RECALL: The normal simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2).$$

$$\Rightarrow Y_i \mid X = X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

$$\Rightarrow f_{Y_i|X_i}(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\beta_0-\beta_1 x)^2}{2\sigma^2}\right\}$$

$$\Rightarrow \ell(\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^{n} \log\left\{\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}\right\}\right\}$$

$$= \sum_{i=1}^{n} -\log\sigma\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2$$

Maximizing this expression is the same as minimizg this expression with sign minus.

$\Rightarrow$ $\boxed{\text{Solving the MLE}}$ $\Longleftrightarrow$ $\boxed{\text{Solving the least squares formulation}}$

The MLE task is to create

$$\left\{ \begin{array}{l} \dfrac{\partial \ell}{\partial \beta_0} = 0 \\[2ex] \dfrac{\partial \ell}{\partial \beta_1} = 0 \\[2ex] \dfrac{\partial \ell}{\partial \sigma^2} = 0 \end{array} \right\}$$

just the normal equations from last lecture

$\Rightarrow$ It will give us again

$$b_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{XX} - \overline{X}\,\overline{X}}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$\ell(\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^{n} -\log \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$= -\sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \log\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

goes away when we take a derivative

$$= -n \log \sigma - 2^{-1}(\sigma^2)^{-1}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$= -n \log\left(\sqrt{\sigma^2}\right) - \frac{1}{2}(\sigma^2)^{-1}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\frac{(\sigma^2)^{-\frac{1}{2}}}{\sqrt{\sigma^2}} + \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 = 0$$

$$\Rightarrow \frac{1}{\cancel{2}\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 = \frac{n}{\cancel{2}}$$

$$\Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}{n}$$

once we plug in $\beta_0 = b_0$ and $\beta_1 = b_1$ from the normal equations.

$$= \frac{\sum_{i=1}^{n} e_i^2}{n} = s_e^2 = \hat{\sigma}^2$$

ISSUE: Last class — and in textbooks — we said that we would use

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Qn: what gives? The issue is that

$$\mathbb{E}\left[\frac{\sum_{i=1}^{n} e_i^2}{n}\right] \neq \sigma^2$$

It needs a bias adjustment: multiply by $\frac{n}{n-2}$

OBJECTIVES:

- Establish the mean and variance of b_1 (or $\hat{\beta}_1$).
- As a byproduct, we will prove the Gauss-Markov Theorem.
- Establish the distribution of a studentization of $\hat{\beta}_1$ (or b_1).

FRAMEWORK: The normal (classical) simple linear regression model is of the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$
$$\rho(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j$$

REMINDER: For the time being, the X_i's are deterministic.

Recall: we showed that

$$b_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{XX} - \overline{X}\,\overline{X}} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

GOAL #1: $\mathbb{E}[b_1] = \beta_1$

regression variance

GOAL #2: $\text{Var}(b_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$

<u>CLAIM</u>: If $b_1 = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$, we can think of b_1 as being linear as a function of the Y_i's, i.e., we can write

$$b_1 = \sum_{i=1}^{n} k_i Y_i ,$$

where $k_i = \dfrac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

Note that

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} - \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \bar{Y}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

b/c $\sum_{i=1}^{n} (X_i - \bar{X}) = 0$, this second term $= 0$

and we have

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad \text{-or-} \quad b_1 = \sum_{i=1}^{n} k_i Y_i ;$$

where $k_i = \dfrac{X_i - \bar{X}}{\sum^{n} (X_i - \bar{X})^2}$

THREE FACTS ABOUT THE k_j'S

FACT #1: $\displaystyle\sum_{i=1}^{n} k_i = 0$

$$\text{well, } \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = 0$$

FACT #2: $\displaystyle\sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

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

$$= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

FACT #3: $\displaystyle\sum_{i=1}^{n} k_i x_i = 1$

left as an exercise: show again these are the same

$$\text{So, } \sum_{i=1}^{n} \frac{(x_i-\bar{x})x_i}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i)}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = \frac{\sum_{i=1}^{n}x_i^2 - \bar{x}\cdot n\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = 1$$

$$\sum_{j=1}^{n}(X_j-\bar{X})(X_j-\bar{X}) = \sum_{j=1}^{n}\left(X_j^2 - 2\bar{X}X_j + \bar{X}^2\right)$$

$$= \sum_{j=1}^{n} X_j^2 - 2\bar{X}n\bar{X} + n\bar{X}^2$$

$$= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

$$\Rightarrow \sum_{i=1}^{n} k_i X_i = 1$$

Getting back to Gauss-Markov:...

$$\mathbb{E}[b_1] = \mathbb{E}\left[\sum_{i=1}^{n} k_i Y_i\right] = \sum_{i=1}^{n} k_i \mathbb{E}[Y_i]$$

$$= \sum_{i=1}^{n} k_i\left(\beta_0 + \beta_1 X_i\right)$$

$$= \beta_0 \cancelto{0}{\sum_{i=1}^{n} k_i} + \beta_1 \underbrace{\sum_{i=1}^{n} k_i X_i}_{=1}$$

FACT #1

by FACT #3

$$= \beta_0 \cdot 0 + \beta_1 \cdot 1 = \beta_1$$

8

Qn: What is $Var(b_1)$?

Since $b_1 = \sum_{i=1}^{n} k_i Y_i$, we want

$$Var\left(\sum_{i=1}^{n} k_i Y_i\right) = \sum_{i=1}^{n} Var(k_i Y_i) = \sum_{i=1}^{n} k_i^2 \, Var(Y_i)$$

okay because the Y_i, Y_j are uncorrelated because $\rho(\varepsilon_i, \varepsilon_j) = 0$

$$= \sum_{i=1}^{n} k_i^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} k_i^2$$

problem: we don't know σ^2 in advance

$$= \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

FACT #2

Let's have a brief moment for interpretation:

1. To lower $Var(b_1)$, lower the regression variance, i.e., make the data adhere to the line better.
2. Also, you can increase n since larger n means $\sum_{i=1}^{n} (X_i - \bar{X})^2$ gets larger.
3. We could also lower b_1 by increasing σ_X^2, the variance of the X_i's.

Last hour, recall that we learned that σ^2 could be estimated by

$$MSE = s^2 = s_e^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

By Slutsky's Theorem, because MSE is unbiased for σ^2, we can replace σ^2 by MSE in $Var(b_1)$ and end up (in particular) with an estimator that is again unbiased.

So, we take

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

but estimate it with

$$s^2(b_1) = \frac{MSE}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

when you use things like lm( ) in R, it uses this.

<u>GOAL:</u> Show that b_1 is BLUE. ☑

best linear unbiased estimator

☑ ☑

"best" in the sense of minimum variance

If there is some other superior linear estimator of β_1, it must be of the form

$$\beta = \sum c_i Y_i$$

For B to be "better" that b_1, it has to be (1) unbiased and (2) have an even smaller standard error.

$$\text{So } \mathbb{E}[B] = \mathbb{E}\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i \, \mathbb{E}[Y_i] = \beta_1$$

put a pin in this; we'll come back to it

$$\text{Recall } \mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i$$

$$\Rightarrow \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i = \beta_1$$

For B to be unbiased, $\sum_{i=1}^{n} c_i = 0$ and

$$\sum_{i=1}^{n} c_i X_i = 1.$$

Qn: What is $Var(B)$?

$$\text{well, } Var(B) = Var\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 \, Var(Y_i)$$

$$= \sigma^2 \sum_{i=1}^{n} c_i^2 .$$

If there are better constants c_i, they are of the for $c_i = k_i + d_i$

$$\Rightarrow Var(B) = \sigma^2 \sum_{i=1}^{n} (k_i + d_i)^2 = \sigma^2 \sum_{i=1}^{n} (k_i^2 + 2k_i d_i + d_i^2)$$

$$= \sigma^2 \sum_{i=1}^{n} k_i^2 + \cancelto{0}{2\sigma^2 \sum_{i=1}^{n} k_i d_i} + \sigma^2 \sum_{i=1}^{n} d_i^2$$

$Var(b_1)$

CLAIM: $\sum_{i=1}^{n} k_i d_i = 0$

Note that $\sum_{i=1}^{n} k_i (c_i - k_i) = \sum_{i=1}^{n} c_i k_i - \sum_{i=1}^{n} k_i^2$

$$= \frac{\sum_{i=1}^{n} (c_i X_i - c_i \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \frac{\underbrace{\sum_{i=1}^{n} c_i X_i}_{=1} - \bar{X} \underbrace{\sum_{i=1}^{n} c_i}_{=0}}{\sum_{j=1}^{n} (X_j - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= 0$$

This means that

$$\mathrm{Var}(B) = \mathrm{Var}(b_1) + \sigma^2 \sum_{i=1}^{n} d_i^2$$

$\Rightarrow$ But $\sigma^2 \sum_{i=1}^{n} d_i^2$ can be no smaller than zero, and even then can only be equal to zero if $d_i \equiv 0$, i.e., $c_i = k_i + 0$

$\Rightarrow$ k_i's produce estimator with the smallest standard error.

<u>THEOREM</u>: Suppose that $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$ and $Z \perp Y$.

$$T = Z \Big/ \sqrt{Y/n} \sim t(n)$$

In English: if you take a standard normal r.v. and divide by the square root of a rescaled independent χ^2 random variable, you end up with a student's t random variable with the same degrees of freedom as the χ^2.

BOTTOM LINE FOR CI's AND HYPOTHESIS TESTS FOR β_1:

$$\frac{b_1 - \beta_1}{s(b_1)} = \frac{b_1 - E[b_1]}{s(b_1)} \sim t(n-2)$$

Recall that $\dfrac{b_1 - \beta_1}{\sqrt{Var(b_1)}} \sim N(0,1)$ because

$E[b_1] = \beta_1$ and b_1 is normal because it is a linear combination of the Y_i's, which are normal because of the ε_i's.

NOTE: we can't really use $\frac{b_1 - \beta_1}{\sqrt{Var(b_1)}}$, but

we can use $\frac{b_1 - \beta_1}{s(b_1)}$!

NOTE:
$$\boxed{\frac{b_1 - \beta_1}{\sqrt{Var(b_1)} \cdot \frac{s(b_1)}{s(b_1)}} = \frac{b_1 - \beta_1}{s(b_1)} \cdot \frac{s(b_1)}{\sqrt{Var(b_1)}}}$$

QN: What is
$$\frac{s^2(b_1)}{\sigma^2(b_1)} = \frac{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{MSE}{\sigma^2}$$

$$= \frac{SSE}{\sigma^2(n-2)}$$

$$= \frac{\chi^2(n-2)}{n-2}$$

When I say SSE, I mean:
$$\boxed{SSE = \sum_{i=1}^{n} e_i^2}$$

$$\Rightarrow \frac{b_1 - \beta_1}{s(b_1)} \sim Z \Big/ \sqrt{\frac{\chi^2(n-2)}{n-2}} = t(n-2)$$

Suppose that you wanted to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

1. Collect $(X_1, Y_1), \ldots, (X_n, Y_n)$.
2. Estimate b_0, b_1, s_e^2.
3. Form $T = \dfrac{b_1 - \cancel{\beta_1}^{\,0}}{s(b_1)}$

   where $s(b_1) = \sqrt{\dfrac{MSE}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$
4. If $|T| > T^*_{\frac{\alpha}{2}, n-2}$, reject H_0 and conclude that β_1 is statistically distinguishable from zero.

Similary, a 95% CI for β_1 is given by

$$b_1 \pm t^*_{\frac{\alpha}{2}, n-2} \, s(b_1)$$

```
myModel <- lm(Y~X, data =myDataFrame)
summary(myModel)
```

$\rightarrow \boxed{}$