Linear Regression

Dustin Johnson
Applied Quantitative Methods

January 14, 2015

INTRODUCTION

The Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where} \quad (i = 1, \ldots, n)$$

- $\beta_0 + \beta_1 x_i$ represents the *systematic* relationship.
- ϵ_i represents the *random error* or for finance folks (CAPM), *idiosyncratic risk*.
- In simple linear regression, we assume that the random errors are normally distributed.
- Simple Linear Regression is a subset of the class of Generalized Linear Models (GLM). GLMs can allow response variables to have error distribution models other than a normal distribution.
    - i.e. logistic regression, Poisson regression, models for counts data, etc.

ASSUMPTIONS

For a simple linear regression model, we assume the following:

$$\epsilon \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

1. $E(\epsilon_i) = 0 \quad \text{for} \quad i = 1, \ldots, n.$
2. $\epsilon_1, \ldots, \epsilon_n$ are statistically independent.
3. $Var(\epsilon_i) = \sigma^2 \quad \text{for} \quad i = 1, \ldots, n$: constant over the observations.
4. ϵ_i is normally distributed for $i = 1, \ldots, n$.

IMPLICATIONS

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where} \quad (i = 1, \ldots, n)$$

If $\epsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, then Y_i is also a random variable with the equivalent assumptions.

$$\epsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \implies Y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

LEAST SQUARES ESTIMATION

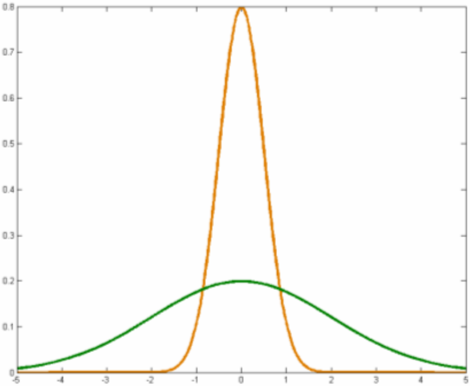
If we make the previous 4 assumptions, various properties of the estimates can be derived.

- In reality, the line $\beta_0 + \beta_1 x_i$ is unknown, hence, so too are the errors ϵ_i.
- We can estimate β_0 and β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.
- Our estimated regression line is therefore $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, at each x_i.
- The least squares criterion chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the residuals $r_i = y_i - \hat{y}_i$ as small as possible (fitted values close to y data values) - specifically, minimize the sum of square residuals w.r.t. β_0 and β_1.

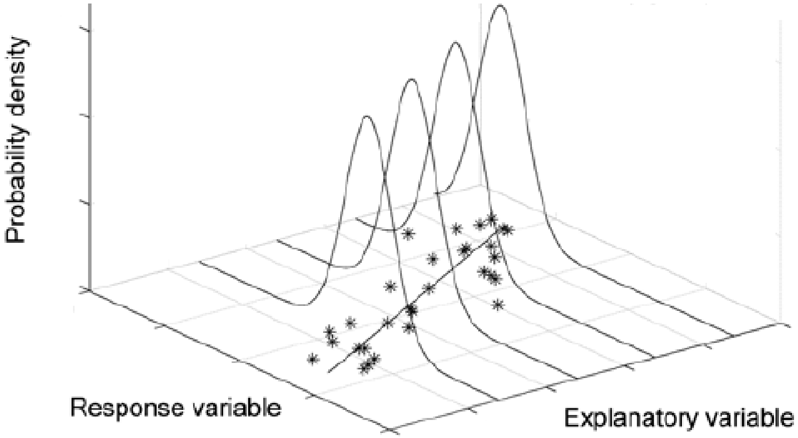
THE NORMAL DISTRIBUTION

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normality



Representation of SLR



SUM OF SQUARED RESIDUALS

$$SS(Residuals) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

Why is this? Ultimately, we would like to solve the optimization problem

$$\min_{\hat{\beta}_0, \hat{\beta}_1} SS(Residual)$$

to obtain the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ (proof and estimates are located in the primer). These estimates are too normally distributed, therefore provide us with the ability to construct confidence intervals for the parameters.

MULTIPLE REGRESSION - THE MODEL

We can easily extend the SLR model to include several explanatory variables as follows:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i \quad \text{for} \quad (i = i, \ldots, n)$$

or in matrix notation,

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)} \boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

ASSUMPTIONS

The assumptions of the simple linear regression still apply to the multiple regression.

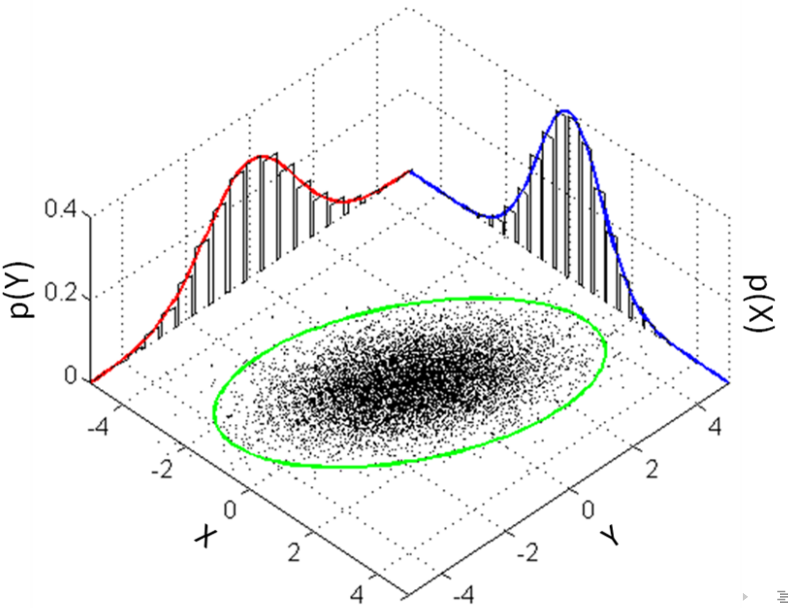
$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \overset{iid}{\sim} \mathcal{MN}(0, \sigma^2) \implies \mathbf{Y} \overset{iid}{\sim} \mathcal{MN}(\mathbf{X}\boldsymbol{\beta}, , \sigma^2\mathbf{I})$$

1. $E(\epsilon_i) = 0 \implies E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{for} \quad i = 1, \ldots, n$
2. $Var(\epsilon_i) = \sigma^2 \implies Var(Y_i) = \sigma^2 \quad (i = 1, \ldots, n).$
3. Independence of ϵ_i and ϵ_j for all $i \neq j$ implies:
   - $Cov(\epsilon_i, \epsilon_j) = 0 \quad (all \quad i \neq j)$
   - $Cov(Y_i, Y_j) = 0 \quad (all \quad i \neq j)$
4. ϵ_i is normally distributed for $i = 1, \ldots, n \implies Y_i$ is also normal (linear function of ϵ).

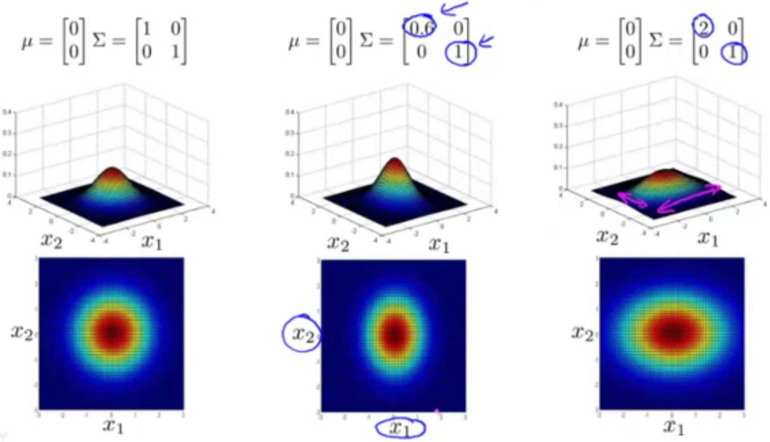
Multivariate Normal Distribution

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

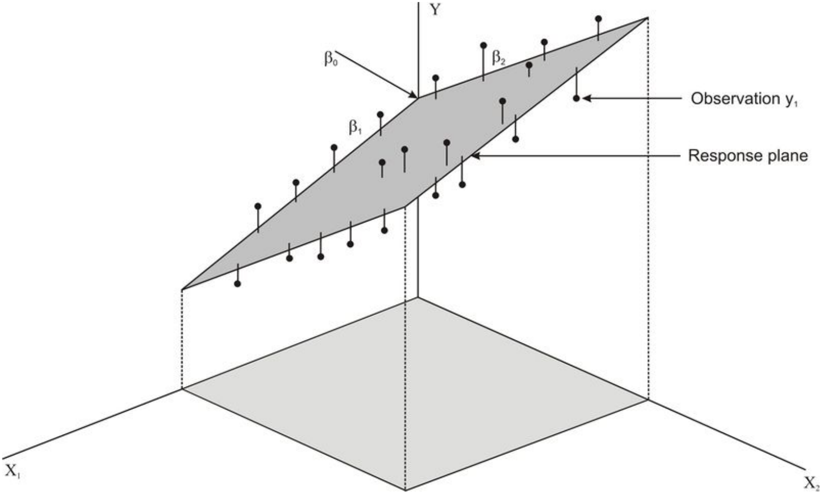
MULTIVARIATE NORMAL DISTRIBUTION



MULTIVARIATE NORMAL DISTRIBUTION



REPRESENTATION OF MULTIPLE REGRESSION



Least Square Estimates

$$SS(Residual) = \sum_{i=1}^{n} r_i^2 = \mathbf{r}^T\mathbf{r} = (\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}}) = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

The $\boldsymbol{\beta}$ that minimizes the SSR is the *least squares estimate*, given by the following explicit expression:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

PROJECTION MATRIX

$$\text{if} \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

and

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad \text{represents the estimated observed} \quad \mathbf{Y}$$

then

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}\mathbf{y},$$

P is called the *Projection Matrix* (why?). It has many interesting properties, so check them out in the primer!

PRECAUTIONS

Collinearity: A linear relationship between two explanatory variables.

Multicollinearity: A linear relationship between more than two explanatory variables.

ANOVA

$$SS_{total} = SS_{regression} + SS_{error}$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} r_i^2$$

Source	df	SS	MS	F
Regression	p	$SS_{regression}$	$SS_{regression}/p$	$\frac{SS_{regression}/p}{SS_{total}/n-1}$
Error	$n-p-1$	SS_{error}	$SS_{error}/n-p-1$	
Total	$n-1$	SS_{total}		

R-SQUARED

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} \quad (0 \leq R^2 \leq 1)$$

- Values close to 1 indicate a good fit.
- How big should it be? ... depends
- Downfall: The more variables we add, the better the R^2, even if they contribute no value to the model.

Improvement: Adjusted R-squared to account for number of parameters added.

$$R^2_{adj} = 1 - \frac{MS_{Residual}}{MS_{Total}}$$

F Test

Fisher's F Test collectively assesses $x_1, \ldots, x_p$ for their explanatory utility. Essentially, it tests the overall regression relationship and asks whether the fitted slopes $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are significantly different from zero. The test statistic is given by:

$$F = \frac{MS(\hat{\beta}_1, \ldots, \hat{\beta}_p)}{MS_{Residual}}$$

- High F Stat (low p-value) enables us to reject the null hypothesis and claim that *at least one of* $\hat{\beta}_1, \ldots, \hat{\beta}_p$ *is nonzero*.

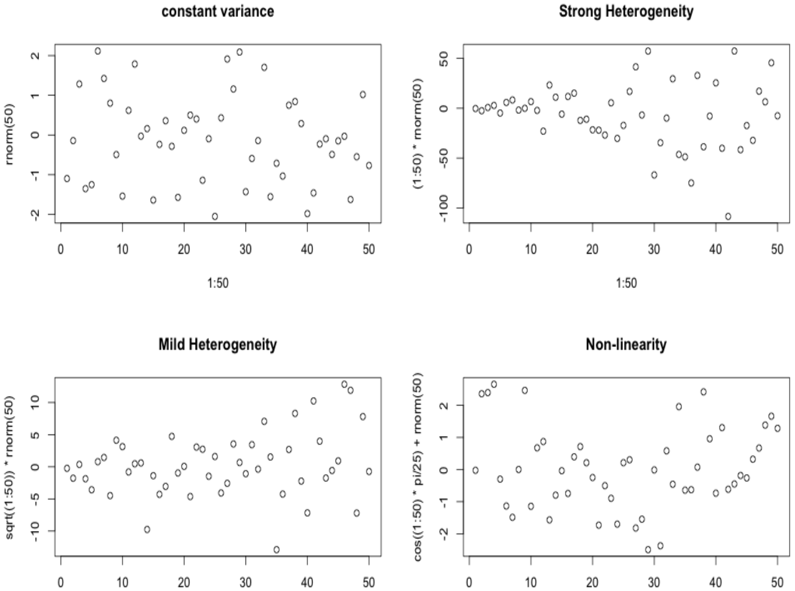
ANOVA Activity in R

1. Import data located on the AQM site.
2. Perform a multiple regression in R.
3. Output an ANOVA table in R.
4. Interpret.

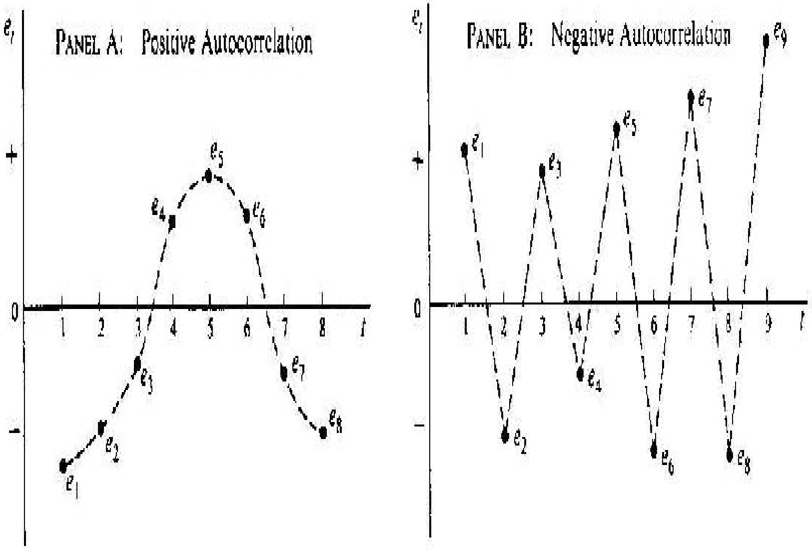
$E(\epsilon_i) = 0$

Plot residuals vs. x_j. If $E(\epsilon) = 0$ is violated, we are assuming that the effect of x_j on $E(Y)$ is linear when it is not, or perhaps an x_j was omitted.

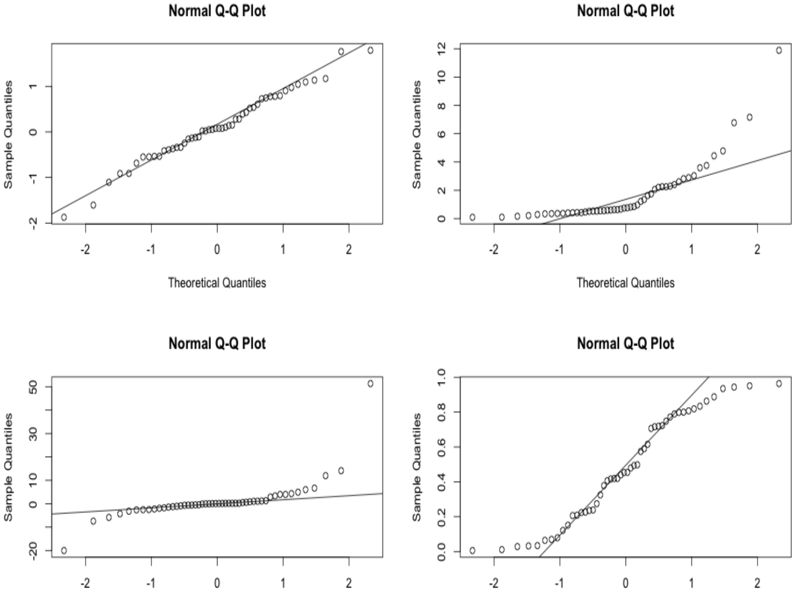
CHECKING FOR CONSTANT VARIANCE: $Var(\epsilon_i) = \sigma^2$



CHECKING FOR UNCORRELATED ERRORS: $Cov(\epsilon_i, \epsilon_j) = 0$



NORMALITY OF RESIDUALS



R ACTIVITY - DIAGNOSTICS

First, look at your model's R^2, F Stat, and parameter estimates and comment on your results. Second, analyze the residual and diagnostic plots for breaches in the model assumptions.

- Do your results indicate a "good" model fit?
- How confident are you?
- Does anything stand-out?
- Are you skeptical? If so, why?