Week 9 Problem Set: Correlation and regression

(30 pts+5 pt Bonus)

For two different combinations of and (of your choosing), fill in the tables below (three columns need to be completed, plus the box for the sum of squared residuals). Try to tweak the parameter values to minimize the sum of squared residuals. This exercise is designed to just re-enforce the mechanics of fitting linear models, that each set of parameter values yields a set of predicted Y values, and that the difference between these predictions and the actual data form the basis of calculating the sum of squared errors.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model matrix  1 | | (predicted y)  1 + x= | (observed y) |  |  |
| 1 | 1 |  | 0 |  |  |
| 1 | 2 |  | 1 |  |  |
| 1 | 3 |  | 7 |  |  |
| 1 | 4 |  | 2 |  |  |
| 1 | 5 |  | 4 |  |  |
| 1 | 6 |  | 10 |  |  |
|  |  |  |  | Sum of = |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model matrix  1 | | (predicted y)  1 + x= | (observed y) |  |  |
| 1 | 1 |  | 0 |  |  |
| 1 | 2 |  | 1 |  |  |
| 1 | 3 |  | 7 |  |  |
| 1 | 4 |  | 2 |  |  |
| 1 | 5 |  | 4 |  |  |
| 1 | 6 |  | 10 |  |  |
|  |  |  |  | Sum of = |  |

Part II

For Part II and III of the problem set, we will analyze a dataset collected by Lucy Donahue and undergraduate Nancy Dong relating to the characteristics of Antarctic passenger vessels over time. This dataset includes several columns you will not need, but we will focus on the columns for vessel length, maximum speed, and engine power.

# Step #1: Using ‘lm’

(4 pts each part)

1. Use the ‘lm’ function to fit the following linear regression model, where is the response and is the covariate:

What is the statistical distribution being modeled here? (Fill in the equation below)

(Note that speeds cannot be negative, but we are going to assume that maximum speed is Normally distributed for the purposes of this problem set. It is a reasonable assumption since the speeds are all far away from zero.) Report the results and plot the data and the best-fit line.

Are there any data points that may be considered outliers? If so, which one(s)? How does the regression slope change if it/they were to be removed from the dataset?

1. Use this model (all the data included) to predict the maximum vessel speed for a vessel that is 100 m in length. What is the confidence interval? What is the prediction interval?
2. Use the ‘lm’ function to fit the following linear regression model:

Report the results and plot the data and the best-fit line.

1. Which covariate “Length” or “EnginePower” explains more of the variation in maximum vessel speed?
2. Use the ‘lm’ function to fit the following linear regression model:

Report the results and plot the data and the best-fit line.

# Step #2: Brute force

Write a function to calculate the negative log likelihood associated with fitting the linear regression model MaxSpeed~Length. (Hint: Your function will need to take as inputs the intercept (), the slope (), and the variance (σ2).) (For full credit, use the ‘dnorm’ function.) (5 pts)

Using the ‘optim’ function in R, minimize the negative log-likelihood to obtain the maximum likelihood regression parameter estimates for ,, σ2. (Hint: You will need starting values for optim; the results of the ‘lm’ calculation would be reasonable starting points.) (5 pts)

Part III (5 point bonus)

A more robust alternative to minimizing mean squared error is to minimize median squared error

There are no analytic formulas for the regression coefficients when using median squared error as the metric of fit. Two possible solutions would be to write a function similar to what was done for Part II and use ‘optim’ to find the maximum likelihood estimators for intercept and slope. Another solution is to simply calculate the median squared error over a two-dimensional grid and find the minimum of the median squared error by brute force ‘grid search’ optimization. Once the median squared error parameter estimates have been found, standard errors can be found using bootstrapping, where bootstrapped samples are generated by sampling X-Y pairs of data with replacement, and finding parameter estimates for each bootstrapped dataset to calculate s.e. for slope and intercept.

Find the intercept and slope parameter estimates and their standard errors for the model

assuming median squared error as the error metric.