**Introduction**

As the whale watching industry grows, many companies are implementing new strategies to attract more customers. One strategy currently employed is providing a 50% refund to customers when no whales are sighted during their trip. However, in order to prevent significant loss of profit, companies will need to be able to predict trips that are most likely to have no whale sightings so that they can be pre-emptively cancelled. In this study, we will take a close look at various factors that may impact the number of whale sightings to answer this important question: Which factor(s) contributes the most to the visibility of whales during trips? Specifically (1) What is the best model for predicting whale sightings and 2) which factors are included in this model?

**Methods**

For analyzing the data, clouds at 8 am, visibility at 8 am, visibility, and rain will be included. The reason for excluding month, year and day is because these factors will not show a linear relationship with whale sightings. Additionally, rain in the morning has been reported by multiple whale watching agencies to not influence the likelihood of seeing a whale.

Here are the predictions for the outcomes of each variable:

|  |  |
| --- | --- |
| **Variable** | **Correlation with Whale Abundance** |
| Clouds at 8 am | Negative |
| Visibility at 8 am | Positive |
| Visibility | Positive |
| Rain | Negative |

Missing factors that are assumed to be controlled include temperature (whale watching trips occur only in certain seasons with little to no fluctuation in temperature) and storms (these trips are usually cancelled).

**Results**

To generate the variables to be included in each model, a best subsets regression was performed (Figure 1). These results where then used to generate the equation for each model (Figure 2). From these equations, the model with the lowest CP value was selected as the best model (Figure 2, Model 4). This model’s equation was y = -4.70292627 + 0.02245406x1 + 0.03670466x2 + 0.10900179x3 + 0.10842267x4 and included the following variables: clouds at 8 am, rain, visibility at 8 am, and visibility (Figure 2). The factor which was included in all 4 models was visibility (Figure 2). When comparing the CP values with the next best 3 models, the differences between models are similar (Figure 2). To check assumptions for the best model, residuals were compiled for all of the factors and plotted (Figure 3). The conclusion from this plot was that the residuals are normally distributed (Figure 3).

Subset selection object

Call: regsubsets.formula(lrgwhale ~ ., data = whalewatch, nvmax = 4)

11 Variables (and intercept)

Forced in Forced out

tripid FALSE FALSE

year FALSE FALSE

month FALSE FALSE

day FALSE FALSE

npass FALSE FALSE

cloud8am FALSE FALSE

rain8am FALSE FALSE

vis8am FALSE FALSE

rain FALSE FALSE

durntot FALSE FALSE

vis FALSE FALSE

1 subsets of each size up to 5

Selection Algorithm: exhaustive

tripid year month day npass cloud8am rain8am vis8am rain durntot vis

1 ( 1 ) " " " " " " " " " " " " " " " " " " " " "\*"

2 ( 1 ) " " " " " " " " " " " " " " "\*" " " " " "\*"

3 ( 1 ) " " " " " " " " " " " " " " "\*" "\*" " " "\*"

4 ( 1 ) " " " " " " " " " " "\*" " " "\*" "\*" " " "\*"

**Figure 1:** The figure above includes a table of the 4 best models from the whale watch data, collected via performing a best subsets regression analysis.

> coef(models, 1)

(Intercept) vis

-4.5246411 0.1252215

> coef(models, 2)

(Intercept) vis8am vis

-4.55500972 0.02850101 0.10413174

> coef(models, 3)

(Intercept) vis8am rain vis

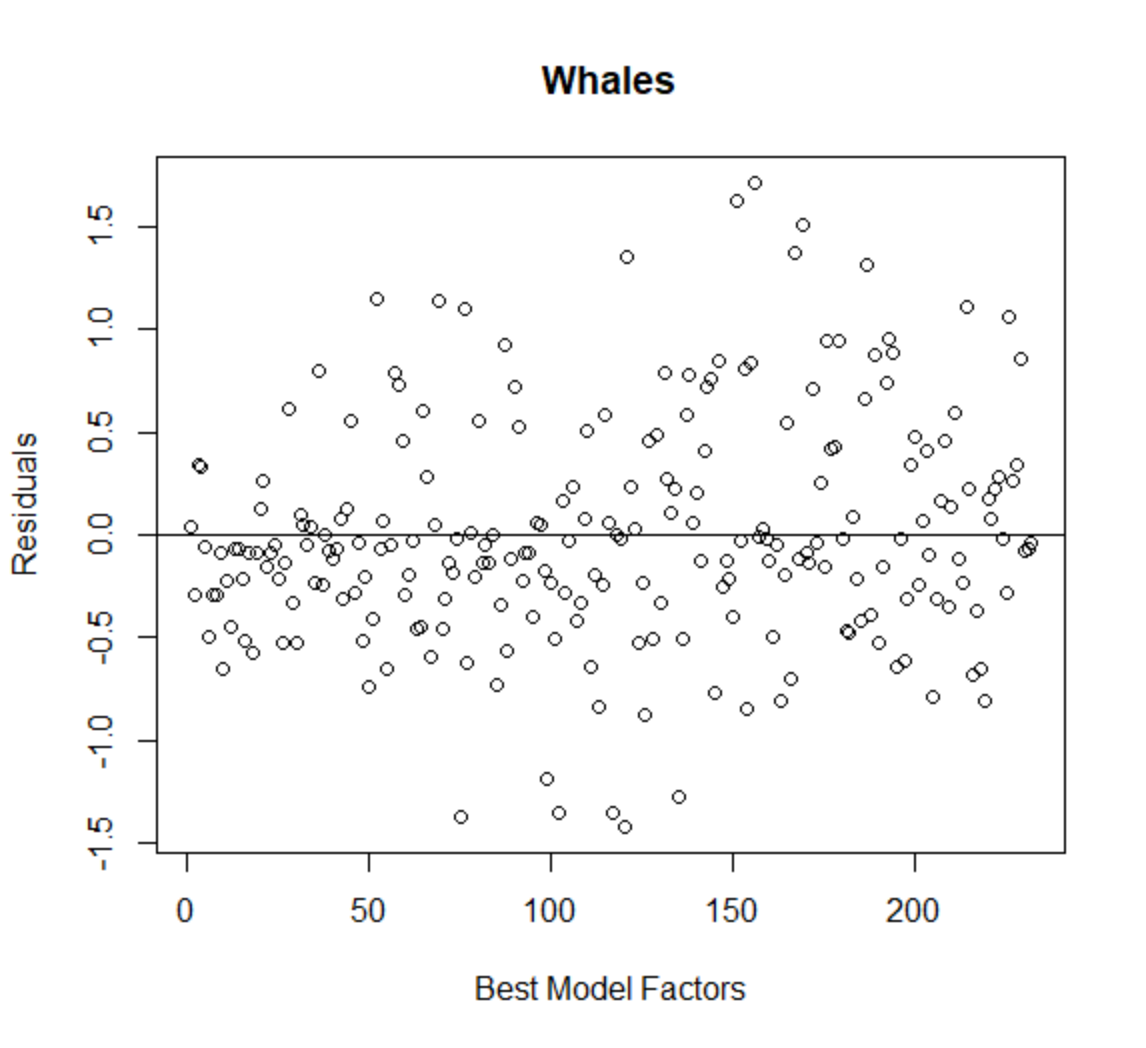
-4.64131515 0.03665241 0.14554839 0.10560035

> coef(models, 4)

(Intercept) cloud8am vis8am rain vis

-4.70292627 0.02245406 0.03670466 0.10900179 0.10842267

**Figure 2:** The figure above contains the coefficients for the equations for the best 4 models. The best model was Model 4 and the equation for this model is y = -4.70292627 + 0.02245406x1 + 0.03670466x2 + 0.10900179x3 + 0.10842267x4.



**Figure 3:** The plot above contains the residuals for all 4 factors, visualized to check the assumptions for regression analysis.

**Discussion**

In conclusion, the factors required to decide whether to cancel a whale watching trip are visibility, visibility at 8 am, rain and clouds at 8 am. This model being the best one could be a result of each factor’s individual impact on whale sightings is no being as strong as when they are paired with each other. To balance the explanatory power and complexity of this model, over-interpretation should be avoided. This includes trying to figure out which factors affect whale sightings the most or the least. These results could be used for the whale watching company by inputting values for each variable and calculating the number of whales that the trip might have. Doing so would allow them to decide whether to cancel a trip or not, thus, saving a lot of lost revenue