1 Distribution Theory for $\hat{Y}$

1.1 Mean and Variance of $\hat{Y}$

- Let Y_h be the (random) value of the response variable for a given (non-random) value X_h of the predictor. Recall that

$$\mathrm{E}(Y_h) \;=\; \beta_0 + \beta_1 X_h$$

and that a **point estimate** of $\mathrm{E}(Y_h)$ is

$$\hat{Y}_h \;=\; b_0 + b_1 X_h$$

- Note that

$$\begin{aligned}\mathrm{E}(\hat{Y}_h) &= \mathrm{E}(b_0 + b_1 X_h) \\ &= \mathrm{E}(b_0) + \mathrm{E}(b_1) X_h \\ &= \beta_0 + \beta_1 X_h \qquad (1)\end{aligned}$$

so $\hat{Y}_h$ is an **unbiased** estimator of $\mathrm{E}(Y_h)$.

- Letting $\boldsymbol{\sigma^2\{\hat{Y}_h\}}$ denote $\mathbf{Var(\hat{Y}_h)}$, we have

$$\begin{aligned}\sigma^2\{\hat{Y}_h\} &= \mathrm{Var}(b_0 + b_1 X_h) \qquad (2)\\ &= \mathrm{Var}(\bar{Y} - b_1\bar{X} \;+\; b_1 X_h) \qquad (3)\\ &= \mathrm{Var}\left(\bar{Y} + (X_h - \bar{X}) b_1\right) \qquad (4)\\ &= \mathrm{Var}(\bar{Y}) + (X_h - \bar{X})^2 \sigma^2\{b_1\} \qquad (5)\\ &= \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)\end{aligned}$$

where line (3) follows from line (2) because (recall) $b_0 = \bar{Y} - b_1\bar{X}$, and line (5) follows from line (4) because (it can be shown) $\bar{Y}$ and b_1 are independent.

- To summarize:

> **Mean and Variance of $\hat{Y}$**: Let Y_h be the (random) value of the response variable for a given (non-random) value X_h of the predictor, and let
>
> $$\hat{Y}_h \;=\; b_0 + b_1 X_h$$
>
> be the **estimate** of $\mathrm{E}(Y_h)$. Then under the simple linear regression model,

with the ϵ_i's independent $N(0, \sigma^2)$, the mean and variance of $\hat{Y}_h$ are

$$\begin{aligned} \mathrm{E}(\hat{Y}_h) &= \beta_0 + \beta_1 X_h \\ \sigma^2\{\hat{Y}_h\} &= \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) \end{aligned} \tag{6}$$

1.2 Normality of $\hat{Y}$

- Because $\hat{Y}_h$ is a **linear combination** of $Y_1, Y_2, \ldots, Y_n$ (since both b_0 and b_1 are), and the Y_i's are normally distributed, so too is $\hat{Y}_h$.

Distribution of $\hat{Y}$: Let Y_h be the (random) value of the response variable for a given (non-random) value X_h of the predictor, and let

$$\hat{Y}_h = b_0 + b_1 X_h$$

be the **estimate** of $\mathrm{E}(Y_h)$. Then under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma^2)$,

$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 X_h,\ \sigma^2\{\hat{Y}_h\}\right), \tag{7}$$

where $\sigma^2\{\hat{Y}_h\}$ is given by (6).

2 Inference for a Mean Response

2.1 Some Background Theory

- Replacing σ^2 in (6) by its estimate MSE and taking the square root gives the (estimated) ***standard error*** of $\hat{Y}_h$, denoted $\boldsymbol{s\{\hat{Y}_h\}}$:

(Estimated) Standard Error of $\hat{Y}$:

$$s\{\hat{Y}_h\} = \sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right)}.$$

- From (7),

$$\frac{\hat{Y}_h - (\beta_0 + \beta_1 X_h)}{\sigma\{\hat{Y}_h\}} \;\sim\; N(0,1)\,. \tag{8}$$

- When we replace $\sigma\{\hat{Y}_h\}$ in (8) by its estimate $s\{\hat{Y}_h\}$, the resulting random variable follows a ***t*** **distibution with *n* − 2 degrees of freedom**, i.e.

> **Fact 2.1** Under the simple linear regression model, with the ϵ_i's independent $N(0,\sigma^2)$,
>
> $$\frac{\hat{Y}_h - (\beta_0 + \beta_1 X_h)}{s\{\hat{Y}_h\}} \;\sim\; t(n-2)\,. \tag{9}$$

2.2 Confidence Interval for a Mean Response

- A $100(1-\alpha)\%$ ***confidence interval for E(Y_h)*** is:

> **Confidence interval for E(Y_h)**: Let Y_h be the (random) value of the response variable for a given (non-random) value X_h of the predictor, and let
>
> $$\hat{Y}_h \;=\; b_0 + b_1 X_h$$
>
> be the **estimate** of $\mathrm{E}(Y_h) \;=\; \beta_0 + \beta_1 X_h$. Then under the simple linear regression model, with the ϵ_i's independent $N(0,\sigma^2)$, $100(1-\alpha)\%$ **confidence interval for E(Y_h)** is
>
> $$\hat{Y}_h \;\pm\; t(\alpha/2, n-2)s\{\hat{Y}_h\} \tag{10}$$
>
> where $t(\alpha/2, n-2)$ is the $100(1-\alpha/2)$th percentile of the $t(n-2)$ distribution.

We can be $100(1-\alpha)\%$ confident that, for a fixed value X_h of the predictor, the interval (10) will contain the true mean response $\mathrm{E}(Y_h) = \beta_0 + \beta_1 X_h$.

2.3 Hypothesis Test for a Mean Response

- We could also use (9) test hypotheses about the value of $\mathrm{E}(Y_h) = \beta_0 + \beta_1 X_h$. See the textbook.

3 Prediction Intervals

- Suppose now we want to **predict** a **new** individual response value for a given value X_h of the predictor. Let $\boldsymbol{Y_{h(\text{new})}}$ denote the new response value we're predicting. $Y_{h(\text{new})}$ will be independent of the observed Y_i's in the data set. The **predicted value** of $Y_{h(\text{new})}$ is

$$\hat{Y}_h \;=\; b_0 + b_1 X_h$$

  (which is also the **estimate** of the true **mean response** $\text{E}(Y_h)$).

- We want an interval that will contain $Y_{h(\text{new})}$ with some specified level of confidence.

- Define the ***prediction error*** to be

$$\text{Prediction error} \;=\; Y_{h(\text{new})} - \hat{Y}_h$$

  Because $\text{E}(Y_{h(\text{new})}) \;=\; \text{E}(\hat{Y}_h) \;=\; \beta_0 + \beta_1 X_h$, we have

$$\text{E}(Y_{h(\text{new})} - \hat{Y}_h) \;=\; 0,$$

  i.e. *on average* the prediction error is zero.

- There are **two** sources of random variation in the prediction error:

  1. Random variation in the predicted value $\hat{Y}_h$.
  2. Random variation in the new response value $Y_{h(\text{new})}$.

  Both sources of random variation must be accounted for in the prediction interval.

- Let $\boldsymbol{\sigma^2\{\text{pred}\}}$ denote $\mathbf{Var}(\boldsymbol{Y_{h(\text{new})} - \hat{Y}_h})$. We have

$$\begin{aligned}
\sigma^2\{\text{pred}\} &= \text{Var}\left(Y_{h(\text{new})} - \hat{Y}_h\right) && (11)\\
&= \text{Var}\left(Y_{h(\text{new})}\right) + \text{Var}\left(\hat{Y}_h\right) && (12)\\
&= \sigma^2 + \sigma^2\{\hat{Y}_h\} && (13)\\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right), && (14)
\end{aligned}$$

  where line (12) follows from (11) since $Y_{h(\text{new})}$ and $\hat{Y}_h$ are independent, and we've used expression (6) for $\sigma^2\{\hat{Y}_h\}$ to obtain line (14) from (13). The two sources of random variation are represented by the two terms in (13).

- Replacing σ^2 in (14) by its estimate MSE and taking the square root gives the ***standard error*** of the prediction error, denoted $\boldsymbol{s\{\text{pred}\}}$:

(Estimated) Standard Error of a Prediction Error:

$$s\{\text{pred}\} \;=\; \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}. \qquad (15)$$

- It can be shown that

Fact 3.1 Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma^2)$,

$$\frac{Y_{h(\text{new})} - \hat{Y}_h}{s\{\text{pred}\}} \;\sim\; t(n-2).$$

- Using the previous fact, a $100(1-\alpha)\%$ ***prediction interval for*** $\boldsymbol{Y_{h(new)}}$ is:

Prediction Interval for $\boldsymbol{Y_{h(\text{new})}}$: Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma^2)$, a $100(1-\alpha)\%$ **prediction interval for $\boldsymbol{Y_{h(\text{new})}}$** is:

$$\hat{Y}_h \;\pm\; t(\alpha/2, n-2)s\{\text{pred}\} \qquad (16)$$

where $s\{\text{pred}\}$ is given by (15) and $t(\alpha/2, n-2)$ is the $100(1-\alpha/2)$th percentile of the $t(n-2)$ distribution.

We can be $100(1-\alpha)\%$ confident that a new observation $Y_{h(\text{new})}$ of the response, at a given value X_h of the predictor, will fall into the interval (16).