Regression Analysis

Aleksandr Savenkov
ols2010@med.cornell.edu

12/19/2017

Outline

- Models
- Linear Regression
  - simple
  - multivariable
- Logistic Regression
- Cox Proportional Hazards Regression

Models

- Mathematical abstraction
  - analogy of real world processes
  - sometimes can be depicted by a graph
  - usually expressed as an equation
- Parameters
  - Y a value you want to predict
  - X one of more variables that you know

Terminology

- dependent variable Y
  - response
  - outcome
  - endpoint
- independent variable X
  - covariate
  - predictor
  - risk factor
  - explanatory variable

Simple Linear Regression

- Can be used when the objective is to model *dependent variable* Y as a linear function of an *independent variable* X.

Examples:

1. Blood pressure as a function of body mass index (BMI)
2. CD4 cell count as a function of HIV RNA analysis

- Make prediction of the response variable for a fixed value of the independent variable

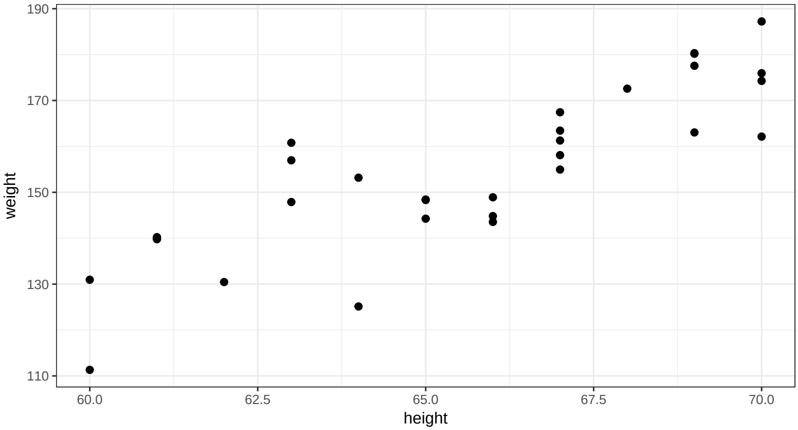
Linear Regression Assumptions

- Simple linear regression model

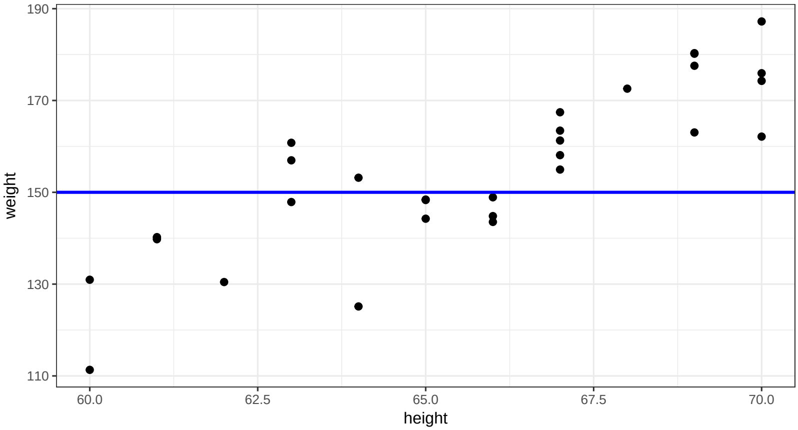
$$Y = \alpha + \beta X + \epsilon$$

- Error terms ϵ are assumed to be:
  - independent
  - have mean 0
  - have common variance
  - normally distributed

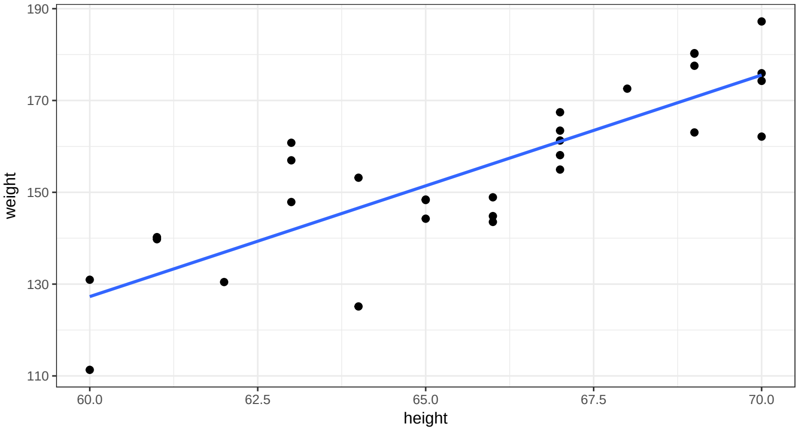
Example



Example (cont.)



Example (cont.)



Example (cont.): best fit line

- Model

$$\text{Weight} = \alpha + \beta \times \text{Height} + \epsilon$$

or

$$Y = \alpha + \beta X + \epsilon$$

- Minimize the vertical distances from data to line

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X = -162 + 4.83X$$

- use line to predict weight for height $= 70$

$$\hat{Y} = -162 + 4.83 * 70 = 176 \text{ lbs.}$$

Interpretation

- α - intercept. The value of Y when X is zero (when within scope)

Ex.: -162 lbs ???

- β - slope. The average change in Y for every one unit change in X

Ex.: 4.83, for every change of an inch in height, there is an average increase in weight of 4.83 lbs

Multiple Regression

- Model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

- No longer a line
- Geometrically is a hyper-plane

Multiple Regression: Ex. (Framingham Offspring Study)

- Association between BMI and SBP
- A total of $n = 3538$ participants were examined
- SBP:
  - mean $= 127.3$
  - sd $= 19.0$
- BMI:
  - mean $= 28.2$
  - sd $= 5.3$

Ex. (cont) Simple Linear Regression

	Regression coefficient	t-statistic	P-value
Intercept	108.28	62.61	0.0001
BMI	0.67	11.06	0.0001

$$\text{SBP} = 108.28 + 0.67 \times \text{BMI}$$

- It appears as though BMI is significantly associated with SBP ($p = 0.0001$). For a one unit increase in BMI, there is a 0.67 mmHG increase in SBP on average.

Ex. (cont) Multiple Linear Regression

- Potential confounders:
  - age (continuous)
  - gender
    - male
    - female
  - hypertension treatment
    - yes (1)
    - no (0)

Ex. (cont) Multiple Linear Regression

	regression coefficient	t-statistic	p-value
intercept	68.15	26.33	<0.0001
BMI	0.58	10.30	0.0001
age	0.65	20.22	<0.0001
male gender	0.94	1.58	0.11
hypertension treatment	6.44	9.74	0.0002

$$\hat{\text{SBP}} = 68.15 + 0.58 * \text{BMI} + 0.65 * \text{age} + 0.94 * \text{gender} + 6.44 * \text{hypertension Rx}$$

Ex. (cont) Interpretation

- BMI

one unit change in BMI increases SPB by 0.57 mmHG on average, holding all other variables constant (significant)

- Age

one year increase in age increases SPB by 0.65 mmHG on average, holding all other variables constant (significant)

- gender

men have a 0.94 mmHG higher SBP on average, holding all other variables constant (not significant)

- Hypertension treatment

hypertension treatment reduces SBP by 6.44 mmHG on average, holding all other variables constant (significant)

Ex. (cont.) Prediction

- SBP prediction (estimate) for
  - female
  - age 50
  - BMI of 25
  - not being treated for hypertension

$$\hat{\mathrm{SBP}} = 68.15 + 0.58 * 25 + 0.65 * 50 + 0.94 * 0 + 6.44 * 1 = 115.15$$

Unadjusted vs Adjusted Variables

- Unadjusted variable
  - simple (univariable) linear regression
  - BMI coefficient $= 0.67$
- Adjusted variable
  - multiple linear regression
  - BMI coefficient $= 0.58$

Interaction

- Definition
  - when a variable has a different effect on the outcome depending on the values of another variable
  - relationship differs between a variable of interest and outcome for different values of another variable
- Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Interaction

- Test $H_0 : \beta_3 = 0$ Vs. $H_a : \beta_3 \neq 0$
- If $p-$value is significant then assume there is a significant interaction
- If an interaction, do a separate analysis in each group (if one of the variables is categorical)

Interaction: some math

- Let X_2 - gender (1 - male, 0 - female)

$$Y = (\beta_0 + \beta_2 X_2) + (\beta_1 + \beta_3 X_2) X_1$$

- If $\beta_3 = 0$ then

$$Y = (\beta_0 + \beta_2 X_2) + \beta_1 X_1$$

Same slope for both genders and different intercepts

- If $\beta_3 \neq 0$ then slope for female is β_1 and $(\beta_1 + \beta_3)$ for males

Multivariable vs Multivariate Regression

- Multivariate regression
  - The simultaneous analysis of multiple endpoints
- Multivariable regression
  - The analysis of multiple explanatory variables in simultaneous association with the outcome variable.

Logistic Regression

- Outcome: binary(0/1)
- Independent variables(predictors/covariates) either continuous and/or categorical
- Examples
  - Mortality (yes/no)
  - Morbidity (yes/no)
  - Success or failure of treatment

Logistic Regression

- We model $P(Y = 1|X = x)$, since $0 \leq P \leq 1$ linear model does not apply
- Transformation

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

- the logit can take values between $-\infty$ and ∞

Logistic regression

- Model:

$$\text{logit}(P(X = x)) = \beta_0 + \beta_1 x$$

or

$$P(X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Odds ratio

- we can rewrite previous equation as follows:

$$\frac{P(X = x)}{1 - P(X = x)} = \exp(\beta_0 + \beta_1 x)$$

- Assume X takes values 0 or 1 (ex. gender)
- Odds ratio

$$\frac{P(X = 1)/(1 - P(X = 1))}{P(X = 0)/(1 - P(X = 0))} = \exp(\beta_1)$$

Odds ratio (cont.)

- Odds ratio
  - one unit change in X yields change in OR
  - OR $= 1$ (no association)
  - OR > 1 (as X increases, the odds increase)
  - OR < 1 (as X increases, the odds decreases)

Ex.

- Data
  - low - birth weight $<$ 2500g
  - age - mother's age
  - lwt - mother's weight in pounds at last menstrual period.
  - race - maternal race
  - smoke - if smoked during pregnancy (1- yes, 0 - no)
  - ptl - premature labor history
  - ht - hypertension history (1 - yes, 0 - no)
  - ui - uterine irritability (y/n)
  - ftv - number of visits to a physician during 1st trimester
  - bwt - birth weight

Data

The data frame has 189 rows and 10 columns.

	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
85	0	19	182	2	0	0	0	1	0	2523
86	0	33	155	3	0	0	0	0	3	2551
87	0	20	105	1	1	0	0	0	1	2557
88	0	21	108	1	1	0	0	1	2	2594
89	0	18	107	1	1	0	0	1	0	2600
91	0	21	124	3	0	0	0	0	0	2622

Simple model

$$\text{logit}(P(Y = 1|\text{smoke})) = \beta_0 + \beta_1 * \text{smoke}$$

term	estimate	std.error	statistic	p.value
(Intercept)	-1.0870515	0.2147338	-5.062322	0.0000004
smoke	0.7040592	0.3196423	2.202647	0.0276196

$\exp(0.704) = 2.02$

In other words, smoking doubled the odds of having a low birth weight baby compared to women who did not smoke during pregnancy

Cox regression: intro

- the idea is similar to linear or logistic regression, but this model deals with survival time data with censoring
- allows one to compare two or more survival profiles (e.g., treatments) while “controlling” for other demographic/prognostic factors of interest.
- different from linear or logistic regression in the sense it models hazard function
- allows to analyze the effect of several risk factors on survival

Hazard

- Fundamental quantity is the **hazard function**

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

The hazard function h(x) is the instantaneous rate at which events occur for subjects who are surviving at time t.

Cox regression model (Cox, 1972)

Semiparametric regression model

$$h(t, X) = h_0(t)\exp(X'\beta) = h_0(t)\exp\left[\sum_{i=1}^{p}\beta_i X_i\right]$$

- $h_0(t)$ - arbitrary unspecified baseline hazard function
- $X = (X_1, X_2, ..., X_p)$ - covariates
    - Ex. X_1 - age, X_2 - gender, etc.
- β vector of unknown parameters
    - $\beta_j > 0$ adverse effect of a covariate X_j on survival
    - $\beta_j = 0$ no effect
    - $\beta_j < 0$ beneficial effect (protective effect)
- baseline hazard: $X_1 = X_2 = ...X_p = 0$

$$h(t, 0) = h_0(t)\exp(0) = h_0(t)$$

Proportional hazards model

- Consider two individuals with vectors of covariates X_1 and X_2.
- The ratio of their hazards is

$$\frac{h(t, X_1)}{h(t, X_2)} = \frac{h_0(t)\exp(X_1'\beta)}{h_0(t)\exp(X_2'\beta)} = \exp\left[\sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i})\right]$$

- Consider one covariate, $X = 1$ - treatment, $X = 0$ - control, then
$$\frac{h(t, X = 1)}{h(t, X = 0)} = \exp(\beta)$$
- individuals in the treatment arm experience event at $\exp(\beta)$ times rate of those individuals in control arm throughout the study period

Cox-PH with One Continuous Covariate

Let *Age* is a covariate of interest. Then hazard is given by:

$$h(t, \text{Age}) = h_0(t)\exp(\beta_1 \text{Age})$$

Hazard for an individual with $Age = a$

$$h(t, \text{Age} = a) = h_0(t)\exp(\beta_1 a)$$

Hazard for an individual with $Age = a + 1$

$$h(t, \text{Age} = a + 1) = h_0(t)\exp\left[\beta_1(a + 1)\right]$$

Hazard ratio for 1 year increase in age

$$\frac{h(t, \text{Age} = a)}{h(t, \text{Age} = a + 1)} = \frac{h_0(t)\exp\left[\beta_1(a + 1)\right]}{h_0(t)\exp(\beta_1 a)} = \exp(\beta_1)$$

Proportional Hazards Assumption (PHA)

Under the proportional hazards assumption, the hazard ratio does not vary with time

- PHA is vital to the interpretation and use PH models
- Evaluating the PHA
  - Observed VS. predicted
  - -log(-log) plot
  - Schoenfeld residuals

Summary

ALWAYS check models assumptions