Regression: Some Review, and Confidence Intervals for Slope

T.Scofield

You may click here to access the .qmd file.

Review of scatterplots, correlation

The **InkjetPrinters** data is available in the `Lock5withR` package, and one of the first data sets used in examples in Chapter 9.

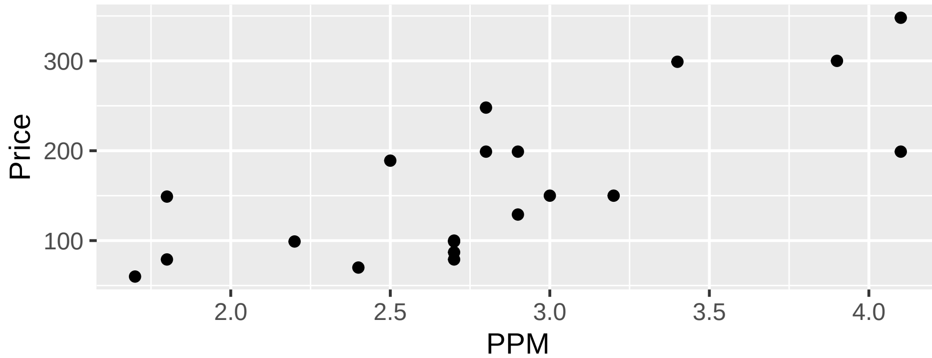
```
head(InkjetPrinters)
```

```
                                 Model PPM PhotoTime Price CostBW CostColor
1 HP Photosmart Pro 8500A e-All-in-One 3.9        67   300    1.6       7.2
2                   Canon Pixma MX882 2.9        63   199    5.2      13.4
3                 Lexmark Impact S305 2.7        43    79    6.9       9.0
4              Lexmark Interpret S405 2.9        42   129    4.9      13.9
5                 Epson Workforce 520 2.4       170    70    4.9      14.4
6                 Brother MFC-J6910DW 4.1       143   348    1.7       7.9
```

A scatterplot involves two quantitative variables, one selected to serve in the role of explanatory variable, and one as response. If we wish to see how `PPM` might serve in explaining `Price`, this scatterplot helps develop the sense that

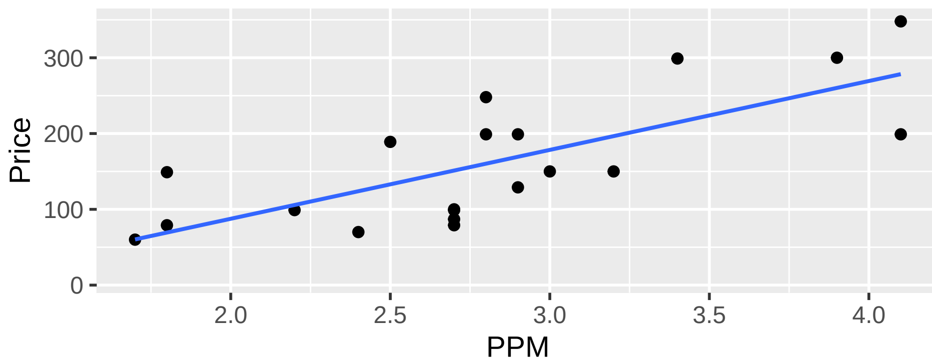
- there is an association
- the association is *positive* (as `PPM` rises, `Price` does, too)
- a line might serve well to describe the relationship

```
gf_point(Price ~ PPM, data=InkjetPrinters)
```



To visualize the best-fit line passing through the data, we can pipe the scatterplot to the `gf_lm()` command:

```
gf_point(Price ~ PPM, data=InkjetPrinters) |> gf_lm()
```



This added line does not perfectly describe the points; the association is not so strong as that (and never is). A measure of the strength of the linear relationship is the **correlation coefficient**:

```
cor(Price ~ PPM, data=InkjetPrinters)
```

```
[1] 0.7396862
```

The stronger the relationship, the closer this correlation coefficient, denoted by r, would be 1 or to (-1).

Coefficients of regression line

The coefficients of the line computed from data are called b_0 (the intercept) and b_1 (the slope). There are formulas for these:

$$b_1 = r\frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x}.$$

So, if we use the correlation $r = 0.7397$ we computed above, and determine the sample mean and sample standard deviations for `PPM` and `Price`, these formulas will give us the slope and intercept.

```
xbar = mean(~PPM, data=InkjetPrinters)
s.x = sd(~PPM, data=InkjetPrinters)
ybar = mean(~Price, data=InkjetPrinters)
s.y = sd(~Price, data=InkjetPrinters)
r = cor(Price ~ PPM, data=InkjetPrinters)

b1 = r * s.y / s.x;  b1
```

```
[1] 90.87807
```

```
b0 = ybar - b1 * xbar;  b0
```

```
[1] -94.22176
```

The more streamlined way to obtain slope and intercept is by using `lm()`.

```
myModel <- lm(Price ~ PPM, data=InkjetPrinters)
coef(myModel)
```

```
(Intercept)         PPM
  -94.22176    90.87807
```

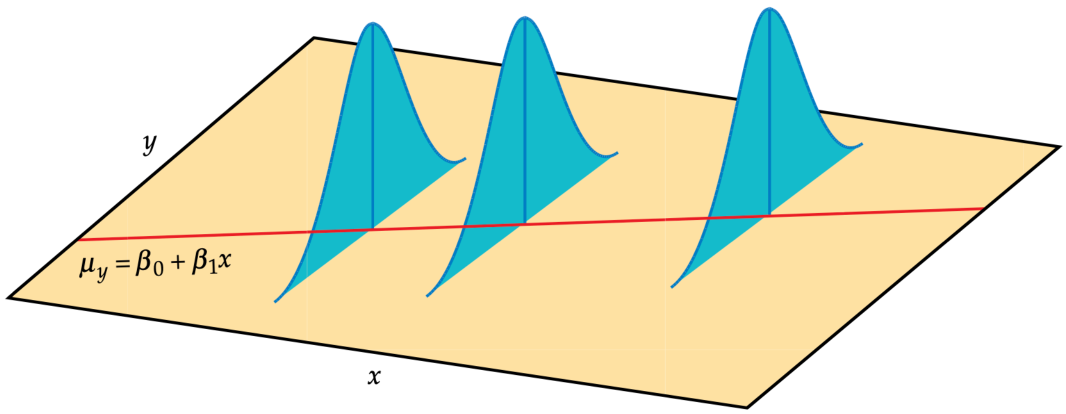
Building a confidence interval for β_1

The idea is that there is some true slope β_1 (also a true intercept β_0). These are population parameters, unknown to us. We have sampled data, and have used it to obtain estimates of slope and intercept, called b_1 and b_0. Slope is particularly important. It speaks to how quickly the response variable changes when the explanatory changes by 1 unit. It would be nice if we could produce an interval of values β_1 is likely to be, not merely a single estimate b_1.

Confidence interval construction is a type of statistical inference, and the conclusions we draw about population values require models that behave a certain way. To draw reliable conclusions about β_1—in this case, to develop a method for 95% confidence interval construction that is successful in enclosing β_1 95% of the time—requires assumptions we refer to as the **simple linear model assumptions**. (The word "simple" is used for cases like ours where there is just *one* predictor variable.) I will have more to say about the **simple linear model**, and about how to investigate whether it holds plausibly, in a future class session. For now, I'll say the model is usually summarized by the mathematical expression

$$Y = \beta_0 + \beta_1 x + \epsilon \qquad \text{where} \qquad \epsilon \sim \text{Norm}(0, \sigma),$$

and can be visualized as corresponding to this picture:



Assuming the model assumptions hold, `summary( lm( ...) )` can be used in tandem with `qt()` to produce a confidence interval for β_1 in the usual way:

$$\text{(point estimate)} \pm \text{(critical value)(standard error)}.$$

Our point estimate and standard error are $b_1 = 90.88$ and $\text{SE}_{b_1} = 19.49$, appearing in the output below:

```
summary(lm(Price ~ PPM, data=InkjetPrinters))
```

```

Call:
lm(formula = Price ~ PPM, data = InkjetPrinters)

Residuals:
   Min     1Q Median     3Q    Max
-79.38 -51.40  -3.49  43.85  87.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -94.22      56.40  -1.671 0.112086
PPM            90.88      19.49   4.663 0.000193 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.55 on 18 degrees of freedom
Multiple R-squared:  0.5471,    Adjusted R-squared:  0.522
F-statistic: 21.75 on 1 and 18 DF,  p-value: 0.0001934
```

The critical value is a t^* value, tailored to the desired level of confidence, and degrees of freedom $df = n - 2$. (Later, when we are using k predictor variables instead of the *one* that is assumed in the simple linear model, the degrees of freedom will be modified to $df = n - 1 - k$.)

For 95% confidence,

```
tstar = qt(0.975, df=18)     # because InkjetPrinters is a sample of size n=20
```

and our 95% CI for β_1 is

```
90.88 + c(-1,1) * tstar * 19.49
```

```
[1]  49.93303 131.82697
```

or $(49.93, 131.83)$. That is, we believe the true slope, with 95% confidence, lies in this interval of numbers.