Motivating Questions

Thursday, August 23, 2012 1:48 PM

Two inter-related, but distinct questions:

1) Is my model doing a good job inside of the sample?
    a. Is it a good fit to the data set?
    b. Is it the best-fitting model among a set of candidate models?
    c. Are its results robust to minor variations in the data?

is my linear model a good fit for this experimental data set?

2) Is my model a good choice for this situation (structure of the dependent/independent variable, correlation structure of the data, etc.)?
    a. Will I recover correct parameters (e.g., beta coefficients)?
    b. Will I recover unbiased, low-variance estimates of substantively meaningful quantities (e.g., marginal effects)?
    c. How would we expect the model to perform under adverse conditions?

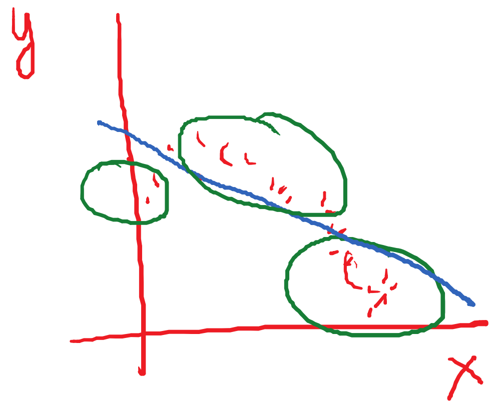
is it ok to use a linear model in this situation? How does my estimator (OLS) perform under these circumstances?

The first question asks us to assess the performance of a **particular model (estimator + sample)** using sample diagnostics, while the second question asks us to assess the characteristics of an **estimator** in different environments

Assessing In-Sample Fit

Thursday, August 23, 2012 2:28 PM

- There are lots of ways we might assess a model's quality, and these assessments presumably vary according to the model's goals
    - Example: are *false positives* or *false negatives* more important to avoid?
- Consider a simple example model: $y = X\beta + \epsilon$
- There are many informal assessment techniques
    - Prediction plots
    - Residual plots



Added Variable Plots

Thursday, August 23, 2012 9:13 PM

$y = \beta_0 + \underline{\beta_1} x + \beta_2 z$

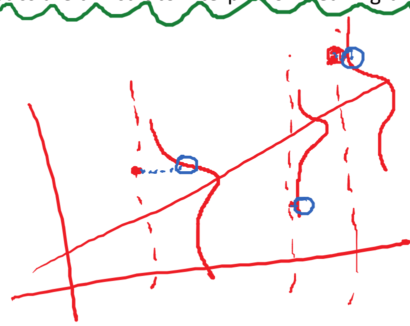
- Problem with a basic scatterplot: omitted variable bias / spurious relationships
- Added variable plots allow an analyst to examine the relationship between the dependent variable and one independent variable, controlling for the other variables in the model
  1. Predict *y* using all the independent variables *z* except *x*, save the residuals → $y = \beta_0 + \beta_2 z \rightarrow r_y$
  2. Predict *x* using all the independent variables *z* except *x*, save the residuals → $x = \alpha_0 + \alpha_2 z \rightarrow r_x$
  3. Plot the residuals from (1) against the residuals from (2); the relationship in this plot (e.g., the estimated coefficient on a regression slope) will be identical to the relationship found between *x* and *y* in a multivariate model including *z* → $r_y \sim r_x$

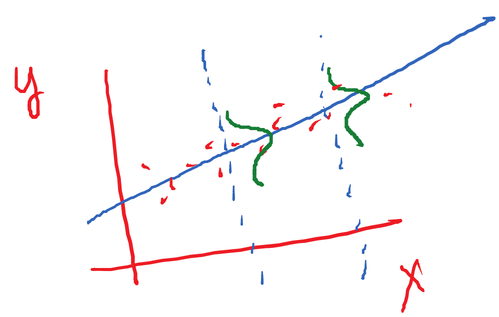
     FWL Theorem
- Allows diagnosis of possible non-linearities and the assessment of marginal contribution to the model

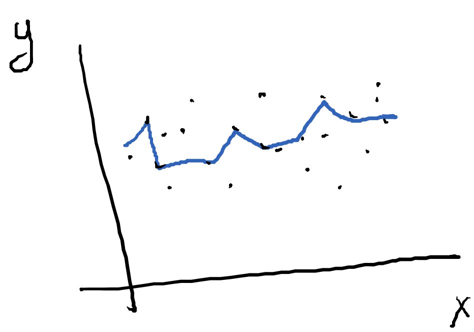
Squared Errors and Likelihood

Friday, August 24, 2012 12:22 PM

- What about more formalized assessments of fit quality?
- One common criterion: how well does the model fit the data?
  - Pathway 1: the goal of a model is to *minimize error in predictions*, $\hat{y} = X\hat{\beta}$ ← $g(X\hat{\beta}) = \hat{y}$
    - Sum of squared errors: $SSE = \boxed{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}$
    - R-squared: $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (\bar{y} - y_i)^2}$ prop. of variance in y explained by model
  - Pathway 2: the goal of a model is to *be consistent with the joint probability of this realization of the dataset*
    std. dev of regression
    - Likelihood: $L = \Pi_{i=1}^{N} \Phi(y_i, \mu = X_i\hat{\beta}, \Sigma = \sigma^2 I_{NxN}) = \sqrt{\frac{1}{2\pi\sigma^2}}\, \Pi_{i=1}^{N} \left(\frac{\exp(-(\hat{y}_i - y_i)^2)}{2\sigma^2}\right)$ OLS / linear
    - For the simple linear model, note that maximizing the likelihood is equivalent to minimizing the sum of squares. pr(data | model)
  - These are two extremely common ways of assessing model quality, but not necessarily the only possible ways.
- We could assess a model's quality by looking at these measures of in-sample fit on an absolute scale and/or comparing them to others
  - The parameters of the model, $\hat{\beta}$, are fitted to maximize a particular model's R-squared / likelihood
- Problems?
  - Susceptible to **over-fitting** ← *
  - Do not account for noise intrinsic to the data
  - Likelihood statistics are difficult to interpret or meaningfully compare

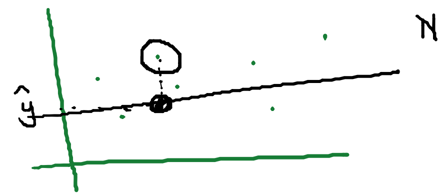






Cross-Validation

Thursday, August 23, 2012 3:50 PM



- Question: is my model being over-fitted? Should I add/remove variables or terms from my model?
- One way of dealing with over-fitting: cross-validation (called the $PRESS$ criterion in the readings)
- Idea:
  1) Drop one observation from the data set
  2) Estimate a model without the dropped observation
  3) Predict $\hat{y}$ for the dropped observation using the estimated model
  4) Replace the dropped observation in the data
  5) Repeat 1-4 for each observation

leave-one-out cross validation

- No chance of over-fitting: the model does not include the fitted observation
- Compare each model's cross-validated prediction error, and choose the one with the **lowest** error
- Computationally demanding for large data sets (N+1 models must be estimated!)

Complexity-Adjusted Criteria

Thursday, August 23, 2012 3:50 PM

proportion of y not explained

- Another approach to over-fitting: *penalize* fit statistics for model complexity, so that adding an arbitrarily large number of terms to the model does not result in fit improvement
  - adjusted-R^2: deflation factor
    $\bar{R}^2 = 1 - (1 - R^2)\left(\frac{\quad}{n-k-1}\right)$ where k is the number of terms in the model
    larger is better

    $\bar{R}^2$

    # variables (incl. constant)

  - * Akaike's Information Criterion:
    AIC = $2k - 2\ln L$
    (in the linear model) = $2k + n \ln SSE - n \ln n$
    smaller is better
    - Note: this is asymptotically equivalent to leave-one-out cross-validation in the linear model, and in some other models!

    $N \to \infty$

    $AIC \to CV_1$

  - * Bayesian Information Criterion:
    BIC = $k \ln n - 2\ln L$
    (in the linear model) $= k \ln n + n \ln SSE - n \ln n$
    smaller is better
- There are many "information criteria" family penalized fit statistics, each with their own theoretical justification; the main difference is in the penalty term
- Can compare non-nested models (i.e., models that contain different terms on the right hand side) as long as they are all estimated on the same dependent variable data

F

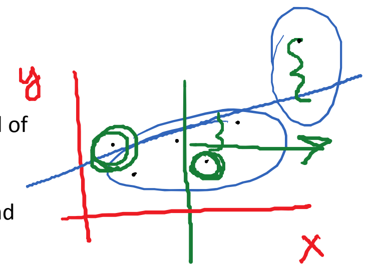
✓ $y = \beta_0 + \beta_1 X$

$y = \beta_0 + \beta_2 X + \beta_3 Z$

$y = \beta_0 + \beta_2 Z$

Outliers and Influential Observations

Friday, August 24, 2012 12:45 PM



- Side topic: occasionally, influential observations can have a significant impact on a model that negatively influences the quality of the overall fit to the data set
- There are informal diagnostics (e.g., scatterplots) that involve looking for observations that have a great deal of *leverage*
  - High-leverage observations are far from the middle of the distribution on the independent variable, and have large error estimates $\hat{\epsilon}_i = \hat{y}_i - y_i$
- There are also formal diagnostics for identifying influential observations
  - DFFITS: standardized change in $\hat{y}_i$ when observation *i* is included vs. deleted when running the regression
  - DFBETAS: standardized change in the $\hat{\beta}$ coefficients when observation *i* is deleted
  - Examine observations with large DFFITS/DFBETAS to consider deletion or reweighting

$$\frac{\hat{\beta} - \hat{\beta}}{se.\hat{\beta}}$$