Linear regression
Class 26, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to use the method of least squares to fit a line to bivariate data.
2. Be able to give a formula for the total squared error when fitting any type of curve to data.
3. Be able to say the words homoscedasticity and heteroscedasticity.

2 Introduction

Suppose we have collected bivariate data (x_i, y_i), $i = 1, \ldots, n$. The goal of linear regression is to model the relationship between x and y by finding a function $y = f(x)$ that is a close fit to the data. The modeling assumptions we will use are that x_i is **not** random and that y_i is a function of x_i plus some random noise. With these assumptions x is called the independent or predictor variable and y is called the dependent or response variable.

Here is a series of examples showing the results of linear regression. We will discuss the details of how to do linear regression in the next section.

Example 1. The cost of a first class stamp in cents over time is given in the following list.

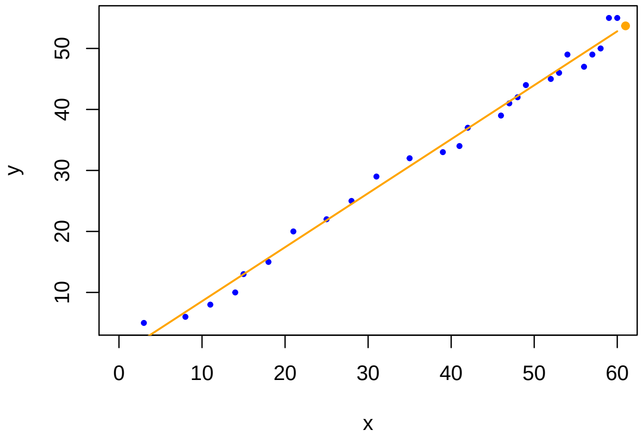
0.05 (1963)	0.06 (1968)	0.08 (1971)	0.10 (1974)	0.13 (1975)	0.15 (1978)	0.20 (1981)
0.22 (1985)	0.25 (1988)	0.29 (1991)	0.32 (1995)	0.33 (1999)	0.34 (2001)	0.37 (2002)
0.39 (2006)	0.41 (2007)	0.42 (2008)	0.44 (2009)	0.45 (2012)	0.46 (2013)	0.49 (2015)
0.49 (2017)	0.50 (2018)	0.55 (2019)				

Using the R function `lm` we found the 'least squares fit' for a line to this data is

$$y = -0.21390 + 0.88203x,$$

where x is the number of years since 1960 and y is in cents.

Using this result we 'predict' that in 2021 ($x = 61$) the cost of a stamp will be 53.6 cents (since $-0.21390 + 0.88203 \cdot 61 = 53.6$).



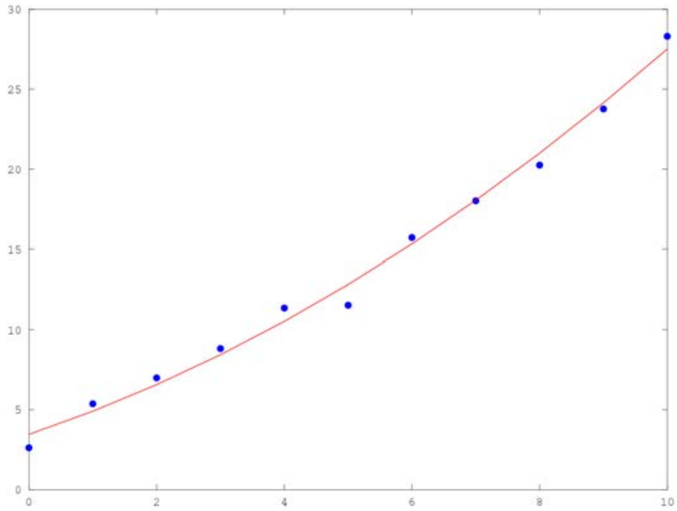
Stamp cost (cents) vs. time (years since 1960). Orange dot is predicted cost in 2021.

Note that none of the data points actually lie on the line. Rather this line has the 'best fit' with respect to all the data, with a small error for each data point.

(Note, the actual cost of a stamp dropped in January 2021 was 55 cents. See https://en.wikipedia.org/wiki/History_of_United_States_postage_rates)

Example 2. Suppose we have n pairs of fathers and adult sons. Let x_i and y_i be the heights of the i^{th} father and son, respectively. The least squares line for this data could be used to predict the adult height of a young boy from that of his father.

Example 3. We are not limited to best fit lines. For all positive d, the method of least squares may be used to find a polynomial of degree d with the 'best fit' to the data. Here's a figure showing the least squares fit of a parabola ($d = 2$).



Fitting a parabola, $ax^2 + bx + c$, to data

Example 4. In fact, we can use linear regression to fit many other types of curves to bivariate data.

3 Fitting a line using least squares

Suppose we have data (x_i, y_i) as above. Our first goal is to find the line

$$y = ax + b$$

that 'best fits' the data. Our model says that each y_i is predicted by x_i up to some error ϵ_i:

$$y_i = ax_i + b + \epsilon_i.$$

So

$$\epsilon_i = y_i - ax_i - b.$$

The method of least squares finds the values $\hat{a}$ and $\hat{b}$ of a and b that minimize the sum of the squared errors:

$$S(a,b) = \sum \epsilon_i^2 = \sum_i (y_i - ax_i - b)^2.$$

Using calculus or linear algebra (details in the appendix), we find

$$\hat{a} = \frac{s_{xy}}{s_{xx}} \qquad \hat{b} = \bar{y} - \hat{a}\,\bar{x} \tag{1}$$

where

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i, \quad s_{xx} = \frac{1}{(n-1)}\sum (x_i - \bar{x})^2, \quad s_{xy} = \frac{1}{(n-1)}\sum (x_i - \bar{x})(y_i - \bar{y}).$$

Here $\bar{x}$ is the sample mean of x, $\bar{y}$ is the sample mean of y, s_{xx} is the sample variance of x, and s_{xy} is the sample covariance of x and y.

Example 5. Use least squares to fit a line to the following data: (0,1), (2,1), (3,4).

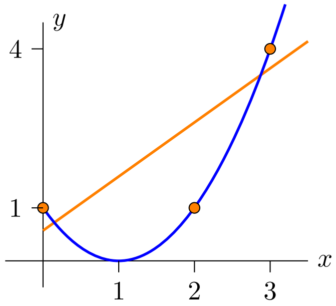
Solution: In our case, $(x_1, y_1) = (0,1)$, $(x_2, y_2) = (2,1)$ and $(x_3, y_3) = (3,4)$. So

$$\bar{x} = \frac{5}{3}, \quad \bar{y} = 2, \quad s_{xx} = \frac{7}{3}, \quad s_{xy} = 2$$

Using the above formulas we get

$$\hat{a} = \frac{6}{7}, \quad \hat{b} = \frac{4}{7}.$$

So the least squares line has equation $y = \frac{6}{7}x + \frac{4}{7}$. This is shown as the orange line in the following figure. We will discuss the blue parabola soon.



Least squares fit of a line (orange) and a parabola (blue)

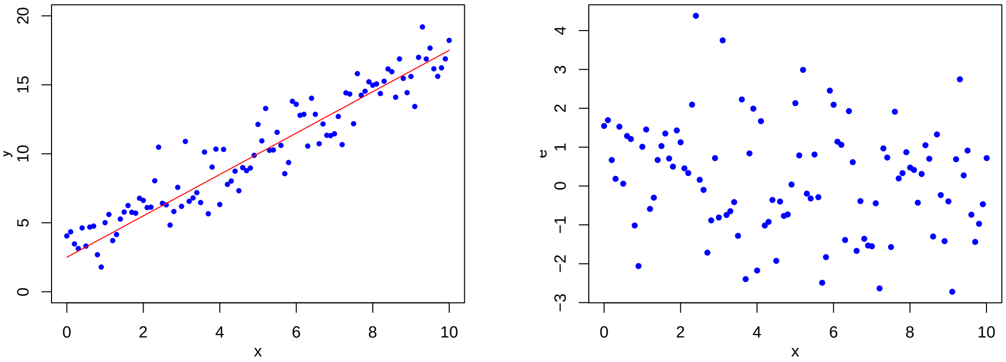
Simple linear regression: It's a little confusing, but the word linear in 'linear regression' does not refer to fitting a line. We will explain its meaning below. However, the most common curve to fit is a line. When we fit a line to bivariate data it is called simple linear regression.

3.1 Residuals

For a line the model is

$$y_i = \hat{a}x + \hat{b} + \epsilon_i.$$

We think of $\hat{a}x_i + \hat{b}$ as predicting or explaining y_i. The left-over term ϵ_i is called the residual, which we think of as random noise or measurement error. A useful visual check of the linear regression model is to plot the residuals. The data points should hover near the regression line. The residuals should look about the same across the range of x.

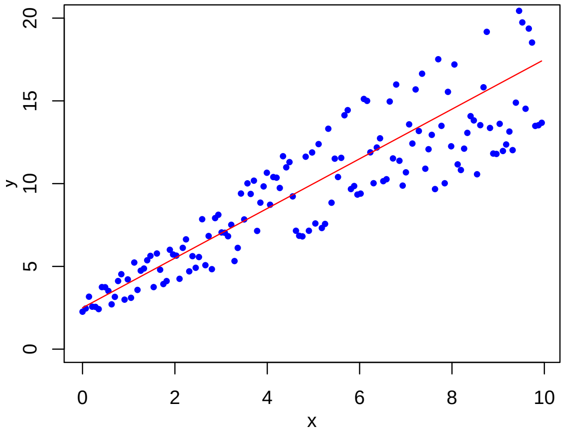


Data with regression line (left) and residuals (right). Note the homoscedasticity.

3.2 Homoscedasticity

An important assumption of the linear regression model is that the residuals ϵ_i have the same variance for all i. This is called homoscedasticity. You can see this is the case for both figures above. The data hovers in the band of fixed width around the regression line and at every x the residuals have about the same vertical spread.

Below is a figure showing heteroscedastic data. The vertical spread of the data increases as x increases. Before using least squares on this data we would need to transform the data to be homoscedastic.



Heteroscedastic Data

4 Linear regression for fitting polynomials

When we fit a line to data it is called simple linear regression. We can also use linear regression to fit polynomials to data. The use of the word linear in both cases may seem confusing. This is because the word 'linear' in linear regression does not refer to fitting a line. Rather it refers to the linear algebraic equations for the unknown parameters.

Example 6. Take the same data as in Example 5 and use least squares to find the best fitting parabola to the data.

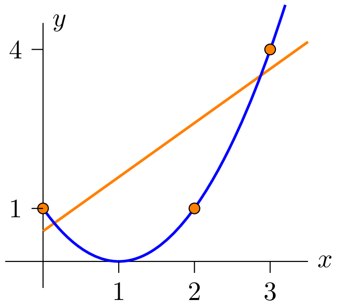
Solution: A parabola has the formula $y = ax^2 + bx + c$. The squared error is

$$S(a, b, c) = \sum (y_i - (ax_i^2 + bx_i + c))^2.$$

After substituting the given values for each x_i and y_i, we can use calculus to find the triple (a, b, c) that minimizes S. With this data, we find that the least squares parabola has equation

$$y = x^2 - 2x + 1.$$

Note that for 3 points the quadratic fit is perfect.



Least squares fit of a line (orange) and a parabola (blue)

Example 7. The pairs (x_i, y_i) may give the age and vocabulary size of a n children. Since we expect that young children acquire new words at an accelerating pace, we might guess that a higher order polynomial would best fit the data.

Example 8. (Transforming the data) Sometimes it is necessary to transform the data before using linear regression. For example, let's suppose the relationship is exponential, i.e. $y = ce^{ax}$. Then

$$\ln(y) = ax + \ln(c).$$

So we can use simple linear regression on the data $(x_i, \ln(y_i))$ to obtain a model

$$\ln(y) = \hat{a}x + \hat{b}$$

and then exponentiate to obtain the exponential model

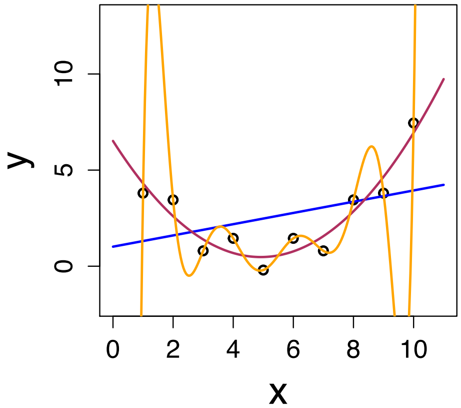
$$y = e^{\hat{b}}e^{\hat{a}x}.$$

4.1 Overfitting

You can always achieve a better fit by using a higher order polynomial. For instance, given 6 data points (with distinct x_i) one can always find a fifth order polynomial that goes through all of them. This can result in what's called overfitting. That is, fitting the noise as well as the true relationship between x and y. An overfit model will fit the original data better but perform less well on predicting y for new values of x. Indeed, a primary challenge of statistical modeling is balancing model fit against model complexity.

Example 9. In the plot below, we fit polynomials of degree 1, 2, and 9 to bivariate data consisting of 10 data points. The degree 2 model (maroon) gives a significantly better fit than the degree 1 model (blue). The degree 10 model (orange) gives fits the data exactly, but at a glance we would guess it is overfit. That is, we don't expect it to do a good job fitting the next data point we see.

In fact, we generated this data using a quadratic model, so the degree 2 model will tend to perform best fitting new data points.



4.2 R function `lm`

As you would expect we don't actually do linear regression by hand. Computationally, linear regression reduces to solving simultaneous equations, i.e. to matrix calculations. The R function `lm` can be used to fit any order polynomial to data. (`lm` stands for linear model). We will explore this in the next studio class. In fact `lm` can fit many types of functions besides polynomials, as you can explore using R help or google.

5 Multiple linear regression

Data is not always bivariate. It can be trivariate or even of some higher dimension. Suppose we have data in the form of tuples

$$(y_i,\, x_{1,i},\, x_{2,i},\, \dots\, x_{m,i})$$

We can analyze this in a manner very similar to linear regression on bivariate data. That is, we can use least squares to fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m.$$

Here each x_j is a predictor variable and y is the response variable. For example, we might be interested in how a fish population varies with measured levels of several pollutants, or we might want to predict the adult height of a son based on the height of the mother and the height of the father.

We don't have time in 18.05 to study multiple linear regression, but we wanted you to see the name.

6 Least squares as a statistical model

The linear regression model for fitting a line says that the value y_i in the pair (x_i, y_i) is drawn from a random variable

$$Y_i = ax_i + b + \varepsilon_i$$

where the 'error' terms ε_i are independent random variables with mean 0 and standard deviation σ. The standard assumption is that the ε_i are i.i.d. with distribution $N(0, \sigma^2)$. So, the mean of Y_i is given by:

$$E[Y_i] = ax_i + b + E[\varepsilon_i] = ax_i + b.$$

From this perspective, the least squares method chooses the values of a and b which minimize the sample variance about the line.

In fact, under the assumption that $\varepsilon_i \sim N(0, \sigma^2)$, the least square estimate $(\hat{a}, \hat{b})$ coincides with the maximum likelihood estimate for the parameters (a, b); that is, among all possible coefficients, $(\hat{a}, \hat{b})$ are the ones that make the observed data most probable.

7 Regression to the mean

The reason for the term 'regression' is that the predicted response variable y will tend to be 'closer' to (i.e., regress to) its mean than the predictor variable x is to its mean. Here closer is in quotes because we have to control for the scale (i.e. standard deviation) of each variable. The way we control for scale is to first standardize each variable.

$$u_i = \frac{x_i - \bar{x}}{\sqrt{s_{xx}}}, \quad v_i = \frac{y_i - \bar{y}}{\sqrt{s_{yy}}}.$$

Standardization changes the mean to 0 and variance to 1:

$$\bar{u} = \bar{v} = 0, \quad s_{uu} = s_{vv} = 1.$$

The algebraic properties of covariance show

$$s_{uv} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \rho,$$

the correlation coefficient. Thus the least squares fit to $v = au + b$ has

$$\hat{a} = \frac{s_{uv}}{s_{uu}} = \rho \quad \text{and} \quad \hat{b} = \bar{v} - \hat{a}\bar{u} = 0.$$

So the least squares line is $v = \rho u$. Since ρ is the correlation coefficient, it is between -1 and 1. Let's assume it is positive and less than 1 (i.e., x and y are positively but not perfectly correlated). Then the formula $v = \rho u$ means that if u is positive then the predicted value of v is less than u. That is, v is closer to 0 than u. Equivalently,

$$\frac{y - \bar{y}}{\sqrt{s_{yy}}} < \frac{x - \bar{x}}{\sqrt{s_{xx}}}$$

i.e., y regresses to $\bar{y}$. Notice how the standardization takes care of controlling the scale.

Consider the extreme case of 0 correlation between x and y. Then, no matter what the x value, the predicted value of y is always $\bar{y}$. That is, y has regressed all the way to its mean.

Note also that the regression line always goes through the point $(\bar{x}, \bar{y})$.

Example 10. Regression to the mean is important in longitudinal studies. Rice (*Mathematical Statistics and Data Analysis*) gives the following example. Suppose children are given an IQ test at age 4 and another at age 5 we expect the results will be positively correlated. The above analysis says that, on average, those kids who do poorly on the first test will tend to show improvement (i.e. regress to the mean) on the second test. Thus, a useless intervention might be misinterpreted as useful since it seems to improve scores.

Example 11. Another example with practical consequences is reward and punishment. Imagine a school where high performance on an exam is rewarded and low performance is punished. Regression to the mean tells us that (on average) the high performing students will do slightly worse on the next exam and the low performing students will do slightly better. An unsophisticated view of the data will make it seem that punishment improved performance and reward actually hurt performance. There are real consequences if those in authority act on this idea.

8 Appendix

We collect in this appendix a few things you might find interesting. You will not be asked to know these things for exams.

8.1 Proof of the formula for least square fit of a line

The most straightforward proof is to use calculus. The sum of the squared errors is

$$S(b,a) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Taking partial derivatives (and remembering that x_i and y_i are the data, hence constant)

$$\begin{aligned} \frac{\partial S}{\partial b} &= \sum_{i=1}^{n} -2(y_i - ax_i - b) = 0 \\ \frac{\partial S}{\partial a} &= \sum_{i=1}^{n} -2x_i(y_i - ax_i - b) = 0 \end{aligned}$$

Summing this up we get two linear equations in the unknowns b and a:

$$\begin{aligned} \left(\sum x_i\right) a + nb &= \sum y_i \\ \left(\sum x_i^2\right) a + \left(\sum x_i\right) b &= \sum x_i y_i \end{aligned}$$

Solving for a and b gives the formulas in Equation (1).

A sneakier approach which avoids calculus is to standardize the data, find the best fit line, and then unstandardize. We omit the details.

For a slew of applications across disciplines see:
https://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression

8.2 Measuring the fit

Once one computes the regression coefficients, it is important to check how well the regression model fits the data (i.e., how closely the best fit line tracks the data). A common but crude 'goodness of fit' measure is the coefficient of determination, denoted R^2. We'll need some notation to define it. The total sum of squares is given by:

$$\text{TSS} = \sum (y_i - \bar{y})^2.$$

The residual sum of squares is given by the sum of the squares of the residuals. When fitting a line, this is:

$$\text{RSS} = \sum (y_i - \hat{a}\, x_i - \hat{b})^2.$$

The RSS is the "unexplained" portion of the total sum of squares, i.e. unexplained by the regression equation. The difference TSS − RSS is the "explained" portion of the total sum of squares. The coefficient of determination R^2 is the ratio of the "explained" portion to the total sum of squares:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}.$$

In other words, R^2 measures the proportion of the variability of the data that is accounted for by the regression model. A value close to 1 indicates a good fit, while a value close to 0 indicates a poor fit. In the case of simple linear regression, R^2 is simply the square of the correlation coefficient between the observed values y_i and the predicted values $ax_i + b$.

Example 12. In the overfitting example (9), the values of R^2 are:

degree	R^2
1	0.3968
2	0.9455
9	1.0000

Notice the goodness of fit measure increases as n increases. The fit is better, but the model also becomes more complex, since it takes more coefficients to describe higher order polynomials.