Simple Linear Regression (Ch 2)
Week 4 – Thursday

Applied Regression Analysis (STAT 757)

Paul J. Hurtado

Thursday, 11 Feb, 2016

Recap: Confidence Intervals

Intercept CI: The distribution of the intercept estimator $\widehat{\beta}_0$ can be computed by *un-standardizing* the following r.v., which follows a *Student's t* distribution with $n-1$ d.f.

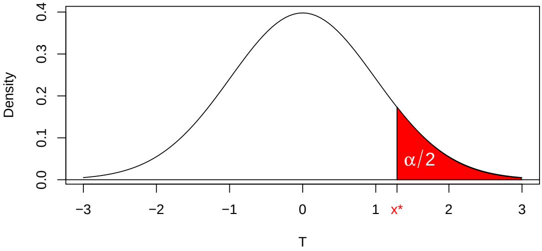
$$T = \frac{\widehat{\beta}_0 - \beta_0}{\text{se}(\widehat{\beta}_0)}$$

Recall $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ and $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, and that the standard error of $\widehat{\beta}_0$ is $\text{se}(\widehat{\beta}_0) = S\sqrt{\frac{1}{n}\frac{\bar{x}^2}{S_{XX}}}$.

CI: 100(1-α)% of the time, *parameter* β_0 is in

$$\left[\widehat{\beta}_0 + qt(\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta}_0),\ \ \widehat{\beta}_0 + qt(1-\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta}_0)\right]$$

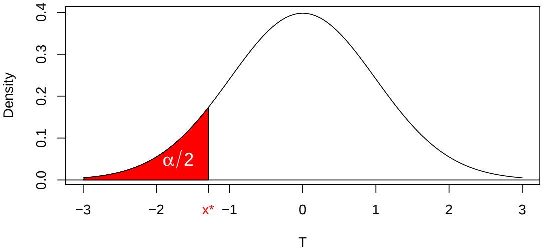
Note: t() vs qt()



The T-score (and Z-score) tables found in classical textbooks tell you the *random variable value* x_* that satisfy

$$P(X > x_*) = \alpha/2$$

Our textbook uses $x_* = t(\alpha/2, n-2)$ to denote these "upper tail" values. **However, in R...**



... we compute these values as **quantiles**. For example, the 25% quantile, x_*, satisfies $P(X \leq x_*) = 0.25$.
Thus, to clarify **textbook** vs **R** notation:

$$-t(\alpha/2, n-2) = \texttt{qt}(\alpha/2, n-2)$$
$$t(\alpha/2, n-2) = \texttt{qt}(1-\alpha/2, n-2)$$

Recap: Confidence Intervals

Slope CI: The distribution of the slope estimator $\widehat{\beta}_1$ can be computed by *un-standardizing* the t_{n-1} distributed r.v.

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\text{se}(\widehat{\beta}_1)}$$

Recall $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ and $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, and that the standard error of $\widehat{\beta}_1$ is $\text{se}(\widehat{\beta}_1) = S/\sqrt{S_{XX}}$.

CI: 100(1-α)% of the time, *parameter* β_1 is in

$$\left[\widehat{\beta}_1 + qt(\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta}_1),\ \ \widehat{\beta}_1 + qt(1-\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta}_1)\right]$$

Recap: Confidence Intervals

Regression Line CI: The distribution of $\widehat{y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*$ (or, $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$) can be computed by *un-standardizing* the t_{n-1} distributed r.v.

$$T = \frac{\widehat{y}_* - (\beta_0 + \beta_1 x_*)}{S\sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}}$$

Thus,

CI: 100(1-α)% of the time, $E(\hat{y}_*) = \beta_0 + \beta_1 x_*$ is in

$$\widehat{y}_* \pm qt(1 - \tfrac{\alpha}{2}, n-2) S\sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}$$

Recap: Prediction Intervals

Prediction Interval: The distribution of $\widehat{Y_*} = \widehat{\beta}_0 + \widehat{\beta}_1 x_*$ (or, $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$) can be computed by *un-standardizing* the t_{n-1} distributed r.v.

$$T = \frac{\widehat{y_*} - (\beta_0 + \beta_1 x_*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}}$$

Note the CI for $E(\widehat{y_*})$ uses se$(\widehat{y_*})$ while the prediction interval uses se$(\widehat{Y_* - y_*})$. Thus,

CI: 100(1-α)% of the time, $E(\hat{y}_*) = \beta_0 + \beta_1 x_*$ is in

$$\widehat{y_*} \pm qt(1 - \tfrac{\alpha}{2}, n-2) S\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}$$

Analysis of Variance

Observe that, if $\beta_1 = 0$ then the SLR model becomes

$$Y = \beta_0 + \epsilon \sim Normal(\beta_0, \sigma)$$

and so $\widehat{\beta_0} = \widehat{y} = \bar{y}$. To test for a significant linear relation ship, we test against this null hypothesis ($H_0 \ : \ \beta_1 = 0$) using

$$T = \frac{\widehat{\beta_1} - 0}{\mathrm{se}(\widehat{\beta_1})} \sim t_{n-2}$$

In multiple regression, however, we need to generalize. This leads us to a different test statistic...

Analysis of Variance

Q: How much of the variation in y_i values comes from the linear component?

$$SST = S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

It can be shown (see next slide) that

$$\overbrace{SST}^{\text{Total variation in Y}} = \overbrace{SS_{reg}}^{\text{SS explained by regression}} + \overbrace{RSS}^{\text{residuals}}$$

Proof that $SST = SS_{reg} + RSS$:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\overbrace{y_i - \widehat{y}_i}^{e_i} + \widehat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} e_i^2 + (\widehat{y}_i - \bar{y})^2 + 2\, e_i(\widehat{y}_i - \bar{y}) = SS_{reg} + RSS + 2\sum_{i=1}^{n} e_i(\widehat{y}_i - \bar{y})$$

However, from our derivation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (see textbook pg. 18), recall that $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i\, e_i = 0$. Thus we see that

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

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

Analysis of Variance

Generalizing the t test above, we can also get a p-value for the hypothesis test

$$H_0 \ : \ \beta_1 = 0 \quad vs \quad H_A \ : \ \beta_1 \neq 0$$

by using a more general test statistic that quantifies how much variation in Y results from the linear trend relative to the random variation determined by the magnitude of σ:

$$F = \frac{SS_{reg}/1}{RSS/(n-2)}$$

F has an $\mathrm{F}_{1,n-2}$ distribution, and will be revisited in Ch. 5.

0-1 Categorical SLR

§2.6-2.7 on Monday