Propulsion Academy

Reading notes

Day 4 - Statistics and Experimental design

Contents

1 Linear regression 2

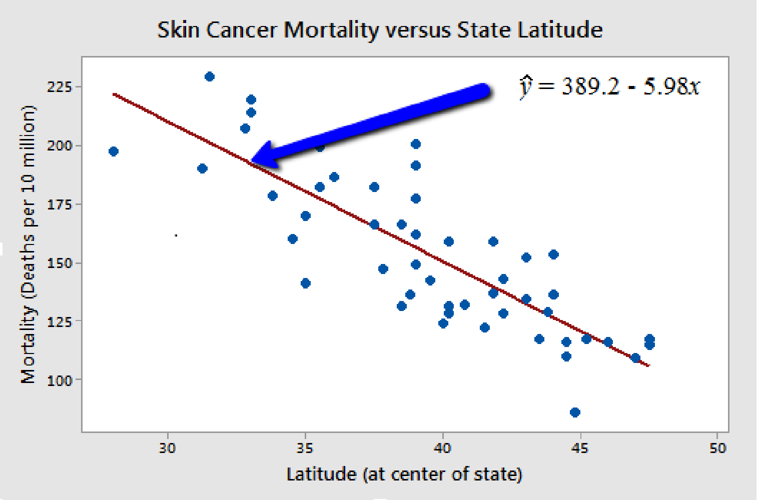
2 Generalized Linear Regression Model (GLM) 8

1 Linear regression

Linear regression is a statistical method that allows us to summarize and study relationships between two types of variables:

- *response*, *outcome*, or *dependent variable*, denoted by y: focus of a question in a study or experiment.
- *predictor*, *explanatory*, or *independent variable*, denoted by x: is one that explains changes in that variable.

Simple linear regression gets its adjective "simple", because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we study later in this course, gets its adjective "multiple," because it concerns the study of two or more predictor variables.
The following figure shows an example of a statistical relationship, which could be of interest. The response variable y is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable x is the latitude (degrees North) at the center of each of 49 states in the U.S. (skincancer.txt)



You might anticipate that if you lived in the higher latitudes of the northern U.S., the less exposed you'd be to the harmful rays of the sun, and therefore, the less risk you'd have of death due to skin cancer. The scatter

plot supports such a hypothesis. There appears to be a negative linear relationship between latitude and mortality due to skin cancer, but the relationship is not perfect. Indeed, the plot exhibits some "trend“, but it also exhibits some "scatter“. It describes a statistical relationship (not a deterministic one).
Some other examples of statistical relationships might include: Height and weight, alcohol consumed and blood alcohol content, vital lung capacity and pack-years of smoking, driving speed and gas mileage.
Since we are interested in summarizing the trend between two quantitative variables, the natural question arises - "what is the best fitting line“.
In order to examine which of the two lines is a better fit, we first need to introduce some common notation:

- y_i denotes the observed response for experimental unit i
- x_i denotes the predictor value for experimental unit i
- $\hat{y_i}$ is the predicted response (or fitted value) for experimental unit i

Simple Regression Model

Then, the equation for the best fitting line is:

$$\hat{y_i} = \beta_0 + \beta_1 x_i$$

The "experimental unit“ denoted by i, is the object or person on which the measurement is made (in the example above: i: "state“).
If we now want to summarize the average skin cancer mortality across the latitude of the state, we can be described by the regression line,

$$\mu_Y = E(Y) = \beta_0 + \beta_1 x,$$

summarizing the trend between the predictor x and the mean of the responses μ_Y. β_0 is the intercept and β_1 the slope of the regression line.
In the example above the trend is described by $\mu_y = 389.2 - 5.98x$. We could also express the average skin cancer mortality for the ith-state:

$E(Y_i) = \beta_0 + \beta_1 x_i$. Of course, the skin cancer mortality will not equal the average $E(Y_i)$. There will be some error. That is, any student's response y_i will be the linear trend $\beta_0 + \beta_1 x_i$ plus some error ϵ_i. So, another way to write the simple linear regression model is $\hat{y}_i = E(Y_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$. For the error terms of the *simple regression model* the following conditions are defined:

- The mean of the response, $E(Y_i)$, at each at each value of the predictor, x_i, is a Linear function of the x_i
- The errors, ϵ_i, are independent.
- The errors, ϵ_i, at each value of the predictor, x_i, are normally distributed.
- The errors, ϵ_i, at each value of the predictor, x_i, have equal variances (denoted σ^2).

A regression line will always contain error terms because, in reality, independent variables are never perfect predictors of the dependent variables. There are many uncontrollable factors in the business world. The error term exists because a regression model can never include all possible variables; some predictive capacity will always be absent, particularly in simple regression.

Least squares

The true regression line is usually not known, however it can be estimated by estimating the coefficients β_0 and $\beta_1 1$ from an observed data set The typical procedure for finding the line of best fit is called the least-squares method. It's approach is to minimizing the sum of the squared deviations between the data and the model. The least squares estimates of the parameters would be computed by minimizing:

$$\begin{aligned} Q &= \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 \\ &= \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1)]^2 \end{aligned}$$

Doing this by taking partial derivatives of Q with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, setting each partial derivative equal to zero, and solving the resulting system of two equations with two unknowns yields the following estimators for the parameters:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Prediction

Once this line is determined, it can be extended beyond the historical data to predict future levels of product awareness, given a particular level of advertising expenditure. The extension of the line of regression requires the assumption that the underlying process causing the relationship between the two variables is valid beyond the range of the sample data. Regression is a powerful business tool due to its ability to predict future relationships between variables such as these.

Interpretation

The interpretation of a simple regression model can be approached as follows. Based on the regression line $\hat{y} = 389.2 - 5.98x$, the coefficient β_1 indicates that if x is increased by one unit, the expected decrease (because of the minus) in response is 5.98.
If we look at the example above, we can argue that for an latitude of 35, we would expect to have a skin cancer mortality of: $389.2 - 5.98 * 35 = 179.9$ Deaths per 10 Million.

Assumptions

1. The dependent variable should be measured on a continuous scale (i.e., it is either an interval or ratio variable). Examples of variables that meet this criterion include revision time (measured in hours),

intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

2. The independent variable (one for simple LM, two or more for multiple LM), which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable).

3. Observations need to be independent of observations (i.e., independence of residuals).

4. There needs to be a linear relationship between the dependent variable and each of your independent variables, and the dependent variable and the independent variables collectively. Testing can be done by visually inspecting these scatter plots and partial regression plots to check for linearity. If the relationship displayed in your scatter plots and partial regression plots are not linear, you will have to either run a non-linear regression analysis or ”transform“ your data.

5. Data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line.

6. There should be no significant outliers, high leverage points or highly influential points. Outliers, leverage and influential points are different terms used to represent observations in your data set that are in some way unusual when you wish to perform a regression analysis. These different classifications of unusual points reflect the different impact they have on the regression line. An observation can be classified as more than one type of unusual point. However, all these points can have a very negative effect on the regression equation that is used to predict the value of the dependent variable based on the independent variables.

7. Finally, you need to check that the residuals (errors) are approximately normally distributed. Two common methods to check this assumption include using: (a) a histogram (with a superimposed normal curve) and a Normal P-P Plot; or (b) a Normal Q-Q Plot of the studentized residuals.

Multiple Linear Regression Model (LM)

In practice we often want to explore the relationship between more than one explanatory variables and a response variable. This can be easily done by fitting a linear equation to observed data using a multiple regression approach.
Formally the model is outlined by,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_p x_{ip} + \epsilon_i, \qquad \text{for} i = 1, 2, \ldots, n, \tag{1}$$

with n observations and p denoting the number of explanatory variables included.
Each β-coefficient represents the change in the mean response, $E(y)$, per unit increase in the associated predictor variable when all the other predictors are held constant.
The intercept term,β_0, represents the mean response, $E(y)$, when all the predictors, are all zero (which may or may not have any practical meaning).
An important assumption for multiple linear regression is that, the data must not show multicollinearity, which occurs when you have two or more independent variables that are highly correlated with each other. This leads to problems with understanding which independent variable contributes to the variance explained in the dependent variable, as well as technical issues in calculating a multiple regression model.

Sources and further Readings

Overview: http://onlinestatbook.com/2/regression/intro.html
Simple linear regression (1):
https://onlinecourses.science.psu.edu/stat501/node/251/
Simple linear regression (2):http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis
Mean squares: https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd431.htm
Assumptions Multiple linear regression: https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php

Further reading: diagnostic plots:
https://data.library.virginia.edu/diagnostic-plots/
Examples: http://statlab.stat.yale.edu/workshops/IntroRegression/StatLab-IntroRegressionFa08.pdf

2 Generalized Linear Regression Model (GLM)

In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. In other words generalized linear models (GLMs) are a framework for modeling response variables that are bounded or discrete. This is used, for example:

- when modeling positive quantities (e.g. prices or populations) that vary over a large scale—which are better described using a skewed distribution such as the log-normal distribution or Poisson distribution (although GLMs are not used for log-normal data, instead the response variable is simply transformed using the logarithm function)

- when modeling categorical data, such as the choice of a given candidate in an election (which is better described using a Bernoulli distribution/binomial distribution for binary choices, or a categorical distribution/multinomial distribution for multi-way choices), where there are a fixed number of choices that cannot be meaningfully ordered

- when modeling ordinal data, e.g. ratings on a scale from 0 to 5, where the different outcomes can be ordered but where the quantity itself may not have any absolute meaning (e.g. a rating of 4 may not be ”twice as good” in any objective sense as a rating of 2, but simply indicates that it is better than 2 or 3 but not as good as 5).

The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

In many statistical computing packages parameter estimation is perfomed based on a maximum-likelihood estimation.

In a generalized linear model (GLM), each outcome Y of the dependent variables is assumed to be generated from a particular distribution in the exponential family, a large range of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. The model is given by a linear perdictor defined by:

$$\eta_i = \beta_0\beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i,$$

which is also sometimes denoted in the matrix notation by $\eta_i = \mathbf{X}\beta$. The mean, μ, of the distribution depends on the independent variables, X, through

$$E(Y_i) = \mu = g^{-1}(\mathbf{X}\beta)$$

where $E(Y)$ is the expected value of Y; $\mathbf{X}\beta$ is the linear predictor, a linear combination of unknown parameters β and g is the link function.

In this framework, the variance is typically a function, V, of the mean:

$$\mathrm{Var}(Y) = V(\mu) = V(g^{-1}(\mathbf{X}\beta)).$$

The unknown parameters, β, are typically estimated with maximum likelihood, maximum quasi-likelihood, or Bayesian techniques. Maximum likelihood estimation (MLE) is used to estimate the parameters (and not ordinary least squares (OLS)), and thus relies on large-sample approximations.

In summary there are three components to any GLM:

- *Random Component:* referring to the probability distribution of the response variable Y. (Remember in LM-Framework we assumed $e_i \sim N(0, \sigma^2)$.
- *Systematic Component:* specifies the explanatory variables $(X_1, X_2, \ldots, X_p)$ in the model, more specifically their linear combination in creating the so called linear predictor $(\beta_0 + \beta_1 X_1, \ldots, \beta_p X_p)$

- *Link Function:* denoted by g, specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; $g(E(Y_i)) = g(\mu_)) = \eta_i$.

In the GLM framework we make the following assumptions:

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.
- The dependent variable Y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables.
- The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure, and overdispersion (when the observed variance is larger than what the model assumes) maybe present.
- Errors need to be independent but NOT normally distributed.

Logistic Regression

In statistics, the logistic model (or logit model) is a statistical model that is usually taken to apply to a binary dependent variable. The model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor. In other words, the goal is to model the probability of a particular outcome as a function of the predictor variable(s): $E(\frac{Y}{n}X) = \pi$.

- Random component: The distribution of Y is Binomial

- Systematic component: Xs are explanatory variables (can be continuous, discrete, or both) and are linear in the parameters $\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$
- Link function: Logit (must map from $(0, 1) \to (-\infty, \infty)$.

$$\eta = \text{logit}(\pi) = \log(\frac{\pi}{1-\pi})$$

Based on the before identified components the model can be written as,

$$\log(\frac{\pi}{1-\pi})\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$$

and

$$\begin{aligned}\frac{\pi}{1-\pi} &= \exp \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} \\ &= \exp \beta_0 * \exp \beta_1 x_{i,1} * \exp \beta_2 x_{i,2} * \cdots * \exp \beta_p x_{i,p}\end{aligned}$$

Binary logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non case.
The result can be interpreted according to value of $\exp \beta_i$, if it is 1 the odds (for success/something being present) is the same at all level of x_i. Is $\exp \beta_i > 1$, the odds are increased by a factor of $\exp \beta_i$ (if $<$ the odds are decreased).

Poisson Regression

In statistics, Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters.
Poisson regression models are generalized linear models with the logarithm as the (canonical) link function, and the Poisson distribution function as

the assumed probability distribution of the response. In other words, the goal is to model the probability of a particular outcome as a function of the predictor variable(s): $E(Y|X) = \lambda$.
In summary, the three components of the Poisson regression can be describe as follows:

- Random component: The distribution of counts Y is Poisson
- Systematic component: Xs are discrete variables used in cross-classification, and are linear in the parameters formula
- Link Function: $log\eta = log\lambda$, because it map from $(0, \infty) \to (-\infty, \infty)$.

Based on the before identified components the model can be written as,

$$\log \lambda = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$$

and

$$\begin{aligned}\lambda &= \exp \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} \\ &= \exp \beta_0 * \exp \beta_1 x_{i,1} * \exp \beta_2 x_{i,2} * \cdots * \exp \beta_p x_{i,p}\end{aligned}$$

The result of fitting such a Poisson Regression Model can be interpreted as follows: the expected counts (outcome) will be increased/decreased by a factor $\exp \beta_i$ when x_i is increased by one unit and all other variables stay constant.
Negative binomial regression is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model.
Poisson regression may also be appropriate for rate data, where the rate is a count of events divided by some measure of that unit's exposure (a particular unit of observation). For example, biologists may count the number of tree species in a forest: events would be tree observations, exposure would be unit area, and rate would be the number of species per unit area. Demographers may model death rates in geographic areas as the count of deaths divided by person-years. More generally, event rates can be calculated as events per unit time, which allows the observation window to

vary for each unit. In these examples, exposure is respectively unit area, person-years and unit time. In Poisson regression this is handled as an *offset*, where the exposure variable enters on the right-hand side of the equation, but with a parameter estimate (for log(exposure)) constrained to 1. The expected outcome to be modeled would be denoted as $E(\frac{Y|X}{\text{offset}})$.

Sources and further Readings

Logistic regression (1):
`https://en.wikipedia.org/wiki/Logistic_regression`
Logistic regression (2): `https://stats.idre.ucla.edu/stata/output/logistic-regression-analysis/` Poisson regression:
`https://en.wikipedia.org/wiki/Poisson_regression`