Exercise 3: Regression

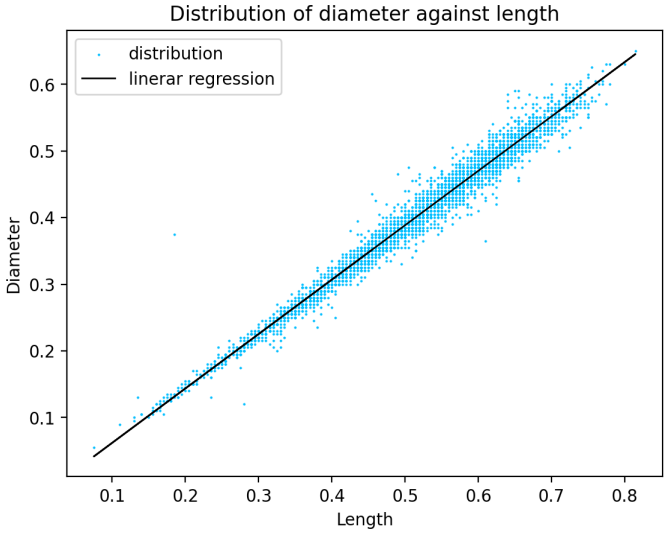
Linear regression, multilinear regression and non-linear regression

1. A simple linear regression model to give diameter as a function of length is:

   $Diameter = 0.8155 \times Length - 0.0194$

   Parameters are 0.8155 and -0.0194, correlation coefficient is 0.9868.

   From the parameters we can infer that diameter is positively related to length, and that length of abalones is typically greater than 0.0238 to ensure diameter is greater than 0.



2. A multilinear model to give the whole weight as a function of different pieces:

   $Whole = 0.9366 \times Shucked + 1.1116 \times Viscera + 1.253 \times Shell - 0.0078$

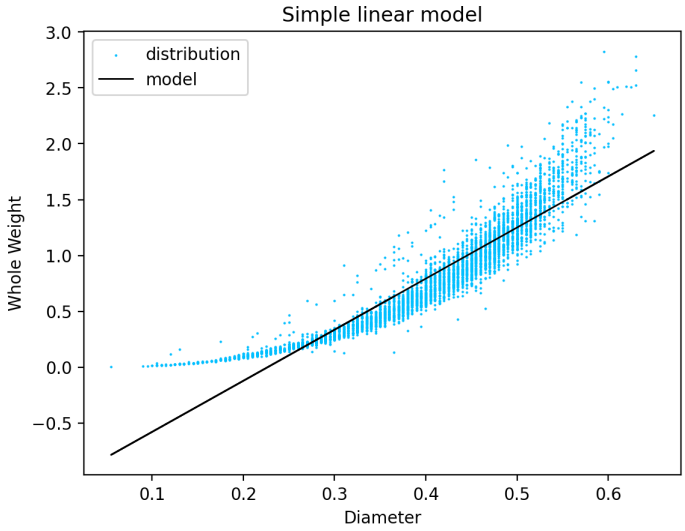
   The correlation coefficient of this multilinear model is: 0.9954.

   If we allow $\pm 5\%$ error, for 2301 among the 4177 abalones, whole weight is not equal to sum ($sum\ of\ pieces < whole\ weight \times 0.95 \vee sum\ of\ pieces > whole\ weight * 1.05$). That's 55.09% of them. If we allow $\pm 10\%$ error, then there will only be 540 (12.93%) abalones falling out of range. Now our common sense may still suggest that the whole weight should be "related", if not exactly equal to the sum of these weights.

   We then calculate the correlation coefficient of the abalones' $Whole\ weight$ versus $Shucked + Viscera + Shell$, which is 0.9951, thought it is a bit lower than that of our fitted multilinear model above, it is high enough for us to say the relation holds.

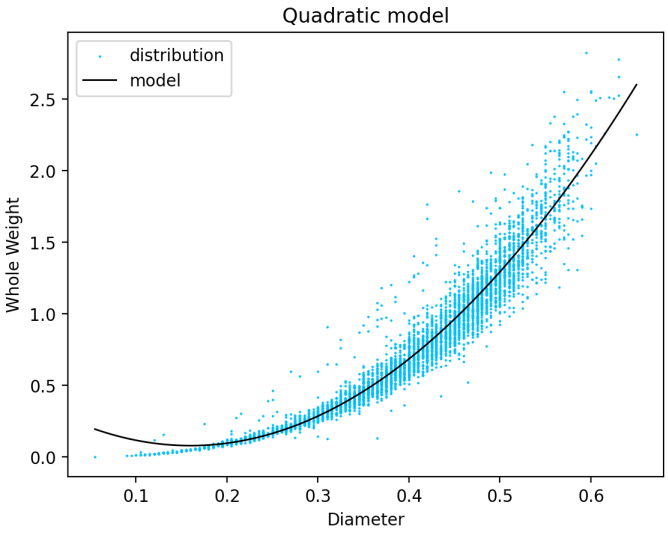
3. Relationship between the whole weight and diameter:

   (a) Simple Linear model: $Whole\ weight = 4.5731 \times Diameter - 1.0365$



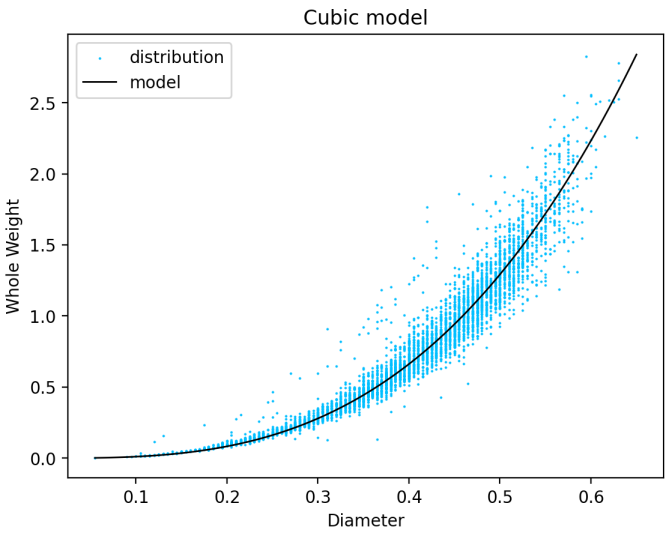
Correlation coefficient: 0.9255.

(b) Quadratic: $Weight = -3.3555 \times Diameter + 10.4968 \times Diameter^2 + 0.3477$



Correlation coefficient: 0.9627.

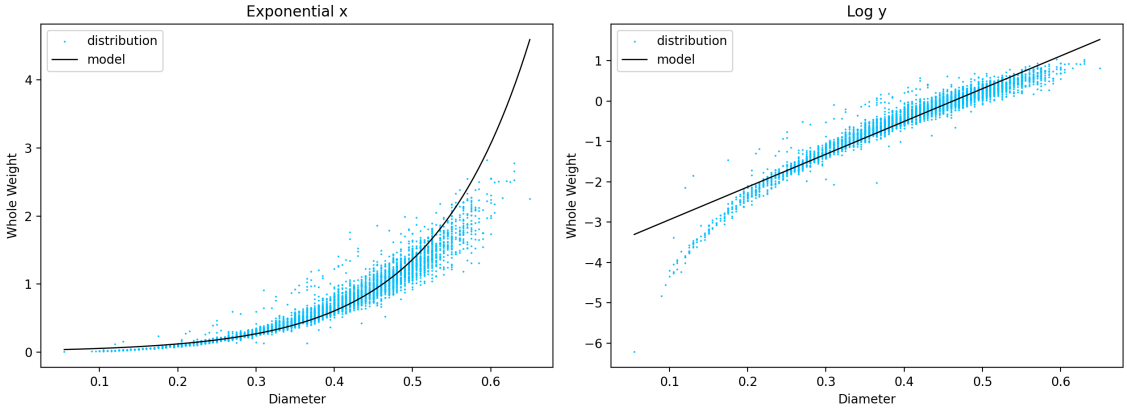
(c) Cubic model without lower or constant terms: $Weight = 10.3376 \times Diameter^3$



Correlation coefficient: 0.9631.

(d) Cubic model: $Weight = e^{8.1167 \times Diameter - 3.7510}$,

or $\log(Weight) = 8.1167 \times Diameter - 3.7510$



Correlation coefficient for the former is 0.9465, the latter is 0.9635.

Based on the plots and the correlation coefficients, we can decide that the cubic model is the most appropriate to represent the dependency. The conclusion also matches our common sense, that if we assume the "density" of these abalones are roughly the same, their weight will linearly depend on their volume. From ***1*** we know that diameter and length are linearly correlated, so from $Volume = Length \times Diameter^2 \times$ constant parameter (depending on the shape) and $Weight = Volume \times Density$ we can derive that $Weight$ linearly depends on $Diameter^3$, a cubic model with no lower order or constant terms.

Logistic Regression

4. Building logistic regression models to tell whether a specimen is infant or adult:

   First we modify the input dataset, changing all nominal values 'M' and 'F' for Sex to 'A' in Weka, for R we further change all A's to 1's and I's to 0's to apply logistic regression.

   (a) Based on $Length$ only: the accuracy is 78.3337%

   (b) $Whole\ Weight$ only: the accuracy is 79.5786%

   (c) $Class\ Rings$ only: the accuracy is 78.8125%

   (d) $Length$, $Whole\ Weight$, and $Class\ Rings$: accuracy is 82.2839%

5. Building logistic regression model for the attribute sex (M/F):

   First we build a model with all the attributes involved, the accuracy is 84.6566%, then we try the same thing with only one attribute absent at a time, the accuracy is recorded as follow:

without	***age***	***workclass***	***fnlwgt***	***education***	***education-num***
accuracy	84.6648%	84.4601%	84.4744%	84.5379%	84.6566%

without	***marital-status***	***occupation***	***relationship***	***race***
accuracy	84.3946%	80.7666%	79.0262%	84.6853%

without	*capital-gain*	*capital-loss*	*hours-per-week*	*native-country*	*class*
accuracy	84.6566%	84.6669%	84.0117%	84.5563%	84.5829%

For attributes age, $education\ num$, $race$, $capital\ gain$ and $capital\ loss$, accuracy remained the same or even became slight higher without them, that is to say, they do not contribute much to the classification, and therefore can be removed safely. The model with these five attributes removed has very good accuracy (84.6648%, even better than model with all attributes, probably the best model we can get) but not enough simplicity.

We notice that removing attributes $occupation$ and $relationship$ would lead to a quite significant (greater than 1%) decrease in the accuracy, so these two are definitely needed for predicting sex. We calculate the model accuracy using only $occupation$ and $relationship$ and get 82.5396%, which seems okay but not that satisfying.

Put other attributes in order according to how much accuracy was affected without them: $house\ per\ week$ > $marital\ status$ > $workclass$ > $fnlwgt$ > $education$ > $native\ country$ > $class$, add them back one by one and keep a record of the accuracy.

1. $house\ pre\ week$ added: accuracy becomes 83.6084%
2. $marital\ status$ added: accuracy becomes 83.85%
3. $workclass$ added: accuracy becomes 84.0977%
4. $fnlwgt$ added: accuracy becomes 84.3004%
5. $eduaction$ added: accuracy becomes 84.458%
6. $native\ country$ added: accuracy becomes 84.5072%
7. $class$ added: accuracy becomes 84.6648%

If we want the difference of the accuracy of our model and the best model (84.6648%) to be smaller than 1%, we may simply stop after step 2; if we want it less than 0.5%, stop after step 4, and step 7 for less than 0.1%, depending on just how accurate or how simple we want the model to be. Here is one possible final model with 4 variables.

Accuracy	*Variables Included*
83.85%	$marital\ status, occupation, relationship, hours\ per\ week$

(a) The attributes age, $workclass$, $fnlwgt$, $education$, $education\ num$, $race$, $capital\ gain$, $capital\ loss$, $hours\ per\ week$, $native\ country$ and $class$ can all be removed from the dataset without affecting the accuracy of the resulting model significantly (by at most 1%), all these attributes do not have high correlation with sex, i.e. they are more independent. For example, it is often the case that two sexes would have similar (at least not that distinct) distributions in different age, race or class groups. Consider conditional probability, having one of these attributes fixed will not significantly change the possibility of a guess (on sex) to be correct or incorrect. Thus they are not very useful to our prediction.

(b) Relationship status is very helpful because it has the values 'Wife' and 'Husband', for most cases the wife would be female and husband would be male, so this information can significantly increase our chance of picking the right (more possible) guess.

(c) Categorical features are represented as 0-1 vectors, one element for each category. Consider the probability p for a person to be female given that person comes from Holland, $odds\ ratio = p/(1-p)$, Because there is only one person from Holland (a female), instinctively we would expect the p derived from this dataset to be very large (close to 1 to a certain extent). The probability of a person from Holland to be female is $p/(1-p)$ times that if the person is not from Holland. Suppose the weight for country=Holland is β, we would expect e^{β} to represent the above "$p/(1-p)$ times" relationship, when p moves closer to 1 the $odds\ ratio$ becomes larger and larger, thus it is natural for β to be "heavy".

Bonus question

If we use quadratic model for every variable except Sex and throw them into multilinear regression we may obtain a coefficient of 0.76, but consider the price we pay for this really small improvement it might not be worthwhile (unfortunately, I failed to come up with better ideas).

Problem Set 3

1. (a) The “No-Free-Lunch” Theorem suggests that there does not exist a model/algorithm to achieve the smallest empirical error on all learning problems.

   (b) For the goal of predicting binary class Y with binary feature X (assume an equal split between the Train and Test sets), there are four possible learning problems:

	x	y (problem 1)	y (problem 2)	y (problem 3)	y (problem 4)
Train	0	0	0	1	1
Test	1	0	1	0	1

   (c) Since there is only one data point in the test set, the empirical error using absolute loss model is $|\hat{y}-y|$ for data point $(1, y)$. Suppose our algorithm predicts $\hat{y}=0$ for $x=1$, then for problems 1 and 3 the empirical error is 0 and the accuracy is 100%, for problems 2 and 4 the empirical error is 1 and the accuracy is 0%. Consider the four problems to have the same probability $1/4$, the general empirical error would be 0.5 and accuracy 50%. Vice versa, if we predict $\hat{y}=1$ for $x=1$, for problems 1 and 3 empirical error is 1, accuracy is 0%, for 2 and 4 empirical error is 0, accuracy is 100%, the general empirical error and accuracy are the same. The average error of all any algorithm is 0.5 and the average accuracy is always 50%.

(d) It is impossible to design a prediction algorithm which yields 100% average accuracy, independent of features, labels and how the data is divided (into train and test set). Suppose we have an algorithm with 100% percent accuracy on some problems (so the mapping from features to labels is fixed), there would always exist other problems where some data points in test set have the same features but different labels. The algorithm will not be 100% correct with these problems, thus average accuracy cannot be 100%. Actually, if we assume all problems are equally likely, the average accuracy of any algorithm would be 50%.

2. Let A_i denote "the perfume is in drawer i", and B_i for "Alice searches in drawer i and doesn't find the perfume". the probability we would like to calculate can be represented as:

   $P(A_i|B_3) = \frac{P(A_i \cap B_3)}{P(B_3)} = \frac{P(A_i)P(B_3|A_i)}{P(B_3)}$ for $i \in \{1,2,3,4,5\}$.

   We already know that $P(A_i) = p_i$ and $P(B_3|A_3) = 1 - s_3$, For $i \in \{1,2,4,5\}$, $P(B_3|A_i) = 1$ (if the perfume is in drawer 1, 2, 4 or 5, surely she'll never find it in No. 3!)

   $P(B_i) = P(\bar{A}_i) + P(A_i)(1 - s_i) = (1 - p_i) + p_i(1 - s_i) = 1 - p_i s_i$ (If Alice fail to find the perfume in drawer i, there are two possible cases: either the perfume is not in the drawer (probability $P(\bar{A}_i)$), or the perfume is in the drawer, but still she does not find it (probability $P(A_i)$ times $1 - s_i$)). Thus we have $P(B_3) = 1 - p_3 s_3$.

   (i) If $i = 3$, $P(A_3|B_3) = \frac{P(A_3)P(B_3|A_3)}{P(B_3)} = \frac{p_3(1-s_3)}{1-p_3 s_3} = \frac{p_3 - p_3 s_3}{1-p_3 s_3}$.

   (ii) If $i \in \{1,2,4,5\}$, $P(A_i|B_3) = \frac{P(A_i)P(B_3|A_i)}{P(B_3)} = \frac{p_i}{(1-p_3 s_3)}$.