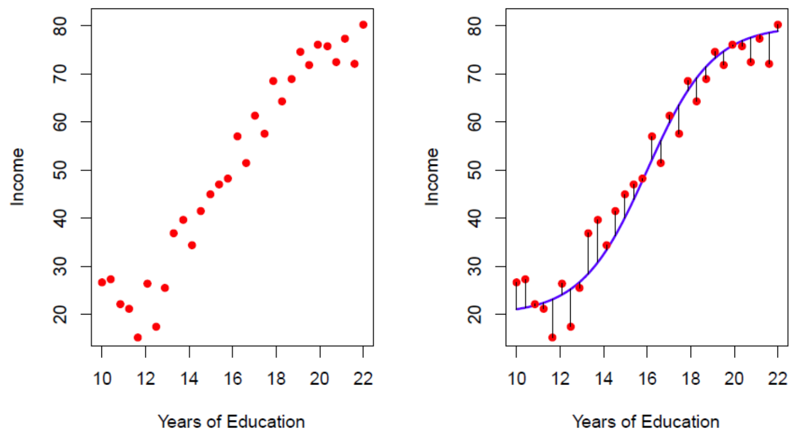
STA 325: Homework 1 (100 points + 10 bonus)

DUE: 11:59pm EST, September 14 (on Sakai)
COVERAGE: ISL Chapters 2–4

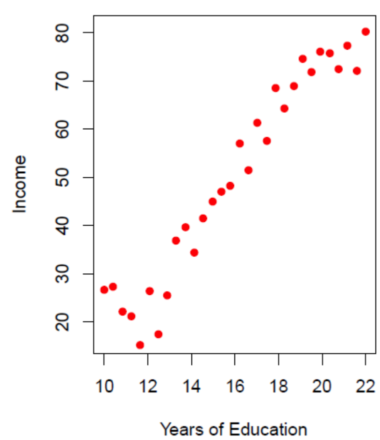
1. **[20 points]** Consider the following non-linear regression fit on annual income (in \$10,000) vs. years of education for $n = 30$ individuals:



(a) Does the model fit the data well? Justify why or why not.

(b) Plot out what the bias, variance, test MSE and training MSE curves may look like as a function of model flexibility. Justify important features in these curves.

(c) Draw out below what the fitted model $\hat{f}(\cdot)$ may look like if we assumed high model flexibility. Use this to justify the test and training MSEs in part (b).



(d) In the first plot, the fitted model $\hat{f}(\cdot)$ suggests significant slope changes at $x = 12$ and $x = 18$. Interpret what this means in terms of the problem. Based on purely income considerations, what advice would you give a graduating high-school student?

2. **[15 points]** For the classification problem (with K classes), we typically adopt the following conditional class probabilities:

$$p_k(x) = P[Y(x) = k], \quad k = 1, \cdots, K,$$

where $Y(x)$ is the discrete response at input predictors x. We discussed in-class the misclassification error measure:

$$\text{MCE}(x) := P[Y(x) \neq g(x)]$$

where $g(\cdot)$ is a chosen classifier function.

   (a) Of the variables $Y(x)$, x and $g(x)$, which are random? Which are not?

   (b) In class, it was claimed that if $p_k(x)$ is known for each k and x, then the optimal (or "Bayes-optimal") classifier which minimizes $\text{MCE}(x)$ is:

$$g(x) = k^*, \quad k^* := \underset{k=1,\cdots,K}{\text{argmax}} \; p_k(x).$$

   Argue why this is true in words (or give a simple proof), justifying each step.

   (c) Explain the intuition behind this classifier in layman's terms (i.e., to someone who is not well-versed in statistics).

   (d) Why is this predictor not that useful in practice?

3. **[10 points]** Suppose you are interested in predicting the number of hours spent on homework by freshmen and seniors. Let the predictor $x = 0$ for freshmen, and $x = 1$ for seniors. Your regression model is $Y(x) = \beta_0 + \beta_1 x + \epsilon$.

   (a) What is the interpretation of β_0?

   (b) What is the interpretation of $\beta_0 + \beta_1$?

   (c) What is the interpretation of β_1?

   (d) Do you expect the R^2 value for this model to be small or large? Why or why not?

4. **[10 points]** In logistic regression, we model the probability $p(x) = P[Y(x) = 1]$ as:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x.$$

   (a) Solve the above equation to get an expression for $p(x)$.

   (b) Is there a linear relationship between x and $p(x)$?

   (c) How do we interpret the effect of a one-unit increase in x on the probability $p(x)$?

5. **[20 points]** ISL Chapter 3, Exercise 9.

6. **[20 points]** ISL Chapter 3, Exercise 14 (omit part g).

7. **[5 points]** ISL Chapter 4, Exercise 8.

8. **[BONUS: 10 points]** Assume the general statistical model:

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

where $Y(x)$ is the response at input predictors x, and ϵ is a random error term. Instead of the MSPE discussed in class, suppose we use a different error measure – the mean absolute predictive error (MAPE):

$$\text{MAPE}(x) := \mathbb{E}[|Y(x) - g(x)|].$$

We wish to find the optimal predictor under this new MAPE error measure.

(a) Of the variables $Y(x)$, x, ϵ and $g(x)$, which are random? Which are not?

(b) Let Z be a continuous random variable with distribution function $F(\cdot)$. For any number m, show that:

$$\mathbb{E}[|Z - m|] = \int_{-\infty}^{m} (m - z)\, dF(z) + \int_{m}^{\infty} (z - m)\, dF(z).$$

(c) Define $m^* = \text{med}(Z)$ as the *median* of Z, satisfying $F(m^*) = 0.5$. Using (b), show that for any number m greater than m^*, we have:

$$\mathbb{E}[|Z - m|] - \mathbb{E}[|Z - m^*|] = (m - m^*)\,[P(Z \leq m^*) - P(Z > m^*)] + 2\int_{m^*}^{m} (m - z)\, dF(z).$$

(d) Using (c), argue that $\mathbb{E}[|Z - m|] - \mathbb{E}[|Z - m^*|] \geq 0$ for any $m > m^*$.

(e) Using (d), show that if $Y(x)$ is known for each x, the optimal predictor minimizing $\text{MAPE}(x)$ is $g(x) = \text{med}[Y(x)]$.

(f) Explain the intuition behind this predictor in layman's terms (i.e., to someone who is not well-versed in statistics).

(g) From (e), the optimal predictor minimizing MSPE (i.e., $g(x) = \mathbb{E}[Y(x)]$) is different from the optimal predictor minimizing MAPE (i.e., $g(x) = \text{med}[Y(x)]$). Give a real-world scenario where the latter predictor may be more preferable to the former.