Statistical inference for the slope and intercept in SLR

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ for\ i = 1, \ldots n$$

Simple Linear Regression Model Parameters

- β_0 is the intercept.
- β_1 is the slope.
- ε_i are independent, normally distributed random errors with mean 0 and variance σ^2,

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

The point estimate of β1 is b_1

- Recall that, $$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

which we can rewrite as, $$= \Sigma c_i(Y_i - \bar{Y}) = \Sigma c_i Y_i - \bar{Y}\Sigma c_i = \Sigma c_i Y_i$$

where $$c_i = \frac{(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}$$

It can be proved that, $E(b_1) = \beta_1 \; and \;\; \sigma^2(b_1) = \sigma^2 \frac{1}{\Sigma(X_i - \bar{X})^2}$, therefore

By replacing the parameter σ^2 with $MSE,$ the unbiased estimator of $\sigma^2\{b_1\}$,

$$s^2\{b_1\} = \frac{MSE}{\Sigma(X_i - \bar{X})^2}$$

$$s\{b_1\} = \sqrt{\frac{MSE}{\Sigma(X_i - \bar{X})^2}}$$

The Sampling Distribution of $\boldsymbol{b_1}$ is Normal $(\beta_1, \sigma^2(b_1))$

$$\frac{\boldsymbol{b_1 - \beta_1}}{\boldsymbol{s\{b_1\}}} \sim \mathbf{t(n-2)}$$

Confidence Interval for β_1

Since $t^* = \frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$

$$P\left\{t\left(\frac{\alpha}{2}; n-2\right) \le \frac{b_1 - \beta_1}{s\{b_1\}} \le t\left(1 - \frac{\alpha}{2}; n-2\right)\right\} = 1 - \alpha$$

Where $t(\frac{\alpha}{2}; n-2)$ denotes the $\left(\frac{\alpha}{2}\right)100$ percentile of the t distribution with $n-2$ degrees of freedom.

Because of the symmetry of the t distribution around its mean 0, it follows that:

$$t\left(\frac{\alpha}{2}; n-2\right) = -t(1 - \frac{\alpha}{2}; n-2)$$

Hence the $1-\alpha$ confidence interval for β_1 are:

$$\boldsymbol{b_1 \pm t\left(1 - \frac{\alpha}{2}; n-2\right) s\{b_1\}}$$

Point estimate $\pm$ Margin error, **where Margin error (denoted by ME) = t * standard error**

Significance Tests for β_1

$$Ho: \beta_1 = \beta_1^* \qquad Ha: \beta_1 \neq \beta_1^*$$

The test statistic $t^* = (b_1 - \beta_1^*)/s\{b_1\} \sim t(n-2)$

For two sided test

Reject H_0 if $|t^*| \geq t_c$, $t_c = t_{n-2}(1-\alpha/2)$

Or, reject H_0 if $p - value \leq \alpha$

For one sided test

Reject H_0 if $|t^*| \geq t_c$, $t_c = t_{n-2}(1-\alpha)$

Or, reject H_0 if $p - value \leq \alpha$

Inference for the intercept, $\boldsymbol{\beta_0}$

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

It can be proved that, $E(b_0) = \beta_0 \; and \; \sigma^2\{b_0\} = \sigma^2[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma\,(X_i - \bar{X})^2}]$

$$s^2\{b_0\} = MSE[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma\,(X_i - \bar{X})^2}]$$

$$s\{b_0\} = \sqrt{MSE[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma\,(X_i - \bar{X})^2}]}$$

Analogous to theorem for b_1, $t^* = (b_0 - \beta_0)/s\{b_0\} \sim t(n-2)$

Confidence Interval for $\boldsymbol{\beta_0}$

$$b_0 \pm t\left(1-\frac{\alpha}{2}; n-2\right)s\{b_0\}$$

Significance Tests for $\boldsymbol{\beta_0}$

$$Ho: \beta_0 = \beta_0^* \qquad Ha: \beta_0 \neq \beta_0^*$$

The test statistic $t^* = (b_0 - \beta_0^*)/s\{b_0\} \sim t(n-2)$

Comments on the inference assumptions

- Both b_1 and b_0 follow *Normal distribution* because they are based on ε which is normally distributed.

- As long as the εs are close to normal, the t-method for the inferences based on b_1 and b_0 is approximately correct, even with small sample sizes.

Comments on the inference assumptions

- Often, the value of the intercept is not of direct interest, so there is no need to calculate CIs or hypothesis tests β_0. Because it is just a single value of Y when X=0 and will be of no much value to predict other Y values.

- Reduce the standard error for estimating the linear impact, β_1, by increasing the dilation in X, i.e., bigger $SSX = \Sigma(X_i - \bar{X})^2$, since $s\{b_1\} = \frac{s}{\sqrt{SSX}}$

One way to do confidence interval for $\boldsymbol{\beta_1}$

```
alpha=0.05
n=48
qt(1-0.5*alpha,n-2)
confint(lm(price~weight, diamond),"weight",level=0.95)
```

$$\alpha = 0.05$$
$$n = 48$$
$$t\left(1-\frac{\alpha}{2}, n-2\right) = t(0.975, 46) = 2.013$$
$$b_1 \pm t\left(1-\frac{\alpha}{2}; n-2\right) s\{b_1\}$$

```
          2.5 %   97.5 %
weight 3556.398 3885.651
```

Conclusion: we are 95% confident that,
the average price will
increase by at least 3556 and at most 3889
when the weight increase by 1 carat .

The other way to do confidence interval for $\boldsymbol{\beta_1}$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.84 on 46 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778 
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

$$b_1 \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) s\{b_1\}$$

$$= 3721 \pm \mathbf{2.013}\,(81.79) = 3556.4\,,\ 3885.65$$

Do hypothesis test for $\boldsymbol{\beta_1}$

$$Ho:\ \beta_1 = 3500\ \ vs \quad Ha: \beta_1 \neq 3500$$

$$The\ test\ statistic:\quad t_s = \frac{b_1 - 3500}{S\{b_1\}} = \frac{3721 - 3500}{81.79} = 2.702$$

$$The\ reject\ region:\quad reject\ H_0, \qquad if\ \ |t_s| > t\left(1 - \frac{\alpha}{2}, n-2\right) = t(0.975, 46) = 2.103$$

$$The\ p\ value = 2Pr(T > 2.702) = 0.00962$$

```
2*(1-pt(2.702,46))
```

```
[1] 0.00962015
```

Conclusion: at a significant level of 5%, when the weight increases by 1 caret, the incensement in the average price is not statistically different from 3500 dollars.

One sided hypothesis test for $\boldsymbol{\beta_1}$

$$Ho:\ \beta_1 = 3500\ \ vs \quad Ha: \beta_1 > 3500$$

$$The\ test\ statistic: \quad t_s = \frac{b_1 - 3500}{S\{b_1\}} = \quad \frac{3721 - 3500}{81.79} = 2.702$$

$$The\ reject\ region: \quad reject\ H_0, \qquad if\ \ |t_s| > t(1-\alpha, n-2) = t(0.95, 46) = 1.679$$

$$The\ p\ value = Pr(T > 2.702) = 0.0048$$

Conclusion: at a significant level of 5%,
when the weight increases by 1 caret,
the incensement in the average price
is not statistically greater than
3500 dollars.