

B565: Homework 1

1. Consider the pairs plot of the “iris” data discussed in class. If we want to predict the iris species using only one of the petal and sepal measurements, based on this plot which one should it be? Explain your choice clearly.
2. A clear jar contains some unknown number of beans. 100 randomly chosen people inspect the jar and make estimates, x_1, \dots, x_{100} of the true number of beans, x_{true} , with no knowledge of the others’ estimates. The sample average, $\bar{x} = \frac{1}{100} \sum_{i=1}^{100} x_i$, is computed.
 - (a) Consider modeling x_1, \dots, x_{100} as a random sample from some distribution with mean μ and variance σ^2 . Is this a reasonable model? Justify your answer clearly.
 - (b) What is the approximate distribution of \bar{x} ?
 - (c) What is the approximate probability that \bar{x} misses μ by more than 2 beans, assuming $\sigma = 10$?
 - (d) What can you say about the actual error, $\bar{x} - x_{\text{true}}$. Make clear any assumptions you are making.
3. Consider the trees data set which you can import to R by

```
> data(“trees”)
```

This data set measures girth (circumference), height, and volume on a small collection of trees. Treat the data set as a matrix, using

```
> X = as.matrix(trees)
```

The features (columns) can be centered to have mean 0 using

```
> X = scale(X, center=TRUE, scale=FALSE)
```

you may wish to use a pairs plot visualize these data.

- (a) Suppose we have data $(x_1, y_1), \dots, (x_n, y_n)$. The *sample correlation* between two variables x and y , $r(x, y)$ is defined to be

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

It can be shown than $|r(x, y)| \leq 1$. Informally, the sample correlation measures the strength of the linear relationship between two variables. That is

- $r(x, y) = 1$ means y can be predicted perfectly as $y_i = ax_i + b$ with $a > 0$
- $r(x, y) = -1$ means y can be predicted perfectly as $y_i = ax_i + b$ with $a < 0$
- $r(x, y) = 0$ means there is no formula $y_i = ax_i + b$ that predicts y better than simply choosing $a = 0$ (not using a).

In the usual case where $|r(x, y)|$ is not exactly 0 or 1 it means that one variable can be partly predicted from the other using a linear model. More precisely, $r^2(x, y)$ is the percent of squared error explained by the best linear model. That is

$$r^2(x, y) = 1 - \frac{\sum_i (y_i - ax_i + b)^2}{\sum_i (y_i - \bar{y})^2}$$

where a, b are the optimal line coefficients.

Approximate all 9 possible correlations using the pairs plot.

- (b) Suppose we consider removing a single variable to simplify this data set. For each variable answer whether or not there is significant loss of information in the data set when the variable is removed. Justify your answers.
- (c) **Using matrix calculations** compute the sample covariance matrix. Do not use an R package to do this.

4. In the “curse of dimensionality” R example done in class we let X be a d -dimensional vector of “unif” variables. This means the components of X are independent $\text{Unif}(0,1)$ random variables. The R example computes the distance of X to the edge of the hypercube by

$$D = \frac{1}{2} - \max_{i=1\dots d} \left| \frac{1}{2} - X_i \right|$$

- (a) The definition of independence says that for X_1, \dots, X_d ,

$$P(X_1 \in A_1, \dots, X_d \in A_d) = \prod_{i=1}^d P(X_i \in A_i)$$

where $X_1 \in A_1, \dots, X_d \in A_d$ means $X_1 \in A_1$ and $X_2 \in A_2$ and \dots , $X_d \in A_d$. What is the probability that a point is at least a away from an edge? That is what is $P(D > a)$ for $0 < a < \frac{1}{2}$.

- (b) Given any distance to edge δ , find d so that the probability of X lying within δ of the edge of the hypercube is at least .9. That is, find d so that $P(D < \delta) \geq .9$.

5. In your book, Chapter 2: problems 16,17, 25.