Linear Models Using R

Suresh,R
Research Scholar
Department of Statistics
Bangalore University
Bangalore-560 056

OVERVIEW

- Correlation Studies
- Regression Analysis
- Analysis of Variance(ANOVA)
- Analysis of Covariance(ANCOVA)

Correlation Analysis

- Correlation
- Correlation coefficient
  1. Pearson correlation coefficient
  2. Spearman rank correlation Rho
  3. Kendal's Tau
- Significance test for correlation coefficient

Correlation

Correlation analysis pertains to the study of interdependence or co-variation of variables.

It will show us the degree to which variables are linearly related.

It yields one number-an index designed to give an immediate picture of how closely two variables move together.

Pearson Correlation Coefficient

$$r = \frac{\operatorname{cov}(x, y)}{\sqrt{s_x^2 s_y^2}}.$$

Spearman's Rank Correlation

$$r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

Kendall's Tau

To describe how to compute tau in a more formal manner, let $K_{ij} = 1$ if the ith and jth pairs of observations are concordant, otherwise $K_{ij} = -1$. Next, sum all of the K_{ij} for which $i < j$, which is denoted by

$$\sum_{i<j} K_{ij}.$$

Then Kendall's tau is given by

$$\hat{\tau} = \frac{2\sum_{i<j} K_{ij}}{n(n-1)}, \tag{13.7}$$

where τ is a lower case Greek tau. In words, Kendall's tau is just the average of the K_{ij} values. (In the notation used here, the number of K_{ij} values is $n(n-1)/2$.) As previously indicated, $\hat{\tau}$ has a value between -1 and 1. If $\hat{\tau}$ is positive, there is a tendency for Y to increase with X—possibly in a nonlinear fashion—and if $\hat{\tau}$ is negative, the reverse is true.

Illustrative Example :

Lets us compute the various correlations measures for this data set

Amount Of Fertilizer	Yield of Pototoes
0	8.34
4	8.89
8	9.16
12	9.5

Significance of Correlation Coefficient

$$H_0 : \rho = 0,$$

$$T = r\sqrt{\frac{n-2}{1-r^2}}.$$

Test Statistic

If the null hypothesis is true and the assumptions are met, then T has a Student's t-distribution with $n-2$ degrees of freedom. That is, if the goal is to have the probability of a Type I error equal to α, reject if $|T| \geq t$, where t is the $1-\alpha/2$ quantile of Student's t-distribution, which is read from table 4 in appendix B.

Regression Analysis

- Regression analysis is used for explaining or modeling the relationship between a single variable ***Y***, *called* the *response, output or dependent variable, and one or more predictor, input, independent or explanatory* variables, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. When $\mathbf{p} = 1$, it is called simple regression but when $\boldsymbol{p} > 1$ it is called multiple regression or sometimes multivariate regression. When there is more than one ***Y***, *then it is called multivariate* multiple regression which we won't be covering here.

- Regression analyses have several possible objectives including

1. Prediction of future observations.
2. Assessment of the effect of, or relationship between, explanatory variables on the response.
3. A general description of data structure.

Regression Analysis

- Simple Linear Regression
  1. Model building
  2. Validation of the fitted model
  3. Inference in regression
  4. Transformation
  5. Missing Observations

- Multiple Linear Regression
  1. Model Building
  2. Multicollinearity
  3. Variable Selection

Simple Regression Model

- Model $$y = \beta_0 + \beta_1 x + \varepsilon,$$

- Assumptions of the errors
  1. Independence,
  2. Constant Variance
  3. Normality of the errors,
- Model Adequacy check: Residual plots

Model Adequacy checks

- Residuals plots
  1. Residuals vs Fitted
  2. Sqrt(Standardized) vs Fitted
  3. Q-Q plot of residuals

- Cooks Distance vs Index-----To Handle Influencing observations

- Hat values vs Index----- Leverage points

Influence and Leverage

- Influential observations - An influential point is one whose removal from the dataset would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of these two properties.

- Leverage - A leverage point is unusual in the predictor space. it has the potential to influence the fit.

Cook's Statistic(To handle Influential Observations)

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{p\hat{\sigma}^2}$$

$$= \frac{1}{p} r_i^2 \left[\frac{h_i}{1 - h_i} \right]$$

residual effect

leverage

Cook's Statistic

The value $4/(n-p-1)$ has been suggested as a rough cut-off for note worthy values of D_i.

Any $D_i > (4/(n-p-1))$ will need some attention.

More on Influence observations

- Use the following commands

model=lm(y~x, data)

Object name=influence.measures(model)

summary(Object name)

influence.measures()- package “stats”

Leverage

- The $h_i = H_{ii}$ are called leverages and are useful diagnostics. Since var $\hat{\varepsilon}_i = \sigma^2(1-h_i)$, a large leverage, h_i , will make var($\hat{\varepsilon}_i$) small. The fit will be "forced" close to y_i

Since $\sum_i h_i = p$,an average value for h_i is $\frac{p}{n}$.

A "rule of thumb" is that the leverages > $2\frac{p}{n}$ should be looked at more closely. Large values of h_i are due to extreme values in *X*.

Autocorrelation

We assume that the errors are uncorrelated, but for temporally or spatially related data this may well be untrue.

Durbin-Watson Test

Test Statistic $DW = \frac{\sum_{i=2}^{n}(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}$ $\underset{\text{Ho}}{\sim} \chi^2$

If you do have correlated errors, you can use generalized least squares.

Tackling Missing Observations problem

Missing data occur when some values of some cases are missing. This is not uncommon. Dealing with missing data is time consuming. Fixing up problems caused by missing data sometimes takes longer than the analysis.

Missing Observations

- **Here are several fix-up methods**

1. Delete the case with missing observations. This is OK if this only causes the loss of a relatively small number of cases. This is the simplest solution.(Will be considered in this discussion)
2. Fill in or *impute the missing values. Use the rest of the data to predict the missing* values. Simply replacing the missing value of a predictor with the average value of that predictor is one easy method.
3. Maximum likelihood methods can be used assuming the multivariate normality of the data. The EM algorithm is often used here.

Inference in Regression

$$H_0 : \beta_1 = \beta_{10}$$

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}$$

Test Statistic

Reject H_0 if $|t_0| > t_{\alpha by 2, n-2}$

Confidence Intervals for β

$$\hat{\beta}_1 - t_{\alpha/2, n-2}\, se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2, n-2}\, se(\hat{\beta}_1)$$

Prediction

- Predicting the future response: A future observation y_o is predicted to be

$$x_0^T \hat{\beta}$$

Where x_0 is the given value of x.

- Confidence intervals for predictions:

100(1–%) % CI for a single future response is:

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

Transformation

Transformation: Transformations of the response and predictors can improve the fit and correct violations of model assumptions.

Variable transformations serve a variety of purposes in data analysis, and are used in particular

- to make distributions more symmetric or normal
- to stabilize spread (variation), and
- to render relationships between variables more nearly linear.

These purposes frequently, but not necessarily, harmonize with one another.

Box-Cox Family of Transformations (1964)

T A B L E 13.1. Values of Certain Power Functions for Five Benchmark Powers

λ	Y^λ	$W = (Y^\lambda - 1)/\lambda$
1	Y	$Y - 1$
$\frac{1}{2}$	$Y^{1/2}$	$2(Y^{1/2} - 1)$
0	1(?)	$\ln Y$
$-\frac{1}{2}$	$Y^{-1/2}$	$2(1 - Y^{-1/2})$
-1	Y^{-1}	$1 - Y^{-1}$

Illustrative Example(Montgomery et.al (2003))

Temperature	Viscosity
24.9	1.133
35	0.9772
44.9	0.8532
55.1	0.755
65.2	0.6723
75.2	0.6021
85.2	0.542
95.2	0.5074

More on BOX-COX transformation

Try the following code to choose lambda using maximum likelihood.

```
boxcox(model,lambda=seq(0.1,0.4,by=0.05),plotit=T)
```

This function is available in the package "Modern Applied Statistics with S" (MASS)

NOTE: Beware of Transformations.....! Unless it's necessary, don't go for it.

Multiple Linear Regression

- Model $$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon.$$

- Assumptions of the errors
  1. Independence
  2. Constant Variance
  3. Normality of the errors

- Model adequacy checks: same as for simple regression.

Illustrative Example (Montgomery et.al (2003))

observation number	Delivery Time	Number of Cases	Distance
1	16.68	7	560
2	11.5	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.5	5	605
11	40.33	16	688
.	.	.	.
.	.	.	.
.	.	.	.
25	10.75	4	150

Multicollinearity

- Often in multiple regression model, several of the independent variables are measures of similar phenomena. This can result in a high degree of correlation among the independent variables. This condition is known as Multicollinearity.
- Collinearity leads
  - Imprecise estimates of β - the signs of the coefficients may be misleading.
  - t-tests will fail to reveal significant factors.
  - Missing importance of predictors.

Illustrative Example(Der and Everitt(2003)-uscrime data 47 states)

obs num	R	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
1	79.1	151	1	91	58	56	510	950	33	301	108	41	394	261
2	163.5	143	0	113	103	95	583	1012	13	102	96	36	557	194
3	57.8	142	1	89	45	44	533	969	18	219	94	33	318	250
4	196.9	136	0	121	149	141	577	994	157	80	102	39	673	167
5	123.4	141	0	121	109	101	591	985	18	30	91	20	578	174
6	68.2	121	0	110	118	115	547	964	25	44	84	29	689	126
7	96.3	127	1	111	82	79	519	982	4	139	97	38	620	168
8	155.5	131	1	109	115	109	542	969	50	179	79	35	472	206
9	85.6	157	1	90	65	62	553	955	39	286	81	28	421	239
10	70.5	140	0	118	71	68	632	1029	7	15	100	24	526	174
11	167.4	124	0	105	121	116	580	966	101	106	77	35	657	170
12	84.9	134	0	108	75	71	595	972	47	59	83	31	580	172
13	51.1	128	0	113	67	60	624	972	28	10	77	25	507	206
14	66.4	135	0	117	62	61	595	986	22	46	77	27	529	190
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
47	84.9	130	0	121	90	91	623	1049	3	22	113	40	588	160

Variance Inflation Factor

- The Variance Inflation Factor (VIF_j) for the j-th variable is given by

$$VIF_j = \frac{1}{(1-R_j^2)}$$

where R_j^2 is the square of the multiple correlation coefficient from the regression of the j-th explanatory variable on the remaining regressors.

There is a universally accepted criterion for establishing the magnitude of a VIF value necessary to identify serious multicollinearity.

It has been proposed that VIF > 10 serve this purpose .

Ridge Regression

The following function will give u a ridge regression estimates .

lm.ridge() ------ package “MASS”

The below function gives u the ridge trace plot.

matplot() ---- package(graphics)

```
matplot (fit$lambda, t(fit$coef), type="l",lty=1,
xlab=expression (lambda), ylab=expression (hat (beta)))
```

Explore these functions and ENJOY ...!

Variable Selection

Backward elimination.

1. Start with all the predictors in the model
2. Remove the predictor with highest p-value greater than alpha critical.
3. Refit the model and goto 2
4. Stop when all p-values are less than alpha critical.

Other methods

- Forward Selection
- Stepwise Regression

Robust Regression

Try out the following functions instead of "lm"

rlm() - package "MASS"

ltsreg() – package "lqs"

A treatment for outliers without removing them, and when errors are nonnormal, particularly when errors are fat or long tailed

ANOVA

Given a factor α occurring at $i = 1,\ldots,I$ levels, with $j = 1,\ldots,J_i$ observations per level. We use the model

Model $\quad y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1,\ldots,I \quad j = 1,\ldots,J_i$

As it stands not all the parameters are identifiable and some restriction is necessary:

Assumptions of ANOVA

Four assumptions underlie all analysis of variance (ANOVA) using the *F* statistic. These are:

(1) Random sampling from the source population;
(2) Independent measures within each sample, yielding uncorrelated response residuals;
(3) Homogeneous variances across all the sampled populations;
(4) Normal distribution of the response residuals around the model.

$$H_0: \mu_1 - \mu_2 = \cdots, = \mu_t,$$

$$H_1: \text{at least one equality is not satisfied}$$

Example

Sl no	Sand	clay	loam
1	6	17	13
2	10	15	16
3	8	3	9
4	6	11	12
5	14	14	15
6	17	12	16
7	9	12	17
8	11	8	13
9	7	10	18
10	11	13	14

Analysis of Covariance

- ANCOVA refers to regression problems where there is a mixture of quantitative and qualitative predictors.

Model $$Y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}, \quad i = 1, \ldots t, \quad j = 1, \ldots, r,$$

Two further assumptions apply to analysis of covariance (ANCOVA):

(5) Repeatable covariate values that are fixed by the investigator;
(6) Linear relation of the response to the covariate.

Some Examples for ANCOVA

(1) In agriculture, an experiment is done to examine the effect of treatments on yield. However, it is known that the density/plot of the plants could affect yield, and this variable is used as a covariate.

(2) In a nutrition experiment where the growth of laboratory rats is tracked, the initial weight of the rats influences their growth rate, and could be used as a covariate.

- In each of the above examples the candidate covariates satisfies two conditions:
  - The covariate is related to the response, and can account for variation in the response.
  - The covariate is not related to the treatment.

Example (just to motivate)

Suppose we are interested in the effect of a medication on cholesterol level. We might have two groups; one receives the medication and the other, the default treatment. We could not treat this as a simple two-sample problem if we knew that the two groups differed with respect to age, as this would affect the cholesterol level. For the patients who received the medication, the mean reduction in cholesterol level was 0% while for those who did not, the mean reduction was 10%. So superficially it would seem that it would be better not to be treated. However, the treated group ranged in age from 50 to 70 while those who were not treated ranged in age between 30 and 50.

References

- *Casella ,G.,* Statistical Design , Springer (2008)
- *Cook,R,D.,* Regression Graphics, Wiley (1998)
- *Cook and Weisberg, S,* Residual and Influence in Regression, Chapman and Hall (1982)
- *Faraway, J.J.,* Practical Regression and ANOVA using R , Wiley (2002)
- *Faraway, J.J.,* Linear Models with R, Chapman and Hall(2005)
- *Everitt, B,S., and Hothron,T.,* A Handbook of Statistical Analysis Using R, ??????????????
- *Montgomery et al.,* Introduction to Linear Regression Analysis, Wiley (2003)

THANK YOU