Simple Linear Regression

October 1, 2017

Motivation

One of the most common goals in statistics is to answer the question

Is the variable X related to the variable Y? If so, what is the relationship and can we use it to predict Y?

- X is called the predictor or explanatory or independent variable or the covariate
- Y is called the response or dependent variable

Linear Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

β_0 is the intercept term

β_1 is the slope of the line

ϵ is a random error term usually assume normally distributed
$\epsilon \sim N(0, \sigma^2)$

Model for a Random Sample

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, \ldots, n$$

ϵ_i independent, identically distributed with mean $= 0$ and variance $= \sigma^2$

$E(Y_i) = \beta_0 + \beta_1 X_i$

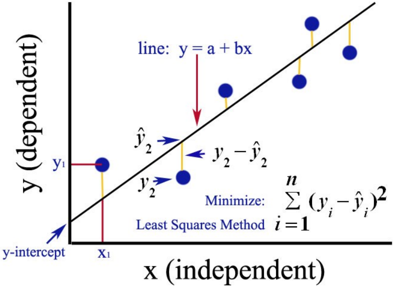
$Var(Y_i) = \sigma^2$

If we assume ϵ_i *iid* $N(0, \sigma^2)$ then Y_i is also normally distributed.

Regression Line

The regression line is a line fit to the (X_i, Y_i) pairs that provides a "best" fit. One method of finding estimates of β_0 and $\beta_1$is to minimize the sum of squared deviances from the observed points to the line. This is called ordinary least squares regression. That is, we minimize

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$



Regression Line

We can also fit a "best" line by minimizing the sum of the absolute value of the differences,

$$\sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 X_i|$$

It is not as straight forward to minimize this function since we are not able to take derivatives. This method is called least absolute value regression.

Finding parameter estimates

OLS - ordinary least squares

The least squares estimates of β_0 and β_1 are found by taking the derivative of Q with respect to β_0 and β_1, setting equal to zero and solving for the parameter estimates.

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}X_i(Y_i - \beta_0 - \beta_1 X_i)$$

Finding parameter estimates

OLS - ordinary least squares

Set the derivatives equal to zero and replace β_0 and β_1 with their estimates b_0 and b_1 we get the normal equations

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

Finding parameter estimates

OLS - ordinary least squares

Solving these equations simultaneously we get

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

These are the least squares estimates for any distribution. We can show these are the maximum likelihood estimates when we assume the error terms are normally distributed.

Fitted Values and Residuals

The fitted values are the predicted values of Y_i.

$$\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \dots, n$$

The residuals are the difference between the observed and fitted Y's

$$e_i = Y_i - \hat{Y}_i \;\; i = 1, \dots, n$$

Clearly if the model fits well, we expect e_i's to be small.

Sum of Squares

SSE $= \sum_{i=1}^{n} e_i^2$ is the sum of squares due to error and is the minimum possible value of Q.

SST $= SSY = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ is the total variation of the observations about their mean.

SSR $= SST - SSE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$is the sum of squares due to regression. It is the amount of variation in the data that is explained by the regression.

Coefficient of Determination

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

is the proportion of variation in the data due to the regression.

$0 \leq r^2 \leq 1$

The bigger r^2 is the more significant the regression is.

Estimation of σ_2

An unbiased estimator of σ^2, the variance of the random error term is

$$s^2 = \frac{SSE}{n-2}$$

Inference on b_0 and b_1

We can show the distribution of b_0 and b_1 are:

$$b_0 \sim N\left(\beta_0, \frac{\sigma^2 sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

We can use these distributions to test hypotheses about the regression coefficients. In particular a test of $H_0 : \beta_1 = 0$ tests whether there is a significant linear relationship between X and Y.

Analysis of Variance - ANOVA

Source of	Sum of			
Variation	Squares	df	MS	F
Regression	SSR	1	MSR=SSR/1	MSR/MSE
Error	SSE	n-2	MSE=SSE/(n-2)	
Total	SST	n-1		

The F value is used to test for significant regression.

Under the null hypothesis $H_0 : \beta_1 = 0$, both MSR and MSE are estimates of σ^2 the random variation.

Under H_0 $F \sim F_{1,n-2}$

Reject H_0 if F is too big.

Prediction

For X^* a value of X not one of the original X_i, we can predict the value of Y associated with X^* as

$$\hat{Y}^* = b_0 + b_1 X_*$$

The mean and variance of the predicted $\hat{Y}^*$ are

$$E(\hat{Y}^*) = \beta_0 + \beta_1 X^*$$

$$var(\hat{Y}^*) = \frac{\sigma^2}{n} + (X^* - \bar{X})^2 \frac{\sigma^2}{SXX}$$

Regression Diagnostics

Residuals

An initial plot of Y versus X should be examined during exploratory data analysis in order to check the linear relationship between them.

The residuals, $e_i = Y_i - \hat{Y}_i$ are estimates of ϵ_i if the model is correct.

- $E(e_i) = 0$ $var(e_i) \approx \sigma^2$
- $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} X_i e_i = 0$ from the normal equations, so the e_i's are NOT independent

Regression Diagnostics

Residual Plots

If the assumed model is correct, the residuals should be randomly scattered around zero with no obvious pattern. We can use plots of the residuals to check the assumptions of the model.

- Checking for Linearity e_i *vs* X_i If this plot deviates from the random scattering around zero, the linear model is not adequate - a more complex model is warranted.
- Checking for constant variance e_i *vs* $\hat{Y}_i$. If this plot deviates from the random scattering around zero it could indicate non-constant variance, i.e. variance is related to X. Look for a funnel pattern.
- Checking the normal assumption A qqplot of the residuals can be used to check the normality assumption.

Violation of Assumptions

For large n the normal assumption is the least critical.

If the model is not linear in X b_0, b_1 and s^2 are biased estimators.

If the variance is not constant $var(b_0)$ and $var(b_1)$ are not as given.

If observations are not independent, $var(b_0)$ and $var(b_1)$ are not as given.

In any of these cases, confidence intervals, t-tests and F-tests may not be valid.

What Can Be Done?

Variance stabilizing transformations can help correct non-constant variance.

Some transformations can improve the normality of the data.

If observations are not independent due to serial or spatial dependence, time series and spatial models should be used rather than regression.

If the linear model is not sufficient, higher order terms may be added.