Analysis of General Linear Model

Dr. Mutua Kilai

Department of Pure and Applied Sciences

2024-01-05

Introduction

In this lecture we shall:

- Define general linear model
- Discuss model building for Simple Linear Model and Multiple Linear Model
- Model Selection and Validation
- Variable selection including stepwise and best subset regression

R Packages and Datasets to use

Motivation for Modeling

- The structural form of the model describes the patterns of interactions or associations in data.
- Inference for the model parameters provides a way to evaluate which explanatory variable(s) are related to the response variable(s) while statistically controlling for the other variables
- Estimated model parameters provide measures of the strength and importance of effects.
- A model's predicted values "smooth" the data - That is, they provide good estimates of the mean of the response variable.

Data for Regression Analysis

- Data for regression analysis may be obtained from non-experimental or experimental studies.
- **Observational data** are data obtained from non-experimental studies. Such studies do not control the explanatory or predictor variable(s) of interest.
- For example, company officials wished to study the relation between age of employee (X) and number of days of illness last year (Y)
- Such data are observational data since the explanatory variable, age, is not controlled.
- A major limitation of observational data is that they often do not provide adequate information about cause-and-effect relationships.

Data for Regression Analysis

- Frequently, it is possible to conduct a controlled experiment to provide data from which the regression parameters can be estimated.
- When control over the explanatory variable(s) is exercised through random assignments, as in the productivity study example, the resulting experimental data provide much stronger information about cause-and-effect relationships than do observational data.

Overview of Steps in Regression Analysis

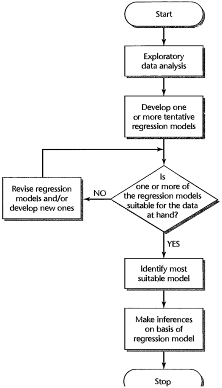


Figure 1: Typical Strategy for regression analysis

General Linear Model

- The term ‘general‘ linear model usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors.
- Two important concepts are mainly described in linear models
- **Dependent variable** The outcome that our model aims to explain usually denoted by Y
- **Independent variable** The variable we wish to use in order to explain the dependent variable. Denoted by X

Simple Linear Model

- The simple regression model can be used to study the relationship between two variables.
- A random experiment is repeated n times under identical conditions. For each trial $i = 1, 2, ..., n$ the value of X_i is known and the response Y_i is recorded.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

- In a simple linear model we have one dependent and one independent variable.

Deriving the OLS

- Given

$$Y_i = \beta_0 + \beta_1 X_i$$

- The sum of squared errors:

$$Q = \sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)$$

- Differentiating w.r.t β_0 and β_1 we have:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i)$$

Cont'd

- $$\frac{\partial Q}{\partial \beta_1} = -2\sum X_i(Y_i - \beta_0 - \beta_1 X_i)$$
- We can expand the equations and have:

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \tag{2}$$

$$\sum X_i Y_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 \tag{3}$$

Cont'd

- Solving the above simultaneously we have:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

- The fitted values of Y is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- The residuals for observation i is the difference between actual and its fitted value.

$$e_i = Y_i - \hat{Y}_i$$

Sum of Squares

- The total sum of squares denoted by SST

$$SST = \sum(Y_i - \bar{Y})^2$$

- The Sum of Squares due to regression:

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2$$

- The Sum of Squares due to errors:

$$SSE = \sum(\hat{e}_i)^2$$

- This implies that:

$$SST = SSE + SSR$$

Goodness of Fit

- The R-squared of the regression sometimes called the coefficient of determination.

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

- R^2 is the fraction of sample variation in Y that is explained by X.
- When interpreting R^2 we multiply by 100. R^2 is the

Gauss Markov Theorem

- Under the assumptions of Simple Linear Regression, the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have a minimum variance among all linear unbiased estimators of β_0 and β_1.
- Thus $\hat{\beta}_0$ and $\hat{\beta}_1$ are said to be BLUE.

Linear Estimators

- The least squares intercept and the slope are linear estimators in the sense that they are linear function of Y_i
- Consider:

$$\hat{\beta}_1 = \frac{\sum(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

- can be written as:

$$\hat{\beta}_1 = \sum m_i Y_i$$

where $m_i = \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$ and $\sum m_i = 0$ and $summ_i x_i = 1$

Unbiasedness of Estimators

- $$E[\hat{\beta}_0] = E[\bar{Y} - \hat{\beta}_1\bar{X}] \\ = \hat{\beta}_0 + \hat{\beta}_1\bar{X} - \hat{\beta}_1\bar{X} \\ = \hat{\beta}_0 \tag{4}$$

- $$E[\hat{\beta}_1] = \sum m_i E[Y_i] \\ = m_i(\beta_0 + \beta_1 X_i) \\ = \beta_1 \tag{5}$$

Variances of Estimators

- $$\begin{aligned} Var(\hat{\beta}_1) &= Var(\sum m_i Y_i) \\ &= \sum m_i^2 Var(Y_i) + \sum\sum k_i k_j cov(Y_i, Y_j) \\ &= \sigma^2 \frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^4} \\ &= \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \end{aligned} \quad (6)$$

Cont'd

- $$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{Y}) + \bar{X}Var(\beta_1) - 2\bar{X}Cov(\bar{Y}, \beta_1) \\ &= \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \\ &= \sigma^2(\frac{1}{n} + \frac{x^2}{\sum(X_i - \bar{X})^2}) \end{aligned} \tag{7}$$

Covariances

- The covariance between β_0 and β_1 is
- $$\begin{aligned} Cov(\beta_0, \beta_1) &= Cov(\bar{Y}, \beta_1) - \bar{X} Var(\beta_1) \\ &= -\frac{\bar{X}\sigma^2}{\sum(X_i - \bar{X})^2} \end{aligned} \tag{8}$$

Inference for β_1

- We test the hypothesis concerning β_1:

$$H_0 : \beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$$

- The sampling distribution of $\hat{\beta}_1$ refers to the different values of $\hat{\beta}_1$ that would be obtained with repeated sampling when the levels of the predictor X are held constant from sample to sample
- An estimate for σ^2 is:

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

thus

$$S^2(\hat{\beta}_1) = \frac{MSE}{\sum(X_1 - \bar{X})^2}$$

Cont'd

- If Y_i are normally distributed then the distribution of $\hat{\beta}_1$ is normal since $\hat{\beta}_1 = \sum m_i Y_i$ and a linear combination of independent random variables are also normally distributed then:

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2})$$

- The $(1-\alpha)100$

$$\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2}),n-2}\sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}}$$

- To test the hypothesis $H_0 : \beta_1 = c$ the test statistic is:

$$t = \frac{\beta_1 - c}{\sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}}}$$

Inference for β_0

- The sampling distribution of β_0 is:

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}))$$

- The $(1-\alpha)100\%$ CI for β_0 is

$$\hat{\beta}_0 \pm t_{(1-\frac{\alpha}{2}),n-2}\sqrt{S^2(\hat{\beta}_0)}$$

- To test the hypothesis To test the hypothesis $H_0 : \beta_0 = c$ the test statistic is:

$$t = \frac{\beta_0 - c}{\sqrt{S^2(\hat{\beta}_0)}}$$

Example 1: US Consumption Expenditure

- In fpp3 package in R, a data set named us_change shows a time series of quarterly percentage changes (growth rates) of real personal consumption expenditure, y and real personal disposable income x for the US from 1970 Q1 to 2019 Q2.

EDA

```
library(plotly)
library(fpp3)
library(tidyverse)
library(knitr)
library(pander)
library(performance)
library(GGally)
us_change %>%
  pivot_longer(c(Consumption, Income), names_to = "Series")
  autoplot(value) + theme_bw()+
  labs(y = "% Change", x = "Time")
```

Cont'd

A scatter plot of consumption changes against income changes is shown in Figure 2.

```
us_change |>
  ggplot(aes(x = Income, y = Consumption)) +
  labs(y = "Consumption (quarterly % change)",
       x = "Income (quarterly % change)") +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```

Cont'd

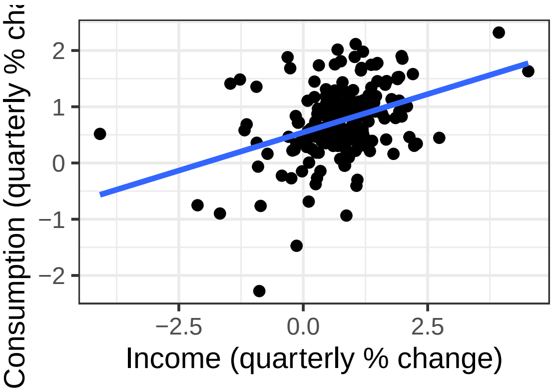


Figure 2: Scatterplot with Fitted Line

Model Fitting

- The model can be fitted using:

```
library(broom)
library(fpp3)
model <- lm(Consumption ~ Income, data = us_change)
kable(tidy(model))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.5445419	0.0540284	10.07881	0
Income	0.2718329	0.0467285	5.81728	0

- The fitted equation is:

$$\hat{Y} = 0.545 + 0.272X$$

ANOVA

```r
library(broom)
library(fpp3)
model <- lm(Consumption ~ Income, data = us_change)
kable(tidy(anova(model)))
```

term	df	sumsq	meansq	statistic	p.value
Income	1	11.80141	11.8014130	33.84075	0
Residuals	196	68.35183	0.3487338	NA	NA

Confidence Interval

```r
library(broom)
library(fpp3)
model <- lm(Consumption ~ Income, data = us_change)
kable(confint(model))
```

	2.5 %	97.5 %
(Intercept)	0.4379903	0.6510935
Income	0.1796776	0.3639881

Labwork One

Consider the data frame named `marketing` in the `datarium` package containing the impact of three advertising medias (youtube, facebook and newspaper) on sales. We want to fit a SLR to see the impact of advertising budget spent on youtube on sales.

i. Create a visualization for the two variables
ii. Fit a SLR model
iii. Obtain the 95% confidence interval and the ANOVA table for the model
iv. Interpret the results

Multiple Linear Regression

- The general multiple linear regression model can be written as:

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

- Where β_0 is the intercept
  $\beta_1, ..., \beta_k$ are the slope parameters associated with $x_1, ..., x_k$
- Consider the following multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{i,p-1} + \epsilon_i$$

- The model can be written using vectors and matrices as:

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

OLS Estimation

- $Y = n \times 1$ vector of response values $\beta = p \times 1$ vector of regression parameters $X = n \times p$ matrix of known constants $\epsilon = n \times 1$ vector of *iid* error terms
- Define the best estimate of β as that which minimizes the SSE $\epsilon'\epsilon$

$$\begin{aligned} \sum \epsilon_i^2 &= \epsilon'\epsilon \\ &= (Y - X\beta)'(Y - X\beta) \end{aligned} \tag{9}$$

- Differentiate w.r.t β and equate to zero and have:

$$Q = (Y - X\beta)'(Y - X\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

Cont'd

- $$\frac{\partial Q}{\partial \beta} = 2X'X\beta - 2Y'X$$

- Equating to zero we have:

$$-2X'Y = -2X'X\beta$$

- Solving for β we get:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Cont'd

- The fitted values are given as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... + \hat{\beta}_{p-1} X_{i,p-1}$$

- Residuals are given by:

$$e_i = Y_i - \hat{Y}_i$$

Inference in MLR

- If the model $Y = X\beta + \epsilon$ is correct, the expectation of Y is $X\beta$ and the expectation of β is:

$$
\begin{aligned}
E[\hat{\beta}] &= [(X'X)^{-1}X']E[Y] \\
&= [(X'X)^{-1}X']X\beta \qquad (10) \\
&= \beta
\end{aligned}
$$

Variance

$$
\begin{aligned}
Var(\beta) &= [(X'X)^{-1}]Var(Y)[(X'X)^{-1}X']' \\
&= [(X'X)^{-1}X']I\sigma[(X'X)^{-1}X']' \qquad (11) \\
&= \sigma^2(X'X)^{-1}
\end{aligned}
$$

ANOVA Table

Source	df	SS	MSS
Regression	p-1	SSR	MSR
Error	n-p	SSE	MSE
Total	n-1	SST	

Coefficient of Multiple Determination

- $$R^2 = \frac{SSR}{SST}$$

  It measures the amount of variation in Y explained by the independent variables.

- The adjusted R^2 is given by:

  $$R_\alpha^2 = 1 - \frac{(n-1)SSE}{(n-p)SST}$$

- It adjusts the R^2 for the number of predictors in the model.

Hypothesis testing for individual regressors

- Determine the null and alternative hypothesis
- Specify the test statistic and its distribution if H_0 is true
- Select α and determine the rejection region
- Calculate the sample value of test statistic and desired p-value
- State your conclusion

The hypothesis is

$$H_0 : \beta_k = 0 \;\; vs \;\; H_1 : \beta_k \neq 0$$

The test statistic is:

$$t = \frac{\beta_k}{se(\beta_k)} \sim t_{n-p}$$

Global test

This is an overall test for the regression model. It investigates the possibility that all the regression coefficients are equal to zero.

$$H_0 : \beta_1 = ... = \beta_k = 0 \ \ vs \ \ H_a : \beta_j \neq 0$$

The test statistic is the F -statistic given by

$$F = \frac{MSR}{MSE}$$

Assumptions of Multiple Linear Regression

Linearity

- There is a linear relationship between the dependent variable and each independent variable
- Linearity may be evaluated by constructing a scatter diagram for each independent variable and examine the diagrams
- Linearity can also be assessed graphically by constructing residual plots. Constructed by plotting residuals against the fitted values and this should exhibit no pattern

Homoscedasticity

- The variation in the residuals is the same for all fitted values of Y.
- The formal test for homoscedasticity is the Breusch Pagan test and the hypothesis is:
  H_0: Constant variance
  H_a: Heteroscedasticity

Normality of residuals

- Residuals are normally distributed with a mean of zero. The assumption is necessary for the validity of the inferences that we make based on the global and individual hypothesis tests
- The formal test for the normality of residuals is the Shapiro-Wilk test. The hypothesis tested is:

  H_0: Normality of residuals H_a: Residuals not normally distributed

Multicollinearity

- This exists when the independent variables are correlated.
- If an independent variable is highly correlated with other variables in the model should be removed.
- To assess the degree to which independent variables are correlated we compute the VIF. A VIF greater than 10 is unsatisfactory.

Autocorrelation

- Successive residuals should be independent implying that there is no pattern in the residuals.
- When successive residuals are correlated we refer to the condition as autocorrelation.
- The formal test is the Durbin Watson test
  H_0: No Autocorrelation
  H_a: Autocorrelation

Example in R

- We fit a multiple linear regression for US consumption given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Where:
- Y is the percentage change in real personal consumption expenditure - X_1 is the percentage change in real personal disposable income - X_2 is the percentage change in industrial production - X_3 is the percentage change in personal savings - X_4 is the change in unemployment rate

Fitting the model

```
library(fpp3)
library(broom)
model2 <- lm(Consumption ~ Income + Production +
                                       Unemployment + Savings,
kable(tidy((summary(model2))))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.2531051	0.0344704	7.342673	0.0000000
Income	0.7405835	0.0401150	18.461493	0.0000000
Production	0.0471726	0.0231420	2.038397	0.0428744
Unemployment	-0.1746853	0.0955107	-1.828959	0.0689490
Savings	-0.0528901	0.0029241	-18.087537	0.0000000

Testing the Assumptions

Normality

```
library(fpp3)
library(broom)
library(performance)
model2 <- lm(Consumption ~ Income + Production +
                               Unemployment + Savings,
check_normality(model2)
```

```
## Warning: Non-normality of residuals detected (p < .001).
```

Autocorrelation

```
library(fpp3)
library(broom)
library(performance)
model2 <- lm(Consumption ~ Income + Production +
                              Unemployment + Savings,
check_autocorrelation(model2)
```

```
## OK: Residuals appear to be independent and not autocorrel
```

Homoscedasticity

```
library(fpp3)
library(broom)
library(performance)
model2 <- lm(Consumption ~ Income + Production +
                              Unemployment + Savings,
check_heteroscedasticity(model2)
```

```
## Warning: Heteroscedasticity (non-constant error variance)
```

Multicollinearity

```
library(fpp3)
library(broom)
library(performance)
model2 <- lm(Consumption ~ Income + Production +
                                   Unemployment + Savings,
check_collinearity(model2)
```

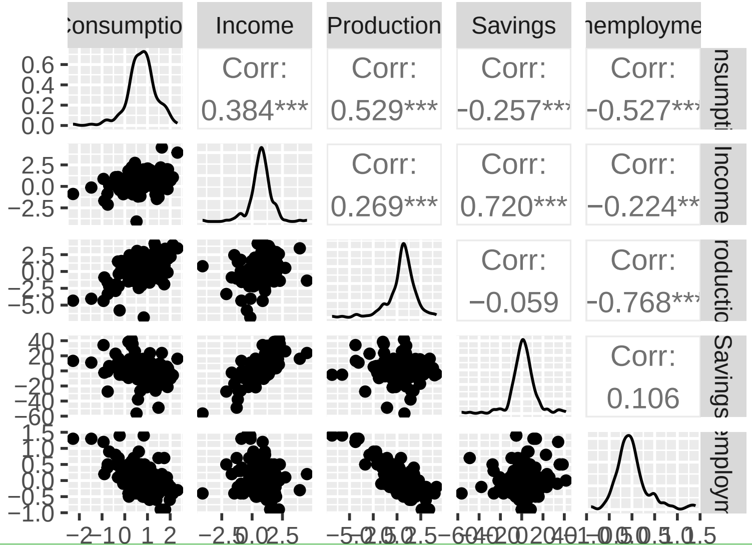
Cont'd

```
## # Check for Multicollinearity
##
## Low Correlation
##
##          Term  VIF   VIF 95% CI Increased SE Tolerance To
##        Income 2.67 [2.18, 3.37]         1.63      0.37
##    Production 2.54 [2.08, 3.19]         1.59      0.39
##  Unemployment 2.52 [2.06, 3.17]         1.59      0.40
##       Savings 2.51 [2.05, 3.15]         1.58      0.40
```

Linearity

```
library(fpp3)
library(GGally)
us_change |>
  ggpairs(columns = 2:6)
```

Cont'd



Income
Production
Savings
Corr:
0.384***
Corr:
0.529***
Corr:
-0.257***
Corr:
-0.527***
Corr:
0.269***
Corr:
0.720***
Corr:
−0.224**
Corr:
−0.059
Corr:
-0.768***
Corr:
0.106
Income
Savings

Thank You!