Multiple Linear Regression Basics

Laura Barnes & Julianne Quinn

University of Virginia
Charlottesville, VA 22904

Agenda

Review of Evidence Informed Systems Enigeering

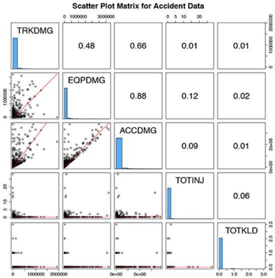
1. Evidence Informed Systems Engineering
   - Problem Description
   - Evidence-Informed Approach
   - Evidence
   - Recommendation
2. Evidence-Informed Approach
   - Hypothesis
   - Visualization or Graphical Analysis
   - Models and Analysis

Review of Data Visualization

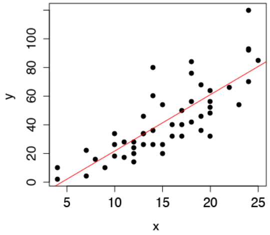
1. Univariate Observation & Visualization
   - Histograms
   - Bar Plots
   - Density Plots
   - Box Plots
   - QQ Plots
2. Multivariate Observation & Visualization
   - Scatter Plots
   - Scatter Plot Matrices
   - Categorical Variable Plots
   - Plots of Principal Components

Are Data Visualization and Simple Statistics Good Enough?

- Simple statistics are not sufficient for most engineering problems
  - Adjusting for confound variables.
  - Multiplicity: even low probability events can show significance if we do enough tests.
- Regression and ANOVA provide analytical tools for understanding, prediction, and control in engineering problems.



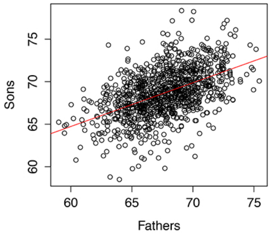
Univariate Linear Regression



- Univariate linear regression reveals the relationship between two variables
- Origin of the name "Regression"? Francis Galton- pioneer of statistics.

Regression to the Mean

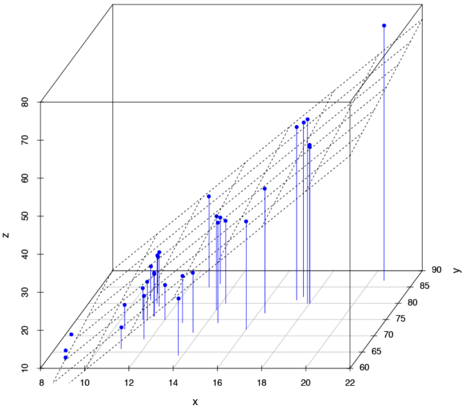
- The original paper by Galton, which regressed sons heights on the heights of fathers, exposed a common fallacy: Regression to the Mean.



- Another example: Israeli Air Force - Kahneman

Multiple Regression Summary

- Multiple Regression: A method for measuring and modeling the relationship between sets of variables.



Multiple Regression Summary

- Examples:
  - Relationship between SAT score and high school grades, gender, preparation courses, ...
  - Relationship between number of crimes and incomes, population, police, ...
  - Relationship between salary and years experience, gender, age, ...
- Relationship does not imply causation.

Linear Regression Models

- Regression models are one type of mathematical model. Models allow us to focus attention to the key elements that describe or predict a system's performance.
- Types of mathematical models:
  - Functional: $y = f(x)$
  - Stochastic: $y = f(x) + \epsilon$ where ϵ is a random variable.
- Linear regression uses stochastic models with two components:
  - Deterministic: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and
  - Stochastic: ϵ
- Matrix notation: $y = f(X) + \epsilon = X\beta + \epsilon$
  - If we have n observations, what are dimensions of y, X, β, and ϵ?

Terminology of Linear Regression Models

Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

- y: *response variable*, *predicted variable*, *regressand*,*dependent variable*, *outcome variable*
- x_i: *explanatory variable*, *predictor variable*, *regressor*, *independent variable*,*input variable*
- β_0: *intercept*
- $\beta_i(i = 1, \cdots, k)$: *regression coefficients*, *effects*
- ϵ: *error term*,*residual*,*noise*

Metric of Goal for Regression

- Regression is the solution to an optimization problem.
- Find a linear fit to the data that minimizes the sum of square errors.
- Why do we use the metric sum of square errors?

Assumptions for Optimization

- The data for the input variables or predictors, $x_1, \cdots, x_k$ are known.
- The predictors are linearly independent.
- The response variable, y, is quantitative.

Assumptions for Inference

- For a sample size of n, the distribution of the $\epsilon_i, i = 1, \cdots, n$ are independent, identical and Gaussian with
- $E(\epsilon_i) = 0$, and
- $Var(\epsilon_i) = \sigma^2$
- What's the distribution of Y_i?
  - The above assumptions imply that Y_i also have Gaussian distributions with $E(Y_i) = X_i\beta$ and $Var(Y) = \sigma^2$.
- Hence, Y is multivariate Gaussian with $E(Y) = X\beta$ and $Var(Y) = \sigma^2$

Least Squares Estimates

- We find optimal estimates for the coefficients β where the criterion is least squares.
- The optimization problem is:

$$\text{minimize} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} - y_i)^2$$

or

$$\text{minimize} (X\beta - y)^T (X\beta - y)$$

- The least square estimate?

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Estimation of Variance

- The estimate for σ^2 where k is the number of predictors or input variables:

$$\hat{\sigma}^2 = \frac{(X\hat{\beta} - y)^T (X\hat{\beta} - y)}{n - k - 1} = \frac{y^T (I - H) y}{n - k - 1}$$

- $H = X(X^T X)^{-1} X^T$ is the hat matrix: $\hat{y} = X\hat{\beta} = Hy$
- So,

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - k - 1}$$

Sum of Squares Decomposition

- The sum of squares has a convenient decomposition:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$
$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2$$

-

Total S.S. $=$ Residual S.S $+$ Model S.S

- This decomposition is important and useful. Why?

ANOVA Table and F Test

- Model Utility Test:

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_A : \beta_i \neq 0, i \in \{1, \cdots, k\}$$

- ANOVA table and F test

Source	Sum of Squares	d.f	Mean Square
Model	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$	k	$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)^2}{k}$
Residual	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-k-1$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k-1}$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	Sample Var.

Source	F	Pr(F)
Model Utility	$\frac{\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)^2}{k}}{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k-1}}$	$F_{(k,n-k-1)}$

F-test

- F-statistic with k and $n-k-1$ degress of freedom

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{k}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}$$

- The larger the F statistic, the more useful the model.

t-tests

- Hypotheses:

$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$

- t-statistic with $n - k - 1$ d.f.:

$$t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

- Check whether the particular X is useful given the presence of other variables.

Example: Equipment Damage in Train Control

- It's easy to estimate linear regression models in R: *lm*
- Predict *EQPDMG* using TEMP + TRNSPD + TONS:

Variable	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	220.3323	3181.6674	0.069	0.9448
TEMP	-95.7136	48.6011	-1.969	0.0489
TRNSPD	2917.5473	64.6640	45.119	<2e-16
TONS	9.3293	0.2346	39.763	<2e-16

- F test results: 1477 on 3 and 40087 d.f, p-value: $< 2.2e - 16$
- Interpret these coefficients and the F test result.