Goal: Confidence Intervals for Regression Lines

This is split into two steps:

- §12.3 Inferences about β_1
  - ↳ Confidence Intervals for Slope
- §12.4 Inferences about $\mu_Y^{(x)}$
  - ↳ Confidence Intervals for Intercept

Recall: Given paired sample data (x_i, Y_i) we assume $Y(x) = (\beta_0 + \beta_1 x) + \varepsilon$

$\varepsilon \sim \text{Normal}(0, \sigma)$

The mean of $Y(x)$ follows along the "Regression Line": $y = \beta_0 + \beta_1 x$

"$\mu_Y^{(x)}$"

We get point estimators for β_0 & β_1 by computing the "least squares best fit" line through the sample points (x_i, y_i)

- $\hat{\beta}_1 = S_{xy}/S_{xx} = \dfrac{\Sigma_i(x_i y_i) - \frac{1}{n}(\Sigma_i x_i)(\Sigma_i y_i)}{\Sigma_i(x_i^2) - \frac{1}{n}(\Sigma_i x_i)^2}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Useful Note: Plugging $\hat{\beta}_0 = (\bar{y} - \hat{\beta}_1 \bar{x})$ into regression line formula gives

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

$$\boxed{\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})}$$

(1) Regression Line ALWAYS goes through $(\bar{x}, \bar{y})$

(2) $\hat{\beta}_1$ = slope determines Regression Line

§12.3 Confidence Intervals for Slope β_1

A bit of theory:

The point estimator for slope β_1 is

$$\hat{\beta}_1 = \frac{\Sigma_i (x_i - \bar{x})(Y_i - \bar{Y})}{\Sigma_i (x_i - \bar{x})^2}$$

Split sum $(Y_i - \bar{Y})$ and factor out $\bar{Y}$

$$= \frac{\Sigma_i [(x_i - \bar{x}) Y_i]}{\Sigma_i (x_i - \bar{x})^2} - \frac{\Sigma_i [(x_i - \bar{x})] \bar{Y}}{\Sigma (x_i - \bar{x})^2}$$

($\Sigma_i [(x_i - \bar{x})] = 0$)

$$= \Sigma_i c_i Y_i \quad \text{where} \quad c_i = \frac{(x_i - \bar{x})}{\Sigma_i (x_i - \bar{x})^2}$$

Plugging in $Y_i = \beta_0 + \beta_1 x - \varepsilon_i$ we can compute:

$$E[\hat{\beta}_1] = \beta_1$$ ← i.e. $\hat{\beta}_1$ is an unbiased est.

$$\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \quad \text{where } \sigma^2 = \text{Var}[\varepsilon]$$

↑ the "within factor" variance

$$= \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{n} \frac{1}{\text{Var}[\bar{x}]}$$

↑ Considering $\bar{x}$ as a random variable...

Note that $\hat{\beta}_1 = \sum_i c_i Y_i$ where each Y_i is normal so $\hat{\beta}_1$ is also normal.

Replacing σ^2 by its estimator the "pooled sample variance" of $Y(x)$ — MSE — yields a random variable with t-distribution:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t(n-2)$$

Note: $\dfrac{MSE}{S_{xx}} = \dfrac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}^2} \Big/ (n-2)$

Thus the $(1-\alpha)$ Confidence Interval for β_1 is:

$$\beta_1 = \hat{\beta}_1 \pm qt\left(\tfrac{\alpha}{2}, n-2\right) \cdot \sqrt{MSE/S_{xx}}$$

This can also be used to perform a t-Test for

$$H_0: \beta_1 = 0$$

But the result will be the same p-value as the F-test in the ANOVA table...

§12.4 Inferences about $\mu_Y^{(x)}$

Recall: Regression line is

$$\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

Data follows distribution

$$Y = \mu_Y + \beta_1 (x - \bar{x}) + \varepsilon$$

$\mu_Y^{(x)}$ ← "True line of means" Regression line of data is "Point Estimate" of this

Just as in Chapter 7, $\bar{y}$ gives CI for μ_Y

$$\text{Var}[\bar{Y}] = \sigma^2/n \quad (\text{where } \sigma^2 = \text{Var}[\varepsilon])$$

So

$$\frac{\bar{Y} - \mu_Y}{\sqrt{MSE/n}} \sim t(n-2)$$

↑ $(n-2)$ degrees of freedom because of "pooled sample var" MSE.

Thus the $(1-\alpha)$ Confidence Interval for μ_Y is:

$$\mu_Y = \bar{y} \pm qt(\alpha/2, n-2) \cdot \sqrt{MSE/n}$$

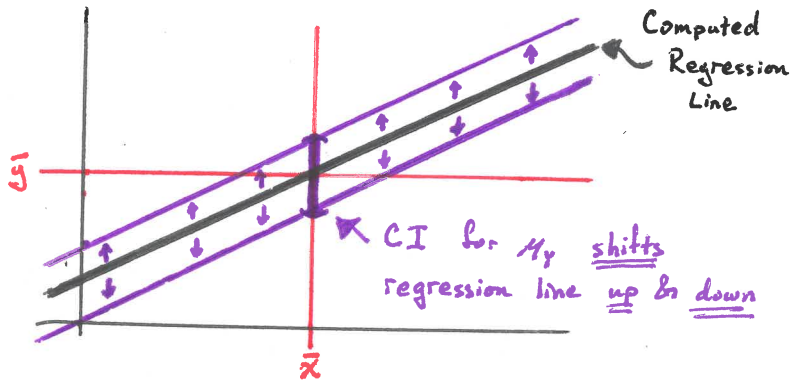
Combining we get an approximate $(1-\alpha)$ CI for $\mu_Y^{(x)}$:

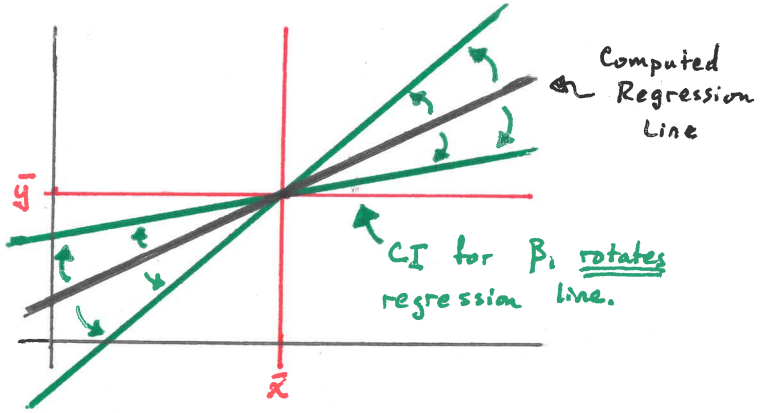
$$\mu_Y^{(x)} = \left(\bar{y} \pm qt(\alpha/2, n-2)\sqrt{MSE/n}\right) + \left(\hat{\beta}_1 \pm qt(\alpha/2, n-2)\sqrt{MSE/S_{xx}}\right)(x-\bar{x})$$

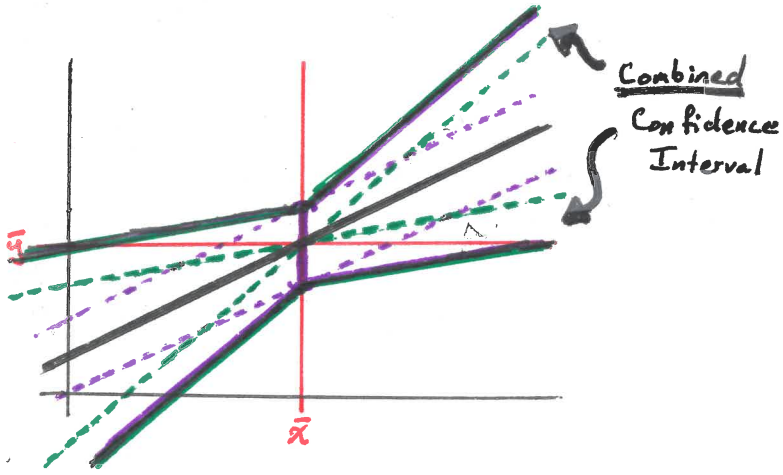
$$= \underbrace{\left(\bar{y} + \hat{\beta}_1(x-\bar{x})\right)}_{\hat{y}} \pm qt(\alpha/2, n-2) \cdot \left(\sqrt{MSE/n} + \underbrace{\sqrt{MSE/S_{xx}}\,(x-\bar{x})}_{\sqrt{MSE \cdot \frac{(x-\bar{x})^2}{S_{xx}}}}\right)$$

$\hat{y}$ – the regression line for sample data

[Note: Width of interval increases as x moves away from mean $\bar{x}$.]







Actually we can do slightly better than this

$$\mu_Y^{(x)} = \underbrace{\left(\bar{y} + \hat{\beta}_1(x-\bar{x})\right)}_{\hat{y}} \pm \underbrace{gt(\alpha/2, n-2)}_{t_{\alpha/2}} \left(\underbrace{\sqrt{MSE/n}}_{\sigma_{\bar{y}}} + \underbrace{\sqrt{MSE \cdot (x-\bar{x})^2 / S_{xx}}}_{\sigma_{\hat{\beta}_1 x}} \right)$$

Recall that std. dev. combines like Pythagorean Thm:

NOT $\left(\sqrt{\sigma_{\bar{y}}^2} + \sqrt{\sigma_{\hat{\beta}x}^2}\right)$ BUT $\sqrt{\sigma_{\bar{y}}^2 + \sigma_{\hat{\beta}x}^2}$

Note that $\sqrt{\sigma_1^2} + \sqrt{\sigma_2^2} > \sqrt{\sigma_1^2 + \sigma_2^2}$

(because hypotenuse is smaller than sum of legs)

so replacing by $\sqrt{\sigma_{\bar{y}}^2 + \sigma_{\hat{\beta}x}^2}$ makes CI smaller

(For Confidence Intervals, smaller is better.)

The $(1-\alpha)$ Confidence Interval for $\mu_Y^{(x)}$ is

$$\mu_Y^{(x)} = \underbrace{\left(\bar{y} + \hat{\beta}_1(x-\bar{x})\right)}_{\hat{y}} \pm \underbrace{gt(\alpha/2, n-1)}_{t_{\alpha/2}} \cdot \underbrace{\sqrt{MSE\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}}_{\sigma_{\hat{y}}}$$

Since $\sigma_{\bar{y}}^2 = \frac{MSE}{n}$ and $\sigma_{\hat{\beta}_1 x}^2 = MSE \frac{(x-\bar{x})^2}{S_{xx}}$ are usually very small this change does not affect the confidence interval graph too much — it makes the interval boundaries become curved.