Homework 1 (Sept. 21)

Deadline: Wednesday, October 5th, at 11:59pm.

Submission: You need to submit one file through Quercus with our answers to Questions 1, 2, 3, and 4, as well as R code and R outputs requested for Question 4. It should be a PDF file titled `hw1_writeup.pdf`. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

Neatness Point: You will be deducted one point if we have a hard time reading your solutions or understanding the structure of your code.

Late Submission: 10% of the total possible marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

- **Problem 1 (5 pts)**
  Assume that we have the regression model

$$Y = f(X) + \epsilon$$

  where ϵ is independent of X and $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = \sigma^2$.

  1. (**2 pt**) Show that
$$f(X) = \mathbb{E}[Y \mid X].$$
  2. (**2 pts**) Prove that
$$\mathbb{E}[Y \mid X] = \operatorname*{argmin}_g \mathbb{E}\Big[(Y - g(X))^2\Big].$$
     Hint: we have the fact that $\mathbb{E}[h(X,Y)] = \mathbb{E}_X \mathbb{E}_{Y|X}[h(X,Y) \mid X]$.
  3. (**1 pt**) Derive that
$$\mathbb{E}\Big[(Y - \mathbb{E}[Y \mid X])^2\Big] = \sigma^2.$$

  Parts (1) and (2) tell us that the best predictor of Y under the mean squared loss is $f(X)$. Part (3) points out why σ^2 is called the irreducible error.

- **Problem 2 (8 pts)**
  Assume that we have the regression model

$$Y = f(X) + \epsilon,$$

  where ϵ is independent of X and $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma^2$. Assume that the training data $(x_1, y_1), ..., (x_n, y_n)$ are used to construct an estimate of f, denoted by $\hat{f}$. Given a new random vector (X, Y) (independent of the training data),

1. (**3 pts**) show that

$$\mathbb{E}\Big[(f(x) - \hat{f}(x))^2 \mid X = x\Big] = \mathrm{Var}\Big(\hat{f}(x)\Big) + \Big[\mathbb{E}[\hat{f}(x)] - f(x)\Big]^2. \tag{0.1}$$

   Hint: You may benefit from adding and subtracting terms, such as

$$f(x) - \hat{f}(x) = f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x).$$

2. (**3 pts**) show that

$$\mathbb{E}\Big[\Big(Y - \hat{f}(x)\Big)^2 \mid X = x\Big] = \mathrm{Var}\Big(\hat{f}(x)\Big) + \Big(\mathbb{E}[\hat{f}(x)] - f(x)\Big)^2 + \sigma^2.$$

3. (**2 pt**) explain whether the expected MSE

$$\mathbb{E}\Big[\Big(Y - \hat{f}(X)\Big)^2\Big]$$

   can be smaller than σ^2 or not?

- **Problem 3 (8 pts)**

Solve Problem 1 on page 52 (Chapter 2.4) in the textbook "Introduction to Statistical Learning". Each sub-question is worth 2 pts.

- **Problem 2 (16 pts)**

This question should be answered using the Carseats data set which is contained in the R package ISLR. Each sub-question is worth 2 pts.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Provide an interpretation of each coefficient in the model. Be careful–some of the variables in the model are qualitative!

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$? Use the significance level 0.05 for the hypothesis test.

(e) On the basis of your response to question (d), fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(f) What are the value of R^2 for models in (a) and (e)? Does larger R^2 mean the model fit the data better?

(g) Using the model from (e), construct the 95 % confidence interval(s) for the coefficient(s).

(h) Fit a linear regression model in (e) with interaction effect(s). Provide an interpretation of each coefficient in the model.

- **Problem 1 (5 pts)**
  Assume that we have the regression model

$$Y = f(X) + \epsilon$$

where ϵ is independent of X and $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = \sigma^2$.

$E_X \; E_{Y|X}[h(X,Y) \mid X]$

1. (**2 pt**) Show that
$$f(X) = \mathbb{E}[Y \mid X].$$
2. (**2 pts**) Prove that
$$\mathbb{E}[Y \mid X] = \underset{g}{\operatorname{argmin}}\, \mathbb{E}\Big[(Y - g(X))^2\Big].$$
   Hint: we have the fact that $\mathbb{E}[h(X,Y)] = \mathbb{E}_X \mathbb{E}_{Y|X}[h(X,Y) \mid X]$.
3. (**1 pt**) Derive that
$$\mathbb{E}\Big[(Y - \mathbb{E}[Y \mid X])^2\Big] = \sigma^2.$$

Parts (1) and (2) tell us that the best predictor of Y under the mean squared loss is $f(X)$. Part (3) points out why σ^2 is called the irreducible error.

1. Since $Y = f(x) + \epsilon$, then we can let $E(Y|x) = E[f(x) + \epsilon]$ and we know $E[\epsilon] = 0$, then since $E(Y|x) = E[f(x)] + E[\epsilon]$, then $E(Y|x) = E[f(x)]$

Then, $f(x) = E[Y|x]$ since ϵ is independent of X

2. Since $Y = f(X) + \epsilon$, and

$E(Y|X) = E[f(X) + \epsilon] = E[f(X)] + E(\epsilon)$

then if we take the conditional expected MSE at any $X = x$, we have any $g(x)$ s.t.

$$E[(Y - g(X))^2] = E[[(Y - E[Y|X]) + (E[Y|X] - g(X))]^2] \Rightarrow$$

$$= E[(Y - E[Y|X])^2] +$$

$$2E[(Y - E[Y|X]) \cdot (E[Y|X] - g(X))] +$$

$E(E[Y|X] - g(X))^2$, then note that

$2\,E_{X,Y}\left[(Y-E[Y\mid X])\cdot(E[Y\mid X]-g(X))\right]$ is equals to $2E_{X,Y}\left[(Y-E[Y\mid X])(E[Y\mid X]-g(X))\right]$

according to the hint we have

$$2E_X E_{Y\mid X}\left[(Y-E[Y\mid X])(E[Y\mid X]-g(X)) \mid X\right]$$

$$=2E_X E_{Y\mid X}\left[Y\cdot E[Y\mid X]-g(X)\times Y\right.$$

$$\left.-(E[Y\mid X])^2+g(X)\times E[Y\mid X] \mid X\right]$$

Then, since we know that the expect value of a constant is constant itself, then the above equation will be $2E_X[E_{Y|X}[Y \times E[Y|X]] - g(x) \times E_{Y|X}[Y]$

$- (E[Y|X])^2 + g(x) \times E[Y|X] \mid X]$,

$$= 2E_X[[E[Y|X] \times E_{Y|X}[Y]] - g(x)E[Y|X] - (E[Y|X])^2 + g(x)E[Y|X] \mid X]$$

$$= 2E_X[\cancel{[E[Y|X]]^2} - \cancel{g(x)E[Y|X]} - \cancel{(E[Y|X])^2} + \cancel{g(x)E[Y|X]} \mid X]$$

$$= E_X[0|X] = 0$$

Therefore, $E[(Y-g(x))^2]$

$= E[(Y-E[Y|X])^2] + E[(E[Y|X] - g(x))^2]$

As we can see, when $g(x) = E[Y|X]$,

then we can minimize the whole equation $E[(Y-g(x))^2]$, therefore, we can know that $E[Y|X] = \arg\min E[(Y-g(x))^2]$

3. Since we know from part a that $E[Y|X] = f(x)$, then $E[(Y-E[Y|X])^2]$

$= E[(Y-f(x))^2] = E[(f(x)+\epsilon-f(x))^2]$

$= E[\epsilon^2] = \sigma^2$

- **Problem 2 (8 pts)**
  Assume that we have the regression model

$$Y = f(X) + \epsilon,$$

where ϵ is independent of X and $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma^2$. Assume that the training data $(x_1, y_1), ..., (x_n, y_n)$ are used to construct an estimate of f, denoted by $\hat{f}$. Given a new random vector (X, Y) (independent of the training data),

1. (**3 pts**) show that

$$\mathbb{E}\Big[(f(x) - \hat{f}(x))^2 \mid X = x\Big] = \mathrm{Var}\Big(\hat{f}(x)\Big) + \Big[\mathbb{E}[\hat{f}(x)] - f(x)\Big]^2. \qquad (0.1)$$

   Hint: You may benefit from adding and subtracting terms, such as

$$f(x) - \hat{f}(x) = f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x).$$

2. (**3 pts**) show that

$$\mathbb{E}\Big[\Big(Y - \hat{f}(x)\Big)^2 \mid X = x\Big] = \mathrm{Var}\Big(\hat{f}(x)\Big) + \Big(\mathbb{E}[\hat{f}(x)] - f(x)\Big)^2 + \sigma^2.$$

3. (**2 pt**) explain whether the expected MSE

$$\mathbb{E}\Big[\Big(Y - \hat{f}(X)\Big)^2\Big]$$

   can be smaller than σ^2 or not?

1. Since we have $E[(f(x) - \hat{f}(x))^2 | X=x]$,

and $f(x) - \hat{f}(x) = f(x) - E[\hat{f}(x)] + E[\hat{f}(x)] - \hat{f}(x)$,

then $E[(f(x) - \hat{f}(x))^2 | X=x] = E[(\underbrace{\hat{f}(x) - E[\hat{f}(x)]}_{\text{part 1}} + \underbrace{E[\hat{f}(x)] - f(x)}_{\text{part 2}})^2 | X=x]$

$= E[(\hat{f}(x) - E[\hat{f}(x)])^2 + 2(\hat{f}(x) - E[\hat{f}(x)])$

$(E[\hat{f}(x)] - f(x))] + (E[\hat{f}(x)] - f(x))^2]$

Note that, $\underbrace{2E[(\hat{f}(x) - E[\hat{f}(x)])}(E[\hat{f}(x)] - f(x)]$,

We can know that $E[\hat{f}(x)]$ is a constant number since the expected value of an estimated function is a constant, and the expected value of

a constant is still a constant, then we can rewrite

the equation as $E[(\hat{f}(x) - E[\hat{f}(x)])]$

$= E(\hat{f}(x)) - E(\hat{f}(x)) = 0$

therefore, $2E[(\hat{f}(x) - E[\hat{f}(x)])(E[\hat{f}(x)] - f(x)]$ equals to 0, and the original equation left with

$$E\left[\left(\hat{f}(x)-E[\hat{f}(x)]\right)^2\right]+E\left[\left(E[\hat{f}(x)]-f(x)\right)^2\right]$$

and according to the def, variance of x

, $Var(x) = E[(x-E(x))^2]$, and as

mention above, $E(E[\hat{f}(x)]) = E[\hat{f}(x)]$

then the final equation equals to

$$Var(\hat{f}(x)) + [E[\hat{f}(x)] - E[f(x)]]^2$$

since we know that $f(x) = B_0 + B_1 x_i$

and B_0, B_1 are unknown but they are constant,

x_i is known constant, so $f(x)$ will be a constant,

and the expected value of a constant is constant

then the final equation will become

$Var(\hat{f}(x)) + [E[\hat{f}(x)] - f(x)]^2$.

2.

? → def

$$\mathbb{E}\left[\left(Y-\hat{f}(x)\right)^2 \mid X=x\right] = \mathrm{Var}\left(\hat{f}(x)\right) + \left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2 + \sigma^2.$$

We can write the equation,

$E[(Y-\hat{f}(x))^2 | X=x]$, since we know

that $E[Y^2] = (E[Y])^2 + Var(Y)$,

then $E[(Y-\hat{f}(x))^2 | X=x]$ will equals

to $(E[Y-\hat{f}(x) | X=x])^2 + Var(Y-\hat{f}(x) | X=x)$

Since we know that $Y = f(x) + \epsilon$, and $\left(E[\,Y - \hat{f}(X) \mid X=x]\right)^2 + \mathrm{Var}(Y - \hat{f}(X) \mid X=x)$

$= \left(E[f(X) + \epsilon - \hat{f}(X) \mid X=x]\right)^2 +$

$\mathrm{Var}(f(X) + \epsilon - \hat{f}(X) \mid X=x)$

$= \left(E[f(X)] + E[\epsilon] - E[\hat{f}(X)]\right)^2 +$

$\mathrm{Var}(f(X) + \epsilon - \hat{f}(X) \mid X=x)$, since

$E[\epsilon] = 0$, and $\underline{E(f(X))}$ explained in the first part

is constant , then the equation will be

$(f(X) - E[\hat{f}(X)])^2 + \mathrm{Var}(f(X) + \epsilon - \hat{f}(X) \mid X=x)$

Since we know $f(x)$ is a constant and $Var(f(x)) = 0$, then $Var(f(x)+\epsilon-\hat{f}(x))$

$= Var(\epsilon - \hat{f}(x))$, and according to the equation $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$

then $Var(\epsilon - \hat{f}(x)) = Var(\epsilon) + Var(\hat{f}(x)) - 2Cov(\epsilon, \hat{f}(x))$

and $Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$, then in this case, $2Cov(\epsilon, \hat{f}(x)) =$

$2E[(\epsilon - E[\epsilon])(\underbrace{\hat{f}(x) - E[\hat{f}(x)]}_{})]$

Since $E[\epsilon]=0$, then $2E[\epsilon(\ )]$

which is also 0 since $E[\epsilon]=0$,

then the original equation will be

$$(E[\hat{f}(x)]-f(x)])^2+Var(\hat{f}(x))+Var(\epsilon)$$

since $Var(\epsilon)=E[\epsilon^2]-(E[\epsilon])^2$, then

since $E[\epsilon]=0$, then $Var(\epsilon)=E[\epsilon^2]$

$$=\sigma^2$$

therefore

the equation will be

$$(E[\hat{f}(x)]-f(x))^2+Var(\hat{f}(x))+\sigma^2$$

$\boxed{\text{M}}$

3. (**2 pt**) explain whether the expected MSE

$$\mathbb{E}\left[\left(Y - \hat{f}(X)\right)^2\right]$$

can be smaller than σ^2 or not?

Since $E[(Y - \hat{f}(X))^2]$

$= Var(\hat{f}(X)) + (E[\hat{f}(X)] - f(X))^2 + \sigma^2$

since we know that $Var(X)$ is always larger than 0. and the $(E[\hat{f}(X)] - f(X))^2$ will always larger than 0 after squared, therefore, the $E[(Y - \hat{f}(X))^2]$ is always larger than

G^2, [illegible].

- **Problem 3 (8 pts)**

  Solve Problem 1 on page 52 (Chapter 2.4) in the textbook “Introduction to Statistical Learning”. Each sub-question is worth 2 pts.

2.4 . Problem 1.

a). Since we have extremely large sample size, and the number of predictors is small, then the flexible method will be better b/c it will fit the data closer, instead of overfitting that is caused by small simple size.

b) The flexible method will be worse due to the overfitting issue caused by small number of observation.

c). Since the relationship is non-linear, then we can say flexible method is better to predict. b/c then inflexible will only work for linear relationship, and flexible method has more degrees of freedom.

d). Since the Var(ϵ) is extremely high, then if we use flexible method, then it will cause the fitting to the error term and instead of the relationship between two variables.