Applied Regression Analysis

Estimation

Earvin Balderama

Department of Mathematics & Statistics
Loyola University Chicago

September 8, 2016
Basic ideas

Sample vs Populations

- Population: The collection of all subjects of interest
- Sample: A representative subset of the population of interest
STOP: This is an important idea

- We’re going to make this distinction between the population and sample many times.
- Throughout this course, remember that there is a distinction between the population mean (commonly $\mu$) and the sample mean (commonly $\bar{x}$), for example.
- Likewise, there is a distinction between the population parameter in a regression model (e.g. $\beta$) and the estimate of a regression coefficient ($\hat{\beta}$).
REGRESSION!

- You are going to hear this term a lot.
- What is regression?
  - **Regression** is a collection of techniques for modeling the relationships and associations between variables.
  - **Linear regression** is a collection of techniques for modeling the \textit{linear} relationship between variables.
  - So what is \textit{non-linear regression}?
Design and Analysis of Scientific Studies

1. Develop the study objectives and the scientific questions of interest.
2. Design a scientific study and data collection protocol to answer the scientific questions of interest.
3. Develop statistical analysis plan.
4. Perform data analysis after the study is completed using appropriate statistical models and techniques accounting for the features of the data and the study design. This stage might involve new method development.
5. Interpret the results and publish your paper!
Overview of Data Analysis

- Data are a (sometime very large) collection of numbers and are of limited use in raw form.
- The first stage of making scientific conclusions from data involved summarizing data using descriptive and graphical techniques: tables, histograms, scatterplots, boxplots.
- The second stage involves statistical modeling and inference leading to actionable conclusions.
- You probably already know some methods of statistical inference (e.g. t-tests, $\chi^2$-test)
Here, we focus on statistical regression analysis for continuous outcomes.

It is important to keep in mind that statistics is both a data science and a data art. Good statistical research and practice need to integrate closely with science and collaborate closely with subject-matter scientists.

Then you can make an impact not only in statistics but also in real world.
Detailed list of course topics

- Simple linear regression: model formulation, least square estimation, the principle of maximum likelihood, tests and confidence intervals for regression parameters, prediction, analysis of variance (ANOVA) approach to regression.
- Multiple linear regression: hypothesis testing and estimation, model interpretation, confounding, and interaction.
- Model checking and diagnostics: residual plots, outliers, and influence
- Model building: criteria for evaluation, model-building methods (forward selection, backward elimination, step-wise selection)
- Generalized Linear Models: logits and logistic transformations, likelihood theory for log odds ratios, testing and confidence interval estimation, interpreting regression coefficients and estimation odds ratios
What is regression analysis?

- **Regression analysis** is a statistical method for investigating *functional relationships* between two or more variables.
  - the values of one variable may depend on the values of other variables.

Some examples:

- What are the effects of air pollution inhalation on lung function among asthmatic children?
- Whether the risk of stillbirth is associated with the use of selective serotonin reuptake inhibitors (SSRIs) during pregnancy?
- Whether cigarette consumption is related to various socioeconomic and demographic variables such as age, education, income, and price of cigarettes?
- What is the effect of a restrictive transfusion in patients with acute gastrointestinal bleeding on risk of death, adjusting for baseline characteristics factors?
What is regression analysis?

- The relationship is expressed in the form of a mathematical equation connecting a response variable with one or more explanatory (or predictor) variables.

- In the cigarette consumption example:
  - The **response variable** is cigarette consumption (measured by cigarette packages sold per capita per year)
  - The **explanatory (or predictor) variables** are the different socioeconomic and demographic variables.
Linear Regression

\[ Y = f(X) + \epsilon \]

- **Response Variable**: Usually denoted \( Y \).
- **Predictor Variable (or Explanatory or Independent Variable)**: Usually denoted as \( X \).
- \( f \) is some unknown function.
  - For linear regression, we constrain \( f \) to be a linear function, i.e., a linear combination of \( X \) and its parameters.
- **Simple Linear Regression**: Only one predictor.
- **Multiple Linear Regression**: More than one predictor.

\[ Y = f(X_1, X_2, \ldots, X_p) + \epsilon \]
Simple Linear Regression: The Model

Each observation, point, or experiment \((X_i, Y_i)\) can be expressed as

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, n. \]

- \(Y_i\) the \(i\)-th observed value of the response variable.
- \(X_i\) the \(i\)-th observed value of the predictor variable.
- \(\beta_0\) and \(\beta_1\) are parameters (or “regression coefficients”). \(\beta_0\) is referred to as the intercept and \(\beta_1\) is the slope.
- \(\epsilon_i\) is the error term.
Simple Linear Regression: Two parts

The model has two parts:

1. **Systematic Part:** \( \mu = \mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i \)
   - Determines the average value of \( Y_i \) given a **fixed** value of \( X_i \).
   - By fitting a “line”, we are essentially modeling the *conditional mean*, \( \mathbb{E}(Y \mid X) \).
   - Sometimes called the *mean function* of \( Y \).

2. **Random Part:** \( \epsilon_i \)
   - Variability around the mean
   - Sources of variability include, for example, measurement error and missing covariates
   - These factors are combined in the error term \( \epsilon_i \).

**Note:** Sometimes referred to the *signal* \(^1\) and the *noise* \(^2\).
Linear Regression: Assumptions

The simple linear regression model is:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

1. The \( X_i \) are considered fixed.
2. The relationship between \( X \) and \( Y \) is linear.
3. \( E[\epsilon_i] = 0 \)
4. \( V(\epsilon_i) = \sigma^2 \) (homoscedasticity)
5. \( \epsilon_i \perp \epsilon_j \) for all \( i,j \in (1, 2, \cdots, n), i \neq j \). (This assumption is violated in longitudinal data, for example)

Note: We have not (YET!) assumed normality of the errors. We don’t need this here, but we will later.
Linearity assumption

Which of the following models are *linear* models?

1. $\mu = \beta_0 + \beta_1 X$
2. $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
3. $\mu = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
4. $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
5. $\mu = \beta_0 + \beta_1 \exp(\beta_2 X)$
Linear Regression: Assumptions

1. **Weak exogeneity**: The predictor variables $X$ are considered fixed (non-random).
2. **Linearity**: The relationship between $X$ and $Y$ is linear. (Partial derivatives of $\mu$ w.r.t. $\beta$’s doesn’t depend on $\beta$’s.)
3. **Expected error is 0**: $\mathbb{E}[\epsilon_i] = 0$ for all $i$. No observation is systematically too high or too low.
4. **Constant Variance**: $\mathbb{V}[\epsilon_i] = \sigma^2$ for all $i$. The strength of the model is the same everywhere.
5. **Independent Errors**: Knowing the error of one observation gives no information about the size of any another error.

**Note**: Constant variance is also called **homoscedasticity**. Non-constant variance is called **heteroscedasticity**.

**Note**: Nothing yet is said about the distribution of the $\epsilon_i$, i.e., we have NOT assumed normality (Yet!)
Estimation of parameters

The simple linear regression model is:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

- The regression line is uniquely defined by \( \beta_0 \) and \( \beta_1 \).
- So fitting the model amounts to finding estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), and also \( \hat{\sigma}^2 \), from available data.
- We want to find a line that “best-fits” the data, i.e., best represents the pattern of the cloud of points \((x_i, y_i)\).
- Basically, the random errors,

\[ \epsilon_i = Y_i - \mu_i = Y_i - \beta_0 - \beta_1 X_i, \]

should be as small as possible; try to **minimize the distance** between the observed values and their expected values.
Least Squares Estimation

- One approach is to **minimize** the **squared error loss function**:

\[
S(\beta_0, \beta_1) = \sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2
\]

with respect to \(\beta_0\) and \(\beta_1\).

- The **least squares estimates** comes from using this squared error loss function, which is easy to work with algebraically, and has many desirable statistical properties.

- Of course, other loss functions could be used, such as
  - **absolute loss** function (but not differentiable at 0),
  - or some other **power** function,
  - or **non-symmetric** loss functions (used if overpredictions or underpredictions are more costly).
Least Squares Estimation

To minimize $S(\beta_0, \beta_1)$, set derivatives to 0:

1. \[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) \overset{set}{=} 0
\]

2. \[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) X_i \overset{set}{=} 0
\]

Rearranging will yield the so-called normal equations:

1. \[n\beta_0 + (\sum X_i)\beta_1 = \sum Y_i\]

2. \[(\sum X_i)\beta_0 + (\sum X_i^2)\beta_1 = \sum X_i Y_i\]

Solve the normal equations to get the least squares estimates of $\beta_0$ and $\beta_1$:

1. \[\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}} = \ldots = \frac{\sum (X_i - \bar{X})Y_i}{\sum (X_i - \bar{X})^2}\]

2. \[\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}\]
Least Squares Estimation

Now we can calculate the fitted values,

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

We can also calculate the residuals, i.e., the difference between observed and fitted values,

\[ e_i = \hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \]

Note:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]
\[ Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \]
Interpretation of regression coefficients

- \( \hat{\beta}_0 \) estimates the expected value of the response variable when the predictor variable is 0.
  - This may not have a sensible interpretation if a value of zero for the predictor is impossible (extrapolation is dangerous!)
- \( \hat{\beta}_1 \) estimates the expected change in the response variable associated with a one unit change in the predictor.
  - We use associated rather than caused as we do not know that the predictor is causing the change in the response variable.
Properties of LSE

- \( \mathbb{E}(\hat{\beta}_0) = \mathbb{E}(\bar{Y} - \hat{\beta}_1 \bar{X}) = \cdots = \beta_0 \)
- \( \mathbb{E}(\hat{\beta}_1) = \mathbb{E}\left( \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} \right) = \cdots = \beta_1 \)
- This means that the LS estimators are **unbiased**.

- \( \mathbb{V}(\hat{\beta}_0) = \mathbb{V}(\bar{Y} - \hat{\beta}_1 \bar{X}) = \cdots = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right] = \cdots = \frac{\sigma^2 \sum X_i^2}{nS_{xx}} \)
- \( \mathbb{V}(\hat{\beta}_1) = \mathbb{V}\left( \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} \right) = \cdots = \sigma^2 \left[ \frac{1}{\sum (X_i - \bar{X})^2} \right] = \frac{\sigma^2}{S_{xx}} \)
- It can be shown that out of all linear unbiased estimators, the LS estimators have **minimum variance**.
Estimation of Variance of Error terms

- \( V(\epsilon_i) = \sigma^2 \) is a population parameter that also needs to be estimated:
  
  \[
  s^2 = \hat{\sigma}^2 = \frac{1}{n-2} \sum(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} \sum e_i^2 = \frac{SSE}{n-2} = MSE
  \]

- It can be shown that \( s^2 \) is an \textit{unbiased} estimator of \( \sigma^2 \).

Why \( n - 2 \) in the denominator?
Distribution of the error term

- No matter what the distribution of the error terms (as long as errors are uncorrelated with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$), the least squares estimators are unbiased and have minimum variance among all unbiased linear estimators; it is the best linear unbiased estimator (BLUE), according to the Gauss-Markov Theorem.

- However, if we want to construct any confidence intervals or prediction intervals, or hypothesis tests, we need to make an assumptions about the form of the distribution of the error terms.

- We will now assume that $\epsilon_i \sim N(0, \sigma^2)$. Why?
  - A normal error term greatly simplifies the theory of regression analysis
  - It is justifiable in many real-world situations where regression analysis is applied.

- What if normality does not hold? We could use other techniques, such as non-parametric models.