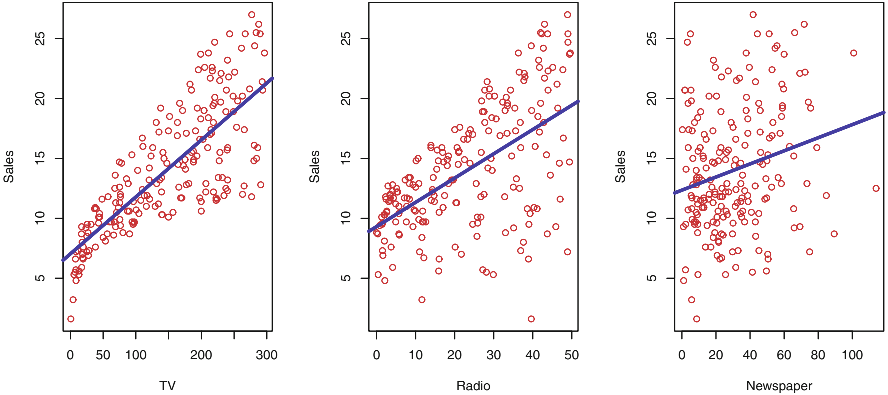
Simple Linear Regression

Rosalba Radice

Analytics Methods for Business

Motivating Example



The Advertising data set. The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively.

Simple linear regression using a single predictor X.

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

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

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The hat symbol denotes an estimated value.

Estimation of the parameters by least squares

Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith residual.
We define the residual sum of squares (RSS) as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2,$$

or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

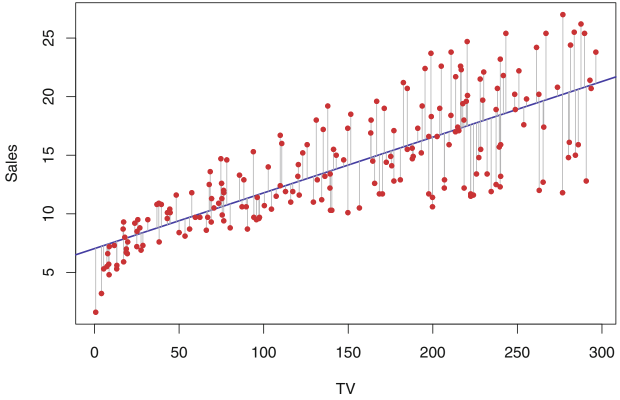
Estimation of the parameters by least squares - continued

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned} \tag{1}$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i$ are the sample means. In other words, (1) defines the least squares coefficient estimates for simple linear regression.

The least squares fit on advertising data



For the `Advertising` data, the least squares fit for the regression of `sales` onto `TV` is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Assessing the accuracy of the coefficient estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

  where $\sigma^2 = \mathrm{Var}(\epsilon)$.
- These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_1 \pm 2 \times \mathrm{SE}(\hat{\beta}_1).$$

Confidence intervals - continued

- That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2\mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2\mathrm{SE}(\hat{\beta}_1)\right] \tag{2}$$

  will contain the true value of β_1.
- Approximately here is for several reasons.
  - The confidence interval relies on the assumption that the errors are Gaussian.
  - Also, the factor of 2 in front of the $\mathrm{SE}(\hat{\beta}_1)$ term will vary slightly depending on the number of observations n in the linear regression.
  - To be precise, rather than the number 2, the confidence interval should contain the 97.5% quantile of a t-distribution with $n-2$ degrees of freedom.
- For the advertising data, the 95% confidence interval for β_1 is

$$[0.042; 0.053]$$

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

  $$H_0\text{: There is no linear relationship between } X \text{ and } Y$$

  versus the alternative hypothesis

  $$H_a : \text{There is some linear relationship between } X \text{ and } Y.$$

- Mathematically, this corresponds to testing

  $$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0,$$

  since if $\beta_1 = 0$ then the simple linear regression reduces to $Y = \beta_0 + \epsilon$, and X is not linearly associated with Y.

Hypothesis testing - continued

- To test the null hypothesis, we compute a t-statistic, given by

$$\mathrm{t} = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

- This will have a t-distribution with $n-2$ degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|\mathrm{t}|$ or larger. We call this probability the p-value.

Results for the advertising data

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

For the `Advertising` data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the `TV` advertising budget is associated with an increase in `sales` by around 50 units (Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars).

Assessing the overall accuracy of the model

- We compute the Residual Standard Error

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-2}\mathrm{RSS}},$$

  where RSS is the residual sum of squares.
- R-squared or fraction of variance explained is

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}},$$

  where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ is the total sum of squares.
- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Advertising data results

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

For the `Advertising` data, more information about the least squares model for the regression of number of units sold on TV advertising budget.