Inference in Simple Linear Regression

**Fact:** Under normal regression model \((b_0,b_1)\) and \(SSE\) are independently distributed and 
\[
\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}, \quad \frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}, \quad SSE \sim \sigma^2 \chi_{n-2}^2.
\]

**Confidence interval for \((\beta_0, \beta_1)\):** \(100(1-\alpha)\%\) (two-sided) confidence interval for 
\[
\beta_i \quad (i=0,1):
\]
\[
(b_i - t(1-\alpha/2;n-2) s(b_i)), \quad (b_i + t(1-\alpha/2;n-2) s(b_i))
\]
for \((i=0,1)\), where \(t(1-\alpha/2;n-2)\) is the \((1-\alpha/2)\) quantile of \(t_{n-2}\) distribution; i.e., \(P(t_{n-2} > t(1-\alpha/2;n-2)) = \alpha/2\).

**Hypothesis tests for \((\beta_0, \beta_1)\):** \(H_0 : \beta_i = \beta_{i0}\) (\(i=0\) or \(1\)).

1. Alternative: \(H_1 : \beta_i > \beta_{i0}\). Reject \(H_0\) at level \(\alpha\) if 
\[
\frac{b_i - \beta_{i0}}{s(b_i)} > t(1-\alpha;n-2).
\]
Or if, P-value = \(P(t_{n-2} > T_i^{observed}) < \alpha\).

2. Alternative: \(H_1 : \beta_i < \beta_{i0}\). Reject \(H_0\) at level \(\alpha\) if 
\[
\frac{b_i - \beta_{i0}}{s(b_i)} < t(\alpha;n-2).
\]
Or if, P-value = \(P(t_{n-2} < T_i^{observed}) < \alpha\).

3. Alternative: \(H_1 : \beta_i \neq \beta_{i0}\). Reject \(H_0\) at level \(\alpha\) if 
\[
\left|\frac{b_i - \beta_{i0}}{s(b_i)}\right| > t(1-\alpha/2;n-2).
\]
Or if, P-value = \(P(|T_i^{observed}| > t(1-\alpha/2;n-2)) < \alpha\).

Inference for mean response at \(X = X_h\)

---

https://stats.libretexts.org/Bookshelves/Computing_and_Modeling/Supplemental_Modules_(Computing_and_Modeling)/Regression/Inference_in_Simple_Linear_Regression

Updated: Fri, 05 Jun 2020 01:08:44 GMT
Powered by
• Point estimate: $\hat{Y}_h = b_0 + b_1 X_h$.

Fact: $(E(\hat{Y}_h) = \beta_0 + \beta_1 X_h) = E(Y_h)$, $\text{Var}(\hat{Y}_h) = \sigma^2(\hat{Y}_h) = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$. Estimated variance is $s^2(\hat{Y}_h) = \text{MSE} \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$.

Distribution: $\frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)} \sim t_{n-2}$.

Confidence interval: $(100(1-\alpha))% confidence interval for $(E(Y_h))$ is $\left((\hat{Y}_h - t(1-\alpha/2;n-2)s(\hat{Y}_h),\hat{Y}_h + t(1-\alpha/2;n-2)s(\hat{Y}_h))\right)$.  

Prediction of a new observation $(Y_{(h(new))})$ at $(X = X_h)$

• Prediction: $\hat{Y}_{(h(new))} = \hat{Y}_h = b_0 + b_1 X_h$.

Error in prediction: $\left(Y_{(h(new))} - \hat{Y}_{(h(new))} = Y_{(h(new))} - \hat{Y}_h\right)$. 

Fact: $(\sigma^2(Y_{(h(new))} - \hat{Y}_h) = \sigma^2(Y_{(h(new))}) + \sigma^2(\hat{Y}_h) = \sigma^2\left[1+\frac{1}{n}+ \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$.

Estimate of $(\sigma^2(Y_{(h(new))} - \hat{Y}_h)) is $(s^2(Y_{(h(new))} - \hat{Y}_h) = \text{MSE} \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$.

Distribution: $\frac{Y_{(h(new))} - \hat{Y}_h}{s(Y_{(h(new))} - \hat{Y}_h)} \sim t_{n-2}$.

Prediction interval: $(100(1-\alpha))% prediction interval for $(Y_{(h(new))}) is $\left((\hat{Y}_h - t(1-\alpha/2;n-2)s(Y_{(h(new))} - \hat{Y}_h),\hat{Y}_h + t(1-\alpha/2;n-2)s(Y_{(h(new))} - \hat{Y}_h))\right)$. 

• Confidence band for the regression line: At $(X=X_h)$ the $(100(1-\alpha))% confidence band for the regression line is given by $((\hat{Y}_h \pm w_{\alpha}s(\hat{Y}_h), \text{qquad where} \sim w_{\alpha} = \sqrt{2F(1-\alpha; 2, n-2)})$.

Here $\sqrt{F(1-\alpha;2,n-2)}$ is the $(1-\alpha)$ upper cut-off point (or, $(1-\alpha)$ quantile) for the $(F_{2,n-2})$ distribution with d.f. $(2,n-2))$.

Example $(\text{Page}Index(1))$: Simple linear regression

We consider a data set on housing price. Here $(Y=\text{\textdollar})$ selling price of houses (in $1000), and $(X=\text{\textdollar}) size of house (100 square feet). The summary statistics are given below:

$(n = 19), (\overline{X} = 15.719), (\overline{Y} = 75.211)$

$(\sum_i (X_i - \overline{X})^2 = 40.805), (\sum_i (Y_i - \overline{Y})^2 = 556.078), (\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = 120.001)$. 

Estimates of $(\beta_1)$ and $(\beta_0)$:

$b_1 = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} = \frac{120.001}{40.805} = 2.941$
\[b_0 = \overline{Y} - b_1 \overline{X} = 75.211 - (2.941)(15.719) = 28.981.\]

- **Fit and Prediction**: The fitted regression line: \(Y = 28.981 + 2.941X\). When \(X = 18.5 = X_h\), the predicted value, that is an estimate of the mean selling price (in $1000) when size of the house is 1850 sq. ft., is \(\hat{Y}_h = 28.981 + (2.941)(18.5) = 83.39\).

- **MSE**: The degrees of freedom (df) \(= n-2 = 17\). \(\text{SSE} = \sum_i(Y_i - \overline{Y})^2 - b_1^2\sum_i(X_i - \overline{X})^2 = 203.17\). So, \(\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{203.17}{17} = 11.95\).

- **Standard Error Estimates**: \(s^2(b_0) = \text{MSE} \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_i(X_i - \overline{X})^2}\right] = 73.00\), \(s(b_0) = \sqrt{s^2(b_0)} = 8.544\). \(s^2(b_1) = \frac{\text{MSE}}{\sum_i(X_i - \overline{X})^2} = 0.2929\), \(s(b_1) = \sqrt{s^2(b_1)} = 0.5412\).

- **Confidence Intervals**: We assume that the errors are normal to find confidence intervals for the parameters \(\beta_0\) and \(\beta_1\). We use the fact that \(\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}\) and \(\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}\) where \(t_{n-2}\) denotes the \(t\)-distribution with \(n-2\) degrees of freedom. Since \(t(0.975;17) = 2.1098\), it follows that 95% two-sided confidence interval for \(\beta_1\) is \((2.941 \pm 2.1098)(0.5412) = (1.80, 4.08)\). Since \(t(0.95;17) = 1.740\), the 90% two-sided confidence interval for \(\beta_0\) is \((28.981 \pm 1.740)(8.544) = (14.12, 43.84)\).

**Contributors**

- Agnes Oshiro

(Source: Spring 2012 STA108 Handout 4)