University of Edinburgh
School of Mathematics

Generalised Regression Models

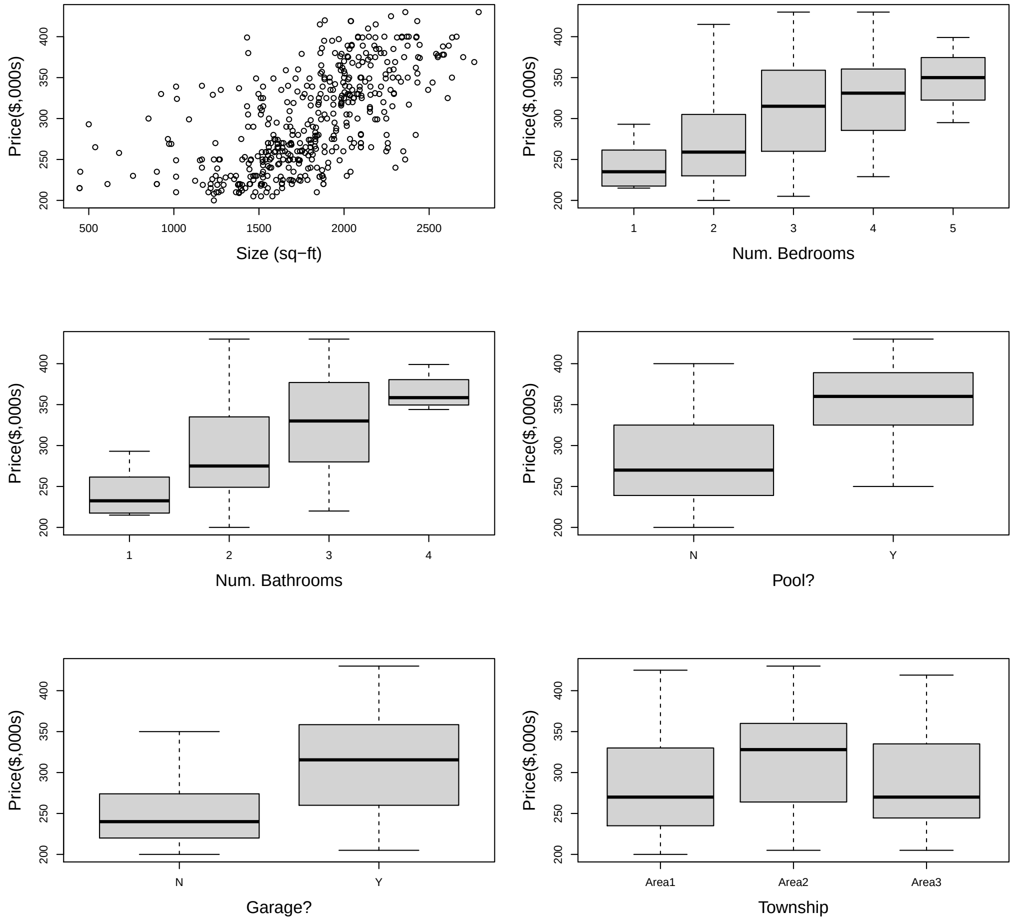
GRM: Case Study — Linear Models **Semester 1, 2022–2023**

FEEDBACK

Explore the data

Load the data and create scatter plots/boxplots of property prices against the other variables:

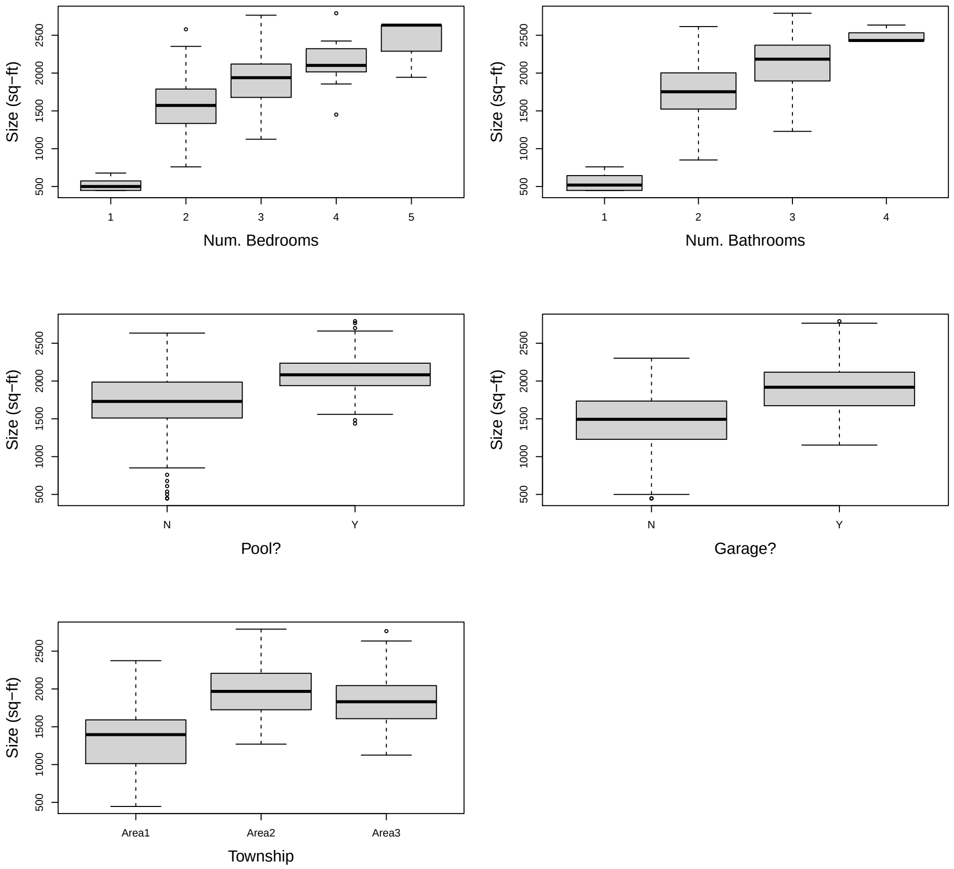
```
RealEstate <- read.csv(file = "RealEstate.csv")
plot(RealEstate[, "Size"], RealEstate[, "Price"])
boxplot(RealEstate[, "Price"] ~ RealEstate[, "Bedroom"])
...
```



Quiz 1: Price is positively related with property size (correlation: 0.629), number of bedrooms and number of bedrooms. On average, properties with a pool have higher prices, and likewise for properties with a garage. Township area 1 and 3 have similar property prices, whilst prices are higher on average for township area 2.

Explore what relationships exist amongst the other variables:

```
boxplot(RealEstate[, "Size"] ~ RealEstate[, "Bedroom"])
boxplot(RealEstate[, "Size"] ~ RealEstate[, "Bathroom"])
...
```



Quiz 2: Properties that have a larger number of bedrooms and bathrooms tend to be larger in size. Larger properties are more likely to have a pool and/or a garage. Properties in township area 1 are typically smaller in size, whilst areas 2 and 3 are similar property sizes on average.

Extra: Relationships between discrete/categorical variables can be examined by tabulating the data:

```
table(RealEstate[,"Pool"], RealEstate[,"Garage"])
      N   Y
  N 103 254
  Y   3  94
```

Only 3 properties have a pool but do not have a garage.

Building the linear regression model

Simple linear regression model for prices based on property size:

```
Model1 <- lm(formula = Price ~ Size, data = RealEstate)
summary(Model1)

Call:
lm(formula = Price ~ Size, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max
-103.870  -35.515   -4.834   34.578  135.867

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.307e+02  9.907e+00   13.20   <2e-16 ***
```

```
Size        9.259e-02  5.378e-03   17.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.05 on 452 degrees of freedom
Multiple R-squared:  0.396,     Adjusted R-squared:  0.3947
F-statistic: 296.4 on 1 and 452 DF,  p-value: < 2.2e-16
```

From the summary it is clear that property size is an important covariate in this model with a very significant estimated coefficient. The fitted regression line is $\mathbb{E}[Y] = 130.7 + 0.09259x$ where x represents the size of the property.

Multiple linear regression model for prices with size and pool covariates:

```
Model2 <- lm(formula = Price ~ Size + Pool, data = RealEstate)
summary(Model2)

Call:
lm(formula = Price ~ Size + Pool, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max
-93.402 -32.520  -4.439  29.903 138.967

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.512e+02  9.617e+00  15.723  < 2e-16 ***
Size        7.610e-02  5.435e-03  14.001  < 2e-16 ***
PoolY       4.289e+01  5.330e+00   8.048 7.53e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.11 on 451 degrees of freedom
Multiple R-squared:  0.4719,    Adjusted R-squared:  0.4695
F-statistic: 201.5 on 2 and 451 DF,  p-value: < 2.2e-16
```

Both property size and pool covariates are significant in this model. The fitted regression line represents two parallel lines of $\mathbb{E}[Y] = 194.1 + 0.0761x$ and $\mathbb{E}[Y] = 151.2 + 0.0761x$ depending on whether there is or is not a pool at the property.

Full linear regression model with all available covariates:

```
ModelAll <- lm(formula = Price ~ ., data = RealEstate)
summary(ModelAll)

Call:
lm(formula = Price ~ ., data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max
-102.712  -26.820   -2.725   26.914  144.984

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   168.299517  10.847095  15.516  < 2e-16 ***
Size            0.087686   0.007409  11.835  < 2e-16 ***
Bedroom        -4.589244   3.823238  -1.200    0.231
Bathroom       -1.356759   5.301431  -0.256    0.798
PoolY          37.506282   5.005961   7.492 3.67e-13 ***
GarageY        26.750508   5.322853   5.026 7.28e-07 ***
TownshipArea2 -40.297300   6.857973  -5.876 8.23e-09 ***
```

```
TownshipArea3 -54.531966   6.237602  -8.742  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.61 on 446 degrees of freedom
Multiple R-squared:  0.5592,    Adjusted R-squared:  0.5523
F-statistic: 80.84 on 7 and 446 DF,  p-value: < 2.2e-16
```

Property size, pool, garage and township are significant covariates in this model in describing the variation we see in the property prices. However the coefficients for the number of bedrooms and bathrooms are not significant. This suggests that we have over-fitted the data and we can obtain a simpler model that is equally as good as describing property prices.

Results from performing stepwise variable selection:

```
ModelBest <- step(ModelAll, trace = 0) #"trace = 0" suppresses printout
summary(ModelBest)

Call:
lm(formula = Price ~ Size + Pool + Garage + Township, data = RealEstate)

Residuals:
   Min     1Q Median     3Q    Max
-99.16 -26.19  -3.58  27.33 146.62

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   161.733755   8.957585  18.056  < 2e-16 ***
Size            0.082696   0.006061  13.644  < 2e-16 ***
PoolY          37.660219   4.932017   7.636 1.37e-13 ***
GarageY        26.666906   5.306996   5.025 7.29e-07 ***
TownshipArea2 -40.061330   6.849435  -5.849 9.56e-09 ***
TownshipArea3 -54.274291   6.230444  -8.711  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.6 on 448 degrees of freedom
Multiple R-squared:  0.5575,    Adjusted R-squared:  0.5525
F-statistic: 112.9 on 5 and 448 DF,  p-value: < 2.2e-16
```

Quiz 3: Covariates are removed/added one at a time and the model is re-evaluated, the model that maximises the Akaike information criterion (AIC) is retained for the next step. The algorithm ends once removing/adding no longer improves the AIC. The AIC is defined as $2k - 2\ell(\hat{\beta})$ where k is the number of parameters, so the best model is determined by that which best describes the data (via the log-likelihood) but is not too complicated (via the number of parameters).

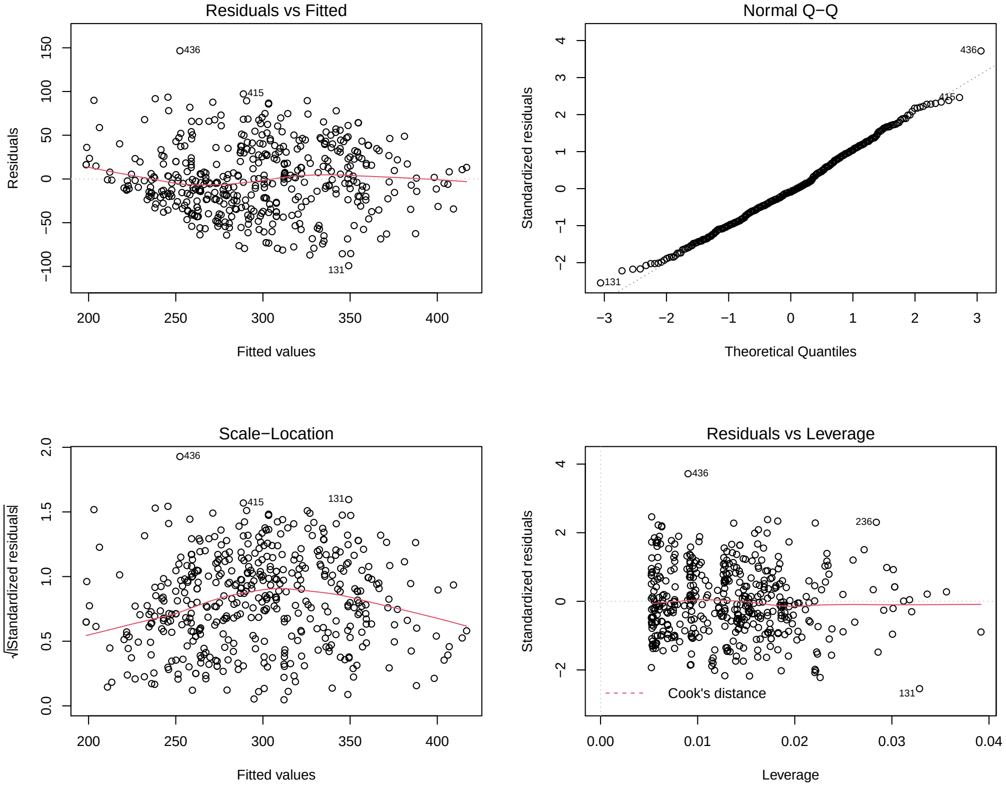
Quiz 4: The kept variables are property size, is there a pool and/or garage, and the township area. The number of bedrooms and bathrooms are the eliminated variables. This sounds counter intuitive as we would have though that these should be important in determining property prices. However, from exploring the data we saw that the number of bedrooms and bathrooms are related to property size, so it is likely to be sufficient to have only one descriptor for how big the property is.

Quiz 5: Both township estimates in the best model are negative, indicating that property prices in area 2 and 3 are lower than in area 1. Therefore township area 1 has higher prices after accounting for property features. This is a different description compared to the earlier discussion (see Quiz 1) which is likely due to the different types of properties within the three areas.

Validating assumptions

Assessing the linear regression assumptions for the best model:

```
par(mfrow = c(2, 2))
plot(ModelBest)
```



Quiz 6: From the plots, the linear regression assumptions appears to be valid:

- The residuals appear to be randomly scattered around the zero line with no obvious deviations. Therefore the zero expectation of the residuals appear to be justified.
- The points in the normal Q-Q plot lie on the 1:1 line with the slight deviation in the tails not being too large. The normality assumption appears to be applicable.
- The trend in the scale-location plot (red line) appears to be larger for mid-range fitted values, however there is sufficient variation in the scatter plot that a constant line through the image could appear to be a reasonable alternative. It is therefore reasonable to assume that the variance of the residuals is constant.
- *Extra:* Leverage gives a measure of how important each datum is in the fitted regression model. As all leverage estimates are close to zero, then there are no entries in the spreadsheet that would result in a substantial change in the model if it was to be removed.