Simple Linear Regression

Dr. Kiah Wah Ong

Inferences concerning the mean response $\beta_0 + \beta_1 x$

Let's say a utility company collected n data $(x_1, y_1), \cdots, (x_n, y_n)$ where x_i is temperature at time i and y_i is the power consumption at x_i. Suppose we have our linear regression model set up to be

$$y = \beta_0 + \beta_1 x + \epsilon$$

We now look at how we can use the data to estimate the mean (average) power consumption for a given temperature that we called x_0.

Notice that the mean response at x_0 is given by

$$\mu_0 = \beta_0 + \beta_1 x_0$$

Inferences concerning the mean response $\beta_0 + \beta_1 x$

If a point estimator for $\mu_0 = \beta_0 + \beta_1 x_0$ is required, $\hat{\mu_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ being an unbiased estimator, is a natural choice.

However, if a confidence interval or hypothesis testing about the mean response is required, then we need to know the probability distribution for the estimator $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Inferences concerning the mean response $\beta_0 + \beta_1 x$

Again from

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) y_i$$

and

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$$

we have

$$\begin{aligned}
\hat{\beta}_0 + \hat{\beta}_1 x_0 &= (\overline{y} - \hat{\beta}_1 \overline{x}) + \hat{\beta}_1 x_0 \\
&= \overline{y} - \hat{\beta}_1 (\overline{x} - x_0) \\
&= \frac{\sum_{i=1}^{n} y_i}{n} - \left(\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) y_i \right) (\overline{x} - x_0) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - c(x_i - \overline{x})(\overline{x} - x_0) \right) y_i
\end{aligned}$$

where $c = 1/S_{xx}$

Inferences concerning the mean response $\beta_0 + \beta_1 x$

From

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - c(x_i - \overline{x})(\overline{x} - x_0) \right) y_i$$

where $c = 1/S_{xx}$ we see that $\hat{\beta}_0 + \hat{\beta}_1 x_0$ can be expressed as a linear combination of independent normal random variables and so it itself also normally distributed.

Since we know

$$E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

we only need to compute $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$.

Inferences concerning the mean response $\beta_0 + \beta_1 x$

Note that

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) &= \sum_{i=1}^{n} \left(\frac{1}{n} - c(x_i - \overline{x})(\overline{x} - x_0) \right)^2 \mathrm{Var}(y_i) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}} \right)\end{aligned}$$

Hence

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}} \right) \right)$$

Inferences concerning the mean response $\beta_0 + \beta_1 x$

Recall from your previous statistics classes:

If Z and X_n^2 are independent r.v. with $Z \sim \mathrm{N}(0,1)$ and X_n^2 with chi-square with n degree of freedom , then the r.v.

$$T_n := \frac{Z}{\sqrt{X_n^2/n}} \sim t_n$$

that is T_n has a t-distribution with n degree of freedom.

Inferences concerning the mean response $\beta_0 + \beta_1 x$

Applying this to

$$\hat{\beta}_0+\hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\right) \text{ and } \frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

and $\hat{\beta}_0 + \hat{\beta}_1 x_0$ being independent of SS_R/σ^2 it follows that

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sqrt{\frac{1}{n} + \frac{(\overline{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}$$

Inferences concerning the mean response $\beta_0 + \beta_1 x$

Once we have this result

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sqrt{\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}} \sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}$$

we can then use it to obtain the following confidence interval of $\hat{\beta}_0 + \hat{\beta}_1 x_0$.

Inferences concerning the mean response $\beta_0 + \beta_1 x$

For any significant level $0 < \alpha < 1$,

$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sqrt{\frac{1}{n} + \frac{(\overline{x}-x_0)^2}{S_{xx}}}\ \sqrt{\frac{SS_R}{n-2}}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

gives the following confidence interval

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm \sqrt{\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}}\ \sqrt{\frac{SS_R}{n-2}}\ \ t_{\alpha/2,n-2},$$

Inferences concerning the mean response $\beta_0 + \beta_1 x$

```r
getwd()
#naming the data as data1
data1<-read.csv("Experiment1.CSV", header=TRUE, sep=",")
x<-data1$x
y<-data1$y
plot(x,y)
model=lm(y~x)
predict(model,data.frame(x=68),interval="confidence", conf.level=0.95)

```

Inferences concerning the mean response $\beta_0 + \beta_1 x$

```
> data1<-read.csv("Experiment1.CSV", header=TRUE, sep=",")
> x<-data1$x
> y<-data1$y
> plot(x,y)
> model=lm(y~x)
> predict(model,data.frame(x=68),interval="confidence", conf.level=0.95)
       fit      lwr     upr
1 168.8177 157.9585 179.677
```

Inferences concerning the Point Prediction

We can use our regression model to predict the value of y for any given value of x. For example, in our regression model based on the data file Experiment1, we have

$$\hat{y} = 18.9446 + 2.2040x$$

We can then use it to predict the value of y when $x = 68$, namely

$$\hat{y} = 18.9446 + 2.2040 * 68 = 168.8166$$

But we know the fitted value $\hat{y}$ can never be accurate because of the error term ϵ. The true model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

hence x is not the only factor that affected y.

Inferences concerning the Point Prediction

Let $y_0 = \beta_0 + \beta_1 x_0 + \epsilon$ be the future response whose input level is x_0 and consider the probability distribution of the response minus its predicted value $\hat{\beta}_0 + \hat{\beta}_1 x_0$. Since

$$y_0 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\right)$$

and

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\right)$$

and that y_0 is independent of the earlier data values $y_1, \cdots, y_n$ that were used to determine $\hat{\beta}_0$ and $\hat{\beta}_1$, it follows that y_0 is independent of $\hat{\beta}_0 + \hat{\beta}_0 x_0$ ...

Inferences concerning the Point Prediction

therefore

$$y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0 \sim \mathrm{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\right)$$

Note that this is because of

$$\begin{aligned}
E(y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) &= E(\beta_0 + \beta_1 x_0 + \epsilon - \hat{\beta}_0 - \hat{\beta}_1 x_0) \\
&= E(\beta_0 + \beta_1 x_0) + E(\epsilon) - E(\hat{\beta}_0) - E(\hat{\beta}_1 x_0) \\
&= \beta_0 + \beta_1 x_0 + 0 - \beta_0 - \beta_1 x_0 \\
&= 0
\end{aligned}$$

while

$$\begin{aligned}
\mathrm{Var}(y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) &= \mathrm{Var}(\beta_0 + \beta_1 x_0 + \epsilon - \hat{\beta}_0 - \hat{\beta}_1 x_0) \\
&= 0 + \mathrm{Var}(\epsilon) + \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)
\end{aligned}$$

Inferences concerning the Point Prediction

From

$$y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0 \sim \mathrm{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\right)$$

we get

$$\frac{y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}}} \sim \mathrm{N}(0, 1)$$

Using the argument laid out before, we obtain

$$\frac{y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0}{\sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}$$

Inferences concerning the Point Prediction

So, for any $0 < \alpha < 1$,

$$P\left(-t_{\alpha/2,n-2} < \frac{y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0}{\sqrt{1 + \frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

ILLINOIS TECH

Inferences concerning the Point Prediction

This gives the prediction interval at $100(1-\alpha)$ % as

$$\hat{\beta}_0 - \hat{\beta}_1 x_0 \pm t_{\alpha/2,n-2}\sqrt{\left(1+\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}\right)\frac{SS_R}{n-2}}$$

Inferences concerning the Point Prediction

```r
getwd()
#naming the data as data1
data1<-read.csv("Experiment1.CSV", header=TRUE, sep=",")
x<-data1$x
y<-data1$y
plot(x,y)
model=lm(y~x)
predict(model,data.frame(x=68),interval="confidence", conf.level=0.95)
predict(model,data.frame(x=68),interval="prediction", conf.level=0.95)
```

Inferences concerning the Point Prediction

```
> data1<-read.csv("Experiment1.CSV", header=TRUE, sep=",")
> x<-data1$x
> y<-data1$y
> plot(x,y)
> model=lm(y~x)
> predict(model,data.frame(x=68),interval="confidence", conf.level=0.95)
       fit      lwr     upr
1 168.8177 157.9585 179.677
> predict(model,data.frame(x=68),interval="prediction", conf.level=0.95)
       fit      lwr      upr
1 168.8177 75.32463 262.3108
> 
```