COMS W4771: Machine Learning (sec:001) - Homework #2

Name: Geraldi Dzakwan (gd2551). Discussants: Cherry (ccc2207), Deka (da2897)

October 9, 2019

Problem 1

a. Let's expand our expectation of the mean squared error. By inserting two expectation of Y of opposite signs that cancel each other out, we can get:

$$\mathbb{E}[(\hat{y}^* - Y)^2] = \mathbb{E}[((\hat{y}^* - \mathbb{E}[Y]) + (\mathbb{E}[Y] - Y))^2]$$

$$\mathbb{E}[(\hat{y}^* - Y)^2] = \mathbb{E}[(\hat{y}^* - \mathbb{E}[Y])^2] + \mathbb{E}[2(\hat{y}^* - \mathbb{E}[Y])(\mathbb{E}[Y] - Y)] + \mathbb{E}[(\mathbb{E}[Y] - Y)^2]$$

There are two things that we can use to simplify the above equation:

1. $\mathbb{E}[\mathbb{E}[Y] - Y]$ is zero because $\mathbb{E}[\mathbb{E}[Y] - Y] = \mathbb{E}[\mathbb{E}[Y]] - \mathbb{E}[Y] = \mathbb{E}[Y] - \mathbb{E}[Y] = 0$
2. $\mathbb{E}[(\mathbb{E}[Y] - Y)^2]$ is the variance of Y $\rightarrow \mathbb{E}[(\mathbb{E}[Y] - Y)^2] = \mathbb{E}[(\mu - Y)^2] = var(Y)$

Thus, we can rewrite our expectation as:

$$\mathbb{E}[(\hat{y}^* - Y)^2] = \mathbb{E}[(\hat{y}^* - \mathbb{E}[Y])^2] + 0 + \mathbb{E}[(\mu - Y)^2] = \mathbb{E}[(\hat{y}^* - \mu)^2] + var(Y)$$

To minimize this expectation, the best way we can do is to take the mean of Y as our optimal prediction, i.e. $\hat{y}^* = \mathbb{E}[Y] = \mu$ and the smallest mean squared error would be the variance, i.e. $var(Y)$. Now let's compute these two in terms of θ.

To compute the mean of a probability density function $p_\theta(y)$, we can take the sum of $yp_\theta(y)$ for all y. In other words, its integral for $0 \leq y \leq \infty$. We can omit for $y < 0$ because the probability is zero within that range. Thus:

$$\mu = \int_0^\infty (y)(\frac{1}{\theta^2} y e^{-y/\theta}) dy = \frac{1}{\theta^2} \int_0^\infty y^2 e^{-y/\theta} dy$$

Using integral by parts, taking $u = y^2$ and $dv = e^{-y/\theta} dy$, we get:

$$\int y^2 e^{-y/\theta} = (y^2)(-\theta e^{-y/\theta}) - \int (-\theta e^{-y/\theta}) 2y dy = -\theta y^2 e^{-y/\theta} + 2\theta \int y e^{-y/\theta} dy$$

Another integral by parts, taking $u = y$ and $dv = e^{-y/\theta}dy$:

$$-\theta y^2 e^{-y/\theta} + 2\theta \int y e^{-y/\theta} dy = -\theta y^2 e^{-y/\theta} + 2\theta(y(-\theta e^{-y/\theta}) - \int (-\theta e^{-y/\theta}) dy)$$

$$= -\theta y^2 e^{-y/\theta} - 2\theta^2 y e^{-y/\theta} + 2\theta^2 \int e^{-y/\theta} dy = -\theta y^2 e^{-y/\theta} - 2\theta^2 y e^{-y/\theta} - 2\theta^3 e^{-y/\theta}$$

We can then calculate the upper and lower bound value:

1. Calculate upper bound:

$$\text{upper\_bound} = -\theta \lim_{y\to\infty} \frac{y^2}{e^{y/\theta}} - 2\theta^2 \lim_{y\to\infty} \frac{y}{e^{y/\theta}} - 2\theta^3 \lim_{y\to\infty} \frac{1}{e^{y/\theta}}$$

   Using L'Hospital theorem a few times, we get:

$$\text{upper\_bound} = -2\theta^2 \lim_{y\to\infty} \frac{y}{e^{y/\theta}} - 2\theta^3 \lim_{y\to\infty} \frac{1}{e^{y/\theta}} - 2\theta^3(0) = -2\theta^3 \lim_{y\to\infty} \frac{1}{e^{y/\theta}} - 2\theta^3(0) - 2\theta^3(0)$$

$$\text{upper\_bound} = -2\theta^3(0) - 2\theta^3(0) - 2\theta^3(0) = 0$$

2. Calculate lower bound:

$$\text{lower\_bound} = -\theta \lim_{y\to 0+} \frac{y^2}{e^{y/\theta}} - 2\theta^2 \lim_{y\to 0+} \frac{y}{e^{y/\theta}} - 2\theta^3 \lim_{y\to 0+} \frac{1}{e^{y/\theta}}$$

$$\text{lower\_bound} = -\theta(\frac{0}{1}) - 2\theta^2(\frac{0}{1}) - 2\theta^3(\frac{1}{1}) = -2\theta^3$$

Finally, in terms of θ, the "optimal prediction" $\hat{y}^*$ is:

$$\boldsymbol{\hat{y}^* = \mu = \frac{1}{\theta^2}(0 - (-2\theta^3)) = 2\theta}$$

Derive a way to compute the variance in which we can reuse our μ:

$$var(Y) = \mathbb{E}[(Y-\mu)^2] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2 - 2Y\mathbb{E}[Y] + \mathbb{E}[Y]^2]$$

$$var(Y) = \mathbb{E}[Y^2] - 2\mathbb{E}[Y[\mathbb{E}[Y]] + \mathbb{E}[\mathbb{E}[Y]^2] = \mathbb{E}[Y^2] - 2\mathbb{E}[Y]\mathbb{E}[Y] + \mathbb{E}[Y]^2$$

$var(Y) = \mathbb{E}[Y^2] - 2\mathbb{E}[Y]^2 + \mathbb{E}[Y]^2 = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \mathbb{E}[Y^2] - \mu^2$ ... Equation 1.1

Calculate $\mathbb{E}[Y^2]$:

$$\mathbb{E}[Y^2] = \int_0^\infty (y^2)(\frac{1}{\theta^2} y e^{-y/\theta}) dy = \frac{1}{\theta^2} \int_0^\infty y^3 e^{-y/\theta} dy$$

Using integral by parts, taking $u = y^3$ and $dv = e^{-y/\theta}dy$, we get:

$$\int y^3 e^{-y/\theta} = (y^3)(-\theta e^{-y/\theta}) - \int (-\theta e^{-y/\theta}) 3y^2 dy = -\theta y^3 e^{-y/\theta} + 3\theta \int y^2 e^{-y/\theta} dy$$

We know from the previous result that $\int_0^\infty y^2 e^{-y/\theta} = 2\theta^3$. We just need to compute the value of $-\theta\, y^3 e^{-y/\theta}$ for $0 \le y \le \infty$.

1. Calculate upper bound:

$$\text{upper\_bound} = -\theta \lim_{y\to\infty} \frac{y^3}{e^{y/\theta}}$$

Using L'Hospital theorem a few times, we get:

$$\text{upper\_bound} = -3\theta^2 \lim_{y\to\infty} \frac{y^2}{e^{y/\theta}} = -6\theta^3 \lim_{y\to\infty} \frac{y}{e^{y/\theta}} = -6\theta^4 \lim_{y\to\infty} \frac{1}{e^{y/\theta}} = 0$$

2. Calculate lower bound:

$$\text{lower\_bound} = -\theta \lim_{y\to 0+} \frac{y^3}{e^{y/\theta}} = -\theta(\frac{0}{1}) = 0$$

Finally, in terms of θ, the smallest mean squared error is:

$$\mathbb{E}[(\hat{y}^* - Y)^2] = var(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$$

$$\boldsymbol{\mathbb{E}[(\hat{y}^* - Y)^2] = \frac{1}{\theta^2}((0-0) + 3\theta(2\theta^3)) - (2\theta)^2 = 6\theta^2 - 4\theta^2 = 2\theta^2}$$

b. For $y \leq 0$, the probability is zero and thus the MLE is also zero. For $y > 0$, the likelihood for $\{y_1, y_2, y_3, ..., y_n\}$ can be stated as:

$$L(\theta|y) = \prod_{i=1}^{n} p_\theta(y_i) = \prod_{i=1}^{n} \frac{y_i e^{-y_i/\theta}}{\theta^2}$$

Take the log likelihood instead so it's easier to expand:

$$\ln L(\theta|y) = \ln(\prod_{i=1}^{n} p_\theta(y_i)) = \sum_{i=1}^{n} \ln(p_\theta(y_i)) = \sum_{i=1}^{n} \ln(\frac{y_i e^{-y_i/\theta}}{\theta^2})$$

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln y_i + \sum_{i=1}^{n} \ln e^{-y_i/\theta} - \sum_{i=1}^{n} \ln \theta^2 = \sum_{i=1}^{n} \ln y_i + \sum_{i=1}^{n} (-y_i/\theta) \ln e - n \ln \theta^2$$

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln y_i - \frac{1}{\theta} \sum_{i=1}^{n} y_i - n \ln \theta^2$$

To minimize this, take its first derivative and find the θ that makes it zero:

$$\frac{d}{d\theta} \ln L(\theta|y) = 0 \to 0 + \frac{1}{\theta^2} \sum_{i=1}^{n} y_i - n\frac{2\theta}{\theta^2} = 0 \to \frac{1}{\theta^2} \sum_{i=1}^{n} y_i - \frac{2n}{\theta} = 0$$

$$\frac{2n}{\theta} = \frac{1}{\theta^2} \sum_{i=1}^{n} y_i \to \theta = \frac{1}{2n} \sum_{i=1}^{n} y_i$$

Since $y > 0$, this MLE formula will always be greater than zero, thus greater than MLE if $y \leq 0$, which is zero. Hence, we can pick this as our MLE formula.

Last check we need to do is to make sure that this $\hat{\theta}$ is the minimizer, that is the second derivative value for this $\hat{\theta}$ is negative. The second derivative is as below.

$$\frac{d}{d\theta^2} \ln L(\theta|y) = -\frac{2}{\theta^3} \sum_{i=1}^{n} y_i + \frac{2n}{\theta^2}$$

This second derivative is negative if and only if:

$$-\frac{2}{\theta^3} \sum_{i=1}^{n} y_i + \frac{2n}{\theta^2} < 0 \rightarrow \theta < \frac{1}{n} \sum_{i=1}^{n} y_i$$

which is true for $\theta = \hat{\theta} := \frac{1}{2n} \sum_{i=1}^{n} y_i$ because $y_1, y_2, ..., y_n$ are all positive. Thus, the MLE formula is:

$$\boldsymbol{\hat{\theta}_{MLE}(y_1, y_2, ...y_n) = \frac{1}{2n} \sum_{i=1}^{n} y_i}$$

Finally, we can plug this estimator $\hat{\theta}_{mle}$ for sample $y_1, y_2, ..., y_n$ into our "optimal estimator" $\hat{y}^* = 2\theta$ that we've figured out before in problem 1a.

$$\boldsymbol{\hat{y}(y_1, y_2, ..., y_n) = 2\hat{\theta}_{MLE}(y_1, y_2, ...y_n) = 2 * \frac{1}{2n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} y_i}$$

which, in other words, is the average of the observed sample $(y_1, y_2, ..., y_n)$.

c. Expand the new squared loss equation:

$$\mathbb{E}[(\hat{Y} - Y)^2] = \mathbb{E}[\hat{Y}^2 - 2\hat{Y}Y + Y^2]$$

$$\mathbb{E}[(\hat{Y} - Y)^2] = \mathbb{E}[\hat{Y}^2] - 2\mathbb{E}[\hat{Y}Y] + \mathbb{E}[Y^2]$$

Since, $Y_1, Y_2, ..., Y_n, Y$ are independent random variables, then $\hat{Y}$ is independent of Y. We then can have $\mathbb{E}[\hat{Y}Y] = \mathbb{E}[\hat{Y}]\mathbb{E}[Y]$ and rewrite our equation into:

$$\mathbb{E}[(\hat{Y} - Y)^2] = \mathbb{E}[\hat{Y}^2] - 2\mathbb{E}[\hat{Y}]\mathbb{E}[Y] + \mathbb{E}[Y^2]$$

We first try to solve the first term, $\mathbb{E}[\hat{Y}^2]$.

$$\mathbb{E}[\hat{Y}^2] = \mathbb{E}[(\frac{1}{n}(Y_1 + Y_2 + ... + Y_n))^2]$$

$$\mathbb{E}[\hat{Y}^2] = \mathbb{E}[\frac{(Y_1 + Y_2 + ... + Y_n)(Y_1 + Y_2 + ... + Y_n)}{n^2}]$$

In this scenario, there would be two cases of multiplications:

1. Multiplications involving dependent random variables, e.g. $(Y_1)(Y_1), (Y_2)(Y_2), ...(Y_n)(Y_n)$. There will be exactly n multiplications and we will have something like:

$$n\mathbb{E}[\hat{Y}\hat{Y}] = n\mathbb{E}[\hat{Y}^2]$$

Notice that it can't be $n\mathbb{E}[\hat{Y}\hat{Y}] = n\mathbb{E}[\hat{Y}]\mathbb{E}[\hat{Y}] = n\mathbb{E}[\hat{Y}]^2$ because both $\hat{Y}$ are dependent of each other.

2. Multiplications involving independent random variables, e.g. $(Y_1)(Y_2), (Y_1)(Y_3), ...(Y_n)(Y_{n-1})$. There will be exactly ${}^nP_2 = n(n-1)$ multiplications as there can exist permutations, for example, Y_1Y_3 and Y_3Y_1. We denote this as:

$$n(n-1)\mathbb{E}[\hat{Y}\hat{Y}] = n(n-1)\mathbb{E}[\hat{Y}]\mathbb{E}[\hat{Y}] = n(n-1)\mathbb{E}[\hat{Y}]^2$$

Then, finally for $\mathbb{E}[\hat{Y}^2]$, we have:

$$\mathbb{E}[\hat{Y}^2] = \mathbb{E}[\frac{n\mathbb{E}[\hat{Y}^2] + n(n-1)\mathbb{E}[\hat{Y}]^2}{n^2}] = \mathbb{E}[\frac{1}{n}\mathbb{E}[\hat{Y}^2] + \frac{n-1}{n}\mathbb{E}[\hat{Y}]^2]$$

$$\mathbb{E}[\hat{Y}^2] = \frac{1}{n}\mathbb{E}[\mathbb{E}[\hat{Y}^2]] + \frac{n-1}{n}\mathbb{E}[\mathbb{E}[\hat{Y}]^2] = \frac{1}{n}\mathbb{E}[\hat{Y}^2] + \frac{n-1}{n}\mathbb{E}[\hat{Y}]^2$$

Substitute this to the whole equation:

$$\mathbb{E}[(\hat{Y} - Y)^2] = \mathbb{E}[\hat{Y}^2] - 2\mathbb{E}[\hat{Y}]\mathbb{E}[Y] + \mathbb{E}[Y^2]$$

$$\mathbb{E}[(\hat{Y} - Y)^2] = \frac{1}{n}\mathbb{E}[\hat{Y}^2] + \frac{n-1}{n}\mathbb{E}[\hat{Y}]^2 - 2\mathbb{E}[\hat{Y}]\mathbb{E}[Y] + \mathbb{E}[Y^2]$$

Last thing to notice is that $\mathbb{E}[\hat{Y}] = \mathbb{E}[Y]$ since $\hat{Y}$ is the average of $Y_1, Y_2, ..., Y_n$ and those random variables come from the distribution of Y. Thus, we get:

$$\mathbb{E}[(\hat{Y} - Y)^2] = \frac{1}{n}\mathbb{E}[Y^2] + \frac{n-1}{n}\mathbb{E}[Y]^2 - 2\mathbb{E}[Y]^2 + \mathbb{E}[Y^2]$$

$$\mathbb{E}[(\hat{Y}-Y)^2] = (1+\frac{1}{n})\mathbb{E}[Y^2] - (2-\frac{n-1}{n})\mathbb{E}[Y]^2 = (1+\frac{1}{n})\mathbb{E}[Y^2] - (\frac{2n-n+1}{n})\mathbb{E}[Y]^2$$

$$\mathbb{E}[(\hat{Y} - Y)^2] = (1 + \frac{1}{n})\mathbb{E}[Y^2] - (\frac{n+1}{n})\mathbb{E}[Y]^2 = (1 + \frac{1}{n})(E[Y^2] - E[Y]^2)$$

From what I've derived in section 1a in Equation 1.1, we know that $\mathbb{E}[(\hat{y}^* - Y)^2] = var(Y) = E[Y^2] - E[Y]^2$. Then, using that, finally it is proved that:

$$\boldsymbol{\mathbb{E}[(\hat{Y} - Y)^2] = (1 + \frac{1}{n})\mathbb{E}[(\hat{y}^* - Y)^2]}$$

d. If $Y = Bern(\theta)$, then:

$$p(y = 1) = \theta$$
$$p(y = 0) = 1 - p(y = 1) = 1 - \theta$$

Using the explanation that I've written in section 1a, the "optimal prediction" for mean squared loss is the expected value of the random variable, i.e. the mean of the distribution. In the case of discrete distribution, the mean is basically the weighted sum of the discrete values, or in other words, the sum of the discrete values time their probability:

$$\mu = \sum_{y \in Y} yp(y) = 0 * p(y = 0) + 1 * p(y = 1) = p(y = 1) = \theta$$

Thus, our optimal prediction is $\boldsymbol{\hat{y}^* = \mu = \theta}$

Again, using the explanation that I've written in section 1a, the smallest mean squared error is basically the variance of the distribution. To compute the variance, we can use the Equation 1.1 that I've already derived in section 1a:

$$var(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \mathbb{E}[Y^2] - \mu^2$$

Compute $\mathbb{E}[Y^2]$:

$$\mathbb{E}[Y^2] = 0^2 * p(y = 0) + 1^2 * p(y = 1) = p(y = 1) = \theta$$

That gives us:

$$var(Y) = \mathbb{E}[Y^2] - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$$

Thus, our smallest mean squared error is $\boldsymbol{\mathbb{E}[(\hat{y}^* - Y)^2] = \theta(1 - \theta)}$

e. Let's generalize our loss function by applying weighted sum to the two conditions using each respective probability:

$$l(\hat{y}, y) = P(\hat{y} \geq y) * l(\hat{y} \geq y, y) + P(\hat{y} < y) * l(\hat{y} < y, y)$$
$$l(\hat{y}, y) = P(\hat{y} \geq y) * (2(\hat{y} - y)^2) + P(\hat{y} < y) * (\hat{y} - y)^2$$
$$l(\hat{y}, y) = 2P(\hat{y} = y) * (\hat{y} - y)^2 + 2P(\hat{y} > y) * (\hat{y} - y)^2 + P(\hat{y} < y) * (\hat{y} - y)^2$$

We can infer about the probability by using the fact that Bernoulli distribution is used for this problem. This means there are only two discrete values possible for y, which are $y = 0$ and $y = 1$. Another implication is that we want our $\hat{y}$ to range inclusively from 0 to 1. Notice that:

1. Because $\hat{y} \in [0, 1]$, then $\hat{y}$ can be greater than y if and only if $y = 0$, i.e. $P(\hat{y} > y) = P(y = 0)$.
2. $\hat{y}$ can be less than y if and only if $y = 1$, i.e. $P(\hat{y} < y) = P(y = 1)$.
3. For $\hat{y} = y$, the loss is zero so we don't need to find $P(\hat{y} = y)$.

We can then use those facts to rewrite our loss function into:

$$l(\hat{y}, y) = 2P(\hat{y} = y) * (y - y)^2 + 2 * P(y = 0) * (\hat{y} - 0)^2 + P(y = 1) * (\hat{y} - 1)^2$$

$$l(\hat{y}, y) = 2P(\hat{y} = y) * 0 + 2(1 - \theta)\hat{y}^2 + \theta(\hat{y} - 1)^2$$

$$l(\hat{y}, y) = (2 - 2\theta)\hat{y}^2 + \theta(\hat{y} - 1)^2 \text{ ... Equation 1.2}$$

To find the optimal predictor, we need to find $\hat{y}$ value such that $d(l(\hat{y}, y))/d\hat{y} = 0$.

$$\frac{d}{d\hat{y}} l(\hat{y}, y) = 0 \rightarrow (4 - 4\theta)\hat{y} + 2\theta(\hat{y} - 1)(1) = 0$$

$$4\hat{y} - 4\theta\hat{y} + 2\theta\hat{y} - 2\theta = 0 \rightarrow \hat{y}(4 - 4\theta + 2\theta) = 2\theta \rightarrow \hat{y} = \frac{2\theta}{4 - 2\theta} = \frac{\theta}{2 - \theta}$$

$$\text{Thus, the optimal prediction is } \boldsymbol{\hat{y}^* = \frac{\theta}{2 - \theta}}$$

To find the smallest expected loss, we can just substitute $\hat{y}^*$ to our loss function in Equation 1.2:

$$l(\hat{y}^*, y) = (2 - 2\theta)\hat{y}^{*2} + \theta(\hat{y}^* - 1)^2$$

$$l(\hat{y}^*, y) = (2-2\theta)(\frac{\theta}{2 - \theta})^2+\theta(\frac{\theta}{2 - \theta}-1)^2 \rightarrow l(\hat{y}^*, y) = (2-2\theta)(\frac{\theta}{2 - \theta})^2+\theta(\frac{2\theta - 2}{2 - \theta})^2$$

$$l(\hat{y}^*, y) = \frac{1}{(2 - \theta)^2}((2-2\theta)\theta^2+\theta(4\theta^2-8\theta+4)) \rightarrow l(\hat{y}^*, y) = \frac{1}{(2 - \theta)^2}(2\theta^2-2\theta^3+4\theta^3-8\theta^2+4\theta)$$

$$l(\hat{y}^*, y) = \frac{1}{(2 - \theta)^2}(2\theta^3 - 6\theta^2 + 4\theta) = \frac{2\theta(\theta - 1)(\theta - 2)}{(\theta - 2)^2} = \frac{2\theta(\theta - 1)}{\theta - 2}$$

$$\text{Thus, the smallest expected loss is } \mathbb{E}[\boldsymbol{l(\hat{y}^*, y)}] = \boldsymbol{\frac{2\theta(\theta - 1)}{\theta - 2}}$$

Problem 2

For the implementation, there are some components I need to define from the data:

1. $X \rightarrow$ the feature matrix from train data. X is a matrix of shape (4920, 8), we exclude the label here.
2. $y \rightarrow$ the label from train data. y is a vector of size 4920.
3. $b \rightarrow$ the params we seek to minimize the sum of squared errors, in other words, the ordinary least squares method. b is a vector of size 9, that includes 8 params (one for every feature) and an intercept.

The main function (*train_params*) then follows this equation:

$$b = (X^T X)^{-1} X^T y$$

To compute that, I use ***numpy*** library. If $X^T X$ is not invertible, i.e. having more than one solution, I pick the least-squares solution using *numpy.linalg.lstsq* function. More on this is explained in the code comments.

a. The feature that I select to be A is ***sex***

b. Squared loss risk of $\hat{\eta}$ on test data is: **0.21874613786114683**

c. Error rate of $\hat{f}$ on test data is **0.3252 (400/1230)**

d. False positive rate of $\hat{f}$ on test data is **0.2056 (133/647)**

e. The performances measures for $\hat{f}$ on the test data with $A = 0$ are as below:

   1. Squared loss risk: **0.2227091215683432**
   2. Error rate: **0.3354 (329/981)**
   3. False positive rate: **0.2598 (126/485)**

f. The performances measures for $\hat{f}$ on the test data with $A = 1$ are as below:

   1. Squared loss risk: **0.20313293699062618**
   2. Error rate: **0.2851 (71/249)**
   3. False positive rate: **0.0432 (7/162)**

g. The performances measures for $\hat{f}_0$ on the test data with $A = 0$ are as below:

   1. Squared loss risk: **0.22202972286135503**
   2. Error rate: **0.3374 (331/981)**
   3. False positive rate: **0.2845 (138/485)**

h. The performances measures for $\hat{f}_1$ on the test data with $A = 1$ are as below:

   1. Squared loss risk: **0.20010343642943937**
   2. Error rate: **0.2691 (67/249)**
   3. False positive rate: **0.0556 (9/162)**

i. False positive rate on the first subpopulation ($A = 0$) is much higher (approx. 22-23% higher) than the false positive rate on the second subpopulation ($A = 1$). This disparate behavior between the two subpopulations still exists even after we train our model separately for each subpopulation, i.e. that 22-23% difference happens in both training scenarios.

   This implies that for everyone in subpopulation where $A = 1$ (it can be either all men or all women, no metadata on this), they are more likely to commit a crime after they are released on parole (higher tendency of recidivism). This is because no matter how we design our training scenario, our false positive rate is always lower when $A = 1$. In other words, our prediction is more accurate to predict positive label if $A = 1$ and less accurate when $A = 0$.

   Another implication that can be drawn is related to the relative utility of the sex feature for predicting the label. Sex can be treated as the "most" important feature to predict the label, i.e. by putting a bigger weight on this feature. A concrete example would be that after we get the coefficients from the linear regression, we can reduce other features coefficient and add those reductions to the sex feature coefficient, of course, proportionately.

   If we do that, when $A = 0$, the multiplication of the sex feature time its coefficient will indeed always be zero. But, since we also reduce the other features coefficient, the inclination of predicting positive label will decrease, i.e. the predictive value is lower than the threshold of 0.5. This can result in lower false positive rate when $A = 0$.

Problem 3

For this problem, I use ***statsmodels.api.OLS*** to do the linear regression using affine function. To add the intercept, ***statsmodels.api.add_constant*** is used. Meanwhile, I use ***sklearn.preprocessing.PolynomialFeatures*** to do the features map expansion (make them quadratic or cubic) before passing them to the OLS.

a. Below are the linear plots. The peachpuff colored lines (49 in total) are the linear regression functions that are fit on their respective dataset. The average curve is depicted by the black line. Clearly, the average curve is also a linear function.

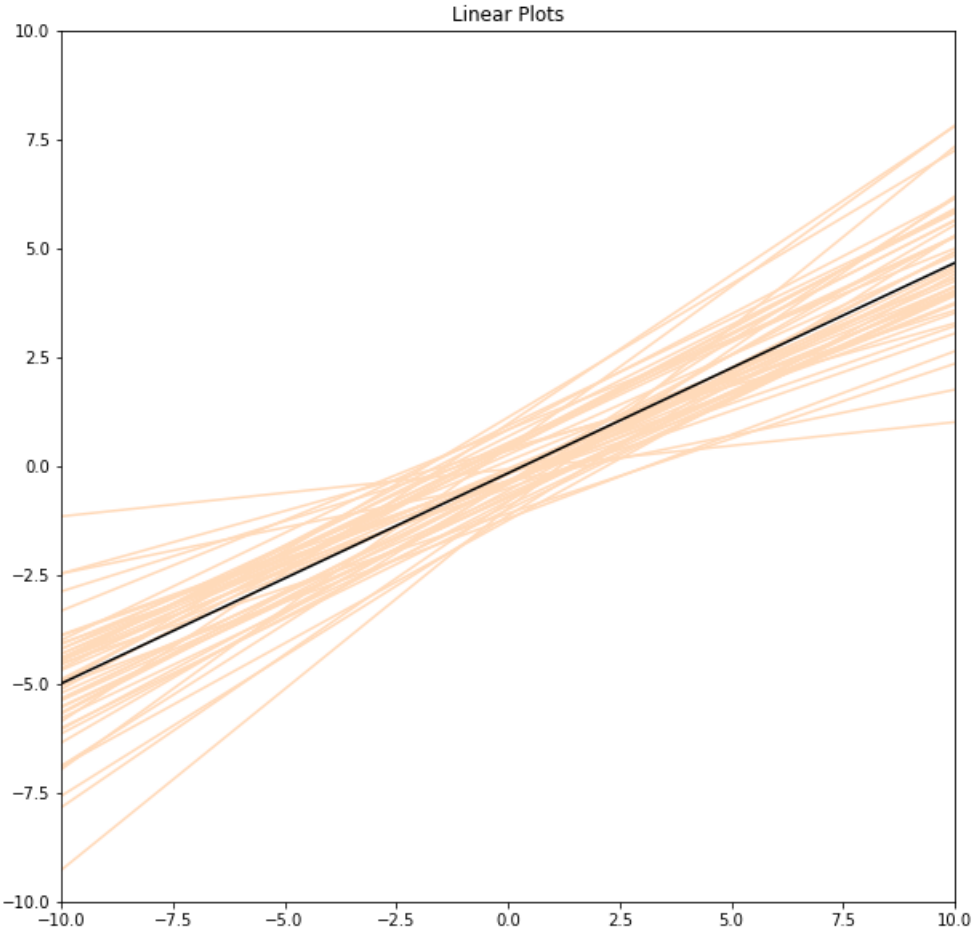

Figure 1: Linear Plots

b. Below are the quadratic plots. The peachpuff colored lines (49 in total) are the linear regression functions with quadratic feature map that are fit on their respective dataset. The average curve is depicted by the black line in the middle. Instead of being a quadratic function, the average curve is roughly a linear function, at least from what is seen in the figure.

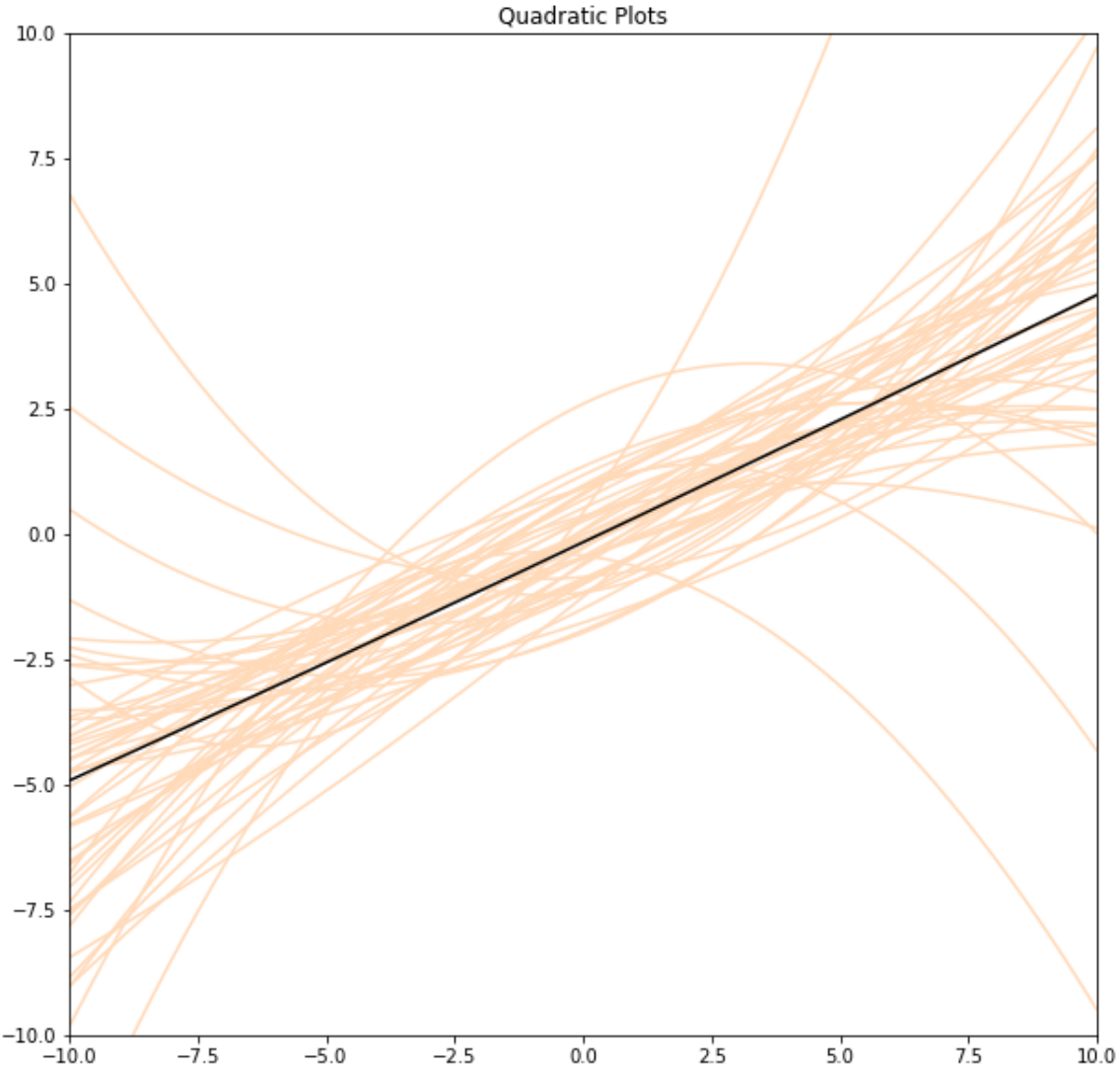

Figure 2: Quadratic Plots

c. Below are the cubic plots. The peachpuff colored lines (49 in total) are the linear regression functions with cubic feature map that are fit on their respective dataset. The average curve is depicted by the black curve in the middle. We can see that, only for this regression, the average curve is not a linear function. It starts to look like a curve towards the end (will be explained more in section 3d).

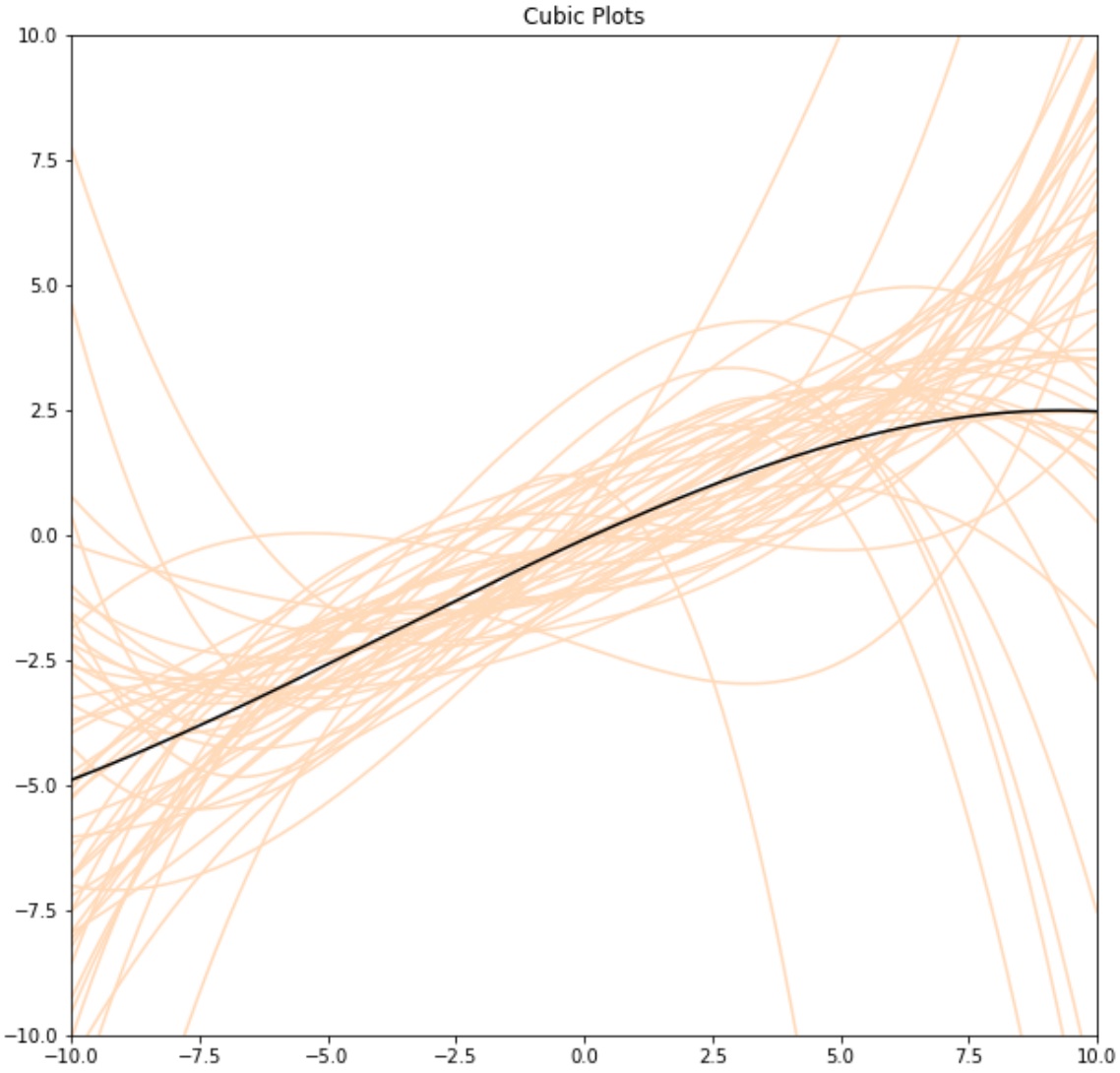

Figure 3: Cubic Plots

d. 1. Cubic Case Explanation

The peculiar result for the cubic case is that for $6 \leq x \leq 10$, the trend of the average curve begins to shift (see Figure 3). The trend is still upward, but the upward trend starts to diminish and the curve will go downward at some point later.

Let's define our affine cubic function as: $b_0 + b_1 x + b_2 x^2 + b_3 x^3$. Suppose we have two groups of affine cubic functions, one that has positive b_3 values and one that has negative b_3 values. The reason there is a trend shift is because the average of b_3 constants in the second (negative) group is significantly higher than the first (positive) group. In my case, in its absolute value, the average of the negative b_3 constants is 0.0066. Meanwhile, the average of the positive b_3 constants is just 0.0048. As x goes higher, the average value of $b_3 x^3$ will go down because the "weight" is more on the negative constants. The b_1 and b_2 constants don't really matter in this case, because as x goes higher, x^3 are much more bigger than x and x^2. Thus, b_3 is what matters the most.

Another reason is that there is a cubic function that goes down very steeply. It reaches the bottom of the plot $\rightarrow y = -10.0$ even before x reach 5.0 (see Figure 3). The b_3 value for that function is -0.0374. Its absolute value is the biggest compared to those of other 48 functions. The average of absolute b_3 values across 49 functions is 0.0057. Thus, it is also 6 times higher than the average and hence no wonder it drags down the average curve.

2. Linear and Quadratic Case Explanation

For the linear and quadratic case, we observe a similar result, that is the average curve is roughly a linear function. Suppose we define our affine function as:

1. $b_0 + b_1 x$ for linear case
2. $b_0 + b_1 x + b_2 x^2$ for quadratic case

For linear case, all b_1 values are positive so that the linear upwards trend is preserved in the average curve. For the quadratic case, suppose we divide the affine functions into two groups like we did in the cubic case before. It turns out that there is only a slight difference between the average of b_2 absolute values in the two groups. For the positive group, the average of b_2 is 0.0276. Meanwhile, for the negative group, the average of b_2 absolute values is 0.0247. The average curve stays linear in this case because the difference of b_0 and b_1 averages in those two groups probably cancel out the slight difference of b_2 averages in those two groups.

Problem 4

The two expressions that we need to compare are as below:

$$\mathbb{E}[\hat{R}(\hat{w})] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i^T w - Y_i)^2]$$

$$\mathbb{E}[R(\hat{w})] = \mathbb{E}[(X^T w - Y)^2]$$

Suppose that in the whole population, there are m examples of (X,Y) that can be drawn from the probability distribution P in which m is much bigger than n ($m >> n$). To denote the difference between the sample and the whole population, I use an apostrophe. Hence, we can rewrite our true risk expression as:

$$\mathbb{E}[R(\hat{w})] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}(X_i'^T w - Y_i')^2]$$

To simplify the empirical risk expression, we can substitute $X_i^T\hat{w}$ in the empirical risk expression with some value, say $\hat{Y}_i$, that is the predicted Y_i. Moreover, because we use the same predictor $\hat{w}$ for both the empirical and true risks, the predicted Y_i is the same for both expressions, e.g. $X_i'^T w$ is also $\hat{Y}_i$. That gives us:

$$\mathbb{E}[\hat{R}(\hat{w})] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2]$$

$$\mathbb{E}[R(\hat{w})] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i')^2]$$

Let's expand the expectation, starting from the empirical risk. To simplify the expansion, we first omit the summation expression. Using the property that $var(A) = E[A^2] - E[A]^2$ (like the one I derive in problem 1a in Equation 1.1), we have:

$$var(\hat{Y}_i - Y_i) = \mathbb{E}[(\hat{Y}_i - Y_i)^2] - (\mathbb{E}[\hat{Y}_i - Y_i])^2$$

$$\mathbb{E}[(\hat{Y}_i - Y_i)^2] = var(\hat{Y}_i - Y_i) + (\mathbb{E}[\hat{Y}_i - Y_i])^2$$

$$\mathbb{E}[(\hat{Y}_i - Y_i)^2] = var(\hat{Y}_i) + var(-Y_i) - 2cov(\hat{Y}_i, Y_i) + (\mathbb{E}[\hat{Y}_i] - \mathbb{E}[Y_i])^2$$

$$\mathbb{E}[(\hat{Y}_i - Y_i)^2] = var(\hat{Y}_i) + var(Y_i) - 2cov(\hat{Y}_i, Y_i) + (\mathbb{E}[\hat{Y}_i] - \mathbb{E}[Y_i])^2 \text{ ... (Equation 4.1)}$$

Notice that $cov(\hat{Y}_i, Y_i)$ is not zero because our estimator $\hat{w}$ is calculated using n pairs of $(X_i, Y_i) \rightarrow (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_N)$, thus $\hat{Y}_i$ depends on Y_i.

Using the same step, expand the expectation for the true risk:

$$\mathbb{E}[(\hat{Y}_i - Y_i')^2] = var(\hat{Y}_i) + var(Y_i') - 2cov(\hat{Y}_i, Y_i') + (\mathbb{E}[\hat{Y}_i] - \mathbb{E}[Y_i'])^2$$

The difference in this case is that our estimator $\hat{w}$ is not based on (derived from) $Y_i^{'}$, hence $cov(\hat{Y}_i, Y_i^{'})$ is zero. Another thing that we can infer is that because Y_i and $Y_i^{'}$ come from the same probability distribution P, both random variables will have the same expectation $\rightarrow \mathbb{E}[Y_i^{'}] = \mathbb{E}[Y_i]$. These imply:

$$\mathbb{E}[(\hat{Y}_i - Y_i^{'})^2] = var(\hat{Y}_i) + var(Y_i) - 2 * 0 + (\mathbb{E}[\hat{Y}_i] - \mathbb{E}[Y_i])^2$$

$$\mathbb{E}[(\hat{Y}_i - Y_i^{'})^2] = var(\hat{Y}_i) + var(Y_i) + (\mathbb{E}[\hat{Y}_i] - \mathbb{E}[Y_i])^2 \text{ ... (Equation 4.2)}$$

Subtract Equation 4.2 with Equation 4.1 to get:

$$\mathbb{E}[(\hat{Y}_i - Y_i)^2] - \mathbb{E}[(\hat{Y}_i - Y_i^{'})^2] = -2cov(\hat{Y}_i, Y_i)$$

$$\mathbb{E}[(\hat{Y}_i - Y_i^{'})^2] = \mathbb{E}[(\hat{Y}_i - Y_i)^2] + 2cov(\hat{Y}_i, Y_i) \text{ ... (Equation 4.3)}$$

Substitute Equation 4.3 to our definition of true risk:

$$\mathbb{E}[R(\hat{w})] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i^{'})^2]$$

$$\mathbb{E}[R(\hat{w})] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i)^2] + \frac{1}{m}\sum_{i=1}^{m}(2cov(\hat{Y}_i, Y_i))$$

We can't compare directly this true risk expression to our empirical risk definition because the population size is different (m versus n). Suppose we just calculate the true risk for n random data from the whole population. Then, there are two cases:

1. The data that we take are exactly the same as the sample $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ that we use to create the estimator $\hat{w}$. In this case, $\mathbb{E}[\hat{R}(\hat{w})] = \mathbb{E}[R(\hat{w})]$.

2. The data that we take are different from the sample used to create the estimator. In this case, we can rewrite our true risk expression as:

$$\mathbb{E}[R(\hat{w})] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2] + \frac{1}{n}\sum_{i=1}^{n}(2cov(\hat{Y}_i, Y_i))$$

.

$$\mathbb{E}[R(\hat{w})] = \mathbb{E}[\hat{R}(\hat{w})] + \frac{1}{n}\sum_{i=1}^{n}(2cov(\hat{Y}_i, Y_i))$$

The covariance value in this case would be some positive value because $\hat{Y}_i$ follows the same trend as Y_i as that is how a linear estimator $\hat{w}$ works.

Hence, for this case, $\mathbb{E}[R(\hat{w})] > \mathbb{E}[\hat{R}(\hat{w})] \rightarrow \mathbb{E}[\hat{R}(\hat{w})] < \mathbb{E}[R(\hat{w})]$.

Thus, based on those two conditions, for the same number of (n) data points involved, $\mathbb{E}[\hat{\boldsymbol{R}}(\hat{\boldsymbol{w}})] \leq \mathbb{E}[\boldsymbol{R}(\hat{\boldsymbol{w}})]$. If we were to calculate the true risk for the whole population (for m data points), the inequality will still hold because we average/normalize the squared error to get the risk so the number of the data points shouldn't matter.

Problem 5

For this problem, I use ***statsmodels.api.OLS*** to compute the affine functions $\hat{f}$, $\hat{h}_1$ and $\hat{h}_2$. To add the intercept, ***statsmodels.api.add_constant*** is used. Meanwhile, specifically for the affine function $\hat{g} = mx + b$, I compute it from scratch by finding m and b that minimize $\hat{R}_{Q_n}(g)$.

a. The test risk of $\hat{f}$ is **4.034974776250005**. The plot of outlier-ness is as below. The black dots are for the first 900 train data and the red dots are for the last 100 train data. The affine function that I got is $\boldsymbol{\hat{f}(x) = 1.065x + 0.2403}$.

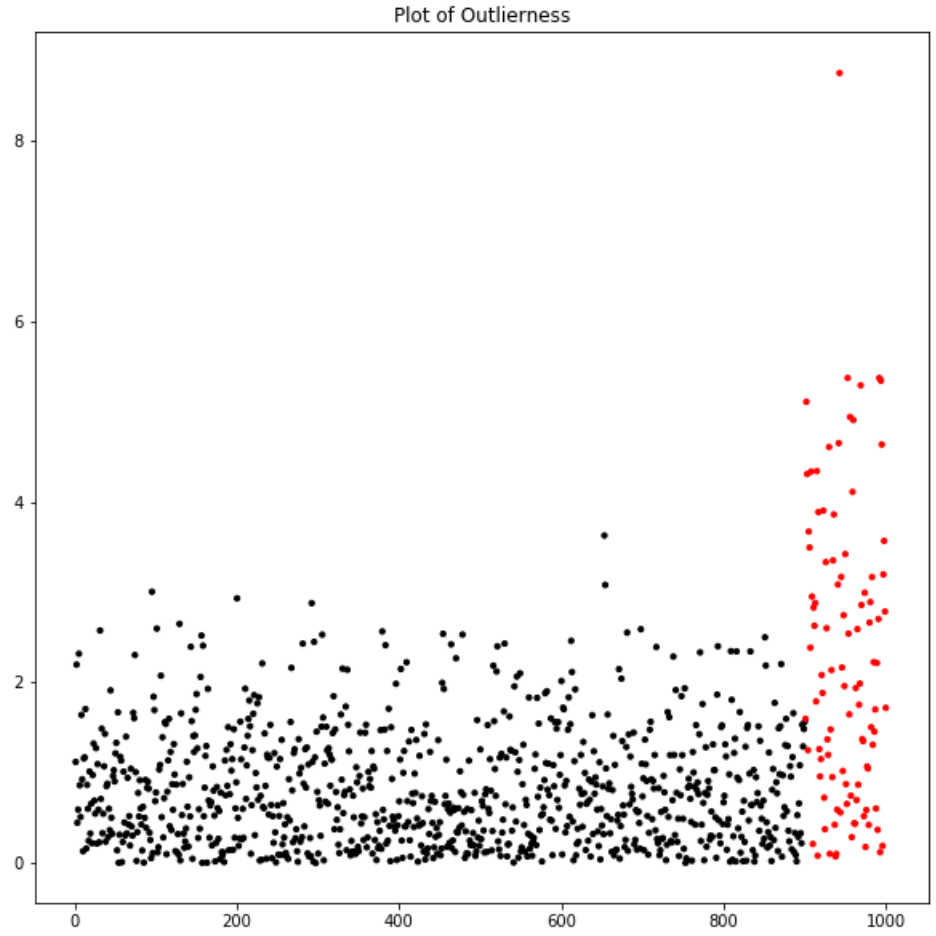

Figure 4: Plot of Outlierness

b. Let's define the affine function $\hat{g}(x) = mx + b$ and expand the risk function.

$$\hat{R}_{Q_n}(g) = \frac{1}{2*900}\sum_{i=1}^{900}(\hat{g}(x_i) - y_i)^2 + \frac{1}{2*100}\sum_{i=901}^{1000}(\hat{g}(x_i) - y_i)^2$$

$$\hat{R}_{Q_n}(g) = \frac{1}{2*900}\sum_{i=1}^{900}(m(x_i) + b - y_i)^2 + \frac{1}{2*100}\sum_{i=901}^{1000}(m(x_i) + b - y_i)^2$$

For readability, let's expand our risk equation for $1 \leq i \leq 900$ first so that the equation won't be too long.

$$\hat{R}_{Q_n}(g)_{1..900} = \frac{1}{2*900}\sum_{i=1}^{900}(m(x_i) + b - y_i)^2$$

$$\hat{R}_{Q_n}(g)_{1..900} = \frac{1}{2*900}\sum_{i=1}^{900}(m^2{x_i}^2 + 2mx_ib + b^2 - 2mx_iy_i - 2by_i + {y_i}^2)$$

$$\hat{R}_{Q_n}(g)_{1..900} = \frac{1}{2*900}(m^2\sum_{i=1}^{900}{x_i}^2 + 2mb\sum_{i=1}^{900}x_i + b^2\sum_{i=1}^{900}1 - 2m\sum_{i=1}^{900}x_iy_i - 2b\sum_{i=1}^{900}y_i + \sum_{i=1}^{900}{y_i}^2)$$

We know that the sum of all elements is the same as the mean of all the elements multiplied by the number of elements. Let's define another term j in which $\overline{x_j}, \overline{y_j}, \overline{x_jy_j}, \overline{x_j^2}, \overline{y_j^2}$ respectively denotes the average of $x_i, y_i, x_iy_i, x_i^2, y_i^2$ for $1 \leq i \leq 900$. That gives us:

$$\hat{R}_{Q_n}(g)_{1..900} = \frac{1}{2*900}(900m^2\overline{{x_j}^2} + 2*900mb\overline{x_j} + 900b^2 - 2*900m\overline{x_jy_j} - 2*900b\overline{y_j} + 900\overline{{y_j}^2})$$

$$\hat{R}_{Q_n}(g)_{1..900} = \frac{1}{2}(m^2\overline{{x_j}^2} + 2mb\overline{x_j} + b^2 - 2m\overline{x_jy_j} - 2b\overline{y_j} + \overline{{y_j}^2})$$

Using the very same steps, we can obtain a similar result for $901 \leq i \leq 1000$. By also defining another term, say k, in which $\overline{x_k}, \overline{y_k}, \overline{x_ky_k}, \overline{x_k^2}, \overline{y_k^2}$ respectively denotes the average of $x_k, y_k, x_ky_k, x_k^2, y_k^2$ for $901 \leq i \leq 1000$, the risk equation would be:

$$\hat{R}_{Q_n}(g)_{901..1000} = \frac{1}{2}(m^2\overline{{x_k}^2} + 2mb\overline{x_k} + b^2 - 2m\overline{x_ky_k} - 2b\overline{y_k} + \overline{{y_k}^2})$$

Finally, summing them would result in:

$$\hat{R}_{Q_n}(g) = \frac{1}{2}(m^2(\overline{{x_j}^2} + \overline{{x_k}^2}) + 2mb(\overline{x_j} + \overline{x_k}) + 2b^2 - 2m(\overline{x_jy_j} + \overline{x_ky_k}) - 2b(\overline{y_j} + \overline{y_k}) + (\overline{{y_j}^2} + \overline{{y_k}^2}))$$

We then want to find m and b such that $\hat{R}_{Q_n}$ is minimum. We can do this by taking the first derivative against m and against b and respectively find m and b value that make each of them zero.

$$\frac{d}{dm}(\hat{R}_{Q_n}(g)) = 0$$
$$m(\overline{x_j^2} + \overline{x_k^2}) + b(\overline{x_j} + \overline{x_k}) - (\overline{x_j y_j} + \overline{x_k y_k}) = 0 \ \text{... (Equation 5.1)}$$

$$\frac{d}{db}(\hat{R}_{Q_n}(g)) = 0$$
$$m(\overline{x_j} + \overline{x_k}) + 2b - (\overline{y_j} + \overline{y_k}) = 0 \ \text{... (Equation 5.2)}$$

We can eliminate b by multiplying Equation 5.2 with $\overline{x_j} + \overline{x_k}$ and by multiplying Equation 5.1 by 2 and then subtract Equation 5.1 with Equation 5.2 (after the multiplications). In mathematical notation, $2*(Eq.5.1)-(\overline{x_j}+\overline{x_k})*(Eq.5.2) = 0$.

$$m(2(\overline{x_j^2} + \overline{x_k^2}) - (\overline{x_j} + \overline{x_k})^2) - 2(\overline{x_j y_j} + \overline{x_k y_k}) + (\overline{x_j} + \overline{x_k})(\overline{y_j} + \overline{y_k}) = 0$$

$$m = \frac{2(\overline{x_j y_j} + \overline{x_k y_k}) - (\overline{x_j} + \overline{x_k})(\overline{y_j} + \overline{y_k})}{2(\overline{x_j^2} + \overline{x_k^2}) - (\overline{x_j} + \overline{x_k})^2}$$

We got our m. We can then substitute m in Equation 5.2 to get our b.

$$b = \frac{1}{2}((\overline{y_j} + \overline{y_k}) - \frac{2(\overline{x_j y_j} + \overline{x_k y_k}) - (\overline{x_j} + \overline{x_k})(\overline{y_j} + \overline{y_k})}{2(\overline{x_j^2} + \overline{x_k^2}) - (\overline{x_j} + \overline{x_k})^2}(\overline{x_j} + \overline{x_k}))$$

Finally, by calculating all the needed variables from each subpopulation and by using those two equations, the affine function that I got is $\boldsymbol{\hat{g}(x) = 1.1258x + 0.8411}$.

c. The test risk of $\hat{g}$ is **3.642645831591768**, better than the test risk of $\hat{f}$.

d. The test risks of each subpopulation of size 500 is as stated below in Table 1.

	Test Subpopulation 1	Test Subpopulation 2
$\hat{f}$	1.14655962195859	6.92338993054142
$\hat{g}$	1.8277141906456527	5.457577472537884

Table 1: Subpopulation Test Risks Table

e. Test risk of $\boldsymbol{\hat{h}_1}$ in test subpopulation 1 is **1.0840055002248241** and the test risk of $\boldsymbol{\hat{h}_2}$ in test subpopulation 2 is **1.015860502522174**. As additional information, $\boldsymbol{\hat{h}_1 = 1.0031x + 0.0526}$ and $\boldsymbol{\hat{h}_2 = -0.9325x + 9.7305}$.