S&DS 242/542: Theory of Statistics

Lecture 21: Models with covariates and response

Predictive modeling

Predicting income from education and seniority (ISLR)

For n individuals, we observe their:

- Annual income
- Years of education
- Seniority

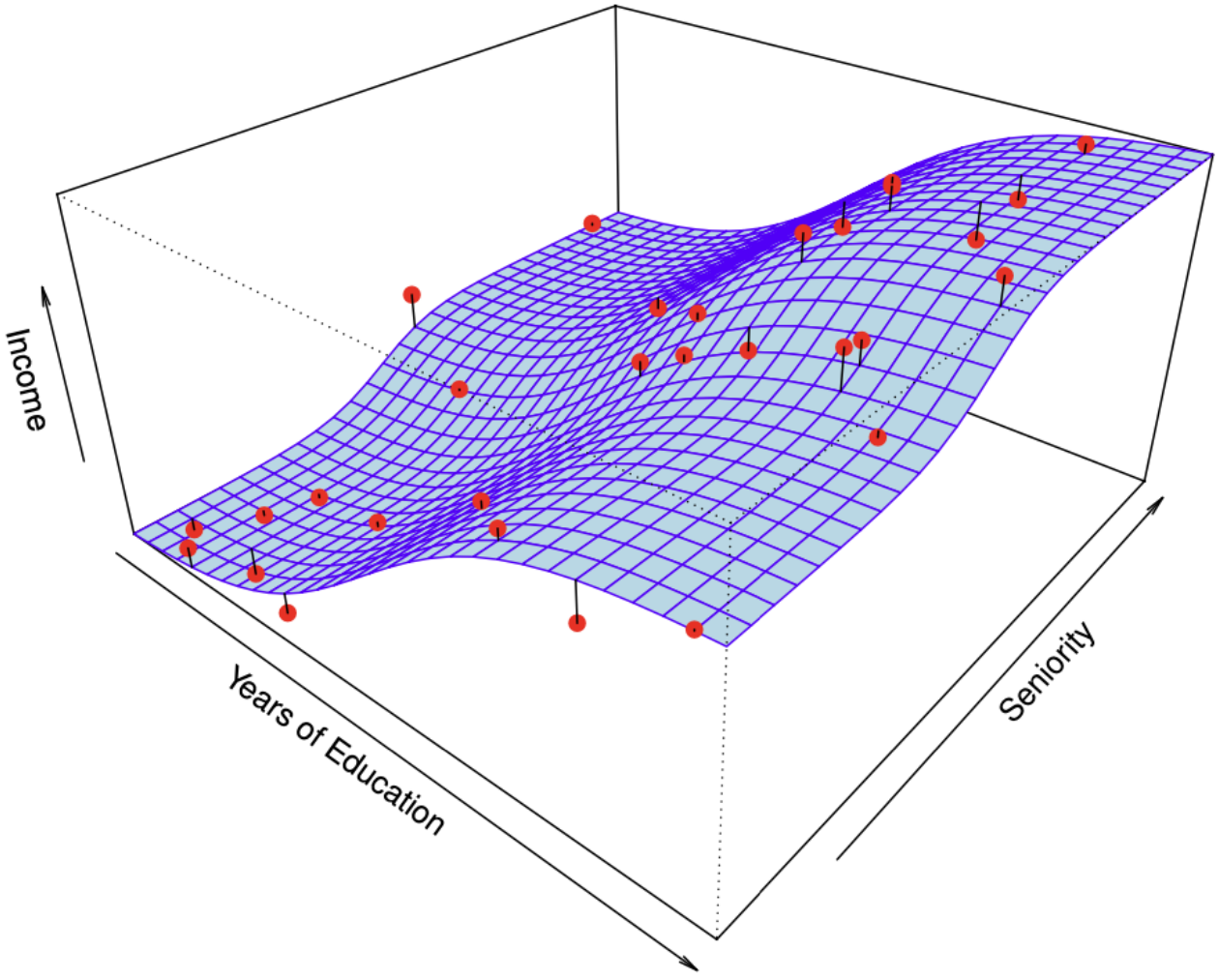
We wish to build a predictive model for income based on years of education and seniority.

Here income is the *output variable*, *response*, or *dependent variable*. We will denote these by $Y_1, \ldots, Y_n$.

Years of education and seniority are the *input variables*, *covariates*, *predictors*, or *independent variables*. We will denote these by $X_1, \ldots, X_n$. Each X_i is the tuple of predictors for the i^{th} individual.

We will call $(X_1, Y_1), \ldots, (X_n, Y_n)$ our *training data*.

Predicting income from education and seniority (ISLR)



Goal: Learn a function $f(\cdot)$ for which $Y \approx f(X)$

Prediction and inference

There are two types of reasons why we may wish to learn the relationship $Y \approx f(X)$.

Prediction:

- Given a new input X_{test}, predict the response Y_{test}
- $f(\cdot)$ is a black box, and we may not be interested in its form

Inference:

- Understand how X is associated to Y: What is the form of the relationship between each predictor and the response? Which predictors actually influence the response? How would perturbing a predictor affect the response?
- We may want $f(\cdot)$ to have an interpretable and simple form

Both goals require learning a "true" relationship between X and Y, rather than some $f(\cdot)$ that only fits well on our training data.

Formulations of the data model

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{\text{test}}, Y_{\text{test}})$ are i.i.d., and let (X, Y) represent the distribution of a single observation.

There are different models we may posit for the data, expressing different assumptions about the distribution of (X, Y):

- The distribution of (X, Y) is arbitrary, and we wish to learn the regression function

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

  Writing $Y = f(X) + \varepsilon$, this means that conditional on $X = x$, ε has mean 0 but its distribution may otherwise depend on x.

- The distribution of X is arbitrary, and

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

  where ε has mean 0 and is independent of X.

Formulations of the data model

- The distribution of X is arbitrary, and

$$Y = f(X \mid \beta) + \varepsilon$$

  where, in additional to the previous assumptions for ε, the function $f(\cdot \mid \beta)$ has a parametric form with parameter β.

  For example, if $X = (x_1, \ldots, x_p)$ has p quantitative predictors, we may consider a linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

  with parameters $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{R}^{p+1}$.

We will oftentimes make a distinction between the model for (X, Y) that we choose to fit, vs. the actual distribution of the data.

Irreducible vs. reducible prediction error

Suppose we predict Y by a function $g(X)$. Fixing a test point x, we may be interested in the mean-squared prediction error

$$\mathbb{E}[(Y_{\text{test}} - g(x))^2 \mid X_{\text{test}} = x]$$

or its average over the distribution of X_{test}.

Suppose $Y = f(X) + \varepsilon$ where ε is mean-0 and independent of X. We have:

$$\mathbb{E}[(Y_{test} - g(x))^2 \mid X_{test} = x]$$

$$= \mathbb{E}[(f(x) + \varepsilon - g(x))^2 \mid X_{test} = x]$$

$$= \mathbb{E}[(f(x) - g(x))^2] + \underbrace{\mathbb{E}[2(f(x) - g(x)) \cdot \varepsilon]}_{=0} + \mathbb{E}[\varepsilon^2]$$

$$= (f(x) - g(x))^2 + \mathrm{Var}[\varepsilon]$$

Irreducible vs. reducible prediction error

Thus

$$\mathbb{E}[(Y_{\text{test}} - g(x))^2 \mid X_{\text{test}} = x] = \underbrace{(f(x) - g(x))^2}_{\text{reducible error}} + \underbrace{\text{Var}[\varepsilon]}_{\text{irreducible error}}$$

- We would incur the irreducible error even if we had perfect knowledge of the true function $f(\cdot)$. This error is non-zero if Y cannot be perfectly predicted given X, e.g. if there are unmeasured variables that also influence Y.
- The reducible error is the error that we may hope to minimize by learning an accurate prediction model $g(\cdot)$. In the best case, $g(x) = f(x)$ and the reducible error is 0.

Bias-variance decomposition of reducible error

In practice, we estimate $f(x)$ by a function $\hat{f}(x)$ that is random and depends on our training data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Still fixing the test point x, let us define the *bias* and *variance* of the prediction at x by

$$\text{Bias} = \mathbb{E}[\hat{f}(x)] - f(x)$$
$$\text{Variance} = \text{Var}[\hat{f}(x)]$$

where $\mathbb{E}[\cdot]$ and $\text{Var}[\cdot]$ are over $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Then $\mathbb{E}[(\text{reducible error})^2] = \mathbb{E}\left[(f(x) - \hat{f}(x))^2\right]$

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

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

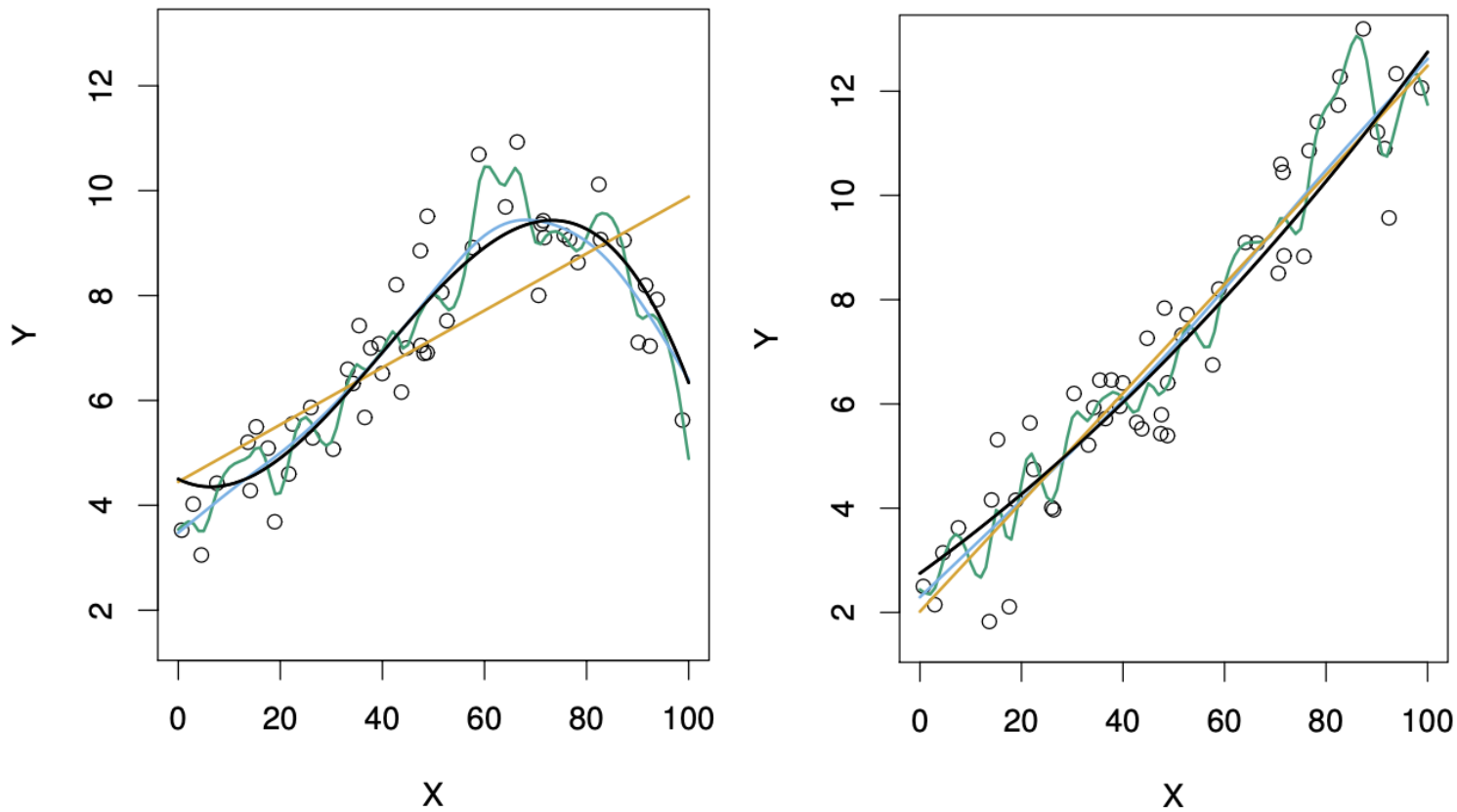
Three-fold decomposition of prediction error

Thus the prediction mean-squared-error at a test point x is

$$\begin{aligned}
&\mathbb{E}[(Y_{\text{test}} - \hat{f}(x))^2 \mid X_{\text{test}} = x] \\
&\qquad = \underbrace{(\mathbb{E}[\hat{f}(x)] - f(x))^2}_{\text{squared bias}} + \underbrace{\text{Var}[\hat{f}(x)]}_{\text{variance}} + \underbrace{\text{Var}[\varepsilon]}_{\text{irreducible error}}
\end{aligned}$$

- Variance refers to the variability of the our prediction $\hat{f}(x)$ under different possible realizations of our training data. Large variance may indicate that we have *overfitted* to the training data.
- Bias refers to a difference between the true regression function $f(x)$ and our prediction $\hat{f}(x)$ averaged over the randomness of the training data. Large bias may indicate that the true $f(x)$ is far from the class of functions represented by our model.

Bias-variance tradeoff in prediction error



Black curve: True regression function $f(x)$

Orange, blue, green curves: Three different estimates $\hat{f}(x)$

Considerations for model complexity

The choice between fitting a simple model vs. a complex model may be determined by two trade-offs:

- *Prediction vs. inference:* To what extent do we care about the predictive accuracy of the model, and to what extent do we care about its interpretability and the meaning it encodes for the relationship between X and Y?
- *Bias vs. variance:* Even if our primary goal is predictive accuracy, what would be the right model complexity that balances bias and variance of our model predictions? This depends on the amount of available data and the complexity of the true regression function $f(x)$.

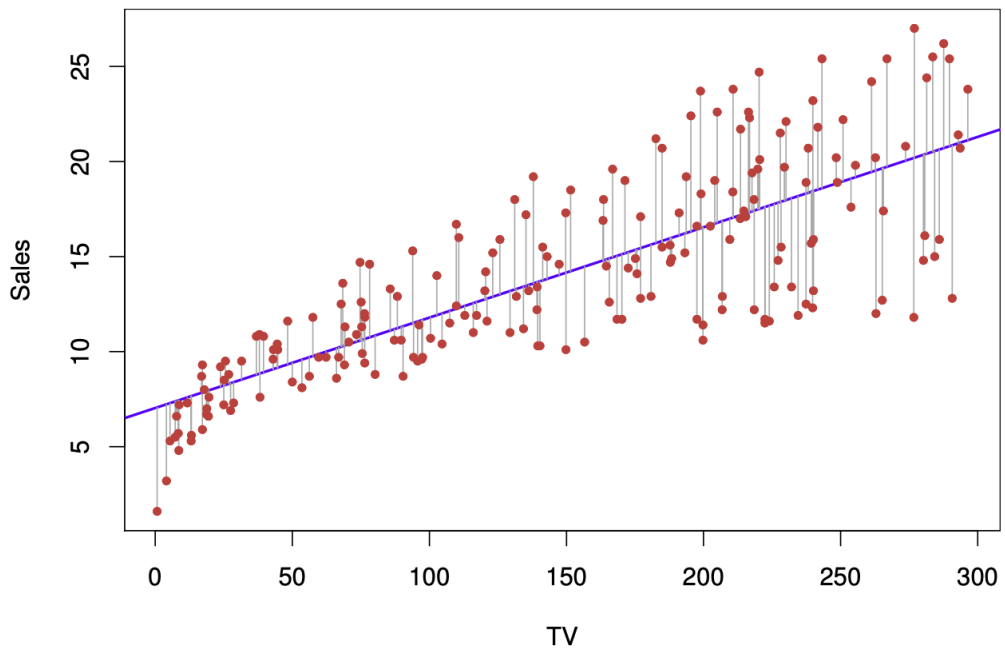
Simple linear regression

Simple linear regression

Consider data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X \in \mathbb{R}$ is a single quantitative predictor. The simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

with parameters β_0, β_1, representing the slope and intercept of the regression line.



The population regression line

Restricting ourselves to this linear model, the parameters we wish to learn may be the *population least-squares coefficients*

$$\beta_0^*, \beta_1^* = \arg\min_{\beta_0,\beta_1} \mathbb{E}[(\beta_0 + \beta_1 X - Y)^2]$$

These parameters minimize the prediction mean-squared-error over the true distribution of data.

These have an explicit form:

$$0 = \frac{\partial}{\partial \beta_0} \mathbb{E}[(\beta_0 + \beta_1 X - Y)^2]$$

$$= \mathbb{E}[2(\beta_0 + \beta_1 X - Y)] = 2\beta_0 + 2\beta_1 \cdot \mathbb{E}X - 2\mathbb{E}Y$$

$$0 = \frac{\partial}{\partial \beta_1} \mathbb{E}[(\beta_0 + \beta_1 X - Y)^2]$$

$$= \mathbb{E}[2X(\beta_0 + \beta_1 X - Y)] = 2\beta_0 \cdot \mathbb{E}X + 2\beta_1 \cdot \mathbb{E}[X^2] - 2 \cdot \mathbb{E}[XY]$$

The population regression line

$$\Rightarrow \beta_0 \cdot \mathbb{E}X + \beta_1 \cdot (\mathbb{E}X)^2 = \mathbb{E}X \cdot \mathbb{E}Y$$

$$\beta_0 \cdot \mathbb{E}X + \beta_1 \cdot \mathbb{E}[X^2] = \mathbb{E}[XY]$$

$$\Rightarrow \beta_1^* = \frac{\mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y}{\mathbb{E}[X^2] - (\mathbb{E}X)^2} = \frac{Cov[X,Y]}{Var[X]}$$

$$\beta_0^* = \mathbb{E}Y - \beta_1^* \cdot \mathbb{E}X$$

Correlation and regression

Suppose X, Y both have mean 0 and variance 1, and their correlation is $\text{Corr}(X, Y) = \rho$. Then

$$\beta_0^* = 0, \quad \beta_1^* = \rho,$$

and the population least-squares regression line is $Y = \rho X$.

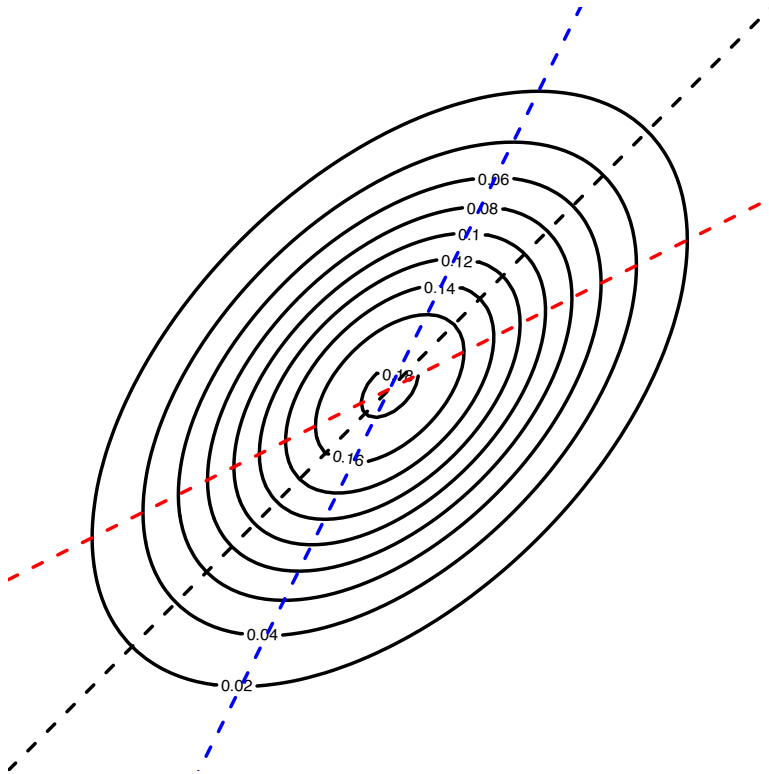
This is *not* symmetric between X and Y. If we were predicting X from Y, then the least-squares line would be $X = \rho Y$, i.e. $Y = \frac{X}{\rho}$.

Interpretation: Let $(X, Y) = (\text{height of father}, \text{height of son})$. Given that the father is 1 std. dev. taller than average, the son is only ρ std. dev. taller than average in expectation.

(Similarly, given that the son is 1 std. dev. taller than average, the father is only ρ std. dev. taller than average in expectation.)

Correlation and regression

Example: Suppose (X, Y) are bivariate normal w/ correlation 0.5



Black: $Y = X$

Red: Least-squares regression line for predicting Y from X

Blue: Least-squares regression line for predicting X from Y

Least squares estimation

We may fit β_0, β_1 by minimizing the average squared error on our training data,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i - Y_i)^2$$

These are the **least-squares estimators** for β_0, β_1.

The same derivation as for the population coefficients β_0^*, β_1^* shows that these estimators have the closed-form expressions

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

A maximum likelihood interpretation of least squares

Suppose that $X_1, \ldots, X_n \overset{IID}{\sim} f_X$, and

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_n \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$ are normal errors independent of $X_1, \ldots, X_n$. Then $(\hat{\beta}_0, \hat{\beta}_1)$ is the MLE for (β_0, β_1):

Joint density of (X, Y): $f_X(x) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\beta_0 + \beta_1 x - y)^2}$

$$\Rightarrow \ell_n(\beta_0, \beta_1) = \sum_{i=1}^{n} \log\left[f_X(x_i) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\beta_0 + \beta_1 x_i - y_i)^2} \right]$$

$$= \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(\beta_0 + \beta_1 x_i - y_i)^2 + \log\left[f_X(x_i) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \right]$$

Maximizing $\ell_n(\beta_0, \beta_1) \Leftrightarrow$ Minimizing $\frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2$

Unbiasedness in a correctly specified model

Theorem

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ *are i.i.d. and distributed according to a true model*

$$Y = \beta_0^* + \beta_1^* X + \varepsilon$$

where ε *has mean 0 and is independent of* X*. Then* $\mathbb{E}[\hat{\beta}_0] = \beta_0^*$ *and* $\mathbb{E}[\hat{\beta}_1] = \beta_1^*$*, i.e.* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ *is unbiased for* $\beta^* = (\beta_0^*, \beta_1^*)$*.*

Proof:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Denote $\mathbf{X} = (X_1, \ldots, X_n)$

$\mathbf{x} = (x_1, \ldots, x_n)$

$$\mathbb{E}[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}] = \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[Y_i - \bar{Y} \mid \mathbf{X} = \mathbf{x}]}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Unbiasedness in a correctly specified model

We have: $\mathbb{E}[Y_i \mid X = x] = \beta_0^* + \beta_1^* x_i$

$\mathbb{E}[\bar{Y} \mid X = x] = \beta_0^* + \beta_1^* \cdot \bar{x}$

$$\Rightarrow \mathbb{E}[\hat{\beta}_1 \mid X = x] = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (x_i - \bar{x}) \cdot \beta_1^*}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1^*$$

$$\Rightarrow \mathbb{E}[\hat{\beta}_1] = \beta_1^*$$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X} \Rightarrow \mathbb{E}[\hat{\beta}_0 \mid X = x]$

$= \mathbb{E}[\bar{Y} \mid X = x] - \mathbb{E}[\hat{\beta}_1 \mid X = x] \cdot \bar{x}$

$= (\beta_0^* + \beta_1^* \bar{x}) - \beta_1^* \bar{x} = \beta_0^* \Rightarrow \mathbb{E}[\hat{\beta}_0] = \beta_0^*$.

(In fact, this only required the assumption $\mathbb{E}[\varepsilon \mid X = x] = 0$ for each $x \in \mathbb{R}$, i.e. ε has mean-0 conditional on X.)

Consistency for the population coefficients

Theorem

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ *are i.i.d., and let*

$$(\beta_0^*, \beta_1^*) = \arg\min_{\beta_0, \beta_1} \mathbb{E}[(\beta_0 + \beta_1 X - Y)^2]$$

Then $(\hat{\beta}_0, \hat{\beta}_1) \to (\beta_0^*, \beta_1^*)$ *in probability as* $n \to \infty$, *i.e.* $\hat{\beta}$ *is consistent for* β^*.

Proof: $\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2}$

As $n \to \infty$, by LLN

- $\frac{1}{n}\sum_{i=1}^n X_i Y_i \to \mathbb{E}[XY]$
- $\bar{X} \to \mathbb{E}X$
- $\frac{1}{n}\sum_{i=1}^n X_i^2 \to \mathbb{E}[X^2]$
- $\bar{Y} \to \mathbb{E}Y$

Consistency for the population coefficients

$$\Rightarrow \hat{\beta}_1 \to \frac{\mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y}{\mathbb{E}[X^2] - (\mathbb{E}X)^2} = \beta_1^*$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

$$\to \mathbb{E}Y - \beta_1^* \cdot \mathbb{E}X = \beta_0^*$$

Decomposition of prediction error

Recall our decomposition of prediction mean-squared-error at a fixed test point x:

$$\begin{aligned}&\mathbb{E}[(Y_{\text{test}} - \hat{f}(x))^2 \mid X_{\text{test}} = x] \\ &\qquad = \underbrace{(\mathbb{E}[\hat{f}(x)] - f(x))^2}_{\text{squared bias}} + \underbrace{\text{Var}[\hat{f}(x)]}_{\text{variance}} + \underbrace{\text{Var}[\varepsilon]}_{\text{irreducible error}}\end{aligned}$$

Suppose the true model is $Y = \beta_0^* + \beta_1^* X + \varepsilon$. Then

- The true regression function is $f(x) = \beta_0^* + \beta_1^* x$.
- The estimated regression function is $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. Since $\hat{\beta}$ is unbiased, $\mathbb{E}[\hat{f}(x)] = f(x)$ so the squared bias is 0.
- As $n \to \infty$, since $\hat{\beta} \to \beta^*$ in probability, we have also $\hat{f}(x) \to f(x)$ in probability. Then we expect $\text{Var}[\hat{f}(x)] \to 0$.

Assuming that the irreducible error $\text{Var}[\varepsilon] > 0$ is fixed for all n, the prediction error is dominated by this irreducible error for large n.

Decomposition of prediction error

More generally, suppose $Y = f(X) + \varepsilon$ where ε is mean-0 and independent of X. Let

$$(\beta_0^*, \beta_1^*) = \arg\min_{\beta_0, \beta_1} \mathbb{E}[(\beta_0 + \beta_1 X - Y)^2]$$

Then, as $n \to \infty$,

- Since $\hat{\beta} \to \beta^*$ in probability, also $\hat{f}(x) \to \beta_0^* + \beta_1^* x$ in probability. Then we expect $\mathbb{E}[\hat{f}(x)] \to \beta_0^* + \beta_1^* x$, so the squared bias approaches $(\beta_0^* + \beta_1^* x - f(x))^2$.
- We still expect $\mathrm{Var}[\hat{f}(x)] \to 0$.

Thus for large n, the prediction error is approximately

$$\mathbb{E}[(Y_{\text{test}} - \hat{f}(x))^2 \mid X_{\text{test}} = x] \approx (\beta_0^* + \beta_1^* x - f(x))^2 + \mathrm{Var}[\varepsilon]$$

the sum of the irreducible error and the squared approximation error of $f(x)$ by the best linear predictor.