Linear Regression

Section 3.1

Dr. Cathy Poliak, cpoliak@uh.edu

University of Houston

Goal of Statistical Learning

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, that is of gaining knowledge, making predictions, making decisions or constructing models from a set of data. This is studied in a statistical framework, that is there are assumptions of statistical nature about the underlying phenomena (in the way the data is generated).

[1]

[1] *Introduction to Statistical Learning Theory, O. Bousquet, S. Boucheron, and G. Lugosi*

mtcars Example

The goal is to predict $Y =$mpg (the response variable) based on 5 possible variables:

- X_1: cyl Number of cylinders
- X_2: disp Displacement (cu.in.)
- X_3: hp Gross horsepower
- X_4: wt Weight (1000 lbs)
- X_5: am Transmission (0 = automatic, 1 = manual)

Questions We Want to Answer

1. Is there a relationship between `mpg` and any of the predictors $X_1, \ldots, X_5$?
2. How strong is the relationship between `mpg` and any of he predictors $X_1, \ldots, X_5$?
3. Is this relationship linear?
4. How accurately can we predict `mpg`?
5. Do we only need a subset of $X_1, \ldots, X_5$ to predict `mpg` well?

General Approach

- *mpg* is the response or output. We refer to the response usually as Y.
- We have 5 input (predictor) variables.
  - X_1: `cyl` Number of cylinders
  - X_2: `disp` Displacement (cu.in.)
  - X_3: `hp` Gross horsepower
  - X_4: `wt` Weight (1000 lbs)
  - X_5: `am` Transmission (0 = automatic, 1 = manual)
- Let $X = (X_1, X_2, \ldots, X_p)^T$ be p different predictors (independent) variables.
- For this example we will have an input vector as $X = (X_1, X_2, X_3, X_4, X_5)^T$
- We assume there is some sort of relationship between X and Y, which can be written in the general form thus our model is

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

- Where ϵ captures the measurement errors and other discrepancies.
- Statistical leaning refers to a set of approaches for estimating f.

Why Estimate $f(X)$?

- If we have a good $f(X)$ we can make predictions of Y at new points where $X = x$.
- We can understand which variables (components) of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y and which ones are irrelevant.
- Depending on the complexity of f, we may be able to understand how each variable X_j of X affects Y.
- Adapted from https://hastie.su.domains/ISLR2/Slides/Ch2_Statistical_Learning.pdf

Types of Problems to Estimate $f(X)$

- **Regression** problem is when we are the response is a *continuous* or *quantitative* output value.
- **Classification** problem is when the response is a *categorical* or *qualitative* output.

Regression or Classification?

In the following examples do would we use Regression methods or Classification methods? Also are we most interested in prediction or inference? What is the sample size (n) and the number of variables (p)?

1. From the `mtcars` data, we want to predict the `mpg` based on the 5 predictors given previously from 32 automobiles.

   Regression, prediction, $n = 32$, $p = 5$ (input) + 1 outp

2. We are considering launching a new product and wish to know whether it will be a success or failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price and 10 other variables.

   Classification, prediction, $n = 20$, $p = 13$ input + 1 output

How Do We Estimate $f(X)$?

- The goal is to apply a statistical learning method to the training data in order to estimate the unknown function of f.
- Using a model-based approach, called **parametric**, with assumptions about the model.
  1. We make an assumption about the function form or shape of f.
  2. We need a procedure that uses the training data to fit or train the model.

  No assumptions about the model is called a **non-parametric** method.
-
  - Non-parametric method seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.
  - **Advantage**: they have the potential to accurately fit a wider range of possible shapes for f.
  - **Disadvantage**: a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Parametric Method

Parametric methods involve a two-step model-based approach.

1. We make an assumption about the functional form, or shape, of f. Then determine a model.
2. After a model has been selected, we need a procedure that uses the *training* data to fit or train the model.
   - The training data are observations used to train or teach our method how to estimate f.
   - Let x_{ij} represent the value of the jth predictor for observation i, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.
   - Let y_i be the response variable for the ith observation.
   - Then the training data consist of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Training, test, and validation sets

- The model is initially fit on a **training data set**, that is a set of observations used to fit the parameters.
- Successively, the fitted model is used to predict the responses for the observations in a second data set called the **validation data set**.
- Finally, the **test data set** is a data set used to provide an unbiased evaluation of a final model fit on the training data set

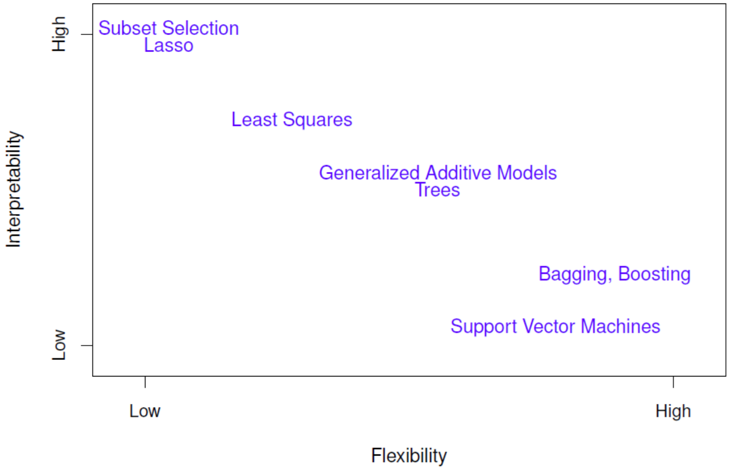
Confusingly the terms test data set and validation data set are sometimes used with swapped meaning. As a result it has become commonplace to refer to the set used in iterative training as the test/validation set and the set that is used for hyper parameter tuning as the **holdout set**.

Flexibility vs. Interpretation

Why choose to use a more restrictive method instead of a very flexible approach?

- If we are mainly interested in inference, then restrictive models are much more interpret-able. For example, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and $X_1, X_2, \ldots, X_p$.
- Very flexible approaches, such as the splines and the boosting methods can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

Flexibility versus Interpretation



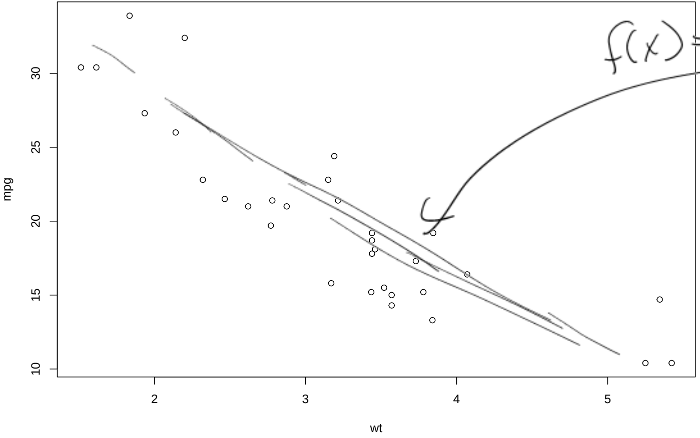
Back to Mtcars Example

- Suppose we want to be able to predict *mpg* based on *wt* from the `mtcars` data frame.
- *mpg* is the response or output. We refer to the response usually as Y.
- We have an input variable or predictor $X = wt$
- We assume there is some sort of relationship between X and Y, which can be written in the general form thus our model is

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

- Where ϵ captures the measurement errors and other discrepancies.
- Our goal is to estimate $f(X)$, by $\hat{f}(X)$ and to see how well $\hat{f}(X)$ can help us determine Y.
- In this example will we have a **regression** or **classification** statistical learning problem?

What would be a good function for $f(X)$?



$f(x) = \beta_0 + \beta_1 X$

Simple Linear Regression Model

- The data are n observations on an explanatory variable x and a response variable y,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- The statistical model for simple linear regression states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \quad \varepsilon_i \sim N(0, \sigma^2)$$

- $\mu_y = \beta_0 + \beta_1 x_i$ is the mean response for y when $x = x_i$ a specific value of x.
- ϵ_i are the error terms for predicting y_i for each value of x_i.
- Notice in our general form that $f(X) = \beta_0 + \beta_1 X$.

Parameters of the Simple Regression Model

- The intercept: β_0.
- The slope: β_1.
- The goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that for each observed y_i, $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$, for $i = 1, 2, \ldots, n$.
- The most common approach is by the minimizing the least squares criterion.

Principle of Least Squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X.
- Then $e_i = y_i - \hat{y}_i$ be the ith residual, the difference between the *ith* observed response value and the ith predicted value by our linear equation.
- The **residual sum of squares** (RSS) is defined by

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates**, are those values that minimize the RSS.

The Least - Squares Estimates $s_x^2 = Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- The method of **least squares** selects estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the **residual sum of squares** (RSS).
- Where the estimate of the slope coefficient β_1 is:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- The estimate for the intercept β_0 is:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \sum_{i=1}^{n} x_i$.

Determining the esitmates of the coefficients

- To minimize $\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2$, we take the partial derivative of the expression with respect to $\hat{\beta}_0$ and also $\hat{\beta}_1$. Then set to zero to determine the minimum.

Determining the esitmates of the coefficients

- To minimize $\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2$, we take the partial derivative of the expression with respect to $\hat{\beta}_0$ and also $\hat{\beta}_1$. Then set to zero to determine the minimum.
- Determining expression for estimate of β_0.

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2 = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)$$

Determining the esitmates of the coefficients

- To minimize $\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2$, we take the partial derivative of the expression with respect to $\hat{\beta}_0$ and also $\hat{\beta}_1$. Then set to zero to determine the minimum.
- Determining expression for estimate of β_0.

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2 = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)$$

$$\begin{aligned}
0 &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1) \\
0 &= \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
\hat{\beta}_0 &= \bar{y} - \bar{x}\hat{\beta}_1
\end{aligned}$$

Estimate of Slope β_1

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1)^2 = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - x_i \hat{\beta}_1 \right)$$

Estimate of Slope β_1

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1)^2 = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - x_i \hat{\beta}_1 \right)$$

$$0 = \sum_{i=1} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Estimate of Slope β_1

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2 = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1\right)$$

$$0 = \sum_{i=1} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$0 = \sum_{i=1}^{n} x_i y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Estimate of Slope β_1

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1)^2 = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - x_i \hat{\beta}_1 \right)$$

$$0 = \sum_{i=1} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$0 = \sum_{i=1}^{n} x_i y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$0 = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i \right)^2 - n \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_1 \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{cov(x,y)}{s_x^2}$$

$$= \frac{cor(x,y)\, s_y s_x}{s_x^2}$$

$$\hat{\beta}_1 = cor(x,y) \frac{s_y}{s_x}$$

MPG Example

- Use the *mtcars* data frame.
- We want to predict *mpg* based on *wt*.
  1. Determine if it is a linear relationship. How can we tell?
  2. Get an estimate of the model.
  3. Is this a good fit for the data?

mean(mpg) = $\bar{y}$ = 20.09062 ; mean(wt) = $\bar{x}$ = 3.21725

sd(mpg) = s_y = 6.02695 ; sd(wt) = s_x = 0.97846

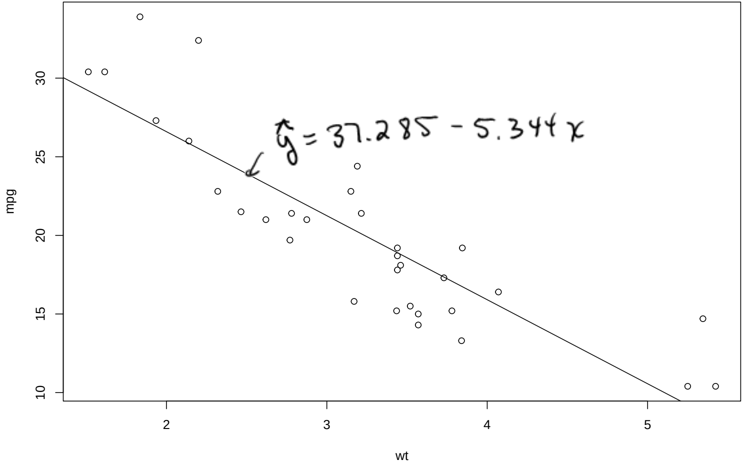
cor(wt, mpg) = r = −0.86765

slope = $\hat{\beta}_1 = -0.86765\left(\frac{6.02695}{0.97846}\right) = -5.3444$

y-intercept = $\hat{\beta}_0$ = 20.09062 − 3.21725(−5.3444) = 37.285

$\hat{y} = \hat{f}(x) = 37.285 - 5.3444x$

Do We Have A Linear Relationship?



The Estimate of the Model

```r
mpg.lm = lm(mpg~wt,mtcars)
summary(mpg.lm)
```

```
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851 = β̂0  1.8776  19.858  < 2e-16 ***
wt           -5.3445 = β̂1  0.5591  -9.559 1.29e-10 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Measuring the Quality of Fit

- We want to quantify the how close the predicted value is to the actual observed value.
- For the regression problem we commonly use the **mean squared error** (MSE) given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the i^{th} observation.

- The MSE will be small if the predicted responses are very close to the true responses.
- We will be more interested in the MSE with the test set than with the training set.

MSE for Predicting `mpg`

- We can assume a linear relationship between `wt` and `mpg` and make an estimate $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 \times wt$.
- We will train on 80% (26) of the data then test based on the remaining 20% (6) of the observations.

```
set.seed(1)
car.train.index = sample(1:32,26)
car.train = mtcars[car.train.index,]
car.test = mtcars[-car.train.index,]
mpg.lm = lm(mpg ~ wt, data = car.train)
car.test.pred = predict(mpg.lm,newdata = car.test)
car.train.pred = predict(mpg.lm)
(mpg.mse.train = 1/26*sum((car.train$mpg - car.train.pred)^2))
```

```
[1] 8.151513
```

```
(mpg.mse.test = 1/6*sum((car.test$mpg - car.test.pred)^2))
```

```
[1] 12.48688
```

Estimators

A statistic $\hat{\theta}$ used to estimate an unknown population parameter θ is called an **estimator**.

- Properties of an estimator $\hat{\theta}$
  - Accuracy - measured by **bias**

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$= E(\theta - E(\hat{\theta}))^2$

  - Precision - measured by its variance, Var$((\hat{\theta})$. The estimated standard deviation of an estimator θ is referred to as its **standard error (SE)**.
  - The **mean squared error (MSE)** combines both measures.

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

- In MATH 3339 we studied estimators for μ and p. In this class we will we will want estimators for $f(X)$.

Example, Estimate of μ

Suppose we take a random sample of 4 from a Normal distribution with $\mu = 10$ and $\sigma = 2$. $E(x_i) = 10$ $\quad var(x_i) = 2^2 = 4$

- Let $\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i$ be an estimator of μ. What is the expected value, bias, variance, and MSE of $\bar{x}$.

$E(\bar{x}) = E\left[\frac{1}{4}\sum_{i=1}^{4} x_i\right] = \frac{1}{4}\sum_{i=1}^{4} E(x_i) = \frac{1}{4}[10+10+10+10] = 10$

$Bias(\bar{x}) = E(\bar{x}) - \mu = 10 - 10 = 0$ "unbiased"

$var(\bar{x}) = var\left[\frac{1}{4}\sum_{i=1}^{4} x_i\right] = \frac{1}{16}\sum_{i=1}^{n} var(x_i) = \frac{1}{16}[4+4+4+4] = 1$

$= \sigma^2/n$

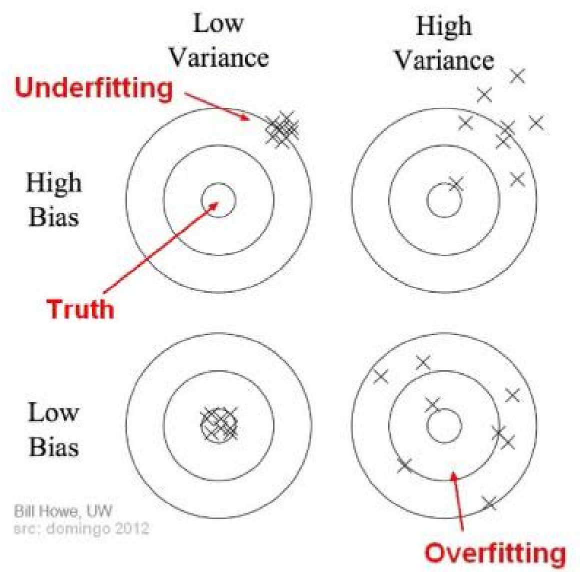
$MSE[\bar{x}] = var(\bar{x}) + Bias(\bar{x})^2 = 1 + 0^2 = 1$

- Let 8 be an estimator of μ. What is the expected value, bias, variance, and MSE of 8?

$E(8) = 8 \quad Bias(8) = 8 - 10 = -2$

$var(8) = 0 \quad MSE(8) = 0 + (-2)^2 = 4$

The Bias-Variance Trade-Off



Bias and Variance for $\hat{f}$

- **Variance** refers to the amount by which $\hat{f}$ would change we estimated it using a different training data set.
- In general more flexible statistical methods have higher variance.
- **Bias** refers to the error that is introduced by approximating a real-life problem by a simpler model.
- In general a more flexible statistical method have lower bias.
- We desire to have low variance and low bias.