POLS0008 Week 9 Seminar

Linear Regression

James Rice

UCL Department of Political Science

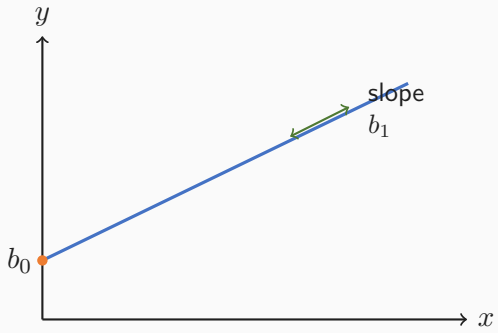
Thursday, March 19, 2026

The regression equation

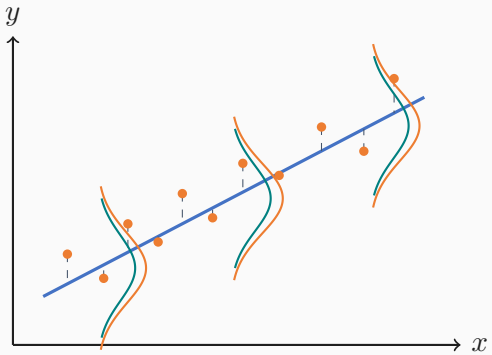
Straight-line model

$$\hat{y} = b_0 + b_1 x$$

- b_0: intercept / constant
- b_1: slope / gradient
- $\hat{y}$: predicted value of the dependent variable



Line of best fit and OLS



- OLS chooses the line that minimises squared residuals.
- The side curves are Gaussian densities for y at selected values of x, centred on the fitted line.
- Residuals: $e_i = y_i - \hat{y}_i$
- e_i: vertical distance from point to line

Interpretation

OLS gives the conditional mean of y for each x; the Gaussian sketches show the remaining spread around that mean.

Coefficients, significance, and prediction

Coefficients

- The **intercept** is the predicted value of y when $x = 0$.
- The **slope** tells us how much y changes for a one-unit increase in x.

Fit and inference

- Small p-value: evidence against a zero effect.
- Higher R^2: more variance explained.
- Avoid extrapolating far beyond the observed range.

OLS in standard form

$$\text{Goal: } \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\text{Where: } \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The basic OLS manipulation is to choose coefficients to minimise squared residuals.

How to read the regression output

What each column means

- **Estimate**: the coefficient itself.
- **Std. Error**: uncertainty around that estimate.
- **t value**: estimate divided by standard error.
- **p value**: evidence against the null that the coefficient is 0.

```
Call:
lm(formula = lab.per ~ Employment, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34938 -0.07958  0.02279  0.09018  0.31974

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7379784  0.0559585  13.188  < 2e-16 ***
Employment  -0.0050459  0.0007354  -6.861 1.64e-11 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.134 on 625 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.07005,   Adjusted R-squared:  0.06856
F-statistic: 47.08 on 1 and 625 DF,  p-value: 1.644e-11
```

Key point:

Write the coefficient interpretation/effect in your own words before discussing significance.

Regression interpretation: constant, slope, and prediction

Constant and slope

$$\hat{y} = b_0 + b_1 x$$

- b_0: predicted Labour share when the explanatory variable equals 0.
- b_1: expected change in Labour share for a one-unit increase in the explanatory variable.

Example

Suppose the estimated equation is

$$\widehat{\text{Labour.per}} = 81.48 - 1.01(\text{Employment})$$

If Employment $= 80$:

$$\widehat{\text{Labour.per}} = 81.48 - 1.01(80)$$

$$= 81.48 - 80.80 = 0.68$$

So Labour vote share is about 0.68%.

Example interpretation:

"When Employment is 80, the predicted Labour vote share is 0.68%, found by substituting 80 into the regression equation for x."

Model fit: what does R^2 add?

Reading R^2

- R^2 is the proportion of variation in the dependent variable explained by the model.
- Bigger R^2 means the line tracks the data more closely.
- In simple regression, it is a convenient way to compare one-predictor models.

What students should not overclaim

- A statistically significant coefficient can still have a low R^2.
- A better-fitting model is not automatically causal.
- Always separate *association*, *prediction*, and *causation*.

Seminar tasks

1. Choose a variable plausibly related to Labour vote share. Produce two histograms and a scatterplot; comment on distribution and relationship.
2. Run a bivariate correlation between `Labour.per` and the chosen variable; justify the method and interpret the result.
3. Fit a simple regression model; interpret the coefficient, t-statistic, p-value, and confidence interval.
4. Interpret the constant and predict Labour vote share at a chosen value of the explanatory variable; say whether prediction at zero is advisable.
5. Fit a second simple regression with a different explanatory variable and decide which model is the better predictor of Labour vote share.

Q1–Q2: exploration and correlation

Q1: what the plots suggest

- `lab.per` is described in the script as approximately normal.
- `Employment` is slightly skewed.
- The scatterplot suggests a roughly linear negative association.

Q2: correlation answer

$$r \approx -0.2646$$

- Moderate negative relationship.
- Higher Employment is associated with lower Labour vote share.
- Pearson is appropriate because the pattern looks approximately linear.

Q3–Q4: regression interpretation and prediction

Q3: simple regression

```
lab.per ~ Employment
```

- The Employment coefficient is negative.
- Interpretation: a one-unit increase in Employment predicts a slightly lower Labour share: $\beta = -0.005, p < 0.001$.

Q4: constant and prediction

$$\widehat{\texttt{lab.per}} \mid \texttt{Employment} = 80 = 0.3343$$

- Predicted Labour vote share is about **33.43%**.
- The constant is the predicted Labour share when Employment = 0.

Q5: which simple model is better?

Model 1

`lab.per ~ Employment`

$$R^2 = 0.0701$$

Explains about 7.01% of the variation.

Model 2

`lab.per ~ British`

$$R^2 = 0.0202$$

Explains only about 2.02% of the variation.

Interpretation:

The Employment model is the better predictor because it explains more variance in Labour vote share.

Key takeaways

- Regression does more than correlation: it gives an equation, a predicted value, and a way to compare models.
- For students, the key interpretive sequence is sign $\rightarrow$ unit change $\rightarrow$ significance $\rightarrow$ confidence interval $\rightarrow$ model fit.
- In the uploaded solutions, Employment is the best single predictor of Labour vote share.
- The most important habits are: inspect plots first, justify the method, interpret the coefficient carefully, and avoid extrapolation.

Questions / discussion