Problem set 1

- **Problem 1 (4 pts)** (This problem is to show you that why f is the best predictor of Y under the mean squared loss.)
  Assume that we have the regression model

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

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

1. (**2 pt**) Prove that for any $x \in \mathcal{X}$,

$$f(x) = \mathbb{E}[Y \mid X = x].$$

2. (**2 pts**) Prove that

$$f = \operatorname*{argmin}_g \mathbb{E}\Big[(Y - g(X))^2\Big].$$

Hint: we have the fact that $\mathbb{E}[h(X, Y)] = \mathbb{E}_X\mathbb{E}_{Y|X}[h(X, Y) \mid X]$.

SOLUTION:

1. Pick any $x \in \mathcal{X}$. We have **Proof.**

$$\begin{aligned}\mathbb{E}[Y \mid X = x] &= \mathbb{E}[f(X) \mid X = x] + \mathbb{E}[\epsilon \mid X = x] \\ &= f(x) + \mathbb{E}[\epsilon] && X \text{ is independent of } \epsilon \\ &= f(x) && \mathbb{E}[\epsilon] = 0.\end{aligned}$$

□

2. **Proof.** For any function g, we have

$$\begin{aligned}\mathbb{E}\Big[(Y - g(X))^2\Big] &= \mathbb{E}_X\mathbb{E}_{Y|X=x}\Big[(Y - g(x))^2 \mid X = x\Big] \\ &= \mathbb{E}_X\Big[\mathbb{E}[Y^2 \mid X = x] - 2g(x)\,\mathbb{E}[Y \mid X = x] + (g(x))^2\Big] \\ &= \mathbb{E}_X\Big[(\mathbb{E}[Y \mid X = x])^2 + \mathrm{Var}(Y \mid X = x) - 2g(x)\mathbb{E}[Y \mid X = x] + (g(x))^2\Big] \\ &= \mathbb{E}_X\Big[(\mathbb{E}[Y \mid X = x] - g(x))^2\Big] + \mathbb{E}_X\Big[\mathrm{Var}(Y \mid X = x)\Big].\end{aligned}$$

Since the second term is independent of g, the minimizer is $g(x) = \mathbb{E}[Y \mid X = x] = f(x)$ for all $x \in \mathcal{X}$.
Alternatively, we have

$$\begin{aligned}\mathbb{E}\Big[(Y - g(X))^2\Big] &= \mathbb{E}\Big[(f(X) + \epsilon - g(X))^2\Big] \\ &= \mathbb{E}\big[(f(X) - g(X))^2\big] + \mathbb{E}[\epsilon^2] + 2\mathbb{E}[\epsilon(f(X) - g(X))] \\ &= \mathbb{E}\big[(f(X) - g(X))^2\big] + \sigma^2 + 2\mathbb{E}[\epsilon]\;\mathbb{E}[f(X) - g(X)] \\ &= \mathbb{E}\big[(f(X) - g(X))^2\big] + \sigma^2,\end{aligned}$$

yielding the desired result. □

- **Problem 2 (6 pts)** (You will derive the Bias-Variance-Tradeoff formula in the lecture.) 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

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

SOLUTION:

1. **Proof.** Since X is independent of $\hat{f}$, we have

$$\begin{aligned}
&\mathbb{E}\Big[(f(x) - \hat{f}(x))^2 \mid X = x\Big] \\
&= \mathbb{E}\Big[(f(x) - \hat{f}(x))^2\Big] \\
&= \mathbb{E}\Big[\Big(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x)\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(f(x) - \mathbb{E}[\hat{f}(x)]\Big)^2\Big] + \mathbb{E}\Big[\Big(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\Big)^2\Big] \\
&\quad + 2\mathbb{E}\Big[\Big(f(x) - \mathbb{E}[\hat{f}(x)]\Big)\Big(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\Big)\Big] \\
&= \Big(f(x) - \mathbb{E}[\hat{f}(x)]\Big)^2 + \mathrm{Var}\Big(\hat{f}(x)\Big) + 2\Big(f(x) - \mathbb{E}[\hat{f}(x)]\Big)\Big(\mathbb{E}[\hat{f}(x)] - \mathbb{E}[\hat{f}(x)]\Big) \\
&= \Big(f(x) - \mathbb{E}[\hat{f}(x)]\Big)^2 + \mathrm{Var}\Big(\hat{f}(x)\Big)
\end{aligned}$$

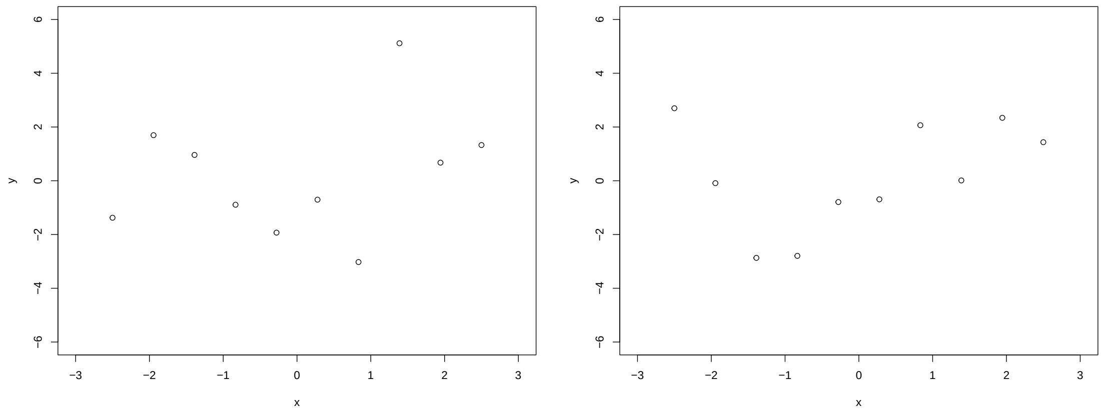
as desired. □

2. **Proof.**

$$\begin{aligned}
&\mathbb{E}\Big[\Big(Y - \hat{f}(x)\Big)^2 \mid X = x\Big] \\
&= \mathbb{E}\Big[\Big(f(x) - \hat{f}(x)\Big)^2 \mid X = x\Big] + \mathbb{E}[\epsilon^2 \mid X = x] + 2\mathbb{E}\Big[\epsilon\Big(f(x) - \hat{f}(x)\Big) \mid X = x\Big] \\
&= \mathbb{E}\Big[\Big(f(x) - \hat{f}(x)\Big)^2 \mid X = x\Big] + \mathbb{E}[\epsilon^2 \mid X = x] + 2\mathbb{E}[\epsilon \mid X = x] \cdot \mathbb{E}\Big[f(x) - \hat{f}(x) \mid X = x\Big] \\
&= \mathbb{E}\Big[\Big(f(x) - \hat{f}(x)\Big)^2 \mid X = x\Big] + \mathbb{E}[\epsilon^2] + 2\mathbb{E}[\epsilon] \cdot \mathbb{E}\Big[f(x) - \hat{f}(x) \mid X = x\Big] \\
&= \mathbb{E}\Big[\Big(f(x) - \hat{f}(x)\Big)^2 \mid X = x\Big] + \sigma^2
\end{aligned}$$

where the penultimate step uses the independence between ϵ and $\hat{f}$ while the last step uses the independence between ϵ and X. The result follows by using part (1). $\square$

- **Problem 3 (6 pts)**
  Suppose we have observed the following data points $(x_i, y_i)_{1\leq i\leq 10}$ shown in the left panel. Further suppose we have observed a different set of data points generated from the same model (see, the right panel). Answer the following questions.



  - (**1 pt**) Draw on both pictures what a fitted linear predictor (trained by using each data set) looks like (use the dotted line type or red).
  - (**1 pt**) Draw on both pictures what an overfitted predictor (trained by using each data set) looks like (use the solid line type or black).
  - (**2 pts**) For both the linear predictor and the overfitted predictor that you drew, state their predicted values for $x = 1.5$. (e.g. you should have two values for each type of predictor trained by using each data set)
  - (**2 pts**) Comment on the predicted values in the previous part as well as the variances of these two types of predictors.

SOLUTION:

  - For linear predictors, they should be two linear lines. As long as the fitted lines are reasonable, you have the full credits.
  - It should represent the overfitted patterns, including interpolation.
  - Should match with the fitted lines they provide in the above two parts.
  - It should be deduced that the variance of the linear predictor is smaller than that of the overfitted one.

- **Problem 4 (8 pts)**

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) (**2 pts**) The sample size n is extremely large, and the number of predictors p is small.

(b) (**2 pts**) The number of predictors p is extremely large, and the number of observations n is small.

(c) (**2 pts**) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

(d) (**2 pts**) The relationship between the predictors and response is highly non-linear.

SOLUTION:

1. The performance of a flexible statistical learning method would be better than an inflexible one. From the bias-variance trade off formula, when we have extremely large n, then the variance term for any model will be closed to zero hence the bias term will dominate. A flexible method will have much smaller bias than an inflexible one since it can approximate any distribution.
2. The performance of a flexible statistical learning method would be worse, since it will probably overfit.
3. If the variance of the noise is extremely large, then both flexible methods and inflexible methods will have bad performance. However, a flexible method is possible to be worse. It will have a risk of overfitting and the model mainly captures the errors in the data. An inflexible method will be more robust to the noise.
4. The performance of a flexible statistical learning method would be better. The flexible method can approximate non-linear dependency between the predictors and response better than an inflexible method such as linear regressions.

- **Problem 5 (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.

  SOLUTION: See the Sol1Q5.pdf.