Introduction to the Linear Model

Charlie Gibbons
University of California, Berkeley
Economics 140

Summer 2011

Outline

Decomposition

We can write

$$y_i = \mathbb{E}[y_i \mid x_i] + \underbrace{(y_i - \mathbb{E}[y_i \mid x_i])}_{\equiv \epsilon_i}.$$

Indicator variables

Let x_i be a *indicator variable*, a variable that is 1 if i has a particular quality or 0 otherwise.

We have

$$\begin{aligned}\mathbb{E}[y_i \mid x_i] &= x_i \times \mathbb{E}[y_i \mid x_i = 1] + (1 - x_i) \times \mathbb{E}[y_i \mid x_i = 0] \\ &= \mathbb{E}[y_i \mid x_i = 0] + \big[\mathbb{E}[y_i \mid x_i = 1] - \mathbb{E}[y_i \mid x_i = 0]\big] x_i \\ &= \beta_0 + \beta_1 x_i\end{aligned}$$

or

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

This is a linear model that requires *no assumptions*.

CAPM

Economic theory gives us the *capital asset pricing model (CAPM)*:

$$r_t^i - r_t^f = \alpha + \beta \left(r_t^m - r_t^f \right) + \epsilon_t^i.$$

- r_t^i is the return to asset i at time t,
- r_t^f is the return to a risk-free asset at time t, and
- r_t^m is the return to a market basket of assets at time t.

Interpretation

β measures the riskiness of an asset.

α measures whether the asset can beat the market without increased risk.

Question: Theory dictates that $\alpha = 0$. Should we include it in our model?

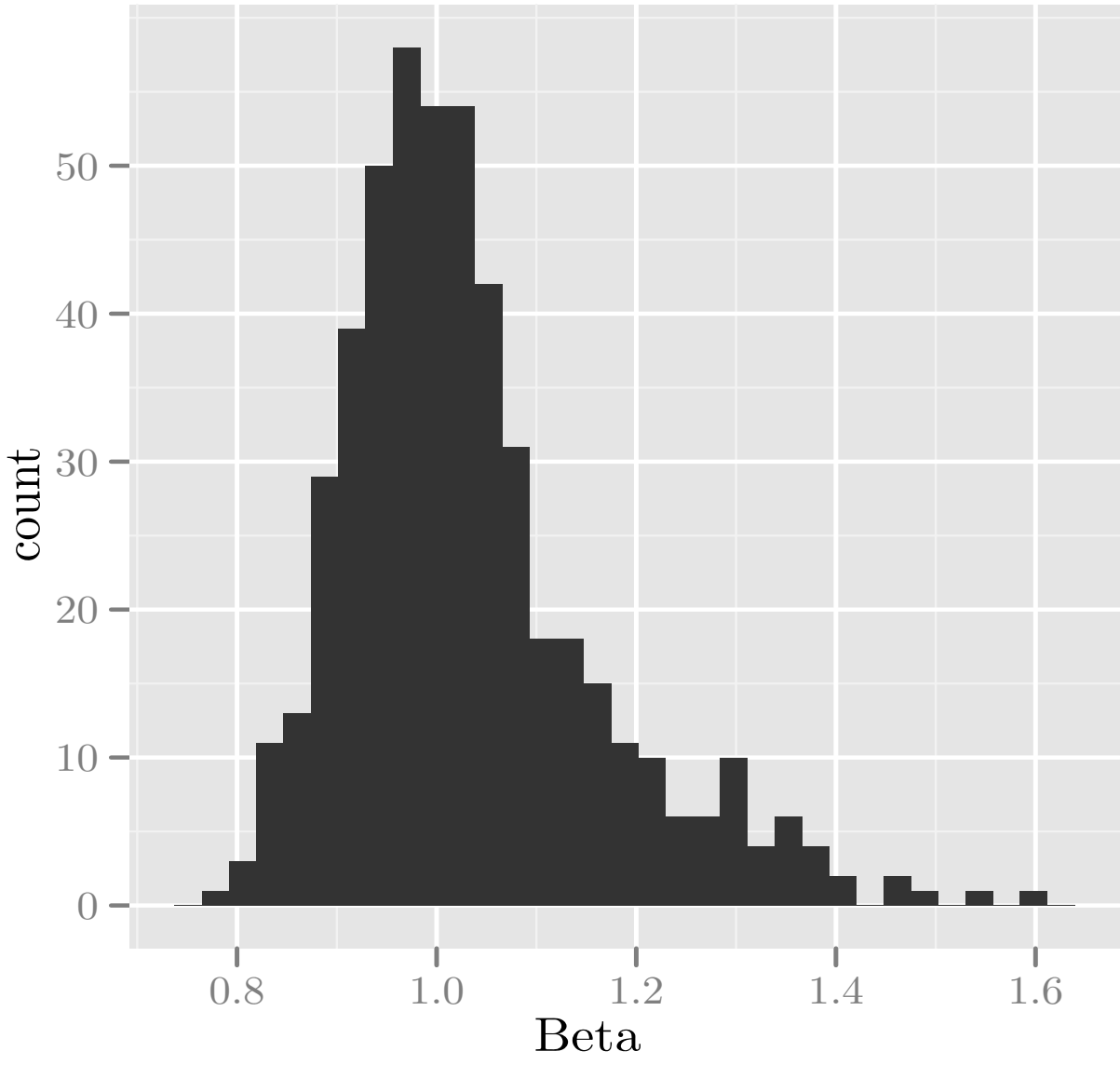


Figure: Betas for all current S & P firms, calculated using 2007–present returns

Covariance

Covariance

Covariance is a measure of linear association between two variables:

$$\begin{aligned}\mathrm{Cov}(x, y) \equiv \sigma_{XY} &= \mathbb{E}\left[\left(X - \mathbb{E}_X(X)\right)\left(Y - \mathbb{E}_Y(Y)\right)\right] \\ &= \mathbb{E}(XY) - \mu_X \mu_Y .\end{aligned}$$

Correlation

Correlation

Correlation normalizes the covariance by dividing by the standard deviations of x and y:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

This makes correlation unit-free and bounds it between -1 and 1. Higher absolute values imply a stronger linear relationship.

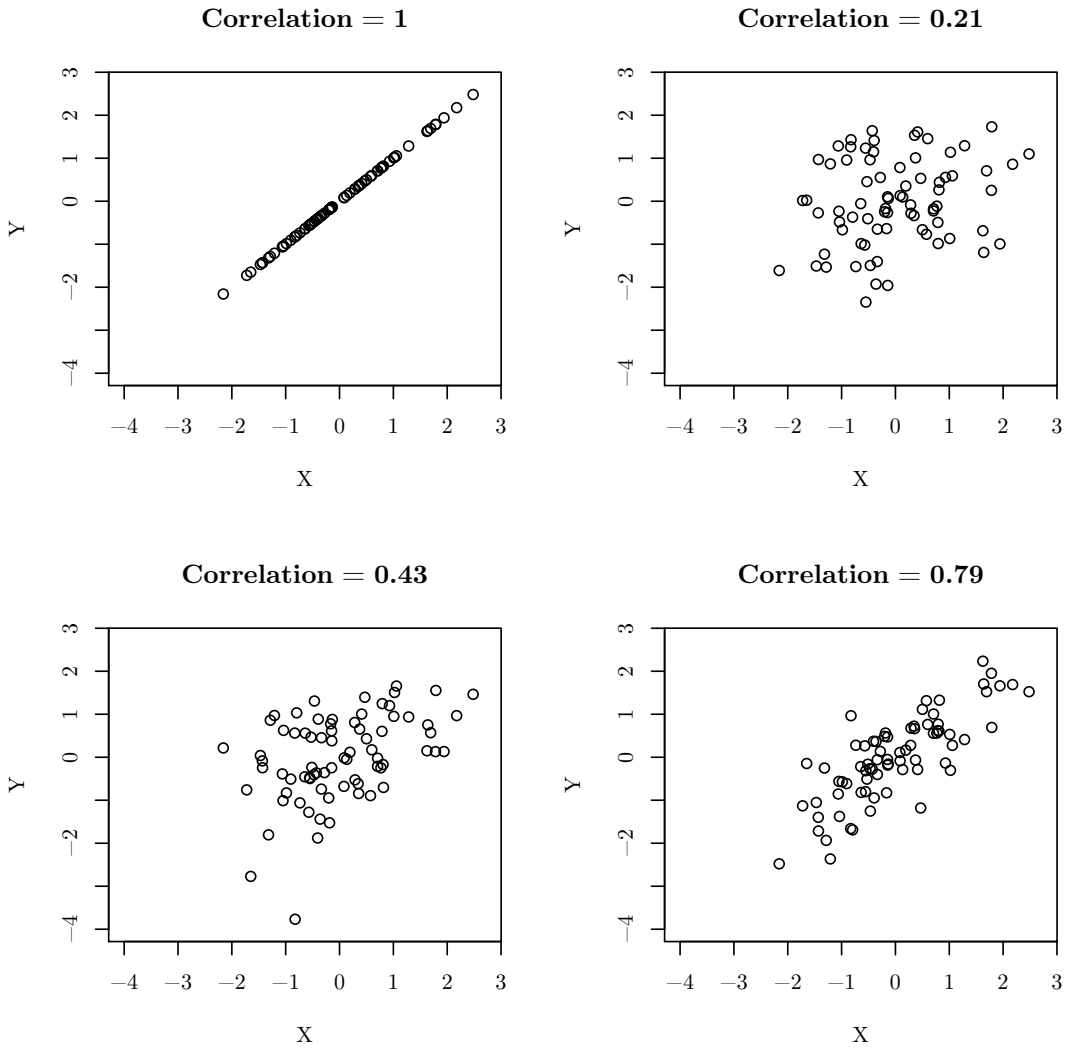
Correlation figures intro

The following pictures show scatterplots of x, $y1$, $y2$, and $y3$ against x. Each illustrates a different degree of correlation.

All variables have been *standardized*—they all have mean 0 and standard deviation of 1 (we'll come back to this topic later in the course).

Check yourself: Given the correlations on the next slide, what are the covariances?

Correlation figures



Linearity assumption

Regression assumes the true relationship

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

and actually estimates

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i.$$

The Greek letter variables are not observed and the two β parameters are to be estimated. β_0 and $\hat{\beta}_0$ are the intercept parameters and β_1 and $\hat{\beta}_1$ are the slope or marginal effects parameters.

Note: β_0 and β_1 are **population parameters** and thus are **not random**.

Exogeneity assumption

Additionally, assume that

$$\mathbb{E}[\epsilon_i \mid x_i] = 0.$$

This is called the *exogeneity* assumption.

Now see that

$$\mathbb{E}[\epsilon_i] = \mathbb{E}_X \left[\mathbb{E}[\epsilon_i \mid x_i]\right] = 0$$

and

$$\mathbb{E}[\epsilon_i x_i] = \mathbb{E}_X \left[\mathbb{E}[\epsilon_i x_i \mid x_i]\right] = \mathbb{E}_X \left[x_i \mathbb{E}[\epsilon_i \mid x_i]\right] = 0.$$

Hence,

$$\mathrm{Cov}(x_i, \epsilon_i) = \mathbb{E}[\epsilon_i x_i] - \mathbb{E}[\epsilon_i]\mathbb{E}[x_i] = 0.$$

The errors are uncorrelated with the predictors.

Residuals

We don't observe the errors, so we need to consider their sample analogues.

Residuals

The residual is the difference between the observed value of y and that predicted using the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

Let's use the residuals and our predictors to create a sample analogue of this result.

Sample analogues

This gives

$$\mathbb{E}[\epsilon_i] \approx \frac{1}{N}\sum_i \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = 0 \text{ and}$$

$$\mathbb{E}[x_i \epsilon_i] \approx \frac{1}{N}\sum_i x_i \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = 0.$$

These are called the *moment conditions*.

Note: We *assume* that the errors are uncorrelated with x, but *force* the residuals to be uncorrelated with x.

First moment condition

Solving the first moment condition for $\hat{\beta}_0$ gives

$$\frac{1}{N}\sum_i y_i - \hat{\beta}_1 \frac{1}{N}\sum_i x_i = \hat{\beta}_0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Second moment condition

See that the second moment condition (multiplied by N) gives

$$\sum_i x_i y_i - \hat{\beta}_0 \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 = 0.$$

Solving the equations

Plugging the result for $\hat{\beta}_0$ into this equation gives:

$$\sum_i x_i y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 = 0$$

$$\sum_i x_i y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) N\bar{x} - \hat{\beta}_1 \sum_i x_i^2 = 0$$

$$\sum_i x_i y_i - N\bar{y}\bar{x} - N\hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \sum_i x_i^2 = 0$$

$$\sum_i (x_i y_i - \bar{y}\bar{x}) - \hat{\beta}_1 \sum_i \left(x_i^2 - \bar{x}^2\right) = 0$$

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_i (x_i - \bar{x})^2 = 0$$

$$\widehat{\mathrm{Cov}}(x_i, y_i) - \hat{\beta}_1 \widehat{\mathrm{Var}}(x_i) = 0$$

Results

We see that

$$\hat{\beta}_1 = \frac{\widehat{\mathrm{Cov}}(x_i, y_i)}{\widehat{\mathrm{Var}}(x_i)} = \hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X};$$

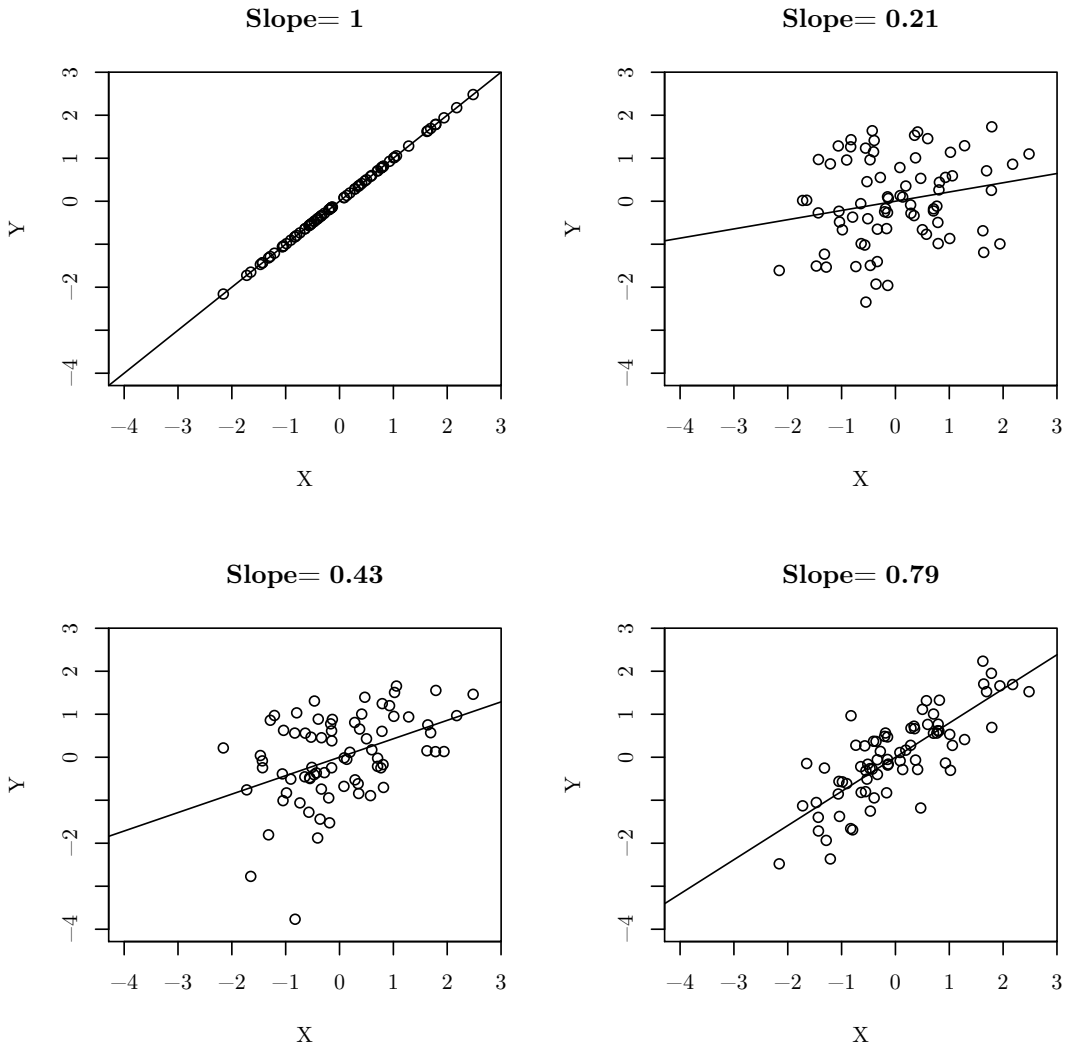
the coefficient on x is the sample correlation between y and x multiplied by the ratio of their standard deviations.

Also note that the intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

mainly serves to put the regression line through the point of means (*i.e.*, the residual at $(\bar{x}, \bar{y})$ is 0). The intercept is difficult to interpret accurately in any other way.

Correlation figures and the linear model



The minimization problem

Linear regression is often called *ordinary least squares* (OLS). It has this name because regression can be solved by minimizing the sum of the squares of the residuals.

OLS solves

$$\min_{b_0, b_1} \sum_i (y_i - (b_0 + b_1 x_i))^2.$$

The *sum of squared residuals* (or *errors*) (SSR/SSE) is simply the sum of squared residuals at the estimated parameter values:

$$\sum_i \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2.$$

Solving a minimization problem

Let's solve this problem. Recall how to solve a minimization problem:

1. Take the derivative of the objective function with respect to each parameter that you are optimizing over (*e.g.*, b_0 and b_1).
2. Set each of those derivatives to 0 and solve.

The derivatives are called the *first order conditions* or *normal equations*.

First normal equation

The first normal equation is

$$\frac{\partial}{\partial b_0}\left[\sum_i \left(y_i - (b_0 + b_1 x_i)\right)^2\right] = 0$$

$$\implies -2\sum_i \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = 0$$

$$\sum_i y_i - \hat{\beta}_1 \sum_i x_i = N\hat{\beta}_0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Second normal equation

The second normal equation is

$$\frac{\partial}{\partial b_1}\left[\sum_i \left(y_i - (b_0 + b_1 x_i)\right)^2\right] = 0$$

$$\implies -2\sum_i x_i \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = 0$$

$$\sum_i x_i y_i - \hat{\beta}_0 \sum_i x_i - \hat{\beta}_1 \sum_i x_i^2 = 0$$

Duality

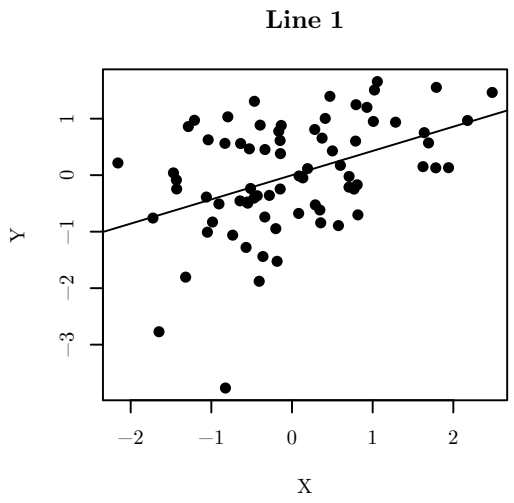
We see that we get the same answer using the moment conditions or by minimizing the squared residuals.

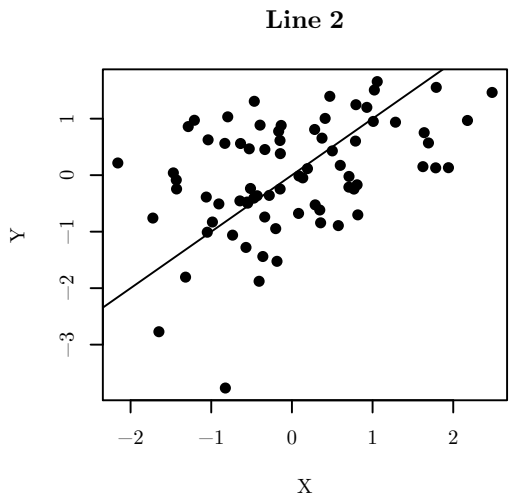
Least squares procedure

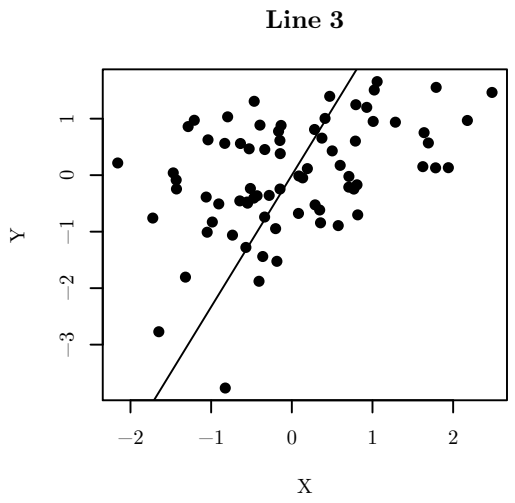
Regression is a mathematical procedure that minimizes the squared differences between the actual and predicted values of y; *i.e.*, the squared vertical distances from the line to the points.

Three proposed regression lines are given on the next slide—which is the true regression?

OLS candidates







Line 1

Assuming a regression line of the form:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i,$$

Line 1 is a regression of y on x that minimizes the sum of the squared vertical distances:

$$\min_{b_0, b_1} \sum_i (y_i - (b_0 + b_1 x_i))^2.$$

Line 3

Assuming a regression line of the form:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i,$$

Line 3 is a regression of x on y that minimizes the sum of the squared horizontal distances:

$$\min_{b_0, b_1} \sum_i \left(x_i - \left(\frac{-b_0}{b_1} + \frac{y_i}{b_1} \right) \right)^2.$$

Line 2

Assuming a regression line of the form:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i,$$

Line 2 is a principal components line that minimizes the sum of the sums of the squared vertical and horizontal distances:

$$\min_{b_0, b_1} \sum_i \left[(y_i - (b_0 + b_1 x_i))^2 + \left(x_i - \left(\frac{-b_0}{b_1} + \frac{y_i}{b_1} \right) \right)^2 \right].$$

Note that this minimizes the square of the shortest distance from the point to the regression line (use the Pythagorean theorem).

Why regress y on x?

So which is "right"?

It depends. Usually we are trying to predict the value of y for some given values of x. For example, what is a person's expected income if he is an Hispanic high school graduate? Then, a regression of y (income) on x (sex, race, and education) gives us the closest prediction of his income (*i.e.*, the sum of squared differences between actual and predicted incomes is minimized).

Why least squares?

Why do we use squares of the residuals rather than something else? What about:

The power 0? Everything is 1, so that doesn't help.

A Negative power? That rewards *big* distances, the opposite of what we want.

Odd positive powers? Large negative residuals can offset large positive residuals.

So why squared, rather than to the powers of 4, 6, or 8?

Residuals raised to other powers

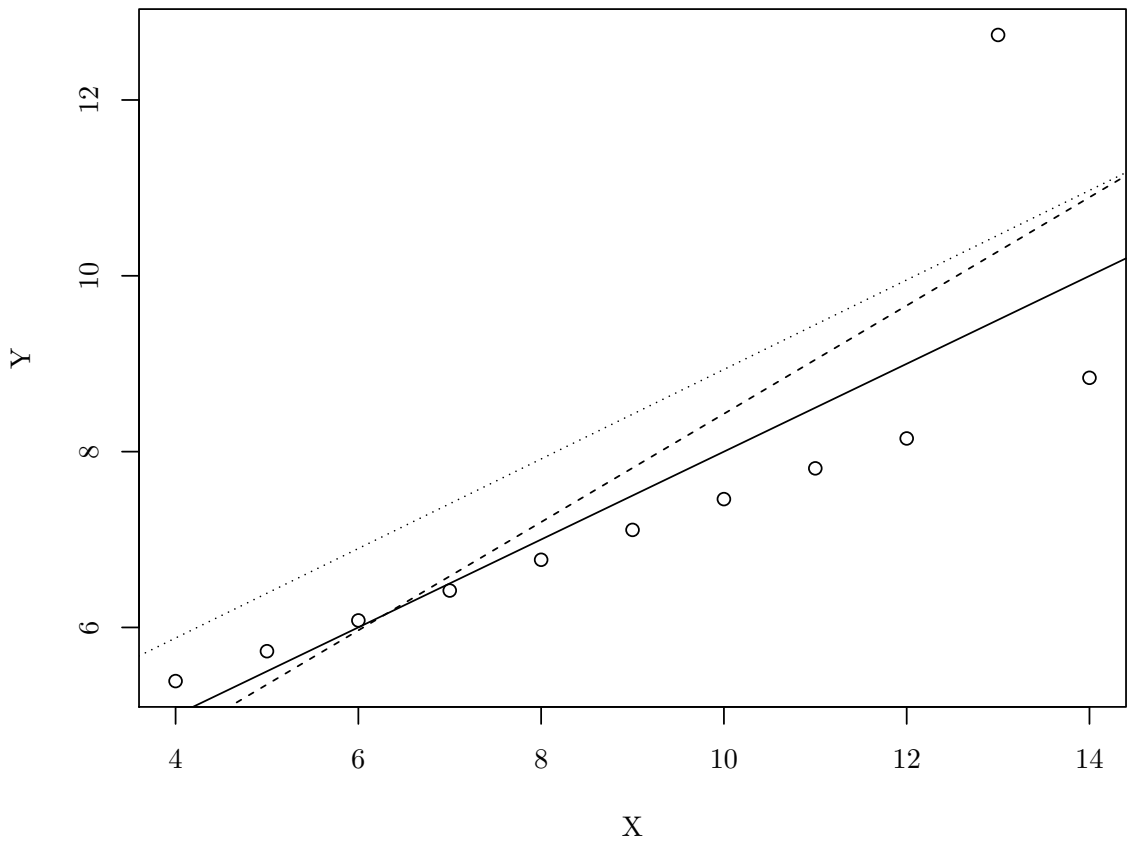


Figure: Minimizing sum of residuals raised to different powers; the solid line is least squares, the dashed uses the quartic, and the dotted line raises to the eighth power.

Leverage

Outliers get high *leverage* when we use higher powers; *i.e.*, they pull the regression line towards themselves.

Other alternatives

Lastly, when we use the sum of squared errors, we essentially minimize the *mean* of squared errors. Why not something else?

Least median of squares

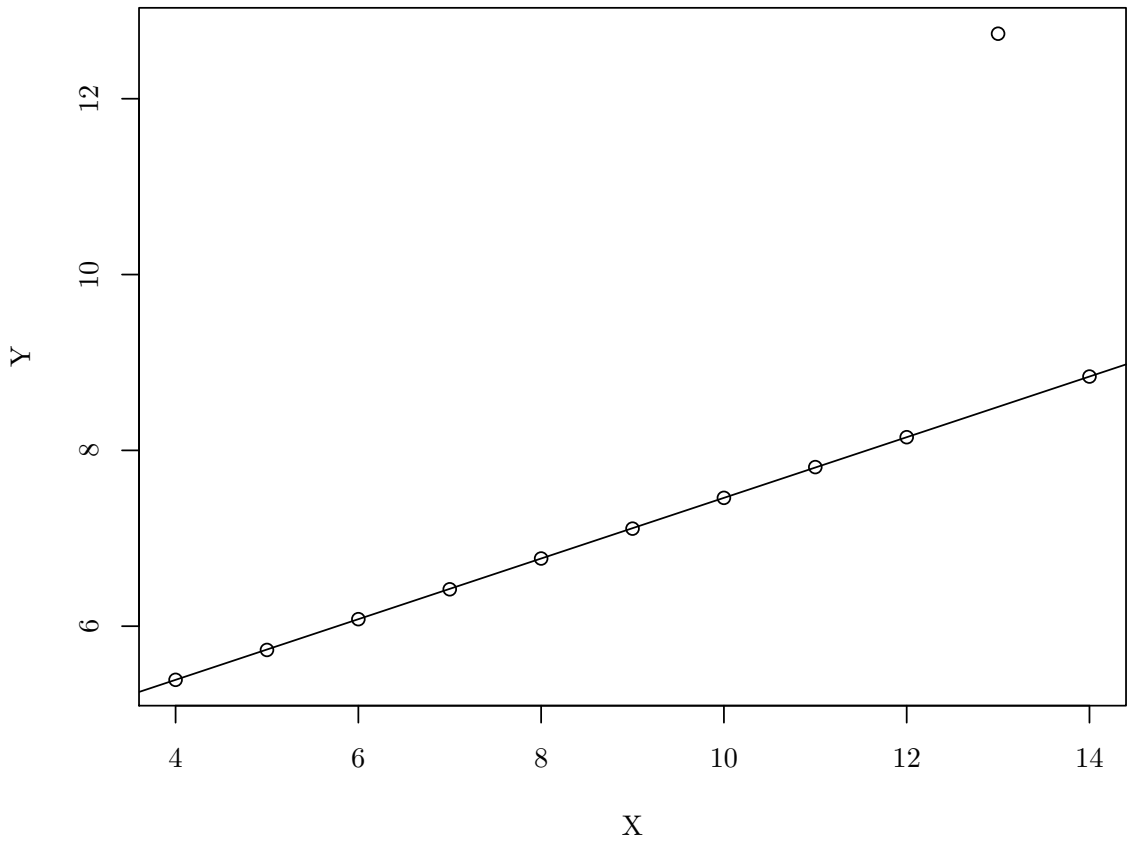


Figure: Least median of squares is an example of *robust* or *resistant* regression.

Linearity

"Linear" in the expression "linear model" refers to the parameters, not the predictors.

Linear:

$$y_i = \beta_0 + \beta_1 e^{x_i} + \epsilon_i$$

Not linear:

$$y_i = e^{\beta_0} + e^{\beta_1} x_i + \epsilon_i$$

Not linear:

$$y_i = \beta_0 \times \beta_1 x_i + \epsilon_i$$

Using logs

In economics, we often use the (natural) logged value for a covariate rather than its actual level. There are two reasons for this:

1. To better approximate an economic model or
2. To facilitate easier interpretation of the results.

Level-level model

Let's consider model 1, a level-level univariate regression:

$$y_i = \beta_0^1 + \beta_1^1 x_i + \epsilon_i^1$$

(the superscripts indicate model 1). We can take the partial derivative of y with respect to x:

$$\beta_1^1 = \frac{\partial y}{\partial x}.$$

Log-level model

Now turn to a log-level regression:

$$\ln(y_i) = \beta_0^2 + \beta_1^2 x_i + \epsilon_i^2.$$

Taking the partial derivative of the left-hand side with respect to x gives

$$\beta_1^2 = \frac{\partial \ln(y)}{\partial x} = \frac{\frac{\partial y}{y}}{\partial x}.$$

Here, β_1^2 measures the proportion change in y for a one unit change in x; $100 \times \beta_1^2$ is the percent change in y for a unit change in x.

Level-log model

Now, a less common level-log model:

$$y_i = \beta_0^3 + \beta_1^3 \ln(x_i) + \epsilon_i^3.$$

Here the partial derivative is:

$$\beta_1^3 = \frac{\partial y}{\partial \ln(x)} = \frac{\partial y}{\frac{\partial x}{x}}.$$

This is the unit change in y for a proportional increase in x; y increases by $\frac{\beta_1^3}{100}$ for a one percent increase in x.

Log-log model

Lastly, we have the log-log model:

$$\ln(y_i) = \beta_0^4 + \beta_1^4 \ln(x_i) + \epsilon_i^4.$$

The partial derivative is:

$$\beta_1^4 = \frac{\partial \ln(y)}{\partial \ln(x)} = \frac{\frac{\partial y}{y}}{\frac{\partial x}{x}} = \frac{\partial y}{\partial x}\frac{x}{y}.$$

This is the percent change in y per percent change in x. In economics, this is an *elasticity*.

Relating these estimates

Note that we can move between these estimates at particular values of y and x:

$$\begin{aligned}\beta_1^2 &= \frac{1}{y} \times \beta_1^1, \\ \beta_1^3 &= x \times \beta_1^1, \text{ and} \\ \beta_1^4 &= \frac{x}{y} \times \beta_1^1.\end{aligned}$$

While these relationships depend upon the particular values of y and x chosen, they can help make estimates comparable across models.

This result also highlights that the marginal effect (*i.e.*, the partial derivative) depends upon the values of x.

Pictoral representation

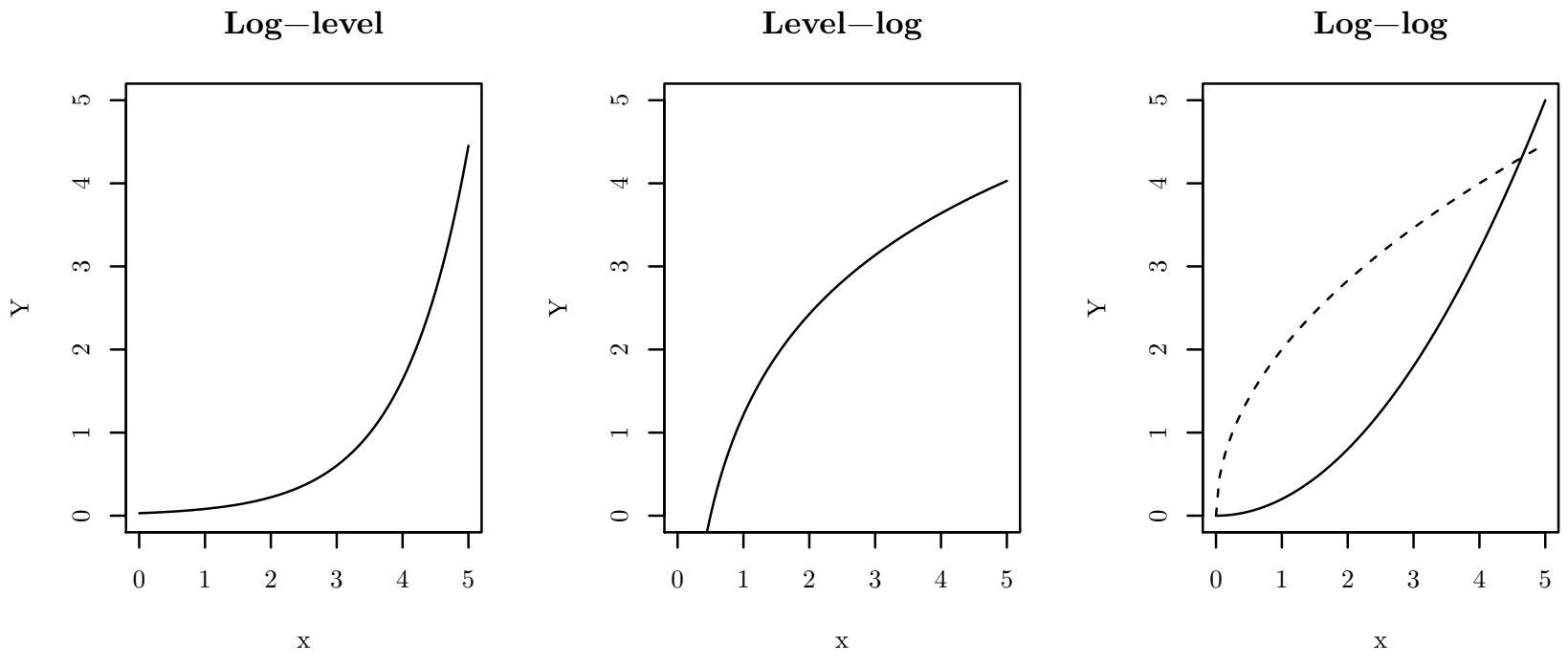


Figure: Plots of log models

Choosing models

Which model do we choose?

We often take the log of something when we think that multiplicative changes matter. For example:

1. If we use the level of income, then a change from \$20,000 to \$30,000 would be treated the same as a change from \$80,000 to \$90,000.
2. If we use the log of income, then a change from \$20,000 to \$30,000 is the same as \$80,000 to \$120,000 (*i.e.*, a 50% increase in both cases).

The model

We begin with the true regression model:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i.$$

When we change the x and y variables, we *cannot* change this relationship.

Scaling predictors

Imagine that x was in dollars in this equation, but we want to change it to thousands of dollars. To do this, we divide x by 1,000. For the true regression relationship to hold, we need

$$\begin{aligned} y_i &= \beta_1 + \frac{1000}{1000}\beta_2 x_i + \epsilon_i \\ &= \beta_1 + (1000 \times \beta_2)\frac{x_i}{1000} + \epsilon_i. \end{aligned}$$

Notice that β_1 doesn't change, but our coefficient on our new x variable is multiplied by 1,000.

Scaling outcomes

Now suppose we change y from dollars to thousands of dollars. Then we have

$$\frac{y_i}{1000} = \frac{\beta_1}{1000} + \frac{\beta_2}{1000} x_i + \frac{\epsilon_i}{1000}.$$

Both of our coefficents are transformed here.

Test statistics

Let's consider the case when we multiply X by c. Then the variance of our esimate is

$$\widehat{\text{Var}}\left(\frac{\hat{\beta}}{c}\right) = \frac{1}{c^2}\widehat{\text{Var}}\left(\hat{\beta}\right).$$

Standardizing this variables yields

$$\frac{\frac{\hat{\beta}}{c} - \frac{\beta}{c}}{\sqrt{\widehat{\text{Var}}\left(\frac{\hat{\beta}}{c}\right)}} = \frac{\frac{1}{c}\left(\hat{\beta} - \beta\right)}{\frac{1}{c}\sqrt{\widehat{\text{Var}}\left(\hat{\beta}\right)}} = \frac{\hat{\beta} - \beta}{\sqrt{\widehat{\text{Var}}\left(\hat{\beta}\right)}}.$$

Hence, test statistics do not change when we scale our variables. This is good—we wouldn't want significance to change if we use pennies instead of dollars, for example.