Monday | Week 16.1 | April 14^{th}

Finishing up simple linear regression: Prediction intervals

Example: Runoff (problem 36, Spring 2019)

$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \leftarrow$ Note: this is identical to $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$\hat{\beta}_0 = -1364$

$\hat{\beta}_1 = 1.08$

$s^2 = 156^2$

$Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

Qu 1: Find $\hat{Y}_0$ for $x_0 = 2000$.

Ans:

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$
$$= -1364 + 1.08 \cdot 2000$$
$$= \boxed{796}$$

Qu 2: Show $E[\hat{Y}_0 - Y_0] = 0$

Ans:

$$E[\hat{Y}_0 - Y_0] = E[\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)]$$
$$= E[\hat{\beta}_0] + E[\hat{\beta}_1] x_0 - E[\beta_0 + \beta_1 x_0]$$
$$= \beta_0 + \beta_1 x_0 - (\beta_0 + \beta_1 x_0) = 0 \checkmark$$

Qu 3: Show that $Var(\hat{Y}_0 - Y_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

Ans:

$$Var(\hat{Y}_0 - Y_0) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0 + \varepsilon))$$

$$= Var(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 - \beta_0 - \beta_1 x_0 - \varepsilon)$$

$$= Var(\bar{Y} + \hat{\beta}_1 (x_0 - \bar{x}) - \varepsilon)$$

$(\bar{Y} \perp \hat{\beta}_1)$ $$= Var(\bar{Y}) + (x_0-\bar{x})^2 Var(\hat{\beta}_1) + Var(\varepsilon)$$

$$= \frac{1}{n}\sigma^2 + (x_0-\bar{x})^2 Var(\hat{\beta}_1) + Var(\varepsilon)$$

$= \dots$ (solve yourself)

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \checkmark$$

Qu 4: Give a 95% prediction interval (PI) for Y_0 where $x_0 = 2000$.

Ans: Since we've assumed the residuals are iid $N(0, \sigma^2)$, by the video lecture for today we have:

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \sim t_{n-2}$$

yielding the PI:

$$\left(\hat{Y}_0 - t_{23,0.025}\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}(x_i-\bar{x})^2}\right)},\ \hat{y}_0 + t_{23,0.025}\sqrt{s^2\left(1+\frac{1}{n}+\dots\right)}\right)$$

✓

Example: Medicine concentration (problem 3, Spring 2016)

$Y_i = bx_i + \varepsilon_i$, $i=1,\dots,10$ $\qquad$ i: patient

Y_i: concentration of medicine

x_i: injection dosage of medicine

ε_i: noise/error.

$\boxed{\text{Qu 1:}}$ How should the residual plot and normal QQ plot be interpretted?

$\boxed{\text{Ans:}}$ The residual plot shows estimated residuals centered at 0 for all dosages. While some residuals for smaller dosages may be slightly larger in magnitude, since n is small there is little to suggest the residual variance varies with x, implying the assumption of homoscedasticity (constant variance) is reasonable.

The QQ plot shows residuals are roughly Gaussian, with perhaps a slight heavy-tailedness, although there is insufficient data to suggest the residuals are non Gaussian.

$\boxed{\text{Qu 2:}}$ Do the model assumptions appear to be reasonable?

$\boxed{\text{Ans:}}$ Yes! For reasons discussed above, and since patients are independent, and dosage is fixed in advance so $x_1, \dots, x_{10}$ are truly fixed.

$\boxed{\text{Qu 3:}}$ Which of the following estimators for b do you prefer when $n=10$?

A) $\tilde{b} = \dfrac{\bar{Y}}{\bar{x}}$ B) $\hat{b} = \bar{Y}$ C) $\hat{\hat{b}} = \dfrac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$.

Assume $E[\hat{b}]=0$ and $Var(\hat{b}) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$, and

$n\bar{x} = 36.5$, $\sum_{i=1}^{n} x_i^2 = 152.25$, & $\sum_{i=1}^{n} x_i Y_i = 331.65$.

$\boxed{\text{Ans:}}$

$$
\begin{aligned}
E[\tilde{b}] &= E\left[\frac{\bar{Y}}{\bar{x}}\right] \\
&= \frac{1}{\bar{x}} E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n\bar{x}} \sum_{i=1}^{n} E[Y_i] \\
&= \frac{1}{n\bar{x}} \sum_{i=1}^{n} b x_i \\
&= \frac{b}{n\bar{x}} \sum_{i=1}^{n} x_i \\
&= b \quad \checkmark
\end{aligned}
$$

$$
\begin{aligned}
Var(\tilde{b}) &= Var\left(\frac{\bar{Y}}{\bar{x}}\right) \\
&= \frac{1}{\bar{x}^2} Var(\bar{Y}) \\
&= \frac{1}{\bar{x}^2} Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{(n\bar{x})^2} \sum_{i=1}^{n} Var(Y_i) \\
&= \frac{n\sigma^2}{(n\bar{x})^2} = \sigma^2 \cdot \frac{10}{(36.5)^2} \approx \sigma^2 \cdot 0.0075
\end{aligned}
$$

$E[\hat{b}] = E[\bar{y}] = b\bar{x} \neq b$

So $\hat{b}$ is a bad estimator for b...

Plugging in for $Var(\hat{b})$:

$$Var(\hat{b}) = \frac{\sigma^2}{\sum_i x_i^2}$$

$$= \sigma^2 \cdot \frac{1}{152.25} \approx \sigma^2 \cdot 0.0066$$

$$< Var(\tilde{b})$$

So $\hat{b}$ is preferred. $\boxed{C)}$

$\boxed{\text{Qu 3:}}$ Find the MLE (SME) for b.

$\boxed{\text{Ans:}}$

$$L(b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - bx_i)^2\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(y_i - bx_i)^2\right\}$$

$$l(b) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - bx_i)^2$$

$$\frac{\partial l}{\partial b} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - bx_i)x_i = 0$$

$$\Rightarrow 0 = \sum_{i=1}^{n} (y_i - b x_i) x_i$$

$$= \sum_{i=1}^{n} x_i y_i - b \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \hat{b}_{MLE} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \hat{b} \quad \checkmark$$

$\boxed{(c)}$

Qu 4: Derive a $(1-\alpha)\times 100\%$ CI for b assuming $\sigma^2 = 2$ and $\alpha = 0.1$

Ans: First, since σ^2 is known and since we assume $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$, $\hat{b} = \frac{\bar{Y}}{\bar{X}}$ is a linear combination of the Gaussian distributed $Y_1, \ldots, Y_n$, so $\hat{b}$ is Gaussian (not t-distributed!). We already know it is unbiased w/ variance $\mathrm{Var}(\hat{b}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$, so a 90% CI is:

$$\left(\hat{b} - z_{.05}\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}},\ \hat{b} + z_{.05}\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}\right)$$

$$= (1.9117,\ 2.4450)$$

$$\boxed{A)}$$