Skipass price: inference and prediction

Alex Costa | Università Cattolica del Sacro Cuore | Applied Linear Models

08/03/2024

1. Dataset description and goals

The purpose of this paper is to study how the price of a ski pass is influenced by the features of the associated ski resort. This information can be used to suggest a price more in line with the resort's features, which could lead to higher revenues. In order to achieve this goal we analyze the "Ski Resorts" dataset, which contains information about several features of 499 ski resorts in 5 different continents in the year 2022. The dataset was taken from Kaggle at the following link: https://www.kaggle.com/datasets/ulrikthygepedersen/ski-resorts. The dataset contains 25 variables:

- ID: resort ID;
- Resort: resort name;
- Latitude: resort latitude;
- Longitude: resort longitude;
- Country: resort country;
- Continent: resort continent;
- Price: ski pass cost for 1 adult for 1 day in the mean season in euro;
- Season: start and end month of the ski season in that resort in 2022;
- Highest point: resort's highest point in meters;
- Lowest point: resort's lowest point in meters;
- Beginner slopes: total length in km of children, blue, and green slopes at the resort;
- Intermediate slopes: total length in km of red slopes at the resort;
- Difficult slopes: total length in km of black, advanced, and expert slopes at the resort;
- Total slopes: total slope length in km;
- Longest run: longest run in km;
- Snow cannons: number of snow cannons;
- Surface lifts: total number of surface lifts, including T-bars, Sunkids lifts, Rope lifts and people movers;
- Chair lifts: number of chair lifts;
- Gondola lifts: number of gondola lifts;
- Total lifts: total number of lifts;
- Lift capacity: number of passengers the resort's lift system can move in an hour;
- Child friendly: the resort is child friendly or not;
- Snowparks: the resort has a snowpark or not;
- Night skiing: the resort offers skiing on illuminated slopes or not;
- Summer skiing: the resort offers skiing during the summer or not.

The number of observations and the number of variables are sufficient to allow us to conduct an adequate analysis. We omit the variables ID and Resort because they are just identifiers, and Country because it would be a categorical variable with 38 levels and we already have the variable Continent. Furthermore, we turn the variable Season into Season duration, a categorical variable with 3 levels: Short (from 1 to 4 months), Medium (5 or 6 months) and Long (from 7 to 12 months). We also group the levels Asia, Oceania and South

America of the variable Continent into one level called Other, since these levels have very small numbers of observations compared to Europe and North America. Finally, we change the unit of measurement of Highest point and Lowest point from meters to kilometers by dividing both by 1000.

2. Exploratory Data Analysis

First of all we perform an exploratory analysis in order to summarize and visualise the main characteristics of the dataset. Some of the most relevant plots are shown below.

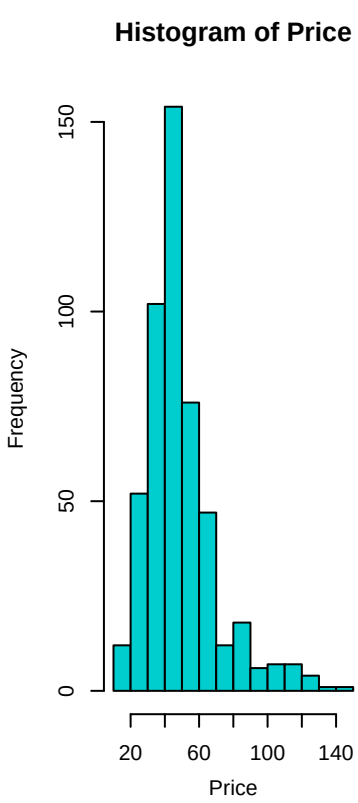


Figure 1

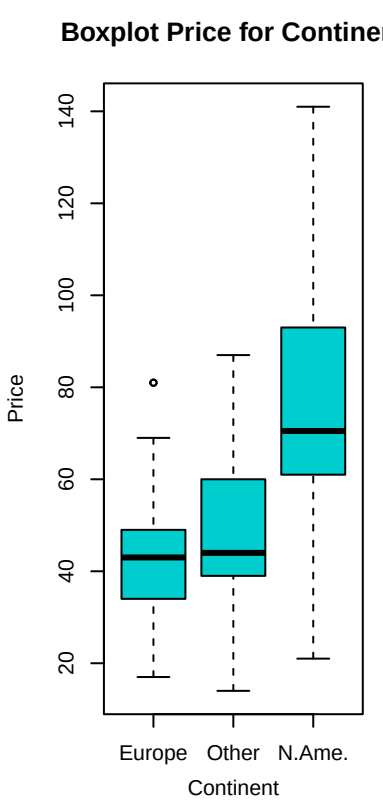


Figure 2

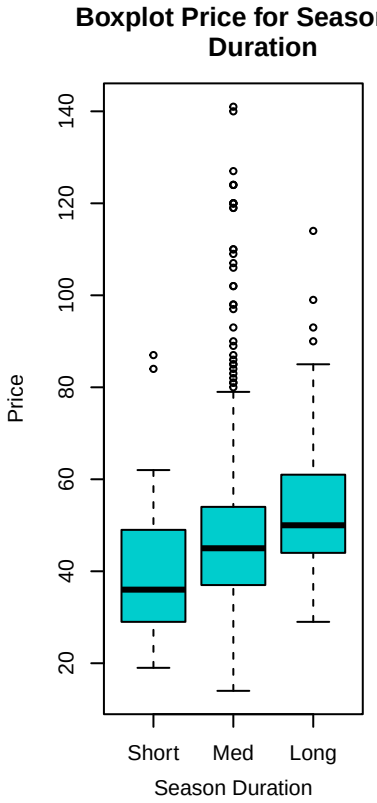


Figure 3

The response variable, Price, has a right-skewed distribution with a median of 45€ and a mean of 49.6€. We can see from the Figures 2 and 3 that Price has different quartiles depending on the Continent and the Season Duration. In particular, ski resorts in North America and with Long season durations have an higher median price compared to the others.

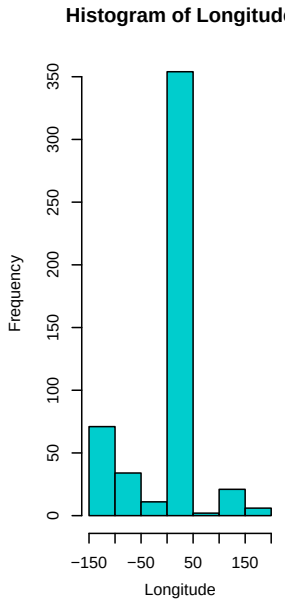


Figure 4

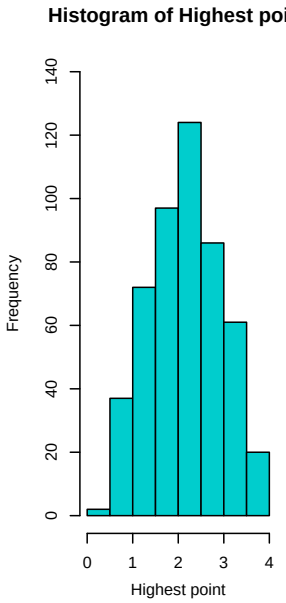


Figure 5

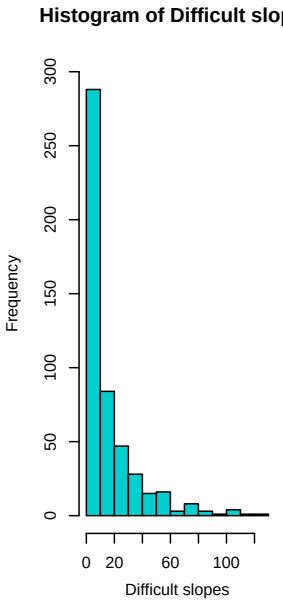


Figure 6

Most of the observations have a Longitude in the interval 0 - 50 (Figure 4). These are almost all the european ski resorts, which indeed are 360 out of 499. The variable Highest point (Figure 5) has an almost symmetrical distribution with a mean of 2161m and a median of 2175m. From Figure 6 we can see that the majority of the ski resorts have at most 20km of Difficult slopes and just 4% of the resorts have more than 60km.

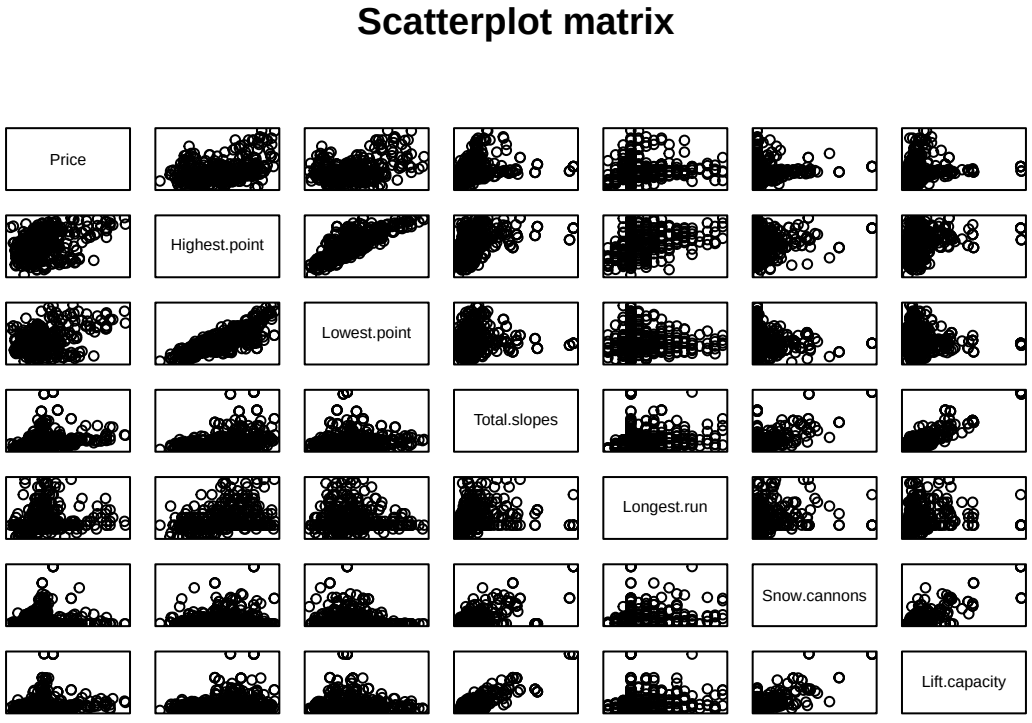


Figure 7

From Figure 7 we can visualize the relationship between the response variable and some of the continuous predictors and between these predictors themselves. Price seems to have a relationship in particular with Highest point, and there is a clear positive linear correlation between Highest point and Lowest point and between Total slopes and Lift capacity. These last two relationships makes total sense: the higher the maximum elevation, the higher the minimum elevation; the more km of slopes there are, the higher the lift capacity must be.

3. Best subset selection

We center all the predictors, that is, we subtract the mean of that predictor from each value, in order to have a more sensible interpretation of the coefficients that we are going to estimate later. Then, we fit a linear regression model with Price as response and all the other variables as predictors and we perform a best subset selection. We don't include Total Slopes because it is the sum, so a linear combination, of Beginner, Intermediate and Difficult slopes, and Total Lifts that is the sum of Surface, Chair and Gondola lifts.

The best subset selection gives as output the best model, so the one with smallest RSS, for each of the possible number of predictors. In our case the best model with 1 predictor contains the variable Continent North America, the best model with 2 predictors contains Continent North America and Difficult Slopes, and so on until the full model with all the 21 predictors.

Now we have to choose the best overall model between the 21 models previously selected. We first look at some measures of goodness of fit: AIC, BIC, Mallow's Cp and Adjusted R^2.

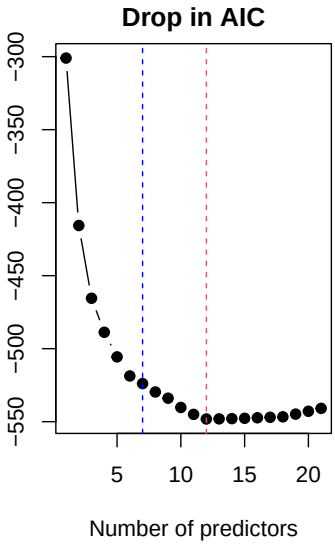


Figure 8

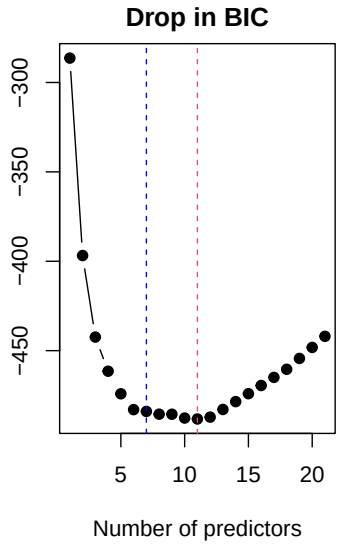


Figure 9

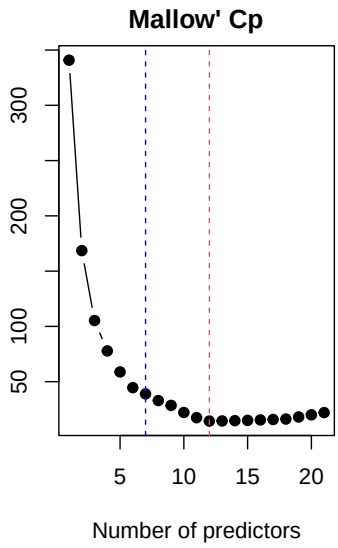


Figure 10

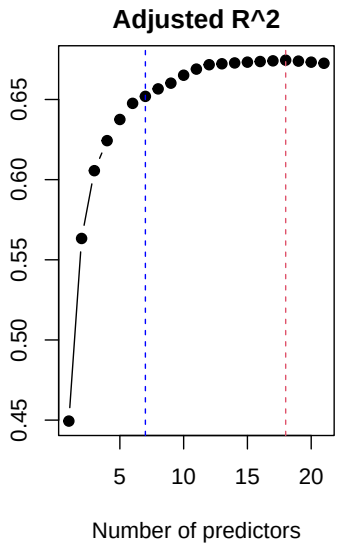


Figure 11

The model with lowest AIC, BIC and Mallow's Cp is the one with 12 predictors, while the one with the highest Adjusted R^2 is the model with 18 predictors (red vertical lines). We also look at a measure of prediction accuracy: the Cross-Validation error. In particular, we consider the Mean Squared Error, computed with the Leave One Out method.

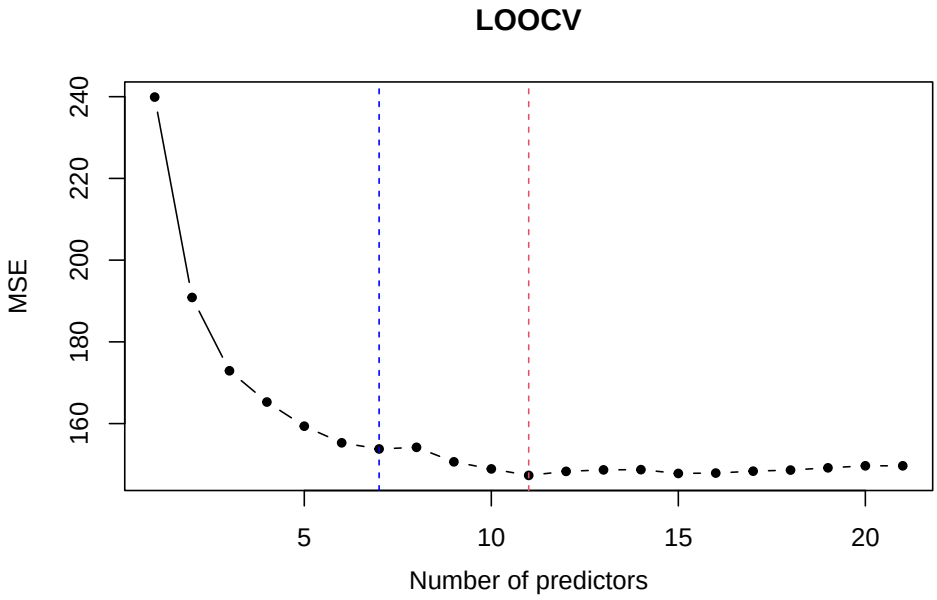


Figure 12

The model with the lowest MSE, and therefore the best in terms of prediction performance, is the model with 11 predictors. Now, applying the Occam's razor principle we choose the model with 7 predictors because it is the number of predictors after which all the analyzed metrics more or less flatten out (blue vertical lines). These 7 predictors, according to the best subset selection are: Longitude, Continent North America, Highest Point, Difficult Slopes, Longest Run, Snowparks and Summer Skiing:

4. Collinearity

We check for potential collinearity issues between the selected predictors. We first look at the correlation matrix and in particular to its associated corrplot to visualize the correlation between the continuous predictors.

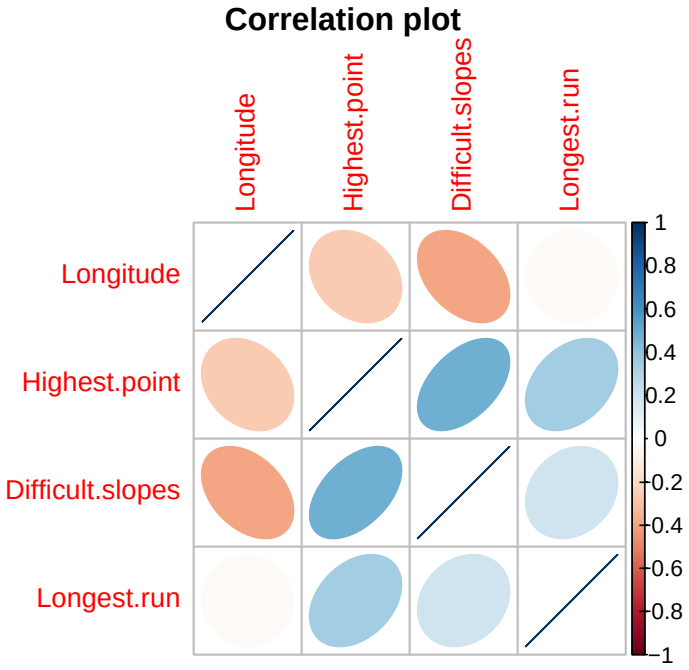


Figure 13

Difficult slopes has a slight correlation with Highest point (cor = 0.48) and Longitude (cor = -0.39). However, these are acceptable levels of correlation that should not cause problems. We check also the VIF (Variance of Inflation Factor), which is another measure of correlation, also suitable for categorical variables.

```
##        Longitude        Continent    Highest.point Difficult.slopes
##         3.503373         3.530294         1.543000         1.615372
##      Longest.run        Snowparks    Summer.skiing
##         1.196104         1.059099         1.087773
```

All the predictors have a VIF much lower than 10, which is the most widespread rule of thumb for deciding if collinearity is high. Hence, we can confirm that we don't have collinearity issues.

5. Diagnostics and unusual observations

We check whether our model respects the assumptions of the normal multiple linear regression model, which, we recall, are: independence, homoscedasticity and normality of the errors and linearity of the model. We start by looking at the scatterplot of residuals vs fitted values.

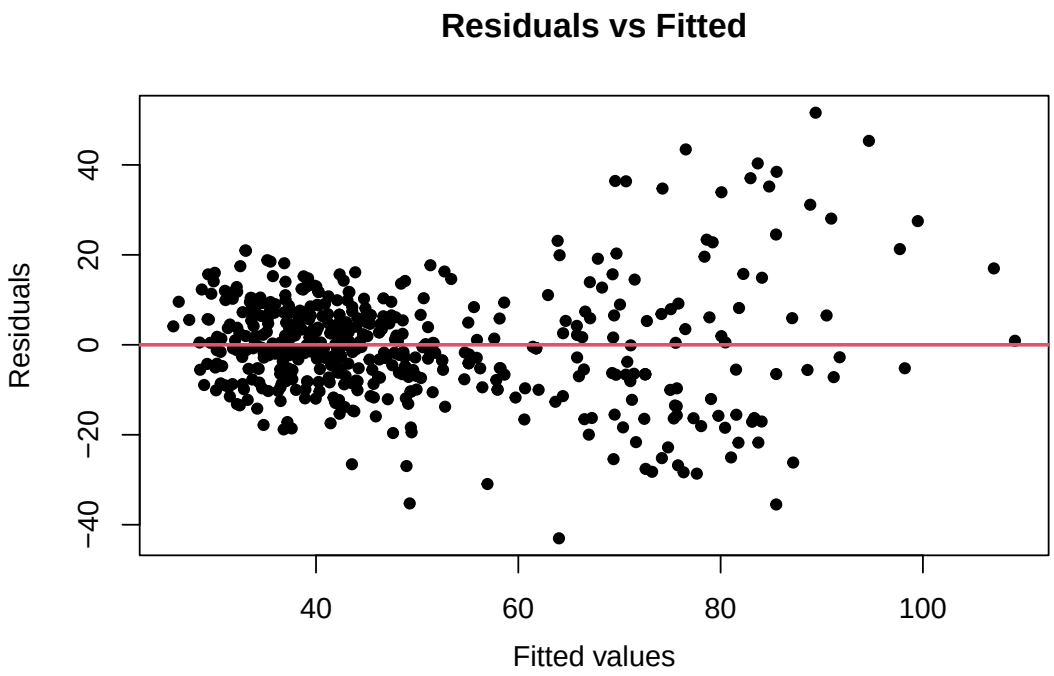


Figure 14

From Figure 14 we can see that the variance of the errors increases as the fitted values increase. Therefore, we have an heteroscedasticity issue. The pattern is also not perfectly linear.

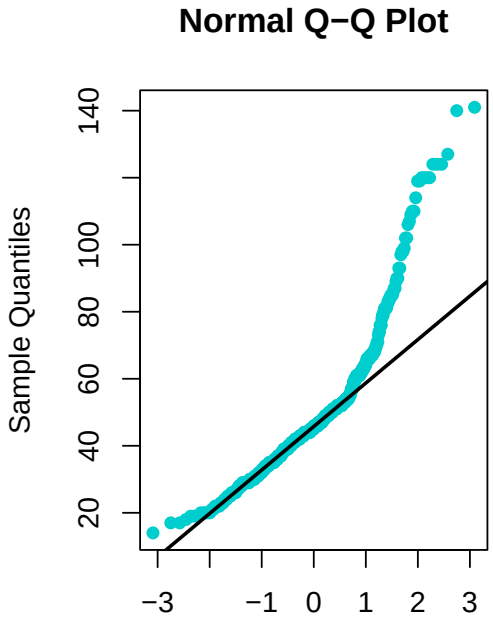


Figure 15

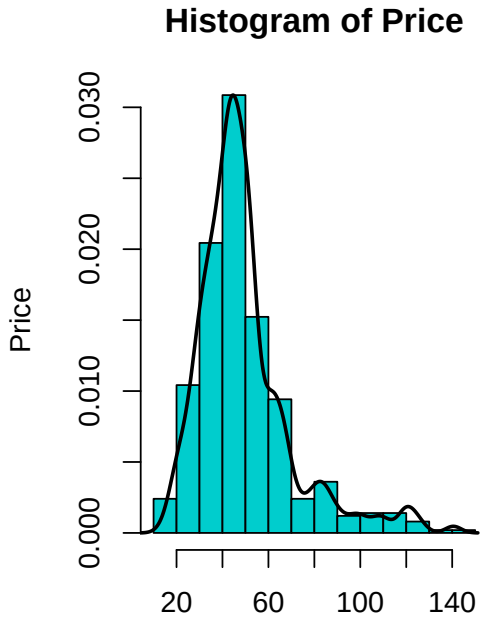


Figure 16

From the Q-Q plot (Figure 15) and the histogram of Price, we can see that the response variable is not

normally distributed, indeed it has a right-skewed distribution, as already mentioned before. This could be a problem because if the assumption of normality of the errors is not met, the reliability of the results in terms of inference (confidence intervals and hypothesis testing) of our model will be limited, so we will try to fix this problem.

Now we check if there are unusual observations: outliers, high leverage points and influential points.

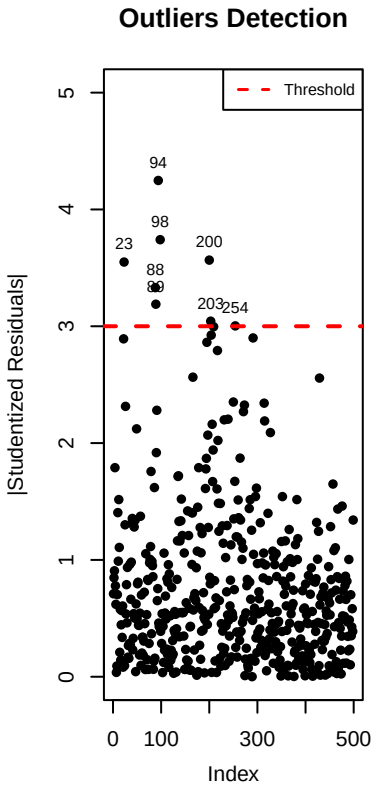


Figure 17

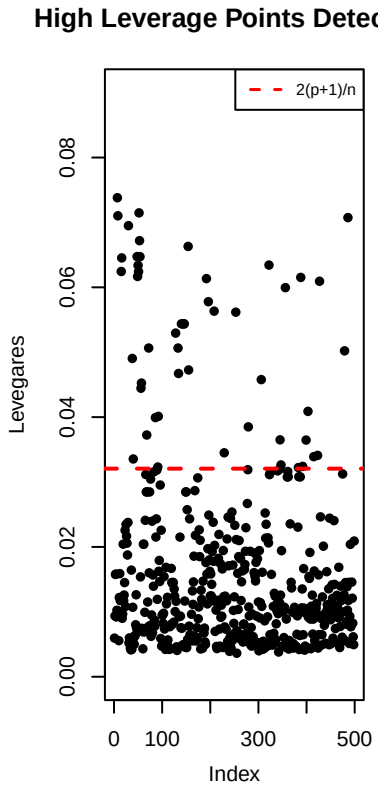


Figure 18

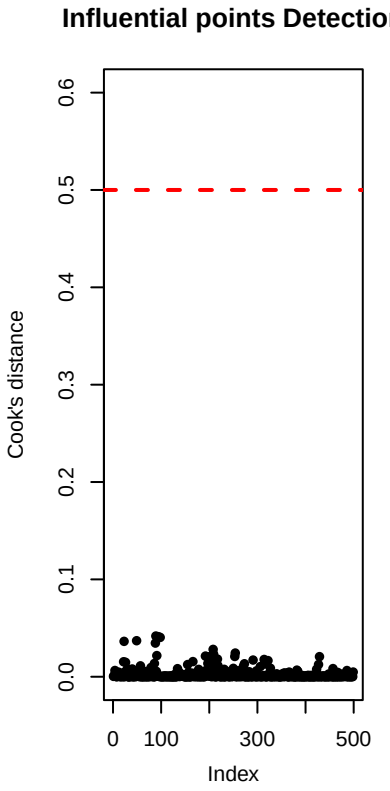


Figure 19

From Figures 17 and 18 we can see that we have quite a few outliers and high leverage points, but none of them are influential (Figure 19), so none of them have a big influence on the fitting and on the inference.

6. Model improvement

We need to fix heteroscedasticity and non normality of the errors. We try some transformations of the response variable that could potentially fix both the issues: $\sqrt{y}$, $\log y$ and $\frac{1}{y}$. The $\log y$ and $\frac{1}{y}$ transformations do not improve heteroscedasticity nor normality, they even make things worse. The square root leads to a very slight improvement for heteroscedasticity:

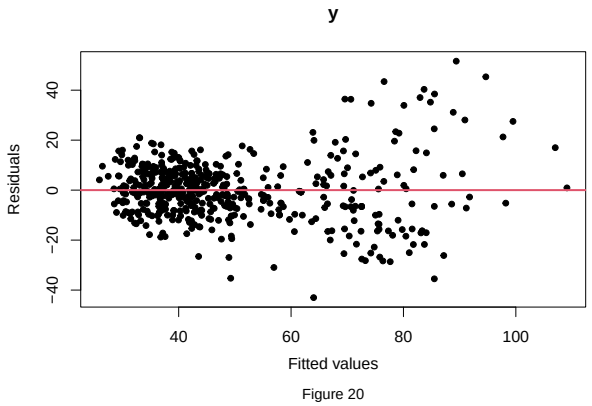


Figure 20

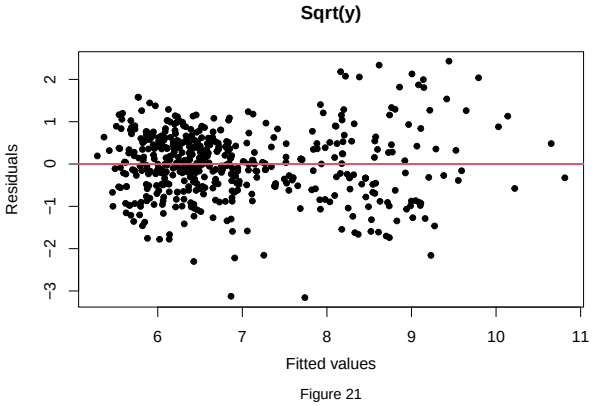


Figure 21

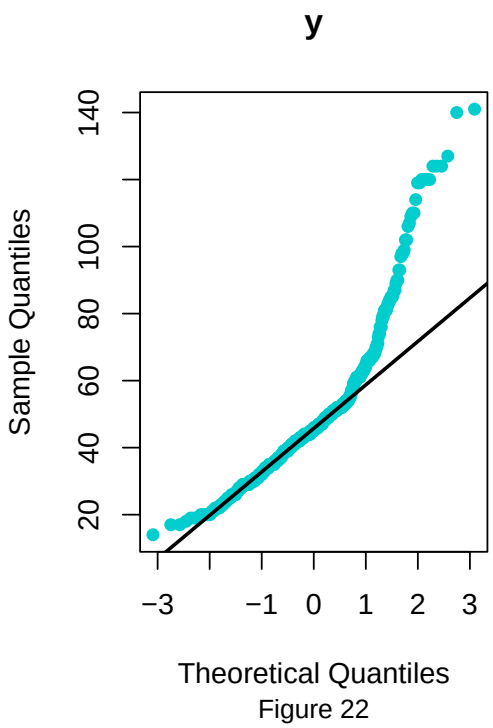


Figure 22

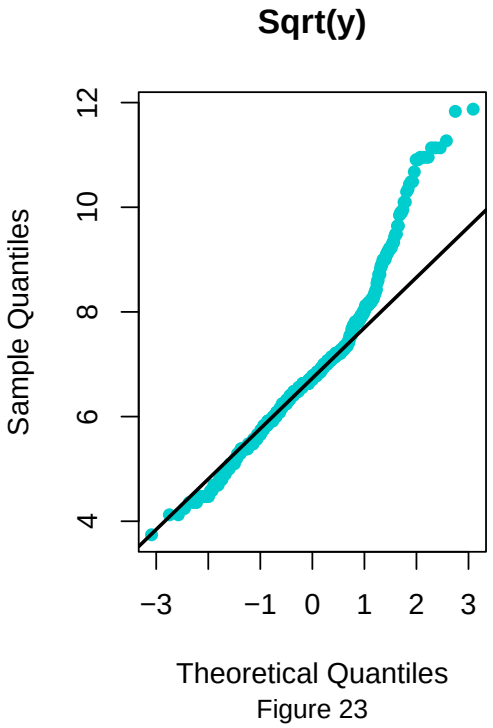


Figure 23

However, these improvements are so small that they do not make a y transformation convenient, so we keep things as they are, aware of the fact that the results in terms of inference will not be completely reliable due to non normality of the response.

7. Coefficients of the best model

We can see below the estimated coefficients of the best model:

```
##          (Intercept)             Longitude ContinentNorth America
##          37.57799263           0.08084752           39.37648484
##       Highest.point      Difficult.slopes            Longest.run
##          4.74382363           0.25472536             0.58976359
```

```
##           SnowparksYes       Summer.skiingYes
##             4.60296867            13.50910369
```

Hence, the best fitted model is:

$$\hat{y}_i = \ 37.58 + 0.08LNi \ + \ 39.38NAi + \ 4.74HPi + 0.25DSi + \ 0.59LRi + \ 4.60SPi + \ 13.51SSi$$

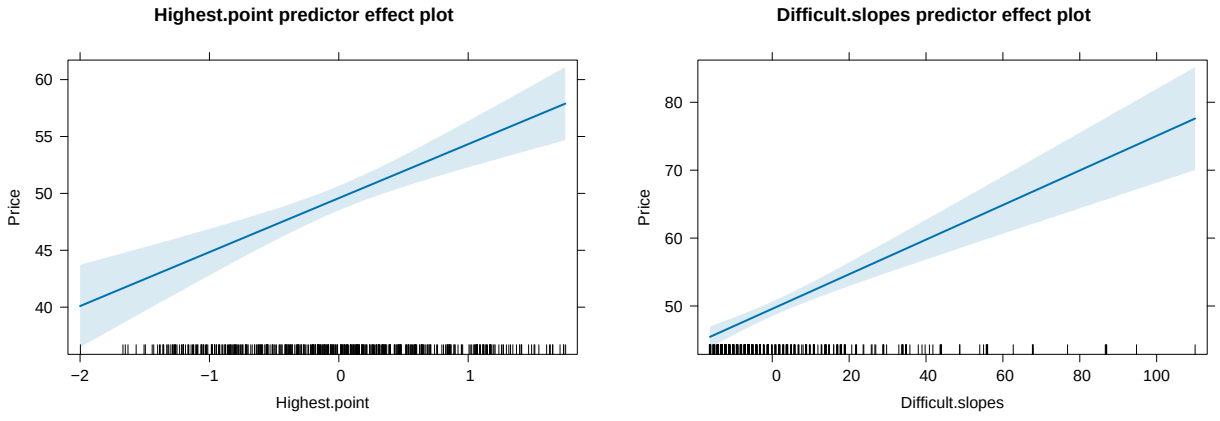
This means that:

- The price of a skipass of a resort which is not in North America, has not a snowpark nor summer skiing, and has longitude, highest point, difficult slopes and longest run equal to their mean values, has an estimated price of 37.58€.
- The price of a skipass is estimated to increase of 0.08€ for every increase of one degree of longitude, all other variables being equal.
- The price of a skipass in North America, all other variables being equal, is estimated to be 39.38€ higher than a skipass in the rest of the world.
- The price of a skipass is estimated to increase of 4.74€ for every increase of 1000m in the highest point, all other variables being equal.
- The price of a skipass is estimated to increase of 0.25€ for every additional km of difficult slopes, all other variables being equal.
- The price of a skipass is estimated to increase of 0.59€ for every additional km in the longest run, all other variables being equal.
- The price of a skipass of a resort that has a snowpark, all other variables being equal, is estimated to be 4.60€ higher than a skipass of a resort that has not a snowpark.
- The price of a skipass of a resort that offers summer skiing, all other variables being equal, is estimated to be 13.51€ higher than a skipass of a resort that has not a snowpark.

We also compute the 95% confidence intervals for the estimated coefficients:

```
##                                2.5 %     97.5 %
## (Intercept)            35.10765601 40.0483292
## Longitude               0.04717366  0.1145214
## ContinentNorth America 34.27710879 44.4758609
## Highest.point           3.01247980  6.4751675
## Difficult.slopes        0.18653395  0.3229168
## Longest.run             0.15673159  1.0227956
## SnowparksYes            2.01400211  7.1919352
## Summer.skiingYes        8.70272389 18.3154835
```

The following plots show the estimated effects of Highest point and Difficult slopes on Price:



The dark blue lines are the fitted lines of two simple linear regression models, one with Highest point as the predictor and the other with Difficult slopes as predictor, while all the others are fixed at their average values, which are all 0 because they have been centered. The light blue areas around the lines are the 95% confidence intervals of the slopes. An interesting thing to notice is the fact that the interval becomes wider as Difficult slopes increases.

8. Tests for the coefficients

We perform both individual and global tests on the significance of the coefficients. We start with the following individual test for each β_j:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

We can find the values of the t static and the corresponding p-values of these tests in the coefficients section of the summary of the model:

```
##                            Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)            37.57799263 1.25729211 29.888037 1.379447e-112
## Longitude               0.08084752 0.01713851  4.717303  3.119948e-06
## ContinentNorth America 39.37648484 2.59535694 15.171896  6.488405e-43
## Highest.point           4.74382363 0.88117746  5.383505  1.134055e-07
## Difficult.slopes        0.25472536 0.03470641  7.339432  8.955854e-13
## Longest.run             0.58976359 0.22039414  2.675950  7.700806e-03
## SnowparksYes            4.60296867 1.31766950  3.493265  5.202956e-04
## Summer.skiingYes       13.50910369 2.44623480  5.522407  5.428563e-08
```

If we set as level of significance the standard 5% we reject all the null hypotheses and we can say that all the coefficients are statistically significant, i.e. different from 0.

We now test all of the β_j (except the intercept) to be 0 simultaneously, which is equivalent to the following test:

$$\begin{cases} H_0 : y = \beta_0 + \epsilon \\ H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon \end{cases}$$

This test is also called Global F-Test and we find the associated F-statistic and p-value in the last row of the summary of the model:

```
##
## Call:
## lm(formula = Price ~ Longitude + Continent + Highest.point +
##     Difficult.slopes + Longest.run + Snowparks + Summer.skiing,
##     data = resorts_centered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.019  -7.067  -0.337   6.535  51.603
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)            37.57799    1.25729  29.888  < 2e-16 ***
## Longitude               0.08085    0.01714   4.717 3.12e-06 ***
## ContinentNorth America 39.37648    2.59536  15.172  < 2e-16 ***
## Highest.point           4.74382    0.88118   5.384 1.13e-07 ***
## Difficult.slopes        0.25473    0.03471   7.339 8.96e-13 ***
## Longest.run             0.58976    0.22039   2.676  0.00770 **
## SnowparksYes            4.60297    1.31767   3.493  0.00052 ***
## Summer.skiingYes       13.50910    2.44623   5.522 5.43e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.26 on 491 degrees of freedom
## Multiple R-squared:  0.6568, Adjusted R-squared:  0.6519
## F-statistic: 134.3 on 7 and 491 DF,  p-value: < 2.2e-16
```

The p-value is much lower than 0.05, hence we reject the null hypothesis and we can say that at least one β_j is statistically significant, i.e. different from 0.

9. Goodness of fit

To evaluate the goodness of fit of our linear regression model we consider the adjusted R^2:

```
## [1] 0.6519436
```

Around 65.2% of the variability of price is explained by the model so the fit is not ideal, it could be better, but still is a good starting point.

10. Prediction

We suppose to have a new observation about a real ski resort which is not present in the dataset. We want to predict the price of the ski pass of the resort "Hoodoo Ski Area", located in Oregon, USA, North America. We find on the web all the information we need about this resort to predict the price using our model:

- Longitude = -121.872
- Continent = North America
- Highest point = 1740m
- Difficult slopes = 3km
- Longest run = 2km
- Snowparks = Yes
- Summer skiing = No

```
##        fit      lwr      upr
## 1 64.92806 40.60988 89.24624
```

Our model predicts the price of the Hoodoo Ski Area's skipass to be around 65€, with a 95% confidence interval: [41€, 89€]. The real medium price is around 70€, so the prediction of the model is not bad.

11. Simulation

We simulate 499 data points from the fitted regression model, assuming the estimated parameters as the true parameters. We then plot the simulated points against the observed points.

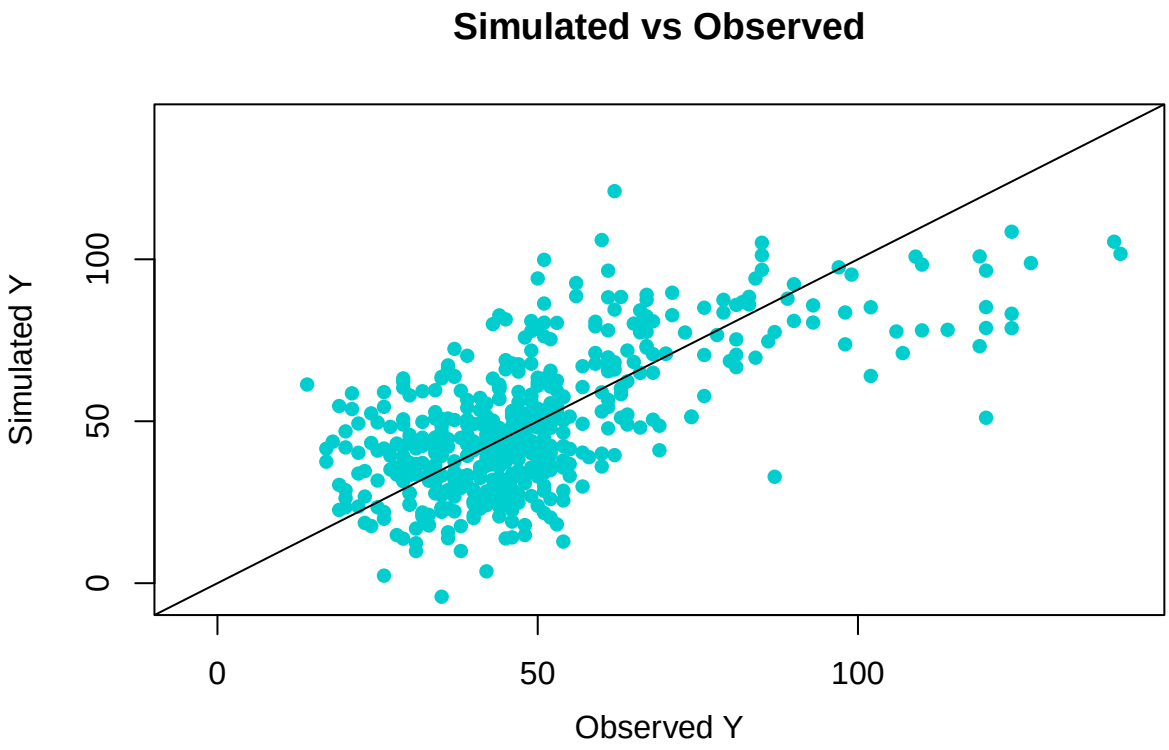


Figure 26

For large values of the observed y, the simulated y do not follow the line. This was expected because we saw from the Q-Q plot (Figure 15) that the large quantiles of the distribution of Price are bigger than the quantiles of a Normal distribution. Indeed, we can see here that simulating those points assuming normal errors leads to having simulated values lower than the observed values.

12. Conclusion

We have built a linear regression model that explains the Price of the ski pass of a resort based on Longitude, Continent, Highest Point, Difficult Slopes, Longest Run, Snowparks and Summer Skiing. The goodness of fit is not ideal but it is a good starting point for possible further analysis. Furthermore, errors do not properly respect the assumption of normality. Despite that, this model can be useful to establish the price of the ski pass for any new ski resorts, or to modify that of existing ski resorts. In the future we could try some kind of non linear regression, compare the results and choose the best model.