathrooms are related to a fully finished basement, since
    a. More bedrooms/bathrooms could be added to the basement.
3. The number of stories is related to the presence of a recreational room, since
    a. The recreational room may be big enough to need its own floor.

This leads us to

lm.price2 <- lm(price ~ lotsize*(driveway + recreation) + bedrooms*(fullbase) + bathrooms*(fullbase) + stories*(recreation) + garage*(driveway))

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -497.3484  2749.2067  -0.181 0.856510
lotsize           3.5959     0.3498  10.279  < 2e-16 ***
bathrooms     14924.1569  1454.2442  10.262  < 2e-16 ***
stories        7128.7990   867.3174   8.219 1.55e-15 ***
drivewayyes    6259.6177  2034.4638   3.077 0.002200 **
recreationyes  4440.4102  1903.1824   2.333 0.020010 *
fullbaseyes    5846.5080  1574.9946   3.712 0.000227 ***
gasheatyes    12949.4428  3223.0822   4.018 6.72e-05 ***
airconyes     12605.9217  1557.9379   8.091 3.98e-15 ***
garage         4355.3216   839.7822   5.186 3.05e-07 ***
preferyes      9431.7782  1671.9235   5.641 2.74e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15450 on 535 degrees of freedom
Multiple R-squared:  0.6712,	Adjusted R-squared:  0.6651
F-statistic: 109.2 on 10 and 535 DF,  p-value: < 2.2e-16
```

Fig1.3: Summary of lm.price2 with associated p-values

Here, we have a much simpler model, taking less relations into account. We observe that, whilst we have less variability being explained (now 66.51%), all our relations (and quantitative variables) are statistically significant ($p < 0.05$). But how do we know which model is 'better'?

To do this, we use the anova function in R to check H_0: all factor relations in lm.price that aren't in lm.price2 are equal to 0.

```
Analysis of Variance Table

Model 1: price ~ (lotsize + bedrooms + bathrooms + stories + garage) *
    (driveway + recreation + fullbase + gasheat + aircon + prefer)
Model 2: price ~ lotsize * (driveway + recreation) + bedrooms * (fullbase) +
    bathrooms * (fullbase) + stories * (recreation) + garage *
    (driveway)
  Res.Df        RSS  Df   Sum of Sq      F    Pr(>F)
1    504 1.0918e+11
2    531 1.4886e+11 -27 -3.9681e+10 6.7841 < 2.2e-16 ***
---
```

Fig1.4: ANOVA test between lm.price and lm.price2

Our corresponding p value is less than 0.05, meaning that we reject H_0 and keep the first model. Let us delve into lm.price and see whether removing the insignificant relations changes anything. This has to be a step-by-step process, removing the most insignificant factor, and plotting the new model repeatedly until we are satisfied, leading to lm.price3;

lm.price3 <- lm(price ~ lotsize*(recreation+fullbase) + bedrooms*(gasheat) + bathrooms*(driveway+aircon+prefer) + stories+ garage*(recreation+gasheat+aircon))

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.419e+04  5.333e+03   2.661 0.008032 ** 
lotsize              3.247e+00  3.981e-01   8.155 2.59e-15 ***
recreationyes        2.032e+04  5.990e+03   3.392 0.000746 ***
fullbaseyes         -3.465e+03  3.625e+03  -0.956 0.339592    
bedrooms             2.317e+03  1.030e+03   2.249 0.024906 *  
gasheatyes          -2.037e+04  1.224e+04  -1.665 0.096541 .  
bathrooms            4.228e+02  3.609e+03   0.117 0.906787    
drivewayyes         -4.385e+03  5.122e+03  -0.856 0.392317    
airconyes           -2.345e+03  3.993e+03  -0.587 0.557174    
preferyes           -6.404e+02  4.379e+03  -0.146 0.883775    
stories              6.571e+03  8.987e+02   7.311 9.94e-13 ***
garage               1.401e+03  1.061e+03   1.321 0.187006    
lotsize:recreationyes -3.408e+00  1.066e+00  -3.195 0.001480 ** 
lotsize:fullbaseyes  1.883e+00  6.583e-01   2.861 0.004396 ** 
bedrooms:gasheatyes  7.671e+03  3.793e+03   2.023 0.043619 *  
bathrooms:drivewayyes 9.463e+03  3.869e+03   2.446 0.014783 *  
bathrooms:airconyes  8.833e+03  2.822e+03   3.130 0.001848 ** 
bathrooms:preferyes  7.086e+03  3.082e+03   2.300 0.021862 *  
recreationyes:garage 5.254e+03  2.095e+03   2.508 0.012441 *  
gasheatyes:garage    1.148e+04  3.967e+03   2.895 0.003949 ** 
airconyes:garage     3.997e+03  1.584e+03   2.523 0.011936 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14650 on 525 degrees of freedom
Multiple R-squared:  0.7101,	Adjusted R-squared:  0.699 
F-statistic: 64.28 on 20 and 525 DF,  p-value: < 2.2e-16
```

Fig1.5: Summary of lm.price3 with associated p-values

Whilst we are happy with the factor relations, we notice that bathrooms is now very insignificant alongside most factor variables. We will leave this in for now as we now have a model explaining 69.9% of the variability, and compare lm.price3 with its predecessor lm.price with H_0: all factor variables in lm.price that aren't in lm.price3 are equal to 0.

```
Analysis of Variance Table

Model 1: price ~ (lotsize + bedrooms + bathrooms + stories + garage) * 
    (driveway + recreation + fullbase + gasheat + aircon + prefer)
Model 2: price ~ lotsize * (recreation + fullbase) + bedrooms * (gasheat) + 
    bathrooms * (driveway + aircon + prefer) + stories + garage * 
    (recreation + gasheat + aircon)
  Res.Df        RSS  Df Sum of Sq      F Pr(>F)
1    504 1.0918e+11                            
2    525 1.1267e+11 -21 -3.49e+09 0.7671 0.7609
```

Fig1.6: ANOVA test between lm.price and lm.price3

We now have a p-value of 0.76, hence failing to reject H_0, so we choose lm.price3 as our main model, since it beats lm.price, which in turn beats lm.price2.

Results and Interpretation

some key results from Figure 1.5 show us that,

- Aircon is present,
  - Each garage increases the price by $3997.
  - Each bathroom increases the price by $8833.
- Gas heating is present,
  - Each garage increases the price by $11480.
- A recreation room is present,
  - Each garage increases the price by $5254.
- The House is in a preferred location,
  - Each bathroom increases the price by $7086.
- A Driveway is present,
  - Each bathroom increases the price by $9463.

Also, a house with nothing has an instant value of $14190, which may be buyer/seller fees incurred once a transaction is made.

Limitations of the Model

Some limitations of lm.price3 is that if the quality of data is poor, the model will be very sensitive to outliers. Prices of Houses are mainly dependent on real estate agents and the buyer/seller. Some homes may have sentimental value to sellers, thus overvaluing their properties. This can clearly be seen by the plot of price to lot size below,

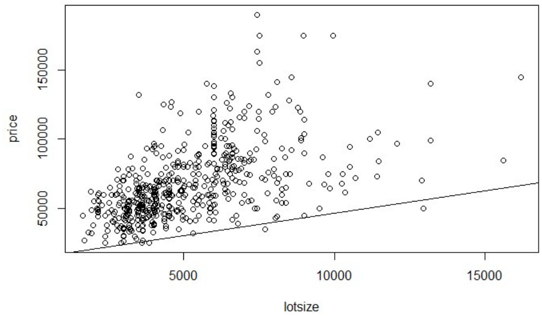

Fig3.1: plot of price to lot size. The black line is lm.price3

Also, this model has been based on a singular city in Canada during July-September 1987. Consequently, this model may not be reproducible for any other city in the world, as House prices can also be based on geographical locations. Furthermore, the data is very outdated, as 36 years ago, the value of the Canadian Dollar was very different, and inflation caused mortgage rates to go up as well, continuing to overvalue properties. This all means that lm.price3 will not represent todays housing markets accurately.

Improvement of Analysis

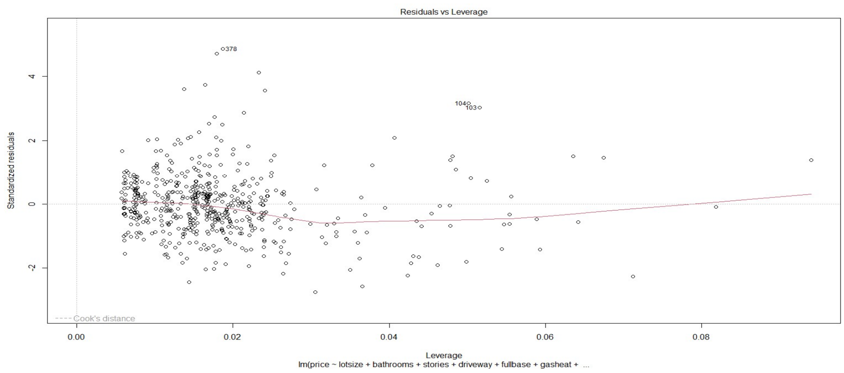

Fig4.1: Graph of Residuals vs Leverage. Cook's distance in red

Whilst lm.price3 is our chosen model, we need to analyse its Cook's Distances. In Figure 4.1, we observe the Cook's distance stays within a [-0.5,0.5] range which indicates that although this is an acceptable model, there are influential points which would be a cause for further examination, especially in the leverage range of (0.02,0.08). This may reveal a more accurate model. We can also observe that from 0.06, the cook's distance starts to linearly increase, which may continue past the graph's limits, so analysing further would also be recommended.

Personally, I believe that the model would be more dependent on the lot size, as properties are priced at dollars/sq feet. Maybe analysing a model that had taken less variables than lm.price3, such as only having lot sizes and bedrooms, could've presented a more true representation. However, this is because Figure 3.1 presents a linear model that shows that house prices are overvalued.