Linear Regression

Benjamin S. Skrainka

June 7, 2016

Objectives

Today's objectives:

- State assumptions of linear regression model
- Estimate a linear regression model
- Evaluate a linear regression model
- Fix common problems which could compromise results

Agenda

Today's plan:

1. Introduce basic model
2. Evaluating model performance
3. Common problems which compromise results

References

References from the machine learning perspective:

- Introduction to Statistical Learning
- Elements of Statistical Learning

References from the econometric perspective:

- Greene's Econometric Analysis
- Wooldridge's Econometric Analysis of Cross-section and Panel Data
- Cammeron & Trivedi's Microeconometrics: Methods and Applications
- Kennedy's A Guide to Econometrics

Modeling with data

Goal of machine learning

Machine learning is a set of tools to learn a very good approximation of the relation between features and a label:

- True model:

$$y = f(x) + \epsilon$$

- Machine learning learns an approximation $\widehat{f}(x)$ of $f(x)$
- Use $\widehat{f}(x)$ to predict y from new values of X
- Determine which features matter:
  - Check model makes sense
  - Know which features drive business and how to move them
- Tune *bias-variance* tradeoff to optimize predictive performance

Goal of regression Analysis

Regression analysis fits a model to perform causal inference:

- Establish causal factors which affect outcome
- Choose no bias at expense of higher variance
- Perform inference on model parameters to establish causal links
- Must demonstrate control for *confounding factors*:
  - Should be *as good as randomly assigned*
  - Must eliminate sources of bias in error term (shock)

Terms

There are many different terms for the same concepts, depending on your background:

- Feature = Covariate = Input = independent variables = regressors = RHS variables = X
- Label = outcome = target = dependent variable = regressand = LHS variable = y
- Train = learn = estimate = fit a model

Types of machine learning models

Two main types of machine learning models:

- Supervised: models a label using features
  - Regression: analyze a continuous outcome, such as price or demand
  - Classification: predict a categorical (discrete) outcome, such as fraud or churn
- Unsupervised: finds patterns or labels for unlabeled data
  - Clustering: hierarchical, k-means
  - Dimension reduction: PCA, SVD, NMF

Types of data

Common types of data:

- Cross-section: x_i
  - One observation per *individual* or *cross-sectional unit*
  - Computed at one point in time
  - Many i, One t
- Time-series: x_t
  - Multiple observations of a quantity over time, e.g., GDP
  - Computed at multiple instants
  - One i, Many t
- Panel-data: x_{it}
  - Observe units over time
  - Example: NLSY (National Logitudinal Survey of Youth)
  - Many i at many t
- Pooled cross-section:
  - An *individual* is observed at either $t = 0$ or $t = 1$
  - Two pools of cross-sectional units

Types of features

- Continuous:
  - Example: price, quantity, sales, tenure
  - May bin using quantiles to model non-linearities better
- Categorical:
  - Takes discrete levels
  - Also called a factor
  - Example: 1/0, Yes/No, Treated/Control, High/Medium/Low
- Text/audio/image
  - May need to generate features

Small, medium, or large data

Size of data affects analysis:

- For causal questions, need to perform inference:
  - Requires large N so that estimator is *asymptotically normal*
  - Requires small N to compute of standard errors
- For prediction, can use truly large data sets:
  - Must check model via cross-validation
  - Can run at scale, but cannot perform inference
  - May need regularization to avoid overfitting if there are a lot of features

Linear Regression

Introduction to linear regression

Regression models the expected value of the outcome, conditional on features:

$$\mathbb{E}[y|x] = x^T \beta$$

or

$$y_i = x_i^T \beta + \epsilon_i, \forall\, i$$

- Linear regression predicts the mean value (mean) of y, holding x fixed
- Model is *linear* in parameters β but features may be non-linear functions of data, such as polynomials or splines
- Other models are possible, such as quantile regression

Notation

Some notation:

- y_i: dependent variable for observation i
- x_i: $K \times 1$ vector of covariates for observation i
- ϵ_i: unobserved shock for observation i
- y: $N \times 1$ vector of y_i
- X: $N \times K$ matrix of covariates, where there are k covariates and each row is x_i^T
- ϵ: $N \times 1$ vector of ϵ_i
- β: parameters (coefficients) to estimate

$$y = X\beta + \epsilon$$

Gauss Markov Assumptions

Often, we assume:

1. Linearity: $y = x^T \beta + \epsilon$
2. Full rank: X has full rank (rank $= K$)
3. *Exogeneity* of regressors: $\mathbb{E}[\epsilon|X] = 0$
4. *Spherical* errors, i.e., *homoscedastic* and not autocorrelated:
    - $Var[\epsilon_i|X] = \sigma^2, \forall\, i$
    - $Cov[\epsilon_i, \epsilon_j|X] = 0, \forall\, i \neq j$
5. Normally distributed errors: $\epsilon|X \sim N(0, \sigma^2 I)$

Can relax many of these assumptions, especially Normality

Exogeneity

An *exogenous* variable is determined outside of the model:

- $\Rightarrow \mathbb{E}[x_i \epsilon_i] = 0$
- If exogeneity fails, then:
  - Estimates for $\hat{\beta}$ will be biased
  - Then x_i is *endogenous*, i.e., determined inside the model
  - x_i is correlated with ϵ_i

Endogeneity

Common causes of endogeneity include:

- Measurement error
  - Example: classical 'errors in variables' $\Rightarrow$ *attenuation bias*
- Omitted variable bias
- Simultaneity
  - Some RHS variable x_j is determined simultaneously with y
  - Example: selection bias because ability is not observed but affects income and education:

$$income_i = \beta_0 + \beta_{edu} \cdot educ_i + \beta_{gender} \cdot gender_i + \cdots + \epsilon_i$$

Homoscedasticity

Homoscedastic errors mean the variance of the shock is constant:

- Often fails in practice because variance of shock depends on covariates
- Can correct by estimating a regression with 'robust' standard errors:
  - Many corrections
  - Example:
- Heteroscedasticity

Classical regression model

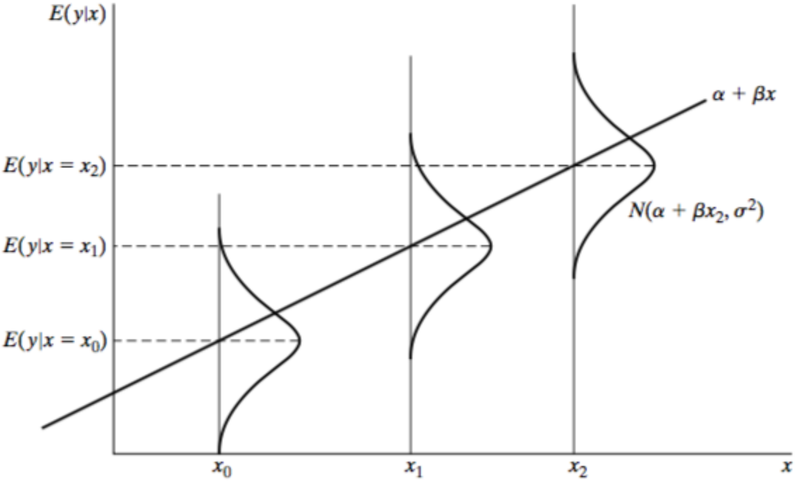

FIGURE 2.2 The Classical Regression Model.

Figure 1: Classical regression model (Greene)

Fitting a regression

Properties

- Parameter estimates: $\hat{\beta}$
- Fitted values (prediction): $\hat{y} = X^T \hat{\beta}$
- Residuals: $\hat{\epsilon}_i = y_i = x_i^T \hat{\beta}$
- Standard error: $s^2 = \dfrac{1}{N-1} \sum_{i=1}^{N} \hat{\epsilon}^2$

Best Linear Unbiased Estimator (BLUE)

Gauss Markov theorem:

- Given classical assumptions:
  - Linearity
  - Full rank
  - Exogenous regressors
  - Homoscedastic, uncorrelated, Gaussian errors
- Then, least squares estimator is:
  - Unbiased
  - Minimum variance estimator

Interpretation of regression

Two interpretations:

- Minimizes MSE
- Projects data onto subspace spanned by X

Interpreting regression results

- Ceteris paribus
- Comparative statics

Example: regression output

Multiple linear regression

Dummy variables

To work with categorical data, use dummy variables:

- Collinear with intercept if saturated model
- Factor = categorical = dummy
- May proxy for unobserved effect

Example: creating dummy variables

Often you need to create a series of dummy variables for a categorical variable which has multiple levels:

```
df = pd.read_csv('amazing_data.csv')
feature_cols = [0, 2, 3, 4, 7]
Xdum = pd.DataFrame(sm.tools.categorical(np.array(df.my_factor
X = pd.concat(df.icol(feature_cols), Xdum], axis=1)
```

Interactions

Displaying results

Display regression results in a table:

- Each row is a feature
- Each column is a model specification
- Quote standard error or p-value:
  - Under each estimate
  - In a separate column
- Do not quote results as an equation with numeric coefficients

Example of regression results

Evaluating a Regression model

R^2

F-statistic

Residuals

Heteroscedasticity

Non-normality

- White noise
- QQ plots
- Normality tests: Jarque-Bera, Shapiro-Wilk

(Serial) correlation

- Temporal
- Spatial or between individuals

Multicollinearity

Model specification: AIC

Cross-validation

Common Problems Applying Regression Models to Data

Common problems

- Non-linearity
- Non-normality

Sources of bias

Some common sources of bias:

- Simultaneity
- Measurement error
  - Classical 'errors in variables'
  - Omitted variable bias
- Attenuation bias

Long-tailed data

- May be better to work in logs
- Use Box-Cox test

Meaning of coefficients for models of $\log(y)$

Outliers

Leverage

To measure influence of an outlier:

- Compute the 'hat matrix': $H = X^T(X^TX)^{-1}X$
- i-th element on the diagonal, $h_{ii} \equiv (H)_{ii}$, is i-th feature's 'leverage'
- An observation with a large residual may not have a lot of influence
- May affect level or slope

Influential points

- Affect slope of regression
- Have large residual, $\hat{\epsilon}_i$ and high leverage, h_{ii}

Advanced Topics

Relaxing the Gauss-Markov assumptions

Generalized Least Squares (GLS)

Non-linearity

- Polynomials
- Splines
- Other basis functions, e.g., $\sin(x)$ and $\cos(x)$ with time-series data

Instrumental variables

Panel data

Time series

Classification or Discrete choice

Local linear regression

Generalized additive models (GAM)

Regularization

Practical tools

Regression tools

- `sm.add_constant()`
- `sm.tools.categorical()`

Plot tools

From Pandas:

- `pd.tools.plotting.scatter_matrix(df, diagonal='kde')`
- `df.boxplot()`

From StatsModels:

- `sm.graphics.influence_plot(results)`
- `sm.graphics.qqplot(ols_fit.resid, line='q')`

Summary