STAT 234 Lecture 24-25
Simple Linear Regression Model (Section 12.1-12.4)

Yibi Huang
Department of Statistics
University of Chicago

Outline

- Simple Linear Regression Models (Section 12.1)
- Least Square Estimate (Section 12.2)
- Hypothesis Tests & Confidence Intervals for β_1 and β_0 (Section 12.3)
- Two Kinds of Conditional Predictions Problems (Section 12.4)

Simple Linear Regression Models (Section 12.1)

Simple Linear Regression Model (Review)

1. The condition mean of Y given $X = x$is a linear function of x, i.e.,

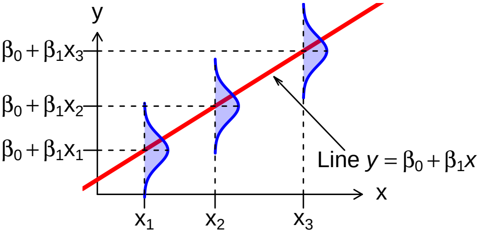
$$\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$$

2. The conditional variance of Y does not change with x, i.e.,

$$\mathrm{Var}(Y \mid X = x) = \sigma^2 \quad \text{for every } x$$

3. (Optional) The conditional distribution of Y given $X = x$ is normal,

$$(Y|X = x) \sim N(\beta_0 + \beta_1 x,\ \sigma^2).$$



Simple Linear Regression Model

Equivalently, the SLR model asserts the values of X and Y for individuals in a population are related as follows

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

- the value of ε, called the **error** or the **noise**, varies from observation to observation, follows a normal distribution

$$\varepsilon \sim N(0, \sigma^2)$$

- In the model, the line $y = \beta_0 + \beta_1 x$ is called the **population regression line**.

Data for a Simple Linear Regression Model

Suppose the data comprised of n individuals/cases randomly sampled from a population.

From case i we observe the response y_i and the predictor x_i:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

The SLR model states that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

How do we estimate intercept β_0 and the slope β_1?

Least-Square Estimates of the Intercept and the Slope (Section 12.2)

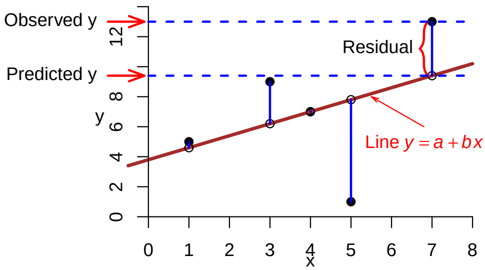
Residuals (Prediction Errors)

If one use the line $y = a + bx$ to predict y from x, the predicted y when $x = x_i$ is

$$\hat{y}_i = a + bx_i.$$

The *residual* (e_i) of the ith observation (x_i, y_i) is

$$\begin{array}{cccc} e_i & = & y_i & - & \hat{y}_i \\ \text{(Residual)} & & \text{(Observed } y\text{)} & & \text{(Predicted } y\text{)} \\ & = & y_i & - & (a + bx_i) \end{array}$$



Residuals are the (signed) **vertical** distances from data points to model line, not the **shortest** distances

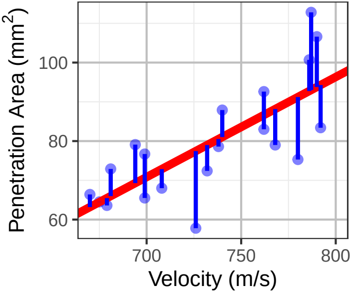
Least-Square Estimates of the Intercept and the Slope

We want a line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ having small residuals:

- Using the line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$, the predicted y when $x = x_i$ is
$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$
- The residual for (x_i, y_i) is $e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$.

For SLR, the least squares estimate $(\widehat{\beta}_0, \widehat{\beta}_1)$ for (β_0, β_1) is the intercept and slope of the straight line with the minimum sum of squared residuals.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$



Solving the Least Squares Problem (1)

To find the $(\widehat{\beta}_0, \widehat{\beta}_1)$ that minimize

$$L(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

one can set the derivatives of L with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to 0

$$\frac{\partial L}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

$$\frac{\partial L}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

This results in the 2 equations below in 2 unknowns $\widehat{\beta}_0$ and$\widehat{\beta}_1$.

$$n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving the Least Squares Problem (2)

$$n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving the Least Squares Problem (2)

$$n\widehat{\beta}_0 + \widehat{\beta}_1 \overbrace{\sum\nolimits_{i=1}^{n} x_i}^{=n\bar{x}} = \overbrace{\sum\nolimits_{i=1}^{n} y_i}^{=n\bar{y}}$$

$$\widehat{\beta}_0 \underbrace{\sum\nolimits_{i=1}^{n} x_i}_{=n\bar{x}} + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving the Least Squares Problem (2)

$$n\widehat{\beta}_0 + \widehat{\beta}_1 \overbrace{\sum_{i=1}^{n} x_i}^{=n\bar{x}} = \overbrace{\sum_{i=1}^{n} y_i}^{=n\bar{y}} \overset{\text{divide by } n}{\Longrightarrow} \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x} = \bar{y} \Rightarrow \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$\widehat{\beta}_0 \underbrace{\sum_{i=1}^{n} x_i}_{=n\bar{x}} + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving the Least Squares Problem (2)

$$n\widehat{\beta}_0 + \widehat{\beta}_1 \overbrace{\sum_{i=1}^n x_i}^{=n\bar{x}} = \overbrace{\sum_{i=1}^n y_i}^{=n\bar{y}} \overset{\text{divide by } n}{\Longrightarrow} \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x} = \bar{y} \Rightarrow \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$\widehat{\beta}_0 \underbrace{\sum_{i=1}^n x_i}_{=n\bar{x}} + \widehat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \Longrightarrow \widehat{\beta}_0 n\bar{x} + \widehat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Solving the Least Squares Problem (2)

$$n\widehat{\beta}_0 + \widehat{\beta}_1 \overbrace{\sum_{i=1}^n x_i}^{=n\bar{x}} = \overbrace{\sum_{i=1}^n y_i}^{=n\bar{y}} \xrightarrow{\text{divide by } n} \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x} = \bar{y} \Rightarrow \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$\widehat{\beta}_0 \underbrace{\sum_{i=1}^n x_i}_{=n\bar{x}} + \widehat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \Longrightarrow \widehat{\beta}_0 n\bar{x} + \widehat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Replacing $\widehat{\beta}_0$ with $\bar{y} - \widehat{\beta}_1 \bar{x}$ in the second equation, we get

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

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

$$\Longleftrightarrow \widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Formulas for the Least Square Estimate for the Slope

Recall the shortcut formulas for sample covariance and variance:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1},$$

$$s_x^2 = s_{xx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}.$$

The LS estimate of the slope is hence

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{s_{xy}}{s_x^2} = \frac{\text{sample covariance of } X \,\&\, Y}{\text{sample variance of } X}.$$

Formulas for the Least Square Estimate for the Slope

Recall the shortcut formulas for sample covariance and variance:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1},$$

$$s_x^2 = s_{xx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}.$$

The LS estimate of the slope is hence

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{s_{xy}}{s_x^2} = \frac{\text{sample covariance of } X \text{ \& } Y}{\text{sample variance of } X}.$$

Another formula:

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \underbrace{\left(\frac{s_{xy}}{s_x s_y}\right)}_{=r}\frac{s_y}{s_x} = r\frac{s_y}{s_x}, \quad \text{where } r = \frac{s_{xy}}{s_x s_y} = \begin{pmatrix}\text{sample}\\ \text{correlation}\end{pmatrix}$$

Properties of the LS Regression Line

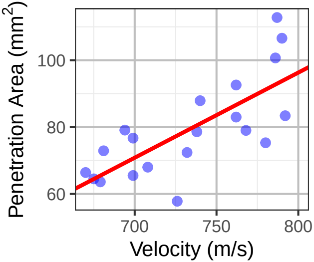
$$\widehat{y} = \overbrace{\widehat{\beta}_0}^{=\bar{y}-\widehat{\beta}_1\bar{x}} + \widehat{\beta}_1 \cdot x$$

$$\Leftrightarrow \widehat{y} - \bar{y} = \widehat{\beta}_1 \cdot (x - \bar{x}) = r\frac{s_y}{s_x}(x - \bar{x})$$

$$\Leftrightarrow \underbrace{\frac{\widehat{y} - \bar{y}}{s_y}}_{z\text{–score of } \widehat{y}} = r \cdot \underbrace{\frac{x - \bar{x}}{s_x}}_{z\text{–score of } x}$$

- The LS regression line *always passes through* $(\bar{x}, \bar{y})$
- As x goes up by 1 SD of x, the predicted value $\widehat{y}$ only goes up by $r \times$ (SD of y)
- When $r = 0$, the LS regression line is horizontal $y = \bar{y}$, and the predicted value $\widehat{y}$ is *always the mean* $\bar{y}$

Example: Armor Strength — Least Square

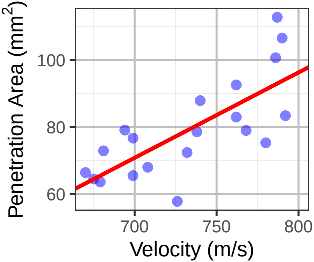


$$n = 20$$
$$\bar{x} = 733.4$$
$$\bar{y} = 79.34$$
$$\sum x_i^2 = 10792614$$
$$\sum y_i^2 = 130028$$
$$\sum x_i y_i = 1172708$$

The LS estimates are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$
$$= \frac{1172708 - 20(733.4)(79.34)}{10792614 - 20(733.4)^2} = \frac{8949}{35103} \approx 0.2549$$
$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 79.34 - 0.2549(733.4) \approx -107.63$$

Example: Armor Strength — Least Square Line (2)



	Velocity (x)	Penetration Area (y)
mean	$\bar{x} = 733.4$,	$\bar{y} = 79.34$
SD	$s_x \approx 42.983$	$s_y = 14.745$
	correlation	$r = 0.7431$

The *slope* and the *intercept* of the least square regression line is

$$\text{slope} = \widehat{\beta}_1 = r\frac{s_y}{s_x} = 0.7431 \times \frac{14.745}{42.983} = 0.2549$$

$$\text{intercept} = \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \cdot \bar{x} = 79.34 - 0.2549(733.4) \approx -107.6$$

The equation of the least square regression line is thus

$$\widehat{y} = -107.6 - 0.2549x.$$

Least Square Regression in R

Regression in R is as simple as `lm(y ~ x)`, in which `lm` stands for "**l**inear **m**odels".

```
armor = read.table(
  "http://www.stat.uchicago.edu/~yibi/s234/ArmorStrength.txt",
  header=TRUE)
lm(penetration.area ~ velocity, data=armor)

Call:
lm(formula = penetration.area ~ velocity, data = armor)

Coefficients:
(Intercept)     velocity
   -107.632        0.255
```

The R output says the least square regression line is

$$\widehat{\text{penetration}}\text{ area }(mm^2) = -107.632 + 0.255 \text{ (firing velocity in m/s).}$$

Our calculation is slightly off due to rounding errors.

Interpretation of Slope

The *slope* indicates how much the response changes *associated* with a unit change in x *on average* (may NOT be *causal*, unless the data are from an experiment).

$$\widehat{\text{penetration area}}\ (mm^2) = -107.632 + 0.255\ (\text{firing velocity in m/s}).$$

- When the firing velocity increases by 1 m/s the penetration area is estimated to increase by 0.255 mm^2 <u>on average</u>.
- When the firing velocity increases by 10 m/s the penetration area is estimated to increase by 2.55 mm^2 <u>on average</u>.

Interpretation of the Intercept

The *intercept* is the predicted value of response when $x = 0$, which might have no practical meaning if $x = 0$ is not a possible value.

$$\widehat{\text{penetration}}\text{ area } (mm^2) = -107.632 + 0.255 \text{ (firing velocity in m/s)}.$$

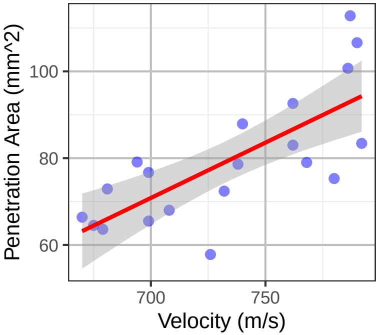
e.g., when the firing velocity is 0 m/s, the predicted penetration area is -107.632 mm^2?

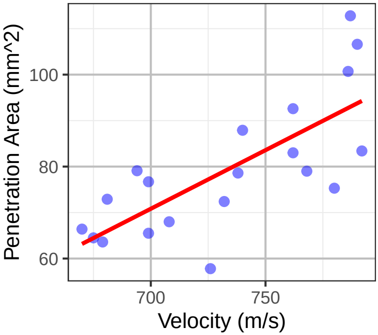
- extrapolation, not reliable

R: Adding the LS Regression Line on the Scatterplot

```
p = ggplot(armor, aes(x=velocity, y=penetration.area)) +
  geom_point(col=rgb(0,0,1,0.5), size=2) +
  xlab("Velocity (m/s)") +
  ylab("Penetration Area (mm^2)")

p + geom_smooth(method='lm', col="red")
p + geom_smooth(method='lm', col="red", se=FALSE)
```



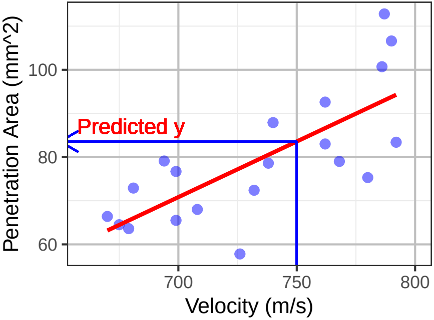


Prediction

One can plug in an x-value to the equation of the least-square regression line to *predict* the response y.

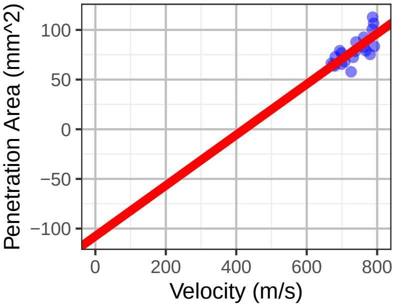
e.g., when the firing velocity is 750 m/s, the predicted penetration area in mm^2 is

$$\widehat{y} = -107.632 + 0.255 \times 750 = 83.57 \text{ mm}^2$$



Extrapolation

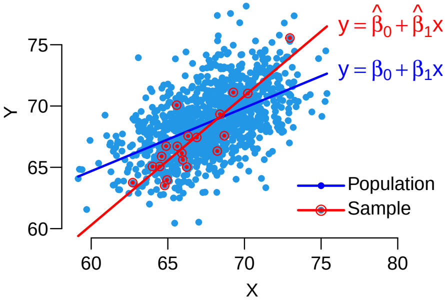
Applying a model estimate to values outside of the realm of the original data is called *extrapolation*.



The variable X = velocity range from 660 to 800 m/s. Prediction made using X-values outside of this range is extrapolation.

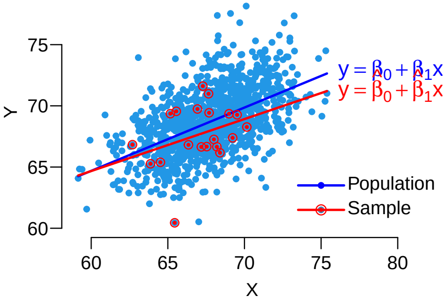
Hypothesis Tests & Confidence Intervals for β_1 and β_0 (Section 12.3)

Sample Regression Line v.s. Population Regression Line



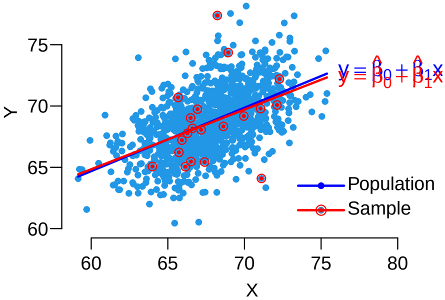
$y = \beta_0 + \beta_1 x$	$y = \widehat{\beta_0} + \widehat{\beta_1} x$
least-square regression line of the population	least-square regression line of the sample
fixed	random, changes from sample to sample
unknown	can be calculated from sample
of interest	not of interest

Sample Regression Line v.s. Population Regression Line



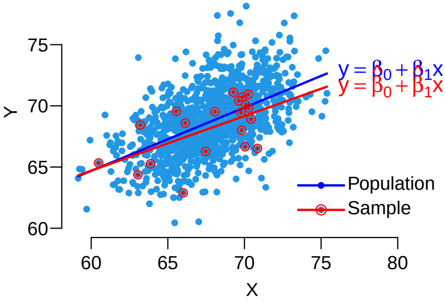
$y = \beta_0 + \beta_1 x$	$y = \widehat{\beta_0} + \widehat{\beta_1} x$
least-square regression line of the population	least-square regression line of the sample
fixed	random, changes from sample to sample
unknown	can be calculated from sample
of interest	not of interest

Sample Regression Line v.s. Population Regression Line



$y = \beta_0 + \beta_1 x$	$y = \widehat{\beta_0} + \widehat{\beta_1} x$
least-square regression line of the population	least-square regression line of the sample
fixed	random, changes from sample to sample
unknown	can be calculated from sample
of interest	not of interest

Sample Regression Line v.s. Population Regression Line



$y = \beta_0 + \beta_1 x$	$y = \widehat{\beta_0} + \widehat{\beta_1} x$
least-square regression line of the population	least-square regression line of the sample
fixed	random, changes from sample to sample
unknown	can be calculated from sample
of interest	not of interest

How Close Is $\widehat{\beta}_1$ to β_1?

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, one can show that

- $\mathrm{E}(\widehat{\beta}_1) = \beta_1$ . . . . $\widehat{\beta}_1$ is an **unbiased** estimate of β_1
- $\mathrm{Var}(\widehat{\beta}_1) = \dfrac{\sigma^2}{\sum(x_i - \bar{x})^2} = \dfrac{\sigma^2}{(n-1)s_x^2}$.

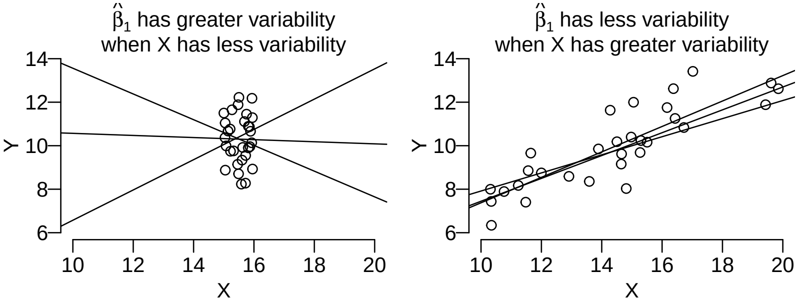
How Close Is $\widehat{\beta}_1$ to β_1?

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, one can show that

- $\mathrm{E}(\widehat{\beta}_1) = \beta_1$ .... $\widehat{\beta}_1$ is an **unbiased** estimate of β_1
- $\mathrm{Var}(\widehat{\beta}_1) = \dfrac{\sigma^2}{\sum(x_i - \bar{x})^2} = \dfrac{\sigma^2}{(n-1)s_x^2}$.

$\widehat{\beta}_1$ will be closer to β_1 if

1) the sample size n is larger, or 2) X has greater variability



Proof of the Unbiasedness of $\widehat{\beta}_1$ (1)

Recall the formula for the LS estimate for slope $\widehat{\beta}_1$ is

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

To find the expected value of $\widehat{\beta}_1$, we will first show an alternative formula for it: $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i - \bar{x})y_i - \sum_{i=1}^n (x_i - \bar{x})\bar{y} \\ &= \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y} \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} \\ &= \sum_{i=1}^n (x_i - \bar{x})y_i\end{aligned}$$

This proves the alternative formula: $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Proof of the Expected Value of $\widehat{\beta}_1$ (2)

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, we know $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i$. Hence,

$$\mathrm{E}[\widehat{\beta}_1] = \mathrm{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^n (x_i - \bar{x})\,\mathrm{E}[y_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The numerator $\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)$ equals

$$\begin{aligned}
\beta_0 \overbrace{\sum_{i=1}^n (x_i - \bar{x})}^{=0} &+ \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i \\
&= \beta_1 \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) + \beta_1 \bar{x} \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} \\
&= \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2
\end{aligned}$$

Putting the numerator back to $\mathrm{E}[\widehat{\beta}_1]$, we get $\mathrm{E}[\widehat{\beta}_1] = \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1$. So $\widehat{\beta}_1$ is an unbiased estimator for β_1.

Proof of Variance of $\widehat{\beta}_1$

For the SLR model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ as x_i's are regarded as fixed numbers and ε_i's are indep. with $\mathrm{Var}(\varepsilon_i) = \sigma^2$, we know that y_i's are also indep. with $\mathrm{Var}(y_i) = \sigma^2$.

Recall that for independent random variables $Y_1, Y_2, \ldots, Y_n$, and fixed numbers $c_1, c_2, \ldots, c_n$ the variance has the property:

$$\mathrm{Var}(c_1 Y_1 + c_2 Y_2 + \ldots + c_n Y_n) = c_1^2 \,\mathrm{Var}(Y_1) + c_2^2 \,\mathrm{Var}(Y_2) + \ldots + c_n^2 \,\mathrm{Var}(Y_n)$$

Apply the property above to the variance of $\widehat{\beta}_1$ with $c_j = \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$, we get

$$\mathrm{Var}(\widehat{\beta}_1) = \mathrm{Var}\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \mathrm{Var}(\sum_{i=1}^n c_i y_i) = \sum_{i=1}^n c_i^2 \,\mathrm{Var}(y_i) = \sigma^2 \sum_{i=1}^n c_i^2$$

$$= \sigma^2 \sum_{i=1}^n \left(\frac{(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^2 = \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Sampling Distribution of the Intercept $\widehat{\beta}_0$

Under the SLR model, the estimate of the intercept

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

is also **unbiased** and (approx.) **normal** with the variance

$$\mathrm{Var}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

What is the intuition here?

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}_0) = \mathrm{Var}(\bar{y} - \widehat{\beta}_1 \bar{x}) &= \mathrm{Var}(\bar{y}) - 2\bar{x}\,\overbrace{\mathrm{Cov}(\bar{y}, \widehat{\beta}_1)}^{=0} + \bar{x}^2\,\mathrm{Var}(\widehat{\beta}_1) \\ &= \frac{\sigma^2}{n} + 0 + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

- $\bar{y}$ and $\widehat{\beta}_1$ are uncorrelated because the slope ($\widehat{\beta}_1$) is invariant if you shift the response up or down ($\bar{y}$).

Covariance of $\widehat{\beta}_1$ and $\widehat{\beta}_0$

The estimates for the slope and the intercept are **negatively correlated** and their covariance is

$$\begin{aligned}
\mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_0) &= \mathrm{Cov}(\widehat{\beta}_1, \bar{y} - \widehat{\beta}_1 \bar{x}) \\
&= \underbrace{\mathrm{Cov}(\widehat{\beta}_1, \bar{y})}_{=0} - \bar{x} \underbrace{\mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_1)}_{=\mathrm{Var}(\widehat{\beta}_1)} \\
&= 0 - \bar{x}\,\mathrm{Var}(\widehat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{aligned}$$

- Usually, if the slope estimate is too high, the intercept estimate is too low

Estimate of σ^2 — Variance of the Errors ε_i.

- A naive estimate of σ^2 is the sample variance of the ε_i

$$\widehat{\sigma}^2 = \frac{\sum(\varepsilon_i - \overline{\varepsilon})^2}{n-1} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

  However, this is not possible as β_0 and β_1 are unknown.

Estimate of σ^2 — Variance of the Errors ε_i.

- A naive estimate of σ^2 is the sample variance of the ε_i

$$\widehat{\sigma}^2 = \frac{\sum(\varepsilon_i - \overline{\varepsilon})^2}{n-1} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

  However, this is not possible as β_0 and β_1 are unknown.

- We can estimate β_0 and β_1 with $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and approximate the errors ε_i with the **residuals**

$$e_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = y_i - \widehat{y}_i$$

  We use the "sample variance" of the residuals e_i to estimate σ^2:

$$\widehat{\sigma}^2 = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2}$$

Estimate of σ^2 — Variance of the Errors ε_i.

- A naive estimate of σ^2 is the sample variance of the ε_i

$$\widehat{\sigma}^2 = \frac{\sum(\varepsilon_i - \overline{\varepsilon})^2}{n-1} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

  However, this is not possible as β_0 and β_1 are unknown.

- We can estimate β_0 and β_1 with $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and approximate the errors ε_i with the **residuals**

$$e_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = y_i - \widehat{y}_i$$

  We use the “sample variance” of the residuals e_i to estimate σ^2:

$$\widehat{\sigma}^2 = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2}$$

- We will show in the next lecture that $\bar{e} = 0$

Estimate of σ^2 — Variance of the Errors ε_i.

- A naive estimate of σ^2 is the sample variance of the ε_i

$$\widehat{\sigma}^2 = \frac{\sum(\varepsilon_i - \overline{\varepsilon})^2}{n-1} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

However, this is not possible as β_0 and β_1 are unknown.

- We can estimate β_0 and β_1 with $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and approximate the errors ε_i with the **residuals**

$$e_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = y_i - \widehat{y}_i$$

We use the "sample variance" of the residuals e_i to estimate σ^2:

$$\widehat{\sigma}^2 = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2}$$

- We will show in the next lecture that $\bar{e} = 0$
- We divide by $n-2$, not $n-1$ or n as We are not able to estimate error unless we have at least 3 observations

Standard Error (SE) of the Slope and the Intercept

The **standard deviation (SD)** of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are the square-root of their variances

$$\mathrm{SD}(\widehat{\beta}_1) = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}, \quad \mathrm{SD}(\widehat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

If the unknown σ is replaced with the estimate

$$\widehat{\sigma} = \sqrt{\frac{\sum(y_i - \widehat{y}_i)^2}{n-2}},$$

The estimated SD's are called the **standard error (SE)**'s:

$$\text{SE}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}, \quad \text{SE}(\widehat{\beta}_0) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}.$$

Sampling Distributions of $\widehat{\beta}_1$ and $\widehat{\beta}_0$

The **sampling distributions** of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are both normal.

$$\widehat{\beta}_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right) \quad \Rightarrow \quad z = \frac{\widehat{\beta}_1 - \beta_1}{\sigma / \sqrt{\sum(x_i - \bar{x})^2}} \sim N(0,1)$$

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)\right) \quad \Rightarrow \quad z = \frac{\widehat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}} \sim N(0,1)$$

This is (approx.) valid

- either if the errors ε_i are i.i.d. $N(0, \sigma^2)$
- or if the errors ε_i are independent and the sample size n is large

If the unknown σ is replaced by $\widehat{\sigma}$, z become the t-statistic with a t-distribution with df $= n - 2$.

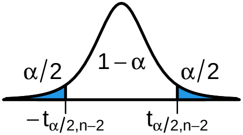
$$T_1 = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma} / \sqrt{\sum(x_i - \bar{x})^2}} = \frac{\widehat{\beta}_1 - \beta_1}{\text{SE}(\widehat{\beta}_1)} \sim t_{n-2},\ T_0 = \frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}} = \frac{\widehat{\beta}_0 - \beta_0}{\text{SE}(\widehat{\beta}_0)} \sim t_{n-2}.$$

Confidence Intervals for β_0 and β_1

The $(1-\alpha)$ **confidence intervals for β_0 and β_1** are respectively

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\text{SE}(\widehat{\beta}_0) \quad \text{and} \quad \widehat{\beta}_1 \pm t_{\alpha/2,n-2}\text{SE}(\widehat{\beta}_1),$$

where $t_{\alpha/2,n-2}$ is the value such that $P(|T| < t_{\alpha/2,n-2}) = 1-\alpha$ for $T \sim t_{n-2}$.



In R, $t_{\alpha/2,n-2}$ = `qt(alpha/2, df=n-2, lower.tail=F)`.

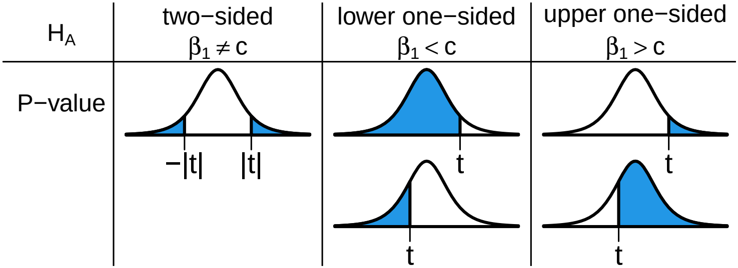
		90% CI $t_{0.1/2,df}$ ↓	95% CI $t_{0.05/2,df}$ ↓		99% CI $t_{0.01/2,df}$ ↓		
α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
ν 1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tests for β_0 and β_1

To **test the hypotheses** H_0: $\beta_0 = c$ or H_0: $\beta_1 = c$ the t-statistic are respectively

$$t = \frac{\widehat{\beta}_0 - c_0}{\text{SE}(\widehat{\beta}_0)} \sim t_{n-2}, \quad \text{and} \quad t = \frac{\widehat{\beta}_1 - c_1}{\text{SE}(\widehat{\beta}_1)} \sim t_{n-2}$$

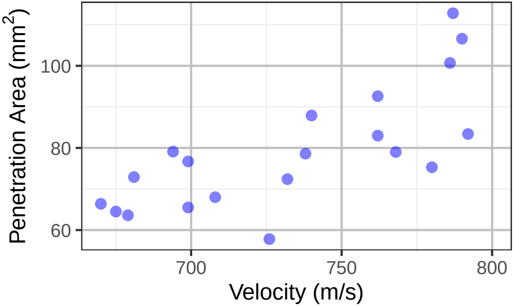
The P-value can be computed based on H_a:



Example: Armor Strength

Soldiers depend on their body armor for protection. Specimens of UHMWPE body armor were shot with a 7.62 mm round at various firing velocities. The penetration areas were recorded[a].

[a] "Testing of Body Armor Materials-Phase III", 2012, by the US Army and the National Research Council



Velocity (m/s)	Penetration Area (mm^2)
670	66.4
675	64.5
679	63.6
681	72.9
694	79.1
699	76.7
699	65.5
708	68.0
726	57.8
732	72.4
738	78.6
740	87.9
762	92.6
762	83.0
768	79.0
780	75.3
792	83.4
786	100.7
790	106.6
787	112.8

Least Square Regression in R

Regression in R is as simple as `lm(y ~ x)`, in which `lm` stands for "**l**inear **m**odels".

```
armor = read.table(
  "http://www.stat.uchicago.edu/~yibi/s234/ArmorStrength.txt",
  header=TRUE)
lm(penetration.area ~ velocity, data=armor)

Call:
lm(formula = penetration.area ~ velocity, data = armor)

Coefficients:
(Intercept)     velocity
   -107.632        0.255
```

```
> lmarmor = lm(penetration.area ~ velocity, data=armor)
> summary(lmarmor)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -107.6324    39.7450   -2.71  0.01440
velocity       0.2549     0.0541    4.71  0.00017

Residual standard error: 10.1 on 18 degrees of freedom
Multiple R-squared:  0.552, Adjusted R-squared:  0.527
F-statistic: 22.2 on 1 and 18 DF,  p-value: 0.000174
```

"`Residual standard error:  10.1`" in the R output is the estimate for $\sigma = \text{SD}(\varepsilon_i)$, $\widehat{\sigma} = 10.1$.

That is, the variance $\sigma^2 = \text{Var}(\varepsilon_i)$ for the noise term ε_i in the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is estimated to be $\widehat{\sigma}^2 = 10.1^2$.

The Summary Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -107.6324    39.7450   -2.71  0.01440
velocity       0.2549     0.0541    4.71  0.00017
```

- The column "**estimate**" shows the LS estimates for the intercept $\widehat{\beta}_0 = -107.6324$ and the slope $\widehat{\beta}_1 = 0.2549$
- The column "**std. error**" gives:

$$\text{SE}(\widehat{\beta}_0) = 39.7450, \quad \text{SE}(\widehat{\beta}_1) = 0.0541$$

R Tests Whether β_1 and β_0 Equal 0 Automatically

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -107.6324    39.7450   -2.71  0.01440
velocity       0.2549     0.0541    4.71  0.00017
```

- The column **t value** shows the t-statistic for testing H$_0$: $\beta_0 = 0$ and H$_0$: $\beta_1 = 0$,
  $$t_0 = \frac{\widehat{\beta_0} - 0}{\text{SE}(\widehat{\beta_0})} = \frac{-107.6350}{37.7450} = -2.71, \ t_1 = \frac{\widehat{\beta_1} - 0}{\text{SE}(\widehat{\beta_1})} = \frac{0.2549}{0.0541} = 4.71$$
  which are simply **the ratios of the first two columns**
- The column "**p.value**" shows the **2-sided** P-values for testing H$_0$: $\beta_0 = 0$ and H$_0$: $\beta_1 = 0$.
- Testing H$_0$: $\beta_1 = 0$ is equivalent to testing whether the penetration area is linearly related to the velocity. The small P-value 0.00017 asserts the relation is significant

Example: Testing Whether β_1 is a Non-Zero Value

To test whether β_1 equal to some non-zero value c_1, one has to calculate the t-statistic and P-value himself.

Ex. To see if the penetration area increases by 0.1 mm^2 on average when the firing velocity increases by 1 m/s i.e., $\beta_1 = 0.1$.

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-107.6324	39.7450	-2.71	0.01440
velocity	0.2549	0.0541	4.71	<0.00017

To test $H_0 : \beta_1 = 0.1$ v.s. $H_A : \beta_1 > 0.1$, the t-statistic is

$$t_1 = \frac{\widehat{\beta_1} - 0.1}{\text{SE}(\widehat{\beta_1})} = \frac{0.2549 - 0.1}{0.0541} \approx 2.863 \text{ with df } = 20 - 2 = 18.$$

The upper one-sided p-value is

`pt(2.863, df = 20-2, lower.tail=F)` ≈ 0.0052.

Conclusion: When the firing velocity increases by 1 m/s, the penetration area increases significantly more than 0.1 mm^2 on average.

Example: Confidence Interval for β_1

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-107.6324	39.7450	-2.71	0.01440
velocity	0.2549	0.0541	4.71	<0.00017

The 95% CI for the slope β_1 is

$$\widehat{\beta}_1 \pm t_{0.05/2,20-2}\text{SE}(\widehat{\beta}_1) = 0.2549 \pm 2.101 \times 0.0541$$
$$= 0.2549 \pm 0.1137 \approx (0.1412, 0.3686).$$

where $t_{0.05/2,20-2} \approx 2.101$ for a 95% CI can be found in R or using t-table for df $= n - 2 = 20 - 2 = 18$.

```
qt(0.05/2, df=20-2, lower.tail=F)
[1] 2.101
qt(1-0.05/2, df=20-2)
[1] 2.101
```

α		0.1	0.05	0.025	0.01	0.005	0.001	0.0005
ν	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922

Interpretation: With 95% confidence, penetration area increases by 0.1412 to 0.3686 mm^2 on average when firing velocity increases by 1 m/s.

Finding CIs for Coefficients Using `confint()`

The `confint()` command can produce confidence intervals for the coefficients β_0 and β_1 for us

```
confint(lmarmor)
                2.5 %   97.5 %
(Intercept) -191.1335 -24.1314
velocity       0.1413   0.3686
confint(lmarmor, level = 0.95)
                2.5 %   97.5 %
(Intercept) -191.1335 -24.1314
velocity       0.1413   0.3686
confint(lmarmor, level = 0.95, "velocity")
          2.5 % 97.5 %
velocity 0.1413 0.3686
confint(lmarmor, level = 0.95, "(Intercept)")
             2.5 % 97.5 %
(Intercept) -191.1 -24.13
```

Two Kinds of Conditional Predictions Problems (Section 12.4)

Two Kinds of Conditional Predictions Problems

There are **two** kinds of conditional prediction problems of the response Y given $X = x_0$ based on a SLR model $Y = \beta_0 + \beta_1 X + \varepsilon$:

- when X is known to be x_0, estimate the **mean response** $\mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0$
- when X is known to be x_0, predict the response for **one specific observation** $Y = \beta_0 + \beta_1 x_0 + \varepsilon$

Two Kinds of Conditional Predictions Problems

There are **two** kinds of conditional prediction problems of the response Y given $X = x_0$ based on a SLR model $Y = \beta_0 + \beta_1 X + \varepsilon$:

- when X is known to be x_0, estimate the **mean response** $\mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0$
- when X is known to be x_0, predict the response for **one specific observation** $Y = \beta_0 + \beta_1 x_0 + \varepsilon$

For the armor strength example, one may want to

- estimate the *average* penetration area on the armor when shot at a firing speed of $x_0 = 700$ m/s, which is $\beta_0 + 700\beta_1$
- predict penetration area on the armor for *one shot* at a firing speed of of $x_0 = 700$ m/s, which is $\beta_0 + 700\beta_1 + \varepsilon$.

Estimation v.s. Prediction

The first one is an **estimation** problem as $\beta_0 + \beta_1 x_0$ only involve fixed parameters β_0, β_1, and x_0.

The second one is a **prediction** problem as $\beta_0 + \beta_1 x_0 + \varepsilon$ involve a random number ε

Estimated Value and Predicted Value

In the first one,

$$\mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0 \quad \text{is estimated by} \quad \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

where the unknown β_0 and β_1 are replaced by their LS estimates.

In the second one,

$$Y = \beta_0 + \beta_1 x_0 + \varepsilon \quad \text{is predicted by} \quad \widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 + 0.$$

The noise ε for a future observation is predicted to be its mean 0 since it's independent of all observed (x_i, y_i)'s. We cannot make a better prediction for ε from the observed (x_i, y_i)'s.

The Two Prediction Problems Differ in the Uncertainty!

For estimating $\mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0$, the variance for the estimate $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$ can be shown to be

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) &= \mathrm{E}[(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - \beta_0 - \beta_1 x_0)^2] \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)\end{aligned}$$

To predict $Y = \beta_0 + \beta_1 x_0 + \varepsilon$, we need to include the extra variability from the noise ε.

$$\begin{aligned}\mathrm{E}[(\widehat{Y} - Y)^2] &= \mathrm{E}[(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - \beta_0 - \beta_1 x_0 - \varepsilon)^2] \\ &= \mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) + \mathrm{Var}(\varepsilon) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) + \sigma^2\end{aligned}$$

As n gets large, $\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0)$ would go down to 0 but $\mathrm{E}[(\widehat{Y} - Y)^2]$ only approaches σ^2.

What Affects the Accuracy of Prediction?

Recall the variances for the two prediction problems are

$$\begin{cases} \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) & \text{for estimating } \mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0 \\ \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) & \text{to predict } Y \text{ when } X = x_0 \end{cases}$$

An accurate prediction (less variance) comes from

- small σ^2 (i.e., small noise ε's)
- large sample size n
- large $\sum_{i=1}^n (x_i - \bar{x})^2$ (more spread in predictors)
- small $(x_0 - \bar{x})^2$

Confidence Intervals and Prediction Intervals

The $100(1-\alpha)\%$ confidence interval for $\beta_0 + \beta_1 x_0$ is

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm t_{\alpha/2,n-2}\, \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

The $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_0 + \varepsilon$ is

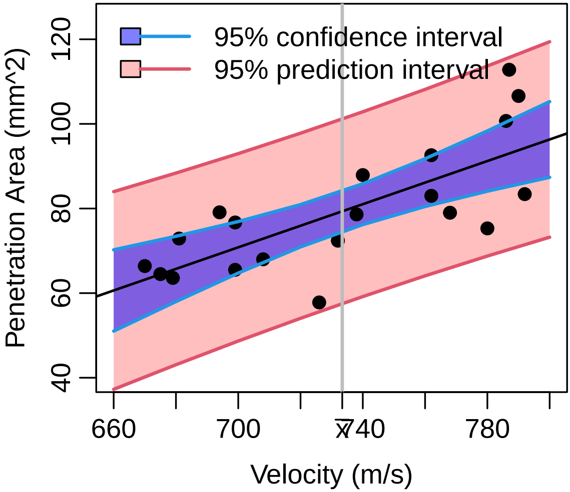
$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm t_{\alpha/2,n-2}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Confidence Intervals and Prediction Intervals in R

```
predict(lmarmor, data.frame(velocity=700), interval="confidence",
        level=0.95)
    fit   lwr   upr
1 70.83 64.73 76.92
predict(lmarmor, data.frame(velocity=700), interval="prediction",
        level=0.95)
    fit   lwr   upr
1 70.83 48.67 92.98
```

- When the firing velocity is 700 m/s, the penetration area is 70.83 mm^2 on average, and the 95% confidence interval is 64.73 to 76.92 mm^2.
- When the firing velocity is 700 m/s, the penetration area for one shot 48.67 to 92.98 mm^2 with 95% confidence.
- The prediction interval for a **single** shot is wider.

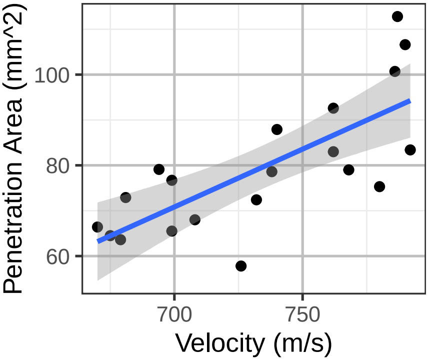
The plot below shows the 95% confidence intervals and the 95% prediction intervals at different values of x_0.



Both the confidence intervals and the prediction intervals are *narrowest when $x_0 = \bar{x}$*.

`geom_smooth(method='lm')` in `ggplot()` by default includes the 95% confidence intervals for estimating $E(y|X = x_0)$.

```
ggplot(armor, aes(x=velocity, y=penetration.area)) +
  geom_point() +
  geom_smooth(method='lm') +
  xlab("Velocity (m/s)") +
  ylab("Penetration Area (mm^2)")
```



Multiple R-Squared = Coefficient of Determination

Properties of Residuals (1)

Recall the LS estimate $(\widehat{\beta}_0, \widehat{\beta}_1)$ that minimizes

$$L(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

is obtained by setting the derivatives of L with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to 0

$$\frac{\partial L}{\partial \widehat{\beta}_0} = \sum_{i=1}^{n} (\underbrace{y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i}_{= y_i - \widehat{y}_i = e_i = \text{residual}}) = 0 \text{ and}$$

$$\frac{\partial L}{\partial \widehat{\beta}_1} = \sum_{i=1}^{n} x_i \overbrace{(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)} = 0,$$

The residuals e_i hence have the properties

$$\underbrace{\sum_{i=1}^{n} e_i = 0}_{\text{Residuals add up to 0.}}, \qquad \underbrace{\sum_{i=1}^{n} x_i e_i = 0}_{\text{Residuals are orthogonal to } x\text{-variable.}}$$

Properties of Residuals (2)

The two properties combined imply that *residuals have 0 correlation with the explanatory variable* X since

$$\mathrm{Cov}(X, e) = \frac{1}{n-1}\Big(\underbrace{\sum_{i=1}^{n} x_i e_i}_{=0} - n\bar{x}\underbrace{\bar{e}}_{=0}\Big) = 0$$

Sum of Squares

Observe that

$$y_i - \bar{y} = \underbrace{(\widehat{y_i} - \bar{y})}_{a} + \underbrace{(y_i - \widehat{y_i})}_{b}$$

Squaring up both sides using the identity $(a+b)^2 = a^2 + b^2 + 2ab$, we get

$$(y_i - \bar{y})^2 = \underbrace{(\widehat{y_i} - \bar{y})^2}_{a^2} + \underbrace{(y_i - \widehat{y_i})^2}_{b^2} + \underbrace{2(\widehat{y_i} - \bar{y})(y_i - \widehat{y_i})}_{2ab}$$

Summing up over all the cases $i = 1, 2, \dots, n$, we get

$$\overbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}^{\text{SST}} = \overbrace{\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2}^{\text{SSR}} + \overbrace{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}^{\text{SSE}} + 2\underbrace{\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})(y_i - \widehat{y_i})}_{=\,0,\ \text{see next page.}}$$

Why $\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})(y_i - \widehat{y}_i) = 0$?

$$\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})(\underbrace{y_i - \widehat{y}_i}_{=e_i}) = \sum_{i=1}^{n}\widehat{y}_i e_i - \sum_{i=1}^{n}\bar{y}e_i$$

$$= \sum_{i=1}^{n}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)e_i - \sum_{i=1}^{n}\bar{y}e_i$$

$$= \widehat{\beta}_0 \underbrace{\sum_{i=1}^{n} e_i}_{=0} + \widehat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i e_i}_{=0} - \bar{y}\underbrace{\sum_{i=1}^{n} e_i}_{=0} = 0$$

in which we used the properties of residuals that $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$.

Interpretation of Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(\overbrace{y_i - \widehat{y_i}}^{=e_i})^2}_{\text{SSE}}$$

- SST = **total sum of squares**
  - total variability of Y
- SSR = **regression sum of squares**
  - variability of Y explained by X
- SSE = **error (residual) sum of squares**
  - variability of Y not explained by the X's

Multiple R-Squared

Multiple R^2, also called the **coefficient of determination**, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variability in } Y \text{ explained by } X$$

which measures the strength of the linear relationship between Y and the X variable

- $0 \leq R^2 \leq 1$

Multiple R-Squared

Multiple R^2, also called the **coefficient of determination**, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variability in } Y \text{ explained by } X$$

which measures the strength of the linear relationship between Y and the X variable

- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y

Interpretation of R-squared

For the Armor Strength data, $R^2 = r^2 = (0.7431)^2 \approx 0.552$, which means — 55.2% of the variability in the penetration area is explained by the firing velocity.

```
> lmarmor = lm(penetration.area ~ velocity, data=armor)
> summary(lmarmor)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -107.6324    39.7450   -2.71  0.01440
velocity       0.2549     0.0541    4.71  0.00017

Residual standard error: 10.1 on 18 degrees of freedom
Multiple R-squared:  0.552, Adjusted R-squared:  0.527
F-statistic: 22.2 on 1 and 18 DF,  p-value: 0.000174
```