University of Edinburgh
School of Mathematics

Generalised Regression Models

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

Problem and Data

The spreadsheet `RealEstate.csv` contains details about 454 properties on the market in Venice on the west coast of Florida, USA. The information about each property is described below:

Variable	Description
Price	The property price (in thousands USD).
Size	Size of the property (in square feet).
Bedroom	Number of bedrooms in the property.
Bathroom	Number of bathrooms in the property.
Pool	Does the property have a pool? (`Y/N`)
Garage	Does the property have a garage? (`Y/N`)
Township	A categorical variable stating which of three regions in the city the property is located. (`Area1`/`Area2`/`Area3`)

The objective of this case study is to investigate what combination of variables can best explain the variation seen in the property prices.

Statistical Model

In this investigation we shall be fitting a linear regression model with property price as the response variable. The linear regression model is defined as:

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \quad \text{for} \quad i = 1, \ldots, n, \quad \text{where} \\ \mu_i &= \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} \end{aligned}$$

Specifically, we assume that the random variables for the property prices (Y_i) are independent normally distributed with common variance, but with unique mean that is a linear function of the covariates $(x_{1,i}, \ldots, x_{p,i})$. For the real estate data, we want to estimate the regression coefficients $(\beta_0, \beta_1, \ldots, \beta_p)$ in order to describe how important each covariate is in explaining the variation seen in the property prices.

R analysis

Begin by downloading and saving the "RealEstate.csv" file and read it into R using:

```
RealEstate <- read.csv(file = "RealEstate.csv")
```

Remark: If the location of the file is not the same as your working directory then you will either have to change the working directory to where you have saved the data file or write the full file path inside the string.

Explore the data

Before fitting the regression models, let's first create some plots to explore what relationships exist within the data. To examine what relationships exist with property prices we can make a scatter plot against size and create a set of boxplots with respect to the bedroom variable:

```
plot(RealEstate[, "Size"], RealEstate[, "Price"])
boxplot(RealEstate[, "Price"] ~ RealEstate[, "Bedroom"])
```

It is clear to see that there is a positive correlation between prices and size, and that prices tend to increase with respect to the number of bedrooms.

Quiz 1: Use boxplots to examine how property prices vary with respect to the other variables. Are prices the same in all township areas? If not, how do they differ?

It is important to also useful to examine if any relationships exists amongst the covariates. This can help in identifying any potential collinearity issues when fitting the regression model.

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

Quiz 2: Describe how property size varies with respect to the number of bedrooms and bathrooms.

Building the linear regression model

To fit a linear regression model we use the `lm()` command. The input are:

- `formula` A symbolic description of the model to be fitted in the form `y ~ x` where `y` is the response variable and `x` is some linear combination of the explanatory variable.
- `data` A data frame containing the variables in the model.

Once the regression model has been fitted, we can use the `summary()` command to examine the coefficient estimates and assess the significance of the covariates.

Let's begin by fitting a simple linear regression model property prices (y) with just property size (x):

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

The fitted regression line is $\mathbb{E}[Y] = 130.7 + 0.09529x$, meaning that expected property price increases by \$95.29 for one additional square foot in size. This is affirms the positive relationship seen earlier.

It is simple to fit a multiple linear regression model by adding extra variable names into the formula:

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

From examining the output we see that there is an estimate for `PoolY` (yes there is a pool), but not for `PoolN`. For categorical variables like 'Pool', 'Garage' and 'Township' there will always be 1 fewer parameters than the number of options. The presented estimates describe the *change* in the regression line from a *baseline case* (without a pool) to the case of interest (with a pool). From the summary, the fitted regression lines are:

$$\mathbb{E}[Y] \quad = \quad \begin{cases} 151.2 + 0.0761x & \text{without pool} \\ 194.1 + 0.0761x & \text{with pool} \end{cases}$$

This represents an increase of \$42,900 expected property prices if the property has a pool. Note that the estimated coefficient for the size of the property has changed after accounting for the existence of a pool.

We can fit a linear regression model with all of the covariates in the real estate data frame by placing a dot (`.`) in the formula:

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

We see from the `Pr(>|t|)` column in the summary table that the p-value for the bathroom coefficient is 0.9001, meaning that (with all else held constant) it is not possible to reject the null hypothesis at the 5% significance level that the bathroom coefficient is zero. Consequently the covariate for the number of bathrooms is not important in the model after accounting for all of the other covariates and so we may consider simplifying the regression model by remove this covariate. After re-fitting, we can then assess the significance of the remaining variables.

To help in performing variable selection, we use the `step()` command on the full. To determine which covariates to keep or remove we use the `step()` command that is a stepwise algorithm that performs variable selection:

```
ModelBest <- step(ModelAll)
summary(ModelBest)
```

Quiz 3: What criteria is `step()` using to perform variable selection? (Hint: run `help(step)`)

Quiz 4: Which covariates are kept and which have been removed? Are the removed covariate what you would expect to be not important for modelling property prices?

Quiz 5: After accounting for property features, which township area(s) has the most expensive properties? Is this what you expected from your earlier boxplots?

Validating assumptions

It is always important to validate that the modelling assumptions. The easiest way is to create series of images using the `plot()` for the best model. With these we can assess the three key assumptions of the residuals of zero expectation (Residuals vs Fitted), normality (Normal Q-Q) and constant variance (Scale-Location).

```
par(mfrow = c(2, 2)) #Sets the plotting window to a 2x2 grid
plot(ModelBest)
```

Quiz 6: Is there evidence to suggest that the linear regression assumptions are satisfied?

Remark: Run `par(mfrow = c(1, 1))` to reset the plotting window for presenting single images.