Lecture

INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: FY BSc

Subject : Probability and Statistics -2

Subject Code: PUSASQ 1.2

Chapter: Unit 4 Chp 1

Chapter Name: Generalised Linear Models

Index

1 Introduction

- Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features').

- The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

- Regression analysis is primarily used for two conceptually distinct purposes.
  - First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.
  - Second, in some situations regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset. To use regressions for prediction or to infer causal relationships, respectively, a researcher must carefully justify why existing relationships have predictive power for a new context or why a relationship between two variables has a causal interpretation. The latter is especially important when researchers hope to estimate causal relationships using observational data.

2 Simple Bivariate Linear Model

Model Specification

- Given a set of n pairs of data (x_i , y_i), i = 1, 2,…,n , the y_i are regarded as observations of a response variable Y_i . For the purposes of the analysis the x_i , the values of an explanatory variable, are regarded as constant.

- The simple linear regression model (with one explanatory variable):

- The response variable Y_i is related to the value x_i by:

$$\boldsymbol{Y_i = \alpha + \beta x_i + e_i \quad i = 1, 2, \dots, n}$$

- where the e_i are uncorrelated error variables with mean 0 and common variance σ^2.

- So $E[e_i] = 0, var[e_i] = \sigma^2, i = 1,2, \dots, n$

- β is the slope parameter, α the intercept parameter.

- This is equivalent to saying that y = mx + c , where m is the gradient or slope and c is the intercept ie where the line crosses the y-axis.

2 Simple Bivariate Linear Model

Fitting the model

- We can estimate the parameters in a regression model using the 'method of least squares'.

- Fitting the model involves:
  (a) estimating the parameters β and α, and
  (b) estimating the error variance σ^2.

- The fitted regression line, which gives the estimated value of Y for a fixed x, is given by:

$$\hat{y} = \hat{\alpha} + \hat{\beta}\, x$$

- where $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$

- These are the equations we use to calculate the 'best' values of α and β. They are given in the Tables.

3 Partitioning the variability of the responses

- To help understand the 'goodness of fit' of the model to the data, the total variation in the responses, as given by $S_{yy} = \sum(y_i - \bar{y})^2$ should be studied.

- Some of the variation in the responses can be attributed to the relationship with x (eg y may tend to be high when x is high, low when x is low) and some is random variation (unmodellable) above and beyond that. Just how much is attributable to the relationship – or 'explained by the model' – is a measure of the goodness of fit of the model.

3 Partitioning the variability of the responses

- We start from an identity involving y_i (the observed y value), $\bar{y}$ (the overall average of the y values) and $\hat{y}_i$ (the 'predicted' value of y).

- Squaring and summing both sides of:

$$\boldsymbol{y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})}$$

- gives:

$$\boldsymbol{\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2}$$

- the cross-product term vanishing.

- The sum on the left is the 'total sum of squares' of the responses, denoted here by SS_{TOT}.

3 Partitioning the variability of the responses

- The second sum on the right is the sum of the squares of the deviations of the fitted responses (the estimates of the conditional means) from the overall mean response (the estimate of the overall mean) – it summarises the variability accounted for, or 'explained' by the model. It is called the 'regression sum of squares', denoted here by SS_{REG}.

- The first sum on the right is the sum of the squares of the estimated errors (response – fitted response, generally referred to in statistics as a 'residual' from the fit) – it summarises the remaining variability, that between the responses and their fitted values and so 'unexplained' by the model. It is called the 'residual sum of squares', denoted here by SS_{RES} . The estimate of σ^2 is based on it – it is $\frac{\boldsymbol{SS_{RES}}}{\boldsymbol{n-2}}$.

3 Partitioning the variability of the responses

- So:

$$\boldsymbol{SS_{TOT} = SS_{RES} + SS_{REG}}$$

- Note that SS_{RES} is often also written as SS_{ERR} ('error').
- For computational purposes $SS_{TOT} = Syy$ and:

$$\boldsymbol{SS_{REG} = \sum\left[\left(\hat{\alpha} + \hat{\beta}x_i\right) - \left(\hat{\alpha} + \hat{\beta}\bar{x}\right)\right]^2 = \hat{\beta}^2 S_{xx} = \frac{S_{XY}^2}{S_{XX}}}$$

- The last step uses the fact that $\hat{\beta} = S_{XY}/S_{XX}$.
- So $\boldsymbol{SS_{RES} = S_{YY} - \frac{S_{XY}^2}{S_{XX}}}$

3 Partitioning the variability of the responses

- It can then be shown that:

$$E[SS_{TOT}] = (n-1)\sigma^2 + \beta^2 S_{XX} \qquad E[SS_{REG}] = \sigma^2 + \beta^2 S_{XX}$$

- from which it follows that $E[SS_{RES}] = (n-2)\sigma^2$.

- Hence:

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n-2}\left(S_{YY} - \frac{S_{XY}^2}{S_{XX}}\right)\right] = E\left[\frac{SS_{RES}}{n-2}\right] = \frac{1}{n-2}E[SS_{RES}] = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$$

- So $\hat{\sigma}^2$ is an unbiased estimator of σ^2 .

3 Partitioning the variability of the responses

- In the case that the data are 'close' to a line (|r| high – a strong linear relationship) the model fits well, the fitted responses (the values on the fitted line) are close to the observed responses, and so SS_{REG} is relatively high with SS_{RES} relatively low.

- r is referring to Pearson's correlation coefficient.

- In the case that the data are not 'close' to a line (|r| low – a weak linear relationship) the model does not fit so well, the fitted responses are not so close to the observed responses, and so SS_{REG} is relatively low and SS_{RES} relatively high.

- The proportion of the total variability of the responses 'explained' by a model is called the coefficient of determination, denoted R^2 . Here, the proportion is:

$$\boldsymbol{R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{S_{XY}^2}{S_{XX}S_{YY}}}$$

- [The value of the proportion R^2 is usually quoted as a percentage].
- R^2 can take values between 0% and 100% inclusive.

4 The full normal model and inference

- The model must be specified further in order to make inferences concerning the response based on the fitted model. In particular, information on the distribution of the Y_i's is required.

- In the full model, we now assume that the errors, e_i , are independent and identically distributed as N(0,σ^2) variables. This will then allow us to obtain the distributions for β and the Y_i's. We can then use these to construct confidence intervals and carry out statistical inference.

- For the full model the following additional assumptions are made:
- The error variables e_i are:
  (a) independent
  (b) normally distributed

- Under this full model, the e_i's are independent, identically distributed random variables, each with a normal distribution with mean 0 and variance σ^2. It follows that the Y_i's are independent, normally distributed random variables, with $E[Y_i] = \alpha + \beta x_i$ and $\text{var}[Y_i] = \sigma^2$.

4 The full normal model and inference

- $\widehat{\mathrm{B}}$, being a linear combination of independent normal variables, itself has a normal distribution, with mean and variance as noted earlier.

- The further results required are:
  (1) $\widehat{\mathrm{B}}$ and $\hat{\sigma}^2$are independent
  (2) $\frac{(\boldsymbol{n}-\mathbf{2})\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}}{\boldsymbol{\sigma}^{\mathbf{2}}}$ has a χ^2 distribution with $\nu = n - 2$.

- Note: With the full model in place the Y_i's have normal distributions and it is possible to derive maximum likelihood estimators of the parameters α , , and (since maximum likelihood estimation requires us to know the distribution whereas least squares estimation does not). It is possible to show that the maximum likelihood estimators of and are the same as the least squares estimates, but the MLE of has a different denominator to the least squares estimate.

4 The full normal model and inference

Inferences on the slope parameter $\boldsymbol{\beta}$

- To conform to usual practice the distinction between $\hat{B}$, the random variable, and its value $\hat{\beta}$, will now be dropped. Only one symbol, namely $\hat{\beta}$ will be used.

- Using the fact that $E(\hat{\beta}) = \beta$ and $var\,(\hat{\beta}) = \sigma^2/S_{XX}$:

- $\boldsymbol{A} = \dfrac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\left(\dfrac{\boldsymbol{\sigma}^2}{\boldsymbol{S_{XX}}}\right)^{1/2}}$ is a standard normal variable

- $\boldsymbol{B} = (\boldsymbol{n} - \mathbf{2})\hat{\boldsymbol{\sigma}}^{\mathbf{2}}/\boldsymbol{\sigma}^{\mathbf{2}}$ is a χ^2 variable with $\nu = n - 2$ degrees of freedom

4 The full normal model and inference

- Now, since $\hat{\beta}$ and $\hat{\sigma}^2$ are independent, it follows that $\frac{\boldsymbol{A}}{\left\{\frac{\boldsymbol{B}}{\boldsymbol{n-2}}\right\}^{\mathbf{1/2}}}$ has a t distribution with $\nu = n - 2$, ie:
- $(\hat{\beta} - \beta)/se(\hat{\beta})$ has a t distribution with $\nu = n - 2 \rightarrow$ (Result A)
- where the symbol $se(\hat{\beta})$ denotes the estimated standard error of $\hat{\beta}$, namely $\left(\frac{\hat{\sigma}^2}{S_{XX}}\right)^{\mathbf{1/2}}$.
- Result (A) can now be used for the construction of confidence intervals, and for tests, on the value of β , the slope coefficient in the model. $H_0: \beta = 0$ is the 'no linear relationship' hypothesis.
- Note that since $\hat{\beta} = \frac{S_{XY}}{S_{XX}}$ and $r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$ if $\hat{\beta} = 0$ then $S_{XY} = 0$ and $r = 0$ too.

5 Estimating a mean response and predicting an individual response

(a) Mean response

- If μ_0 is the expected (mean) response for a value x_0 of the explanatory variable (ie $\mu_0 = E[Y \mid x_0] = \alpha + \beta x_0$), μ_0 is estimated by $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} x_0$, which is an unbiased estimator.
  The variance of the estimator is given by:

$$\mathbf{var}(\hat{\boldsymbol{\mu}}_0) = \left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\}\sigma^2$$

- The distribution actually used is a t distribution - the argument is similar to that described earlier:

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0)/\mathbf{se}[\boldsymbol{\mu}_0] \text{ has a } t \text{ distribution with } v = n - 2 \text{ – Result A}$$

- where $\mathrm{se}[\hat{\mu}_0]$ denotes the estimated standard error of the estimate, namely:

$$\mathbf{se}[\hat{\boldsymbol{\mu}}_0] = \left[\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\}\hat{\sigma}^2\right]^{\frac{1}{2}}$$

- Result A can be used for the construction of confidence intervals for the value of the expected response when $x = x_0$.

5 Estimating a mean response and predicting an individual response

(b) Individual response

- Rather than estimating an expected response $E[Y \mid x_0]$ an estimate, or prediction, of an individual response y_0 (for $x = x_0$) is sometimes required. The actual estimate is the same as in (a), namely:

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$$

- but the uncertainty associated with this estimator (as measured by the variance) is greater than in (a) since the value of an individual response y_0 rather than the more 'stable' mean response is required. To cater for the extra variation of an individual response about the mean, an extra term σ^2 has to be added into the expression for the variance of the estimator of a mean response.

- In other words, the variance of the individual response estimator is:

$$\mathrm{var}(\hat{y}_0) = \left\{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\}\sigma^2$$

5 Estimating a mean response and predicting an individual response

- The result is:

$$(\hat{\boldsymbol{y}} - \boldsymbol{y_0})/\mathbf{se}[\hat{\boldsymbol{y}}_{\mathbf{0}}] \text{ has a } t \text{ distribution with } v = n - 2 \text{ – Result B}$$

- where $\mathrm{se}[\hat{y}_0]$ denotes the estimated standard error of the estimate, namely:

$$\mathbf{se}[\hat{\boldsymbol{y}}_{\mathbf{0}}] = \left[\left\{\mathbf{1} + \frac{\mathbf{1}}{\boldsymbol{n}} + \frac{(\boldsymbol{x_0} - \bar{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{s_{xx}}}\right\}\boldsymbol{\sigma}^{\mathbf{2}}\right]^{\mathbf{1/2}}$$

- Result B can then be used for the construction of confidence intervals (or prediction intervals) for the value of a response when $x = x_0$.
- The resulting interval for an individual response y_0 is wider than the corresponding interval for the mean response μ_0.
- Recall that for an individual response value we have $y_i = \alpha + \beta x_i + e_i$, which is the regression line $\alpha + \beta x_i$ plus an error term, e_i. Since $e_i \sim N(0, \sigma^2)$ an individual point is on the regression line on average - hence we have the same estimate $\hat{\alpha} + \hat{\beta} x_0$ as for the mean response. However, we can see that there is an additional σ^2 for the variance.

6 Checking the model

- The residual from the fit at x_i is the estimated error, the difference between the response y_i and the fitted value ie:

$$\text{residual at } x_i \text{ is } \hat{e}_i = y_i - \hat{y}_i$$

- By examining the residuals it is possible to investigate the validity of the assumptions in the model about (i) the true errors e_i (which are assumed to be independent normal variables with means 0 and the same variance σ^2), and (ii) the nature of the relationship between the response and explanatory variables.
- Plotting the residuals along a line may suggest a departure from normality for the error distribution. The sizes of the residuals should also be looked at, bearing in mind that the value of σ estimates the standard deviation of the error distribution.
- Ideally we would expect the residuals to be symmetrical about 0 and no more than 3 standard deviations from it. So skewed residuals or outliers would indicate non-normality.
- Alternatively a quantile-quantile (Q-Q) plot of the residuals against a normal distribution should form a straight line. They are far superior to dotplots.
- Scatter plots of the residuals against the values of the explanatory variable (or against the values of the fitted responses) are also most informative. If the residuals do not have a random scatter - if there is a pattern - then this suggests an inadequacy in the model.

7 Extending the scope of the linear model

- In certain 'growth models' the appropriate model is that the expected response is related to the explanatory value through an exponential function

$$\boldsymbol{E[Y_i \mid x_i] = \alpha \exp(\beta x_i).}$$

- In such a case the response data can be transformed using $w_i = \log y_i$ and the linear model:

$$\boldsymbol{W_i = \eta + \beta x_i + e_i(\text{ where } \eta = \log \alpha)}$$

- is then fitted to the data (x_i, w_i). The fact that the error structure is additive in this representation implies that it plays a multiplicative role in the original form of the model. If such a structure is considered invalid, different methods from those covered in this chapter would have to be used.

8 The multiple linear regression model

Introduction

- We will now extend our linear regression model. Previously we examined the relationship between Y , the response (or dependent) variable and one explanatory (or independent or regressor) variable X . We now look at k explanatory variables, $X_1, X_2, \dots, X_k$.
- There are many problems where one variable can quite accurately be predicted in terms of another. However, the use of additional relevant information should improve predictions. There are many different formulae used to express regression relationships between more than two variables. Most are of the form:

$$\boldsymbol{E[Y|X_1, X_2, \dots X_k] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

- As with the simple linear regression model discussed earlier Y is a random variable whose values are to be predicted in terms of given data values $x_1, x_2 , \dots, x_k$.
- $\beta_1, \beta_2, \dots, \beta_k$ are known as the multiple regression coefficients. They are numerical constants which can be determined from observed data.

8 The multiple linear regression model

Fitting the model

- As for the simple linear model, the multiple regression coefficients are usually estimated by the method of least squares.
- The response variable Y_i is related to the values $x_{i1}, x_{i2}, \dots\ x_{ik}$ by

$$\boldsymbol{Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \qquad i = 1, \dots, n}$$

- and so the least squares estimates of $\alpha, \beta_1, \beta_2, \dots, \beta_k$ are the values $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ for which:

$$q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [\boldsymbol{Y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}]^{\boldsymbol{2}} \quad \textit{is minimised}$$

- As for the simple linear model, to find the estimates the above is differentiated partially with respect to α and $\beta_1, \beta_2, \dots, \beta_k$ in turn and the results are equated to zero.

9 R^2 in the multiple regression case

- In the bivariate case we noted that the proportion of the total variation of the responses 'explained' by a model, called the coefficient of determination, denoted R^2 , was equal to the square of the correlation coefficient between the dependent variable Y and the single independent variable X.
- In the case of multiple regression with a single dependent variable, Y, and several independent variables, $x_1, x_2, \dots, x_k$, R^2 measures the proportion of the total variation in Y 'explained' by the combination of explanatory variables in the model.
- The value of R^2 lies between 0 and 1. It will generally increase (and cannot decrease) as the number of explanatory variables k increases. If R^2 =1 the model perfectly predicts the values of Y:
- 100% of the variation in Y is "explained" by variation in $x_1, x_2 , \dots, x_k$.
- Because R^2 cannot decrease as more explanatory variables are added to the model, if it is used alone to assess the adequacy of the model, there will always be a tendency to add more explanatory variables. However, these may increase the value of R^2 by a small amount, while adding to the complexity of the model. Increased complexity is generally considered to be undesirable.

9 R^2 in the multiple regression case

- To take account of the undesirability of increased complexity, computer packages will often quote an 'adjusted R^2 ' statistic. This is a correction of the R^2 statistic which is based on the mean square errors (ie the residual mean sum of squares, MSS_{RES}) and takes account of the number of predictors, k, and the number of data points the model is based on. If we have k predictors, and n observations:

$$\boldsymbol{Adjusted\ R^2 = 1 - \frac{MSS_{RES}}{MSS_{TOT}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)}$$

- So MSS_{RES}/MSS_{TOT} would give a measure of how much variability is explained by the residuals (or errors) and takes values between 0 and 1. Hence 1 - MSS_{RES}/MSS_{TOT} would give a measure of how much variability is explained by the regression model. Therefore it is similar measure to the original coefficient of determination, R^2.

- Recall that the mean sum of squares (MSS) is the sum of squares divided by the degrees of freedom. So $MSS_{RES} = SS_{RES}/(n - k - 1)$ and $MSS_{TOT} = SS_{TOT}/(n - 1)$.

- The model which maximises the 'adjusted R^2 ' statistics can be regarded in some sense as the 'best' model. Note, however, that the 'adjusted R^2 ' cannot be interpreted as the proportion of the variation in Y which is 'explained' by variation in the $x_1, x_2, \ldots, x_k$.

10 The full normal model and inference

The full normal model

- Again, to make inferences concerning the responses based on the fitted model, we need to specify the model further. We make the same assumptions as for the linear model:
- In the full model, we now assume that the errors, e_i , are independent and identically distributed $N(0, \sigma^2)$ random variables. This will then allow us to obtain the distributions for β and the Y_i 's.
- We can then use these to construct confidence intervals and carry out statistical inference. The error variables e_i are: (a) independent, and (b) normally distributed.
- Under this full model, the e_i 's are independent, identically distributed random variables, each with a normal distribution with mean 0 and variance σ^2. It follows that the Y_i 's are independent, normally distributed random variables, with:

$$\boldsymbol{E[Y_i] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad and \quad var[Y_i] = \sigma^2}$$

- This mimics the bivariate linear regression model but with the mean dependent on k explanatory variables.

10 The full normal model and inference

Testing hypotheses on individual covariates

- In multiple regression the coefficients $\beta_1, \beta_2, \dots, \beta_k$ describe the effect of each explanatory variable on the dependent variable Y after controlling for the effects of other explanatory variables in the model.
- Each coefficient β_j measures the increase in the value of the response variable y for a corresponding increase in the value of x_j independent of the other covariates.
- As in the bivariate case, hypotheses about the values of $\beta_1, \beta_2, \dots, \beta_k$ can be tested, notably the hypothesis $\beta_i = 0$ which states that, after controlling for the effects of other variables, the variable x_i has 'no linear relationship' with Y.
- Recall that in the bivariate case a hypothesis of $\beta = 0$ was equivalent to $\rho = 0$.
- Generally speaking, it is not useful to include in a multiple regression model a covariate x_i for which we cannot reject the hypothesis that $\beta_i = 0$.

11 Estimating a mean response and predicting an individual response

Mean response

- The whole point of the modelling exercise is so that we can estimate values of the response variable Y given the input variables $x_1, x_2, \dots, x_k$.
- Mean response
- As with the linear model we can estimate the expected (mean) response, μ_0, for a multiple linear regression model given a vector of explanatory variables, $\underline{x}_0$.

$$\boldsymbol{\mu_0 = E\left[Y \middle| \underline{x}_0\right] = \alpha + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k}}$$

- μ_0 is estimated by $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$ which is an unbiased estimator.
- Recall that our multivariate linear regression model stated that the Y_i's are independent, normally distributed random variables, with $E[Y_i] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + B_k x\_ik$. We have simply used this expected value to obtain an estimated mean response corresponding to the vector x_0.
- We are using vector notation here:

$$\boldsymbol{x_0 = (x_{01}, x_{02}, \dots, x_{0k})}$$

11 Estimating a mean response and predicting an individual response

Individual response

- Similarly, we could predict an individual response y_0 (for $\underline{x} = \underline{x}_0$) using the same estimate $\hat{y}_0 = \hat{\alpha} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$ but with an extra σ^2 in the expression for the variance of the estimator compared to the mean response.
- Recall that for an individual response value we have $\mathrm{y_i} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + B_k x_{ik} + e_i$. Each individual response value is associated with an error term from the regression line. Since $e_i \sim \mathrm{N}(0,\sigma^2)$ an individual point is on the regression line on average – hence we have the same estimate $\hat{\alpha} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$ as for the mean response. However, there is an additional σ^2 for the variance.

Thank You!