1 Regression Through the Origin

- If it's **known** that $\boldsymbol{\beta_0 = 0}$ (i.e. that $\mathrm{E}(Y) = 0$ when $X = 0$), we may want to enforce this in the fitted model. The appropriate model to be fitted is

$$Y_i = \beta_1 X_i + \epsilon_i \tag{1}$$

  where the ϵ_i's are iid $N(0, \sigma^2)$.

1.1 Least Squares Estimation of $\boldsymbol{\beta_1}$

- The **least squares estimator $\boldsymbol{b_1}$** of β_1 is the minimizer (with respect to β_1) of

$$Q(\beta_1) = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 .$$

  To find b_1, solve:

$$\frac{dQ}{d\beta_1} = 0 ,$$

  i.e.

$$-2 \sum_{i=1}^{n} X_i (Y_i - \beta_1 X_i) = 0 \tag{2}$$

  for β_1. It's easy to see that the solution to (2) is

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} ,$$

  and b_1 is the least squares estimator of β_1.

- The **fitted regression model** is now

$$\hat{Y} = b_1 X .$$

- The **fitted values** are obtained by plugging the observed X_i's into the regression equation:

$$\hat{Y}_i = b_1 X_i$$

- The **residuals** are now

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 X_i$$

  and **they no longer sum to zero**.

- The **error sum of squares**, **SSE**, is

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

but now the **degrees of freedom** for the SSE is

$$\text{df for SSE} = n - 1$$

since now there's only one parameter in the model. (Recall that the df for SSE always equals n minus the number of parameters).

- The **mean squared error** is now

$$\text{MSE} = \frac{\text{SSE}}{n-1}$$

and MSE is still an **unbiased** estimator of σ^2, i.e.

$$E(\text{MSE}) = \sigma^2.$$

- We can perform a test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, calculate a confidence interval for β_1, a confidence interval for $\text{E}(Y)$, and a prediction interval for $Y_{h(\text{new})}$ under the model (1). See the textbook for details.

1.2 Some Cautions About Performing Regression Through the Origin

- When the regression line is forced to go through the origin:

  ▷ The **residuals no longer sum to zero**.

  ▷ The constraint $\sum_i e_i X_i = 0$ **still holds**.

  ▷ The error sum of squares **SSE may exceed** the total sum of squares **SSTO**, in which case $R^2 = 1 - \frac{\text{SSE}}{\text{SSTO}} < 0$, and so R^2 has no clear interpretation.

For these reasons, it's often advisable to fit a model that includes an intercept, even when it's known that $\beta_0 = 0$. In this case, the estimate b_0 will likely be approximately zero.