

# Summary of ALR4

Hengrui Luo

ABSTRACT. This is a brief summary of *[S.Weisberg]Applied Linear Regression, 4ed, Wiley, 2014*.

This is based on a course which is offered at graduate level by Prof.M.Pratola at OSU during 2016 Spring, who gave a series of nice lectures and notes.

In this summary, Chap.11 and part of Chap.12 corresponding to the text are included in another full course, so they are omitted here. All other materials are neatly re-organized to facilitate understandings and summarized from the problem-solving. And some of the materials should be understood in terms of *[A.Dean&D.Voss] Design and Analysis of Experiments, Springer, 1999*. I also wrote a guide to that title.

I hope this summary will help the reader. I personally felt it can be summarized in one sentence, with so many fancy techniques we merely achieved.

*"Truth that is naked is the most beautiful, and the simpler its expression the deeper is the impression it makes."*

*-On Authorship and Style, Arthur Schopenhauer.*

All errors are mine.

Hengrui Luo

2016-05-15

# Contents

	<b>Page</b>
Chapter 1. Scatter plots and Regression	5
SUMMARY GRAPH	
SCATTER PLOT	
MEAN FUNCTIONS AND VARIANCE FUNCTIONS	
SCATTER PLOT MATRIX	
MISCELLANEA	
Chapter 2. Simple Linear Regression	8
SIMPLE LINEAR REGRESSION	
MISCELLANEA	
Chapter 3. Multiple Linear Regression	11
MULTIPLE LINEAR REGRESSION	
COEFFICIENT OF DETERMINATION	
MISCELLANEA	
Chapter 4. Interpretation of Main Effects	15
PARAMETER ESTIMATES	
COLINEARITY	
MISCELLANEA	
Chapter 5. Complex Regressors	19
FACTORIAL MODEL	
POLYNOMIAL MODEL	
Chapter 6. Testing and ANOVA	25
Chapter 7. Variances	29
WEIGHTED LEAST SQUARES	
GENERALIZED LEAST SQUARES	
MISCELLANEA	
Chapter 8. Transformations	32

	POWER TRANSFORMATIONS	
	SCALED POWER TRANSFORMATIONS	
	BOX-COX MODIFIED POWER TRANSFORMATIONS	
	YEO-JOHNSON TRANSFORMATIONS	
	MISCELLANEA	
Chapter 9.	Regression Diagnostics	35
	LINEARITY	
	ROBUSTNESS	
	INDEPENDENCE	
	HOMOSCEDASTICITY	
	NORMALITY	
	LURKING	
	SIZE	
Chapter 10.	Variable Selection	42
	SELECTION CRITERIA	
	ALGORITHMS	
	MISCELLANEA	
Chapter 11.	Nonlinear Regression	45
Chapter 12.	Binomial and Poisson Regression	46
	LOGISTIC REGRESSION	
	POISSON REGRESSION	

## CHAPTER 1

# Scatter plots and Regression

### 1.1 Summary graph

- (1) Summary graph: scatter plot of response versus regressors. It is a special case of scatter plot.
- (2) Marginal plot: scatter plot of response versus one regressor among many regressors, sometimes referring to the scatter plots between two regressors. Attend that from high dimensional data to low dimensional data we always lose some information and there is no rescue. <sup>1</sup>
- (3) Effect plot: The plot of the *estimate* mean function with all but one regressors fixed<sup>2</sup>. We usually plot a confident interval around the estimated mean function, NOT prediction interval. This plot is a generalization of interactive plot for the linear model.  
An important feature of effect plot for factor model is that since it is not in terms of a particular regression, so it treats all levels equally.
- (4) Box plot: Boxes showing the range and the quantiles(inter-quantiles)<sup>3</sup> of a dataset, with each outliers shown individually. Box plot is a useful tool for comparing different levels of a factor.
- (5) Histogram plot: empirical cumulative binnings, which can be used for normality check. Unimodal/Skewness checks.
- (6) Q-Q plot: empirical quantiles versus theoretical quantiles, which can be used for normality check.

### 1.2 Scatter plot

- (1) Null plot: A null plot has a horizontal straight line as its mean function, constant variance function and no separated points.
- (2) Influential points
  - (a) Outliers (response scale): Cook's distance can be used to check this.
  - (b) Leverage (regressor scale): Hat matrix of a regression determines the leverage of a specific point.
  - (c) Sensitivity: We should distrust the analysis relying heavily on a single case.

---

<sup>1</sup>If the  $R^2$  summary is provided along, we have to explain it too because scatterplot is a more complicated version of correlation matrix.

<sup>2</sup>Fixed at their sample mean if they are factors and there are more than one level, if the data is unbalanced we usually used a weighted mean.

<sup>3</sup>IQR is robust against non-normality while the mean is not, generally speaking the quantiles are all robust against non-normality.

- (3) Over-plotting: Over-plotting often occurs when the observations are taken at only a few regressor values. The solution is to take a jittered plot.
- (4) Tools for looking at scatter plots
  - (a) Size/re-scaling
  - (b) Transformations
  - (c) Smoother: Nonparametric estimators for mean function<sup>4</sup>.
- (5) General formula (RAM for scatter plot)
  - (a) **Relation:** Are there any obvious *linear relation* OR *nonlinear relation* and which regressors are to be included in this model? On the contrary, are the horizontal axis and the vertical axis seem to be independent?
  - (b) **Abnormality:** Are there any clustering? And on the contrary, are there any influential points?
  - (c) **Meaning:** What do the observations in (a) (b) mean in the specific background?

### 1.3 Mean functions and Variance functions

- (1) Mean functions:  $\mathbb{E}(\mathbf{Y}|\mathbf{X})$ , which should always be linear.
- (2) Variance functions:  $\text{Var}(\mathbf{Y}|\mathbf{X})$ , which should preferably be constant. If it is not constant, just use WLS/GLS or other remedies.
- (3) Formal way of writing down a regression model: “REMVM” format.
  - (a) **Regression function:**  $Y_i = \beta_0 + \beta_1 X_i + e_i$
  - (b) **Error term:**  $e_i \stackrel{i.i.d}{\sim} N(0, \sigma^2), \sigma^2 > 0$
  - (c) **Mean function:**  $\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$
  - (d) **Variance function:**  $\text{Var}(Y_i|X_i) = \sigma^2$
  - (e) **Meaning of each parameter.**

### 1.4 Scatter plot matrix

- (1) Scatter plot matrix<sup>5</sup>
  - (a) The joint relationship between response and each regressors
  - (b) The inter-relationship between regressors
  - (c) The marginal relationship between response and each individual regressor
- (2) Correlation matrix
  - (a) Correlation matrix is a classical numerical YET less informative analogy to scatter plot matrix. It does not allow us to detect outliers from observing the correlation matrix. Hence if there are outliers in the

<sup>4</sup>Smoother or fitted lines are not trustworthy at the edge of data, especially when we are doing an extrapolation. That is to say, the correctness of the linear model also depends on the range of regressors, which is usually determined by the data observed.

<sup>5</sup>Projected/Marginal plot is good do not imply that the joint scatter plot is good. especially we should be very careful about the information loss from dimension reduction.

$$\begin{cases} \text{Plots good} & \rightarrow \text{still possible wrong} \\ \text{Plot wrong} & \rightarrow \text{must be wrong} \end{cases}$$

middle of range, then we might not trust the conclusion drawn from a scatter plot matrix.

- (b) Correlation matrix is equivalent to the variance-covariance matrix once we know the variance of each component.

### 1.5 Miscellanea

- (1) P-value: The probability of the occurrence of such an observation under the null hypothesis, when it is close to zero we should reject the null hypothesis.
- (2) Confidence interval: The meaning of a  $100(1 - \alpha)\%$  CI is that if we collect data  $(X_i, Y_i)$  and construct the CI, repeat the *whole* collect-construct procedure several times, in the long run  $100(1 - \alpha)\%$  CIs will contain the actual value of the parameter function. It based on  $\mathbf{Y}|\mathbf{X}$  distribution. Margin of error (MOE) is the perimeter of the confident region. And do not forget that CIs as well as estimators all have *units* in a practical setting.
- (3) Prediction interval: The meaning of a  $100(1 - \alpha)\%$  PI is that if we collect data  $(X_i, Y_i)$  and construct the PI, repeat the *whole* collect-predict procedure several times, in the long run  $100(1 - \alpha)\%$  PI will contain the actual value of the parameter function. It based on  $(\mathbf{Y}, \mathbf{X})$  distribution.

## CHAPTER 2

# Simple Linear Regression

### 2.1 Simple linear regression

SLR		OLS	Inference with normality
Model	$Y_i = \beta_0 + \beta_1 X_i + e_i$	$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i$ $SSE(\beta) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$	$Y_{new} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{new} + e_{new}$
Mean func- tion	$\beta_0 + \beta_1 X$	$\left\{ \begin{array}{l} \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \\ \widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \end{array} \right. \quad \left\{ \begin{array}{l} \mathbb{E}(\widehat{\beta}_0   \mathbf{X}) = \beta_0 \\ \text{Var}(\widehat{\beta}_0   \mathbf{X}) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{S_{XX}} \\ \mathbb{E}(\widehat{\beta}_1   \mathbf{X}) = \beta_1 \\ \text{Var}(\widehat{\beta}_1   \mathbf{X}) = \frac{\sigma^2}{S_{XX}} \\ \text{Cov}(\widehat{\beta}_1, \widehat{\beta}_0   \mathbf{X}) = -\frac{\bar{X} \sigma^2}{S_{XX}} \end{array} \right.$	$\left\{ \begin{array}{l} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_0 + \widehat{\beta}_1 X \end{array} \right. \quad \left\{ \begin{array}{l} \frac{\widehat{\beta}_0 - \beta_0}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}} \sim N(0, 1) \quad \sigma^2 \text{known} \\ \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\frac{SSE}{n-2}} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}} \sim t(n-2) \quad \sigma^2 \text{unknown} \\ \frac{\widehat{\beta}_1 - \beta_1}{\sigma \sqrt{\frac{1}{S_{XX}}}} \sim N(0, 1) \quad \sigma^2 \text{known} \\ \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{SSE}{n-2}} \sqrt{\frac{1}{S_{XX}}}} \sim t(n-2) \quad \sigma^2 \text{unknown} \\ \frac{\widehat{\beta}_0 + \widehat{\beta}_1 X - (\beta_0 + \beta_1 X)}{\sigma \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}} \sim N(0, 1) \quad \sigma^2 \text{known} \\ \frac{\widehat{\beta}_0 + \widehat{\beta}_1 X - (\beta_0 + \beta_1 X)}{\sqrt{\frac{SSE}{n-2}} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}} \sim t(n-2) \quad \sigma^2 \text{unknown} \\ \frac{\widehat{\beta}_0 + \widehat{\beta}_1 X - (\beta_0 + \beta_1 X)}{\sqrt{\frac{SSE}{n-2}} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}} \sim F(2, n-2) \quad \sigma^2 \text{unknown} \\ (\text{Simultaneous}) \end{array} \right.$
Variance func- tion	$\sigma^2$	$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta})}{n-2}$ $\mathbb{E}[\widehat{\sigma}^2] = \sigma^2$	$\left\{ \begin{array}{l} \frac{(n-2)\sigma^2}{\sigma^2} \sim \chi^2(n-2) \\ sefit(Y X_{new}) = \widehat{\sigma} \left( \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{S_{XX}} \right)^{1/2} \\ sepred(Y X_{new}) = \widehat{\sigma} \left( 1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{S_{XX}} \right)^{1/2} \end{array} \right.$
Error	$e_i \sim N(0, \sigma^2)$	$\widehat{e}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$	

1 2 3 4 5 6 7

<sup>1</sup> $S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$ ;  $S_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \sum X_i \sum Y_i$ ;  $SSE = S_{YY} - \widehat{\beta}_0 \sum Y_i - \widehat{\beta}_1 \sum X_i Y_i$

<sup>2</sup>Weighted scheme interpretation:  $\widehat{\beta}_1 = \sum_{ij} \frac{(x_i - x_j)^2}{2nS_{XX}} \cdot \frac{y_i - y_j}{x_i - x_j}$ ,  $\sum_{ij} (x_i - x_j)^2 = 2nS_{XX}$ ;  $\sum_{ij} (x_i - x_j)(y_i - y_j) = 2nS_{XY}$ . So if a pair of points are far away from each other, then they have more influence on the parameter estimates.

<sup>3</sup>Due to Cochran's theorem, the decomposed SS as bilinear forms of the independent observations are mutually independent. So the residuals are independent of  $S_{XX}, S_{XY}$  and hence the OLS of parameters.

<sup>4</sup>In SLR,  $BLUE \Leftrightarrow OLE \stackrel{Normality}{\Leftrightarrow} MLE$

<sup>5</sup>In SLR, the regression line always goes through  $(\bar{X}, \bar{Y})$ . The further we goes from the center, the greater variance of the prediction value will be. But it never goes below  $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$

<sup>6</sup>Uncorrected estimates to variances are always under-estimate. We use homogeneous correction instead of heterogeneous correction because the latter depends on an unknown quantity.

<sup>7</sup>p-value approach is preferred than the critical value approach because it gives a more sensitive strength of the evidence. P-value can be used for comparison of different results from different



## 2.2 Miscellanea

- (1) Are the assumptions satisfied? (**LRIHNLS**, “Loppy Reality Is Hard Not to Laugh at, Sir.”)
  - (a) **Linearity**: Are there nonlinear trend in the mean function?
  - (b) **Robustness**: Are there too many influential points(outlier OR leverage) which greatly affect our regression analysis? Is this model robust against these points?
  - (c) **Independence**: Are each observation independent of each other?
  - (d) **Homoscedasticity**: Are the variances constants?
  - (e) **Normality**: Are the errors looking like normally distributed? Are they mutually independent t?<sup>8 9</sup>
  - (f) **Lurking**: Are there additional covariates should be included?
  - (g) **Size**: Are there enough data? What consequence will a large sample cause?
- (2) Kinds of variables
  - (a) Quantitative: continuous quantity
  - (b) Ordinal: rank or order number
  - (c) Qualitative: variables that indicate which of several categories a variable falls in or which of several quantities a variable has.
- (3) Special SLRs
  - (a) Zero-intercept model
 
$$Y_i = \beta_1 X_i, \epsilon_i \sim NID(0, \sigma^2)$$

$$\widehat{\beta}_1 = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}, \widehat{\sigma}^2 = \frac{SSE}{n-1}$$
  - (b) Centered model: centered model calculation is less oscillatory because it unifies the scale.
 
$$Y_i = \beta_0 + \beta_1 (X_i - \bar{X}), \epsilon_i \sim NID(0, \sigma^2)$$

$$\widehat{\beta}_0 = \bar{Y}, \widehat{\beta}_1 = \frac{SXY}{SXX}$$
  - (c) Log-linear model
 
$$\log Y_i = \beta_1 \log X_i + \beta_0 + \epsilon_i, \epsilon_i \sim NID(0, \sigma^2)$$

The reason of making a log transformation:

    - (i) (L) Obtain more linearity
    - (ii) (H) Fix the nonconstant variability
    - (iii) (N) Fix the skewed nature of the data
    - (iv) (S) Fix the concentrated data

samples. Specially it depends on the sample size since for a large sample almost all tests become significant that is the reason why we want to include sample size as LM assumptions.

<sup>8</sup>This can be verify using autocorrelation plot in time series, however, it is usually obvious in the residual plot too.

<sup>9</sup>Note that residuals are NOT error, residuals are estimates of errors. And while the error term do have constant variance, the residuals might have distinct variances; moreover we notice that residuals are actually correlated because  $\sum_i \hat{\epsilon}_i = 0$ .

(d) Group indicator model

$$Y_i = \beta_1 x_i + \beta_0 + \epsilon_i, \epsilon_i \sim NID(0, \sigma^2), x_i = \begin{cases} 0 & i = 1, \dots, m_1 \\ 1 & i = m_1 + 1, \dots, m_1 + m_2 \end{cases}$$

(i) Invariance

Under the nondegenerate linear transformation  $Z = aX + b, a \neq 0$ , the  $SSE$  and the t-tests do not change

(ii) The test of  $H_0 : \beta_1 = 0$  in this model is exactly the same as two-sample t-test for comparing two groups with assumptions of equal within group mean.

# CHAPTER 3

## Multiple Linear Regression

### 3.1 Multiple linear regression

MLR		OLS	Inference with normality
Model	$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$	$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} + \mathbf{e}$ $SSE(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$	$\mathbf{Y}_{\text{new}} = \mathbf{X}_{\text{new}}\hat{\beta} + \mathbf{e}_{\text{new}},$
Mean func- tion	$\mathbf{X}\beta$	$\begin{cases} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\mathbf{Y}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{cases}$	$\begin{cases} \hat{\beta} & \begin{cases} \left[ \hat{\beta} - \beta \right] \cdot [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1/2} \sim N(\mathbf{0}, \mathbf{I}) & \sigma^2 \text{ known} \\ \left[ \hat{\beta} - \beta \right]^2 \cdot [p \cdot MSE \cdot (\mathbf{X}'\mathbf{X})^{-1}]^{-1} \sim F(p, n-p) & \sigma^2 \text{ unknown} \end{cases} \\ \beta_i & \begin{cases} \frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{((\mathbf{X}'\mathbf{X})^{-1})}} \sim N(0, 1) & \sigma^2 \text{ known} \\ \frac{\hat{\beta}_i - \beta_i}{\sqrt{MSE} \sqrt{((\mathbf{X}'\mathbf{X})^{-1})}} \sim t(n-p) & \sigma^2 \text{ unknown} \end{cases} \\ \mathbf{Y} & \begin{cases} \left[ \hat{\mathbf{Y}} - \mathbf{X}\beta \right] \cdot [\sigma^2 \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']^{-1/2} \sim N(0, 1) & \sigma^2 \text{ known} \\ \left[ \hat{\mathbf{Y}} - \mathbf{X}\beta \right]^2 \cdot [p \cdot MSE \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']^{-1} \sim F(p, n-p) & \sigma^2 \text{ unknown} \end{cases} \end{cases}$ <p><i>Bonferroni</i></p>
Variance func- tion	$\sigma^2 \mathbf{I}$	$\begin{aligned} \hat{\sigma}^2 &= \frac{SSE(\hat{\beta})}{n-p} \\ p &= \text{rank} \mathbf{X} \end{aligned}$	$\begin{cases} \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p) \\ sefit(\mathbf{Y} \mathbf{X}_{\text{new}}) = \hat{\sigma} \sqrt{\mathbf{x}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}'} \\ sepred(\mathbf{Y} \mathbf{X}_{\text{new}}) = \hat{\sigma} \sqrt{1 + \mathbf{x}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}'} \\ se : para.function \quad \begin{cases} se(\mathbf{a}'\beta \mathbf{X}_{\text{new}}) = \hat{\sigma} \sqrt{\mathbf{a}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}'} \\ se(\mathbf{g}(\beta) \mathbf{X}_{\text{new}}) \quad \begin{cases} \text{delta method} \\ \text{Bootstrap} \end{cases} \end{cases} \end{cases}$
Error	$\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$	$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$	$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right) \Rightarrow$ $\mathbf{X}_1   \mathbf{x}_2 \sim N \left( \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}[\mathbf{x}_2 - \mu_2], \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$ $\xrightarrow{\text{univariate}} X_1   x_2 \sim N \left( \mu_1 + \rho_{12} \frac{\sigma_1}{\sigma_2} [x_2 - \mu_2], \sigma_1^2 (1 - \rho_{12}^2) \right)$

1 2 3 4

### 3.2 Coefficient of determination

- (1) Variability explanation

*Proportion Explained*

$$\max(R_{X_1}^2, R_{X_2}^2) \leq \uparrow \text{normality} \leq \max(1, R_{X_1}^2 + R_{X_2}^2) \text{ for two } R_{X_1, X_2}^2$$

regressors case. The first equality holds when the two regressors are uncorrelated; the second equality holds when they are linearly dependent. The coefficient of determination is calculated from marginal regression on each of the regressors only.

Generally the value of coefficient of determination in a normal error situation will be interpreted as the proportion of variability explained by the regressors included in the model.

- (2) SLR:  $R^2 := 1 - \frac{RSS}{RSS + SS_{reg}} = \frac{SS_{reg}}{RSS + SS_{reg}} = \frac{(SXY)^2}{SXX \cdot SYY} = \rho^2(\text{SLR correlation})$ <sup>56</sup>

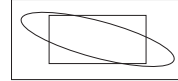
7

$$\text{Adjusted } R^2 := 1 - \frac{RSS/df_{residual}}{(RSS + SS_{reg})/df_{total}}$$

- (3) MLR:  $R^2 := \frac{\hat{\beta}' \mathbf{X}' \mathbf{X} \hat{\beta}}{(\hat{\mathbf{y}} - \mathbf{y})'(\hat{\mathbf{y}} - \mathbf{y}) + \hat{\beta}' \mathbf{X}' \mathbf{X} \hat{\beta}}$

The coefficient of determination is the maximum of the two-variate correlation coefficient between the response and the linear combination of regressors in the mean function. The linear combination which attains this maximum is the OLS  $\mathbf{X}\hat{\beta}$ . This can be deduced ONLY under the

<sup>1</sup>Product interval, Exact interval and Bonferroni adjusted interval:



<sup>2</sup>**Theorem.** For a partitioned matrix

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \\ = \begin{pmatrix} \mathbf{I} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I} \end{pmatrix}$$

, its inverse is given by

$$\begin{pmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{pmatrix} \\ = \begin{pmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21})^{-1} & -(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}\mathbf{A}_{12})^{-1} \end{pmatrix}$$

<sup>3</sup>Confident intervals follows ESQ format: Estimate  $\pm$  Standard deviation of estimate  $\times$  Quantile of estimate distribution under  $H_0$ .

<sup>4</sup>R function “vcov” gives the estimated variance-covariance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  directly, so when we calculate some contrast, we simply take the “vcov” and multiply by  $\mathbf{a}'\mathbf{vcov}\mathbf{a}' = \mathbf{a}'\Sigma\mathbf{a}'$  and square-root it to yield a confident interval for a contrast. When constructing confident interval, the Bonferroni correction only divides the dimension of the parameter space at maximum.

<sup>5</sup>This is often provided in the output and can be used for calculating the sums of squares of each regressor.

<sup>6</sup>To test the population correlation  $H_0 : \rho = 0$  in SLR is equivalent to testing  $H_0 : \beta_1 = 0$ .

<sup>7</sup>Avoid using this as a measure of variability explained when the model is through origin.

normality assumption where we can actually regard the model as a conditional multivariate normal distribution.

- (4) Added-variable plot: The  $R^2$  in an added-variable plot is equal to the square of the partial correlation between the response and the regressor, adjusted for the other regressors in the mean function.
- (5) Normal sample
  - (a) When the data is a *random sample from multivariate normal distribution*, then  $R^2$  will estimate the population multiple correlation and hence as an indicator of percentage of variability explained.
  - (b) When the data is NOT a *random sample from multivariate normal distribution*, then  $R^2$  will depend on the underlying population as well as the sampling plan. For example, if we sample from a wide range then  $R^2$  will be closer to one.
  - (c) Small  $R^2$  issue.
    - (i) Check the normality assumption, if that is broken, we should not trust  $R^2$  on this LOF model. Thus, before we try to use variability explanation, we should always do diagnostic in order to decide if  $R^2$  is a useful summary.
    - (ii) Check the  $\hat{\sigma}^2 = MSE$ , if the variability is large, which is often the case when the sample size is small, then small  $R^2$  may not indicate a bad model.
    - (iii) Check the sample range. The  $R^2$  also depends on the sampling range of the data, when the range is wide the coefficient will be close to 1.
      - (A) If we randomly drop observations depending on *regressors*, then coefficient estimates will not change and will yield estimates closer to estimates using all data, but  $R^2$  depends on the sampling.
      - (B) If we randomly drop observations depending on *response*, then coefficient estimates will change and faraway samplings will yield estimates closer to estimates using all data, but  $R^2$  depends on the sampling.

### 3.3 Miscellanea

- (1) Adding regressors
  - (a) Mean function: With additional terms coming from the added regressor, the coefficients of older terms already in the model will change unless the newly added term is completely independent with the existing term in the model.
  - (b) Variance function: With additional terms coming from the added regressor, the coefficients of older terms already in the model will

change unless the newly added term is completely independent with the existing term in the model.

- (c) When we are to add regressors?
  - (i) Lack of fit
  - (ii) Lurking variable
  - (iii) Diagnostic failed
- (d) Marginal plot  $Y - X_i$  versus Added-variable plot  $Res(Y|X_1) - Res(X_2|X_1)$ 
  - (i) Test of slope in added-variable plot is equivalent to the test of significance of corresponding coefficient. In fact the estimate slope in an added-variable plot is the *same as* the coefficient of itself in the full model iff the regressors are *completely linearly uncorrelated*.
  - (ii) The coefficient of determination in this plot is the proportion of variability in  $Y$  explained by  $X_2$  after adjusting for  $X_1$ . Like what we did for Type III SS.
  - (iii) When is it suitable to add a regressor? When the added-variable plot shows stronger linear relationship than the marginal plot.
  - (iv) Alternative approach to decide whether adding regressor is needed: Response surface.

(2) Predictors and regressors

- (a) Interactions: Tukey's multiplicative form
- (b) Factors: Often the categorical variables which takes only finitely many values. A categorical predictor with  $n$  values can be expressed as  $n - 1$  factorial variable.
- (c) Spline: The base functions are chosen to be the polynomials.
- (d) Principle components.
- (e) Observing the scatter plot matrix, we can decide whether to extend the model to include more regressors or to combine some of the regressors to make the model simpler.

## CHAPTER 4

# Interpretation of Main Effects

### 4.1 Parameter estimates

- (1) **Significance:** Are all these variables significant? And will the conclusion be affected by the sample size?
- (2) **Rate:** Increase one regressor with one unit, *leaving all other regressors fixed*, how will the *expected response/mean function* change. This interpretation actually assumes that the regressor can be changed without affecting the other regressors in the mean function and that the available data will apply when the predictor is changed so. If necessary, yield a confident interval for the parameter we are interpreting as an *uncertainty characterization*.<sup>1</sup>

*Effect plot* is the scatter plot  $\mathbb{E}(\mathbf{Y}|\mathbf{X}) - \mathbf{X}$  which shoes the change of one regressor while retaining other regressors fixed.

- (3) **Sign:** The sign of a parameter estimate indicates the direction of the relationship between the regressor and the response after adjusting for all other regressors in the model. In fact, all the estimates in a MLR are adjusted for all the other regressors in the mean function.
- (4) **Log-Scale**
  - (a) log-Scale of regressor: It allows fitted effect that change most rapidly when the regressor is small and less rapidly when the regressor is large.
  - (b) log-Scale of response: When the error are of the form “plus or minus 5%”, then we might want to use a strictly positive log scale to fit it as a linear model. For any increase in regressor by 1 unit it will multiply the mean of response by  $\exp(\beta)$  units.
  - (c) Using log-scale in regressor captures the changing rate in the response along with regressors;  
Using log-scale in response captures the *multiplicative error* with the changing rate  $\exp(\beta)$ .
- (5) **Variability.**

### 4.2 Colinearity

Colinearity is the phenomena of coefficient of determination being close to one.

---

<sup>1</sup>To interpret the interaction effect regressor, fix only one explanatory variable and explain the expected mean function containing other regressors.

- (1) Interaction: This is reflected via the fact that the only way of changing one regressors is to change other regressors simultaneously. This is also explained under the name of aliased regressors OR over-parameterized mean function<sup>2</sup>. The value of a parameter estimate not only depends on the other regressors in a mean function, but it can also change if the other regressors are replaced by linear combination of the other regressors. And such a linear combination cannot be regarded as simple summing effect since we must account for the correlations between the regressors.<sup>3</sup>

A general way of interpreting the interaction term in the model is:

- (2) Rank-deficiency: Rank-deficiency implies colinearity. Sometimes the design matrix might not have full column rank, so this is the quantitative evidence of interaction of regressors. Rank-deficiency focus on the consistency of a system while the colinearity focus on the compatibility of a system. When the system is a linear one, these two notions are actually the same one. And while rank-deficiency described an exact linear relation between regressors; colinearity only approximately judges it.

$$\mathbf{X}'\mathbf{X} \text{ singular} \begin{cases} \text{Too many redundant regressors} \\ \text{Bad choice of predictor values} \end{cases}$$

- (3) Colinearity phenomena
- (a) Correlated regressors will cause the *estimate coefficients* change when adding/dropping regressors into/from the linear model. Also there will be wider CI and large off-diagonal entries in the estimated variance-covariance matrix.
  - (b) Correlated regressors will cause the phenomenon that the model does not show LOF but *coefficients corresponding to correlated regressors* are insignificant, in that case PCA will solve some problem. Don't misunderstand this sentence, high colinearity does **NOT** mean that the explanatory variable must be insignificant. A good example is the polynomial regression.<sup>4</sup>
  - (c) Correlated regressors will cause the *t-test of the coefficients* change when there are different regressors in the model.
  - (d) Correlated regressors will cause *simultaneous change in values of regressors*, thus the main-effect interpretation will be more difficult.
  - (e) Correlated regressors and response is good, for otherwise there will be no difference between regressing Y on X AND regressing X on Y.
- (4) Methods of detecting colinearity.

<sup>2</sup>In Bayesian modeling, sometimes we intentionally introduce some over-parameterization(over-dispersed model) to get a better model.

<sup>3</sup>If the units of the regressors are different, taking linear combination of them may lead to uninterpretable estimates.

<sup>4</sup>In that case the correlated regressors still explained a large amount of variability.



- (a) The added variable plot: If the partial correlation  $R_{Y,X_1|X_2}^2 \approx 1$ , then we say these two predictors are NOT correlated, on the plot it should be almost the same plot as plotting  $Y$  against  $X_2$ .
- (b) Partial correlation coefficient: The partial correlation coefficient  $R_{Y,X_1|X_2}^2 := R_{Y,(X_1,X_2)}^2 - R_{Y,X_2}^2$ . So if there is a colinearity between these two predictors, then  $R_{Y,X_1|X_2}^2 \approx 0 \iff \frac{R_{Y,X_1|X_2}^2}{R_{Y,X_1}^2} \approx 1 \iff VIF_1 \approx 1$ .
- (c) The variance inflation factor (VIF)

$VIF_1 := \frac{1}{1-R_{Y,X_1|X_2}^2}$  is called the variance inflation factor of  $X_1$ . In general,  $VIF_k := \frac{1}{1-R_k^2}$  where  $R_k^2$  is the coefficient of determination of regressing  $X_k = X_1 + \dots + X_{k-1} + X_{k+1} + \dots + X_p + \epsilon$ , or the partial coefficient of determination of  $X_k$ . VIF tells us how much each explanatory variable is explained by a linear combination of all the other explanatory variables already in the model. For instance, the  $VIF_1$  above tells how much  $X_1$  is explained by a linear combination of  $X_2, X_3, \dots, X_p$ .<sup>5</sup>

These three methods are equivalent, all based on one criterion:  $R_k^2 \leq 1$ . So if  $VIF_k > 10$ , then it is a strong evidence that  $X_k$  has a strong correlation with some *linear combination of the rest regressors*. It is known as variance inflation factor because  $Var(\widehat{\beta}_k^*) = \hat{\sigma}^* VIF_k$  where  $\widehat{\beta}_k^*, \hat{\sigma}^*$  are estimates based on data with  $k$ -th deletion.

If we use the following pooled VIF:  $\overline{VIF} := \frac{1}{p} \sum_{k=1}^p VIF_k$ , then comparing whether it is greater than one is equivalent to comparing  $R_k^2 \leq 1$ .

### 4.3 Miscellanea

#### (1) Dropping regressors

- (a) Mean function:  $\mathbb{E}(Y|X_1 = \mathbf{x}_1) = \mathbb{E}[\mathbb{E}(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1] = \beta_0 + \beta_1' \mathbf{x}_1 + \beta_2' \mathbb{E}(X_2|X_1 = \mathbf{x}_1)$
- (b) Variance function:  $Var(Y|X_1 = \mathbf{x}_1) = \mathbb{E}[Var(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1] + Var[\mathbb{E}(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1] = \sigma^2 + \beta_2' Var(X_2|X_1 = \mathbf{x}_1) \beta_2$
- (c) When we are to drop regressors?
  - (i) Colinearity(Overparameterized)
  - (ii) Complexity
  - (iii) Assumption broken: But be careful that after dropping the regressors, unlike general case for adding regressors, the assumptions of linear model may NOT necessarily hold.

#### (2) Interaction term

<sup>5</sup>We expect large VIF if that is what we try to introduce into our model. For example, the interaction effect in ANOVA and polynomial regressors are expected to have large VIF and that will NOT be our concern.

- (a) When *the regressors are independent*, we can vary only one regressor keeping all other regressors fixed (conditional expectation). This reduces to interpreting a main effect.
- (b) When *the regressors are dependent*, we must consider more complicated interpretation.
- (3) Experiments and observations
  - (a) Causation: The following setups can be used to establish causation.
    - (i) Controlled experimental design
    - (ii) Observing the same association in many different types among different groups. (Randomization)
    - (iii) Observing the same association after accounting for effects of possible lurkings. (Lurking)
  - (b) Correlation: Observational studies
 
$$\left\{ \begin{array}{l} \text{Experiments} \\ \text{Observational data} \end{array} \right\} \left\{ \begin{array}{l} \text{Randomized} \\ \text{Non-randomized} \end{array} \right. \quad (\text{Casual effect})$$

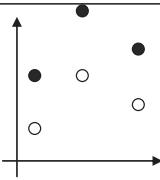
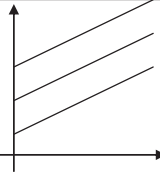
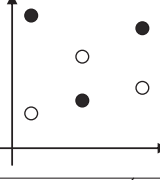
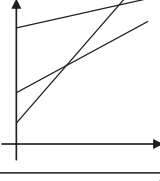
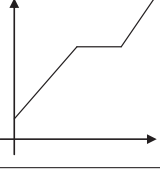
$$\left\{ \begin{array}{l} \text{Experiments} \\ \text{Observational data} \end{array} \right\} \left\{ \begin{array}{l} \text{Common response} \\ \text{Confounding (Aliased)} \end{array} \right. \quad (\text{ONLY association})$$
  - (c) Randomization
    - (i) allows us to generalize the results deduced from a sample to a larger population. (possible causation)
    - (ii) avoids systematic error from the sampling procedure.
    - (iii) guarantees that the correlation between the regressors in the mean function and any lurking variable is close to 0.
- (4) Errors: Two possible explanations about the normal error.
  - (a) The error term is exactly distributed as a normal distribution.
  - (b) The cumulative effects of various kinds of error behaves as if a normal distribution.

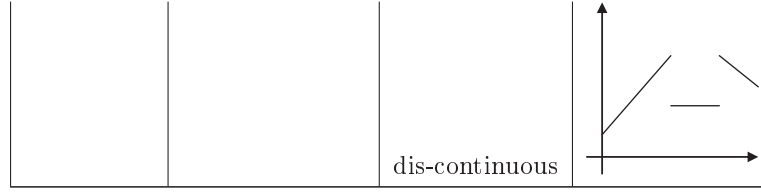
# CHAPTER 5

## Complex Regressors

### 5.1 Factorial model

$$\text{Factorial} \left\{ \begin{array}{l} \text{Main-effect} \\ \text{Interaction} \\ \text{Piecewise linear} \end{array} \right. \left\{ \begin{array}{l} \text{without continuous variable ANOVA} \\ \text{with continuous variable ANCOVA} \\ \text{without continuous variable factorial design} \\ \text{with continuous variable group indicator} \\ \text{with continuity} \\ \text{without continuity} \end{array} \right.$$

Name		Diagram
Main-effect	ANOVA	
	ANCOVA	
Interaction	Factorial design	
	Group indicator	
Piecewise linear		



- (1) Main-effect model: Those model includes *continuous variables* as well as *categorical regressors*, represented by the dummy variables.

$$Y = \tau_1 + \tau_2 + \mu + \mathbf{X}\beta + \epsilon^1$$

A  $d$ -category variables requires  $(d - 1)$  dummy variables to represent.

An  $n$ -factor model with  $c_i, i = 1, \dots, n$  levels for the  $i$ -th factor requires  $\sum_{i=1}^n (c_i - 1)$  dummy variables to represent.

Since dummy variables can take only two values, whilst what value they take does not affect the result of analysis, they are either strictly collinear or independent. Different setup of dummy variables can lead to different coding of levels. If we take the usual setup of 0 and 1; then the level corresponding to all zeroes are called “base level” and there should not be more than one dummy variable taken 1 at each level.

- (2) Interaction model: Besides the categorical regressors, we consider their *products*’ effects too. The number of factors in the highest order interaction term is known as the order of the model. Main-effect model is simpler than interaction model, because the effect of the continuous regressor is the same for all levels of factors.

*Principle of marginality*: a lower order interaction is never tested unless all of higher ordered interactions are all insignificant.

- (3) ANOVA model: A main-effect model with only categorical regressors.

$$Y = \tau_1 + \tau_2 + \mu + \epsilon$$

- (4) ANCOVA model: A main-effect model whose primary interest is in differences due to the levels of factor. (“main-effect with continuous levels”)

$$Y = \tau_1 + \tau_2 + \mu + \mathbf{X}\beta + \epsilon$$

Procedure of effect plot in ANCOVA:

- (a) Compute a fitted value for each level of the factor.
  - (b) Use a *weighted average* of these fitted values, with the weights determined by the sample size at each level.
  - (c) When we vary one variable, we fixed other regressors at their own weighted average value
- (5) Group indicator: With different levels of a factor we can fit different fitted lines for the same continuous predictor, dummy variables are used

<sup>1</sup>In this chapter we generally assume that there are equal number of observations at each level/level combinations, we can also adjust the estimate accordingly if there are unbalanced data. However, we want to point out that that setup is just like the

to indicate different groups. (“continuous effect with breakpoints”)

$$Y = \tau_{group1} \cdot (\mathbf{X}\beta_{\mathbf{group1}}) + \tau_{group2} \cdot (\mathbf{X}\beta_{\mathbf{group2}}) + \tau_{group3} \cdot (\mathbf{X}\beta_{\mathbf{group3}}) + \epsilon$$

- (6) Piece-wise linear model: This is a generalization of the group-indicator model. This is also a special case of spline regression with linear splines.

$$\text{Define } Z_1 = \begin{cases} 1 & X > w_1 \\ 0 & \text{otherwise} \end{cases} \text{ and } Z_2 = \begin{cases} 1 & X > w_2 \\ 0 & \text{otherwise} \end{cases}, w_1, w_2 \text{ are known}$$

as breakpoints/knots.

$$Y = \mathbf{X}\beta_{\mathbf{common}} + Z_1 [(\mathbf{X} - w_1)\beta_1 + 0] + Z_2 [(\mathbf{X} - w_2)\beta_2 + 0] + \epsilon (\text{with continuity at knots})$$

$$Y = \mathbf{X}\beta_{\mathbf{common}} + Z_1 [(\mathbf{X} - w_1)\beta_1 + \alpha_1] + Z_2 [(\mathbf{X} - w_2)\beta_2 + \alpha_2] + \epsilon (\text{without continuity at knots})$$

## 5.2 Polynomial model

- (1) Polynomials

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_d X^d + \epsilon$$

Polynomials are *global* fitting models. They may not capture the local features and not provide an understandable model. However, polynomials are a special case of splines, which are usually local.

- (a) Quadratic model: the simplest curve that can approximate a mean function with extreme.

- (i) The extreme location  $\frac{-\hat{\beta}_1}{2\hat{\beta}_2}$  of the quadratic model  $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ . Assuming there is *no colinearity between regressors* so that the maximization can be done independently.

The mean function could lead to unreasonable prediction when:

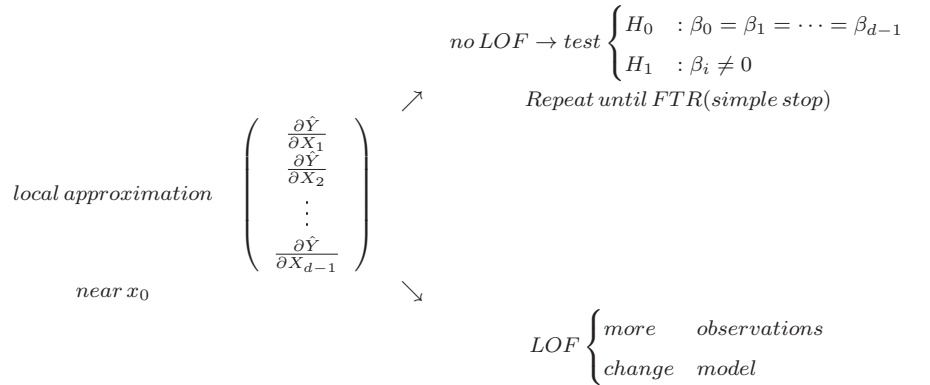
- (A) the predictors are out of the range of the observed data.  
(B) the model has a great amount of lack of fit.

- (ii) Methods of estimations for non-linear models

(A) Delta method  $Var(g(\hat{\beta})) = [\nabla_{\beta} g(\beta_{TRUE})]' Var(\hat{\beta}) [\nabla_{\beta} g(\beta_{TRUE})]$

(B) Bootstrap method  $Var(g(\hat{\beta})) = \text{the variance of the empirical distribution of } g(\hat{\beta}).^2$

- (iii) Optimization: Rapidest gradient descendant.



<sup>2</sup>Bootstrap can also be used for constructing an confident interval for a (nonlinear) estimate. The bootstrap is a sampling with replacement.

- (b) Many regressors  $\Leftrightarrow$  Response surface: The regressors in a polynomial model are highly collinear, so we generally assume it is an over-fit model and drop insignificant regressors.

- (i) Number of interaction terms

- (A) A general setup for  $d$ -th order polynomial regression

$$Y = \sum_{j=1}^d X^j + \epsilon$$

- (B) A general setup for a  $d$ -th order response surface of  $k$ -predictors

$$Y = \sum_{j=1}^d \sum_{l_1+\dots+l_k \leq j} X_1^{l_1} \dots X_k^{l_k} + \epsilon$$

With  $k$  predictors, the full  $d$ -th order polynomial regression will include  $2^d = \binom{k}{d} + \binom{k}{d-1} + \dots + \binom{k}{1}$  regressors in the mean function. These interaction terms in response surface model is called  $j$ -th order general interaction effect.

- (ii) Number of observations

- (A) Fit the model

To fit a polynomial of degree  $d$ , we need at least  $d + 1$  observations.

To fit a response surface of order  $d$  with  $k$  predictors, we need at least  $\binom{k+d}{d}$  observations.

- (B) LOF test

To carry out a LOF test for a polynomial of degree  $d$ , we need at least  $d + 2$  observations.

To carry out a LOF test for a response surface of degree  $d$  with  $k$  predictors, we need at least  $\binom{k+d}{d} + 1$  observations.

- (C) Estimate SSPE

To estimate SSPE we need some more replications.

- (iii) Number of extremes:

For a polynomial of degree  $d$ , we have at most  $d - 1$  extremes. And if we are fitting a one-factor model with  $c$  levels, then we need a polynomial of degree  $c - 1$  to yield an exact fit. Higher order polynomial regression will not improve the fit.

For a response surface of order  $d$  with  $k$  predictors, we can only detect those trends of order  $d' \leq d$ .

- (c) Orthogonal polynomials: Interpretability as trends/Non-collinearity/Numerical stability

- (i) Coded level. Coded level of  $U_j = \frac{X_j - \bar{X}}{c}$  where the levels  $X_i$  of the factor  $X$  are equally spaced with  $c$  units apart.

- (ii) Orthogonality. A set of polynomials on polynomial space  $\sum_{j=1}^K \xi_s(U_j) \cdot \xi_r(U_j)$ , then these  $\xi_s$  form an orthogonal family.
- (iii) Trend. The vector  $(\xi_i(U_1), \dots, \xi_i(U_K))$ ,  $i = 1, \dots, N$  is called the  $i$ -th order trend at the  $K$  coded levels. The rows of trend table are these vectors. Note that  $i \leq K - 1$  due to the fundamental theorem of algebra.  $\xi_0 \equiv 1$

	Coded level $U_j$ (1,2,3,4,5)
Order $i$ of the regressor $\xi_i$ (linear/quadratic/cubic)	Value of regressors at coded levels $\xi_i(U_j)$

Example for quadratic orthogonal polynomials at five levels.

	1	2	3	4	5	
0	1	1	1	1	1	$\Rightarrow \begin{cases} A_0 = \frac{[\sum Y_{i1}] \cdot 1 + [\sum Y_{i2}] \cdot 1 + [\sum Y_{i3}] \cdot 1 + [\sum Y_{i4}] \cdot 1 + [\sum Y_{i5}] \cdot 1}{1^2 + 1^2 + 1^2 + 1^2 + 1^2} \\ A_1 = \frac{[\sum Y_{i1}] \cdot (-2) + [\sum Y_{i2}] \cdot (-1) + [\sum Y_{i3}] \cdot 0 + [\sum Y_{i4}] \cdot 1 + [\sum Y_{i5}] \cdot 2}{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2} \\ A_2 = \frac{[\sum Y_{i1}] \cdot (-1) + [\sum Y_{i2}] \cdot 2 + [\sum Y_{i3}] \cdot 0 + [\sum Y_{i4}] \cdot (-2) + [\sum Y_{i5}] \cdot 1}{(-1)^2 + 2^2 + 0^2 + (-2)^2 + 1^2} \end{cases}$
1	2	-1	0	1	2	
2	-1	2	0	-2	1	

- (iv) Coefficient. The coefficient of the polynomial regression model  $E[Y_{ij}] = A_0 + A_1 \xi_1(U_j) + \dots + A_{p-1} \xi_{p-1}(U_j)$ ,  $i = 1, \dots, N$ ,  $j = 1, \dots, K$ , where  $Y_{ij}$  is the  $i$ -th observation at  $X_j$  is given by  $A_m = \frac{\sum_{s=1}^K \sum_{t=1}^N Y_{st} \xi_m(U_s)}{N \sum_{s=1}^K \xi_m^2(U_s)}$ .  $K$  is the number of unique levels of  $X$ ;  $N$  is the number of observations taken at each level of  $X$ .

## (2) Splines

Splines are **usually** local fitting models. They may not capture the global trends, and hence they might neither predict well nor provide an interpretable model. Splines connect the factorial model and the non-linear regression.

$$Y = \mathbf{X}_{n \times p} \beta_{p \times 1} + f_{p \times 1}(\mathbf{X}_{n \times p}) + \epsilon_{p \times 1}.$$

The form of  $f = (f_j)$  is written as  $f(\mathbf{X}) = (f_j)$ ;  $f_j(\mathbf{X}_j) = \gamma_j \cdot R(\mathbf{X}_j, \text{Knot}_j)$ , of course the knot should also be a  $p$ -vector and  $\mathbf{X}_j$  is the  $j$ -th column of the design matrix representing the  $j$ -th covariates in our model.

- The spline model with different number of basis are not nested, we cannot perform F-test to detect significance of the model.
- The t-test is not applicable to test if each spline basis is significant.
- To choose an appropriate number of spline basis is therefore a variable selection problem. Conversely, we can regard the testing for significance of each regressor in MLR setting as a special case of variable selection procedure w.r.t. least square loss.

## (3) Principal components analysis.

- PCA often greatly reduces the collinearity problem yet provides non-interpretable regressors (as a linear combination of predictors).
- PCA often reduces the number of regressors so it is a good method from the perspective of parsimony.

- (c) PCA is a scale-dependent method, which means that change of scale will affect the result of PCA. PCA gives more weights on larger scale data.
- (d) PCA is highly dependent on the sampling plan since we need to yield a precise variance-covariance estimate.



## CHAPTER 6

# Testing and ANOVA

(1) State the model

Always state the model in format of “REMVM”, especially the ANOVA model and side conditions before doing any test.

(2) F-test

The nested hypothesis is the necessity that we carry out this F-test for *nested* model significance

(a) Simple case

$$H_0 : \mathbf{E}(Y|X_1 = x_1, X_2 = x_2) = x_1\beta_1 \text{ v.s. } H_1 : \mathbf{E}(Y|X_1 = x_1, X_2 = x_2) = x_1\beta_1 + x_2\beta_2$$

The testing statistic is  $F := \frac{SSE_{H_0} - SSE_{H_1}}{df_{H_0} - df_{H_1}} / \frac{SSE_{H_1}}{df_{H_1}} = \frac{SS_{reg}}{df_{reg}} / \hat{\sigma}^2 \stackrel{H_0}{\sim} F(df_{H_0} - df_{H_1}, df_{H_1})$ . Here we should always put the “research hypothesis” into  $H_1$  in order to prove it, we cannot prove  $H_0$  using statistical results.

(b) General case

$$H_0 : \mathbf{E}(Y|X_1 = x_1 \cdots X_q = x_q) = (\mathbf{x}_1, \cdots \mathbf{x}_q) \beta \text{ v.s. } H_1 : \mathbf{E}(Y|X_1 = x_1 \cdots X_p = x_p) = (\mathbf{x}_1, \cdots \mathbf{x}_p) \beta, p \geq q$$

$$\text{The testing statistic is } F := \frac{SSE(x_1, \cdots x_q) - SSE(x_1, \cdots x_p)}{q - p} / \frac{SSE(x_1, \cdots x_p)}{p - 1} = \frac{SS_{reg}}{df_{reg}} / \hat{\sigma}^2 \stackrel{H_0}{\sim} F((q - 1) - (p - 1), p - 1)$$

**Caution:** Since F-tests simply compare the (weighted) means, if two sets of data have too many uncontrolled variables then such a comparison might not work very well. In such a case we use confident interval to carry out comparison, see if the simultaneous CIs overlap. In ANOVA, only the main effects are independent, can we compare

(3) The ANOVA

(a) Principle of marginality<sup>1</sup>

If we take this principle, we usually compare the highest order terms and no other lower terms if there is no significant difference.

- (i) Type I ANOVA: Type I ANOVA is the sequential analysis of variance, which fits models according to the order that the regressors entered into the mean function. This tests the main effect of factor A, followed by the main effect of factor B after the main effect of A, followed by the interaction effect AB after the main effects.

---

<sup>1</sup>When the design is balanced, three types of sums are equal.

- (ii) Type II ANOVA: Type II ANOVA is the ANOVA table obtained under the principle of marginality. This tests each main effect after the other main effect. Note that no significant interaction is assumed (in other words, you should test for interaction first ( $SS(AB|A, B)$ ) and only if AB is not significant, continue with the analysis for main effects). Therefore it makes test from the highest order interactive term.
  - (iii) Type III ANOVA: Type III ANOVA test is computed for every regressor with adjustment to all other regressors including one. This ANOVA depends on the parameterization of the model and hence breaks the principle of marginality.
- (b) Principle of parsimony. When we try to reduce a model, always remember to do LOF and F-test for reduced model.

When the sample size is large, we prefer to use indicator model instead of single regression in order to capture the *great variability of the whole sample*. It is not always a good idea to reduce the model when the sample size is large. In a large sample, there are three main ways of introducing variability.

- (i) Find more lurking explanatory variables from EDA.
  - (ii) Create interaction/higher order terms from existing explanatory variables.
  - (iii) Use the factor/indicator variables to replace continuous variables.
- (c) ANOVA versus Regression. ANOVA is simply use *indicator variables* for each level in the data where regression usually use *a single (continuous) variable* to model. One thing to be mentioned is that if we use the actual value, then we can use regression to smooth the mean function and predict the response anywhere within the range of predictors. If we use factorial variables, prediction is only available at certain levels but we let the data itself to determine the spacing and ordering of the levels.
- (d) Comparison with factorial design.

Usually ANOVA will assume that all the regressors are uncorrelated with other regressors, which were true for most designs of experiments. However, if the regressors happen to be collinear, the ANOVA may yield weird results. And even in an experimental design setting, we still need all the effects to be orthogonal in order to pool some insignificant factors into error to yield enough degree of freedom to estimate  $\sigma^2$ .

More precisely,

- (i) If we are considering a design, then we implicitly assumes that each effect is uncorrelated with other effect even if they have

F-test are LRT for normal distributions or asymptotically normal distributions.

## CHAPTER 7

### Variances

#### 7.1 Weighted least squares

WLS		OLS $RSS_W(\beta) = (\mathbf{Y} - \mathbf{X}\beta)' \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta)$	Inference with normality
Model	$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$	$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} + \mathbf{e}$	$\mathbf{Y}_{\text{new}} = \mathbf{X}_{\text{new}}\hat{\beta} + \mathbf{e}_{\text{new}}$
Mean function	$\mathbf{X}\beta$	$\begin{cases} \hat{\beta} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \\ \hat{\mathbf{Y}} &= \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \\ \mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W} \end{cases}$	$\begin{cases} \begin{cases} \hat{\beta} &\sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}) \\ \hat{\mathbf{Y}} &\sim N(\mathbf{X}\beta, \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}') \end{cases} & \sigma^2 \text{ known} \\ Use \begin{cases} Bonferroni CI \\ Scheffe CI \end{cases} & \sigma^2 \text{ unknown} \end{cases}$
Variance function	$\sigma^2\mathbf{W}^{-1}$ $\mathbf{W} = \text{diag}(w_i)$ $w_i, \text{weights}$	$\begin{aligned} \hat{\sigma}^2 &= \frac{RSS_W(\hat{\beta})}{n-p} \\ p &= \text{rank}\mathbf{X} \end{aligned}$	$\begin{cases} \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p) \\ sefit(\mathbf{Y} \mathbf{X}_{\text{new}}) &= \hat{\sigma}\sqrt{\mathbf{x}'_{\text{new}}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}} \\ sepred(\mathbf{Y} \mathbf{X}_{\text{new}}) &= \hat{\sigma}\sqrt{\frac{1}{\mathbf{w}_{\text{new}}} + \mathbf{x}'_{\text{new}}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}} \end{cases}$
Error	$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{W}^{-1})$	$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$	

#### 7.2 Generalized least squares

GLS		OLS $RSS_G(\beta) = (\mathbf{Y} - \mathbf{X}\beta)' \Sigma^{-1} (\mathbf{Y} - \mathbf{X}\beta)$	Inference with normality
Model	$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$	$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} + \mathbf{e}$	$\mathbf{Y}_{\text{new}} = \mathbf{X}_{\text{new}}\hat{\beta} + \mathbf{e}_{\text{new}}$
Mean function	$\mathbf{X}\beta$	$\begin{cases} \hat{\beta} &= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y} \\ \hat{\mathbf{Y}} &= \mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y} \\ \mathbf{H} &= \mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1} \end{cases}$	$\begin{cases} \begin{cases} \hat{\beta} &\sim N(\beta, \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}) \\ \hat{\mathbf{Y}} &\sim N(\mathbf{X}\beta, \sigma^2\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}') \end{cases} & \sigma^2 \text{ known} \\ Use \begin{cases} Bonferroni CI \\ Scheffe CI \end{cases} & \sigma^2 \text{ unknown} \end{cases}$
Variance function	$\Sigma$	$\begin{aligned} \hat{\sigma}^2 &= \frac{RSS_G(\hat{\beta})}{n-p} \\ p &= \text{rank}\mathbf{X} \end{aligned}$	$\begin{cases} \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p) \\ sefit(\mathbf{Y} \mathbf{X}_{\text{new}}) &= \hat{\sigma}\sqrt{\mathbf{x}'_{\text{new}}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}} \\ sepred(\mathbf{Y} \mathbf{X}_{\text{new}}) &= \hat{\sigma}\sqrt{\frac{1}{\mathbf{w}_{\text{new}}} + \mathbf{x}'_{\text{new}}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}} \end{cases}$
Error	$\mathbf{e} \sim N(\mathbf{0}, \Sigma)$	$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$	

#### 7.3 Miscellanea

##### (1) Weights

- (a) When the weights increase, the variance estimates are under-estimates, the WLS of coefficients are over-estimated, the hypothesis tests of parameter functions have higher type II error.

- (b) When the weights decrease, the variance estimates are over-estimates, the WLS of coefficients are under-estimated, the hypothesis tests of parameter functions have higher type I error.
  - (c) The definition of residuals are very tricky in WLS setup because the sum of squares of these residuals will NOT equal the residual sum of squares, so in order to maintain the equality we have to use Pearson residual  $\hat{e}_i = \sqrt{w_i}(y_i - \hat{\beta}x_i)$  where the  $w_i$  is the weight of the  $i$ -th case.
- (2) Samplings
- (a) Finite population approach: Treat the population as if a finite sample, the regression fitting the data is understood as an approximation to the truth.
  - (b) Superpopulation approach: Treat the population as if it were a random sample from a theoretical superpopulation, the regression fitting the data is understood as usual.
  - (c) Where the weights come from?
    - (i) Adjusting for nonconstant variance in the data.
    - (ii) Reflect sampling weights in sampling.  
 Assume the inclusion probability  $\pi_i$  determined by each sampling procedure gives  $w_i = \frac{1}{\pi_i}$  as the sampling weight of each sample. Therefore the LS can be used to capture the sampling procedure.
  - (d) *Random sampling* allows one to generalize what one sees in the cases actually observed to the population from which they were chosen from.
- (3) Misspecified variances: Use the OLS to fit data generated from WLS, the variance of parameter estimation will inflate. These methods are ONLY for variance correction, do NOT refit WLS using these variances as weights.

These *sandwich estimators* are conservative.

- (a) Huber-White correction:  $Var(\hat{\beta}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\text{diag}(\hat{e}_i)\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$
  - (b) HC3 correction:  $Var(\hat{\beta}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\text{diag}\left(\frac{\hat{e}_i}{(1-h_{ii})^2}\right)\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{-1}$
- (4) General correlation structure

If we have  $n$  observations and  $\Sigma$  is completely unknown, then the total number of parameters is the number of regression coefficients  $p$  plus  $n$  variances on the diagonal of  $\Sigma$  plus  $\frac{n(n-1)}{2}$  covariances on the off-diagonals of  $\Sigma$ , i.e.  $p + n + \frac{n(n-1)}{2} > n$ . So the only solution to fit a GLS model is to assume some kind of correlation structure which will reduce the number of all unknown parameters but not violating the homoscedasticity too much.

- (a) Compound symmetry  $\Sigma_{CS} := \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}_{n \times n}$
- (b) Auto-regressive  $\Sigma_{AR} := \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{n-1} & \cdots & \rho & 1 \end{pmatrix}_{n \times n}$
- (c) Block-diagonal

## CHAPTER 8

# Transformations

### 8.1 Power transformations

- (1) Definition:  $\psi_\lambda(Y) = \begin{cases} Y^\lambda & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases}$
- (2) Usage: Transform the response/regressors simultaneously, each variable might have its own parameter value  $\lambda$ . This is used for linearity.
- (3) Attention
  - (a)  $Y$  should be strictly positive.
  - (b) **The log rule:** If the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable with its log.
  - (c) **The range rule:** If the range of a variable is considerably less than one order of magnitude, then any transformation of that variable is unlikely to be helpful.

### 8.2 Scaled power transformations

- (1) Definition:  $\psi_{S,\lambda}(Y) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases}$
- (2) Usage: Transform the response/regressors simultaneously, each variable might have its own parameter value  $\lambda$ . This is used for transformation selection.
- (3) Attention
  - (a)  $Y$  should be strictly positive.
  - (b) The scaled power transformations preserves the direction of relationship.
  - (c) Due to the fact that it preserves the correlation, The scale power transformations are used for selecting the corresponding power transformations usually.
  - (d)  $\lambda$  is chosen to minimize the residual sums of squares in the transformed model.

---

<sup>1</sup>Any transformations should be performed only on continuous random variables.

### 8.3 Box-Cox modified power transformations

- (1) Definition:  $\psi_{M,\lambda}(Y) = \begin{cases} \text{GeometricMean}(Y)^{1-\lambda} \cdot \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \text{GeometricMean}(Y) \cdot \log Y & \lambda = 0 \end{cases}$
- (2) Usage: Transform the response/regressors simultaneously, each variable might have its own parameter value  $\lambda$ .<sup>2</sup>
- (3) Attention
  - (a)  $Y$  should be strictly positive.
  - (b) The scaled power transformations preserves the direction of correlation.
  - (c) The scale power transformations are used for selecting the corresponding power transformations usually.
  - (d) If used for *transformation for linearity*,  $\lambda$  is chosen to minimize the residual sums of squares in the transformed model. Under normality assumption this is equivalent to maximizing the likelihood function. Velilla's criterion:  $\lambda$  is chosen to minimize the log of determinant of variance-covariance matrix  $\mathbf{V}(\lambda)$  in the transformed response.
  - (e) If used for *transformation for normality*,  $\lambda$  is chosen to make the residuals from the transformed model as close to normal as possible.

### 8.4 Yeo-Johnson transformations

- (1) Definition:  $\psi_{YJ,\lambda}(Y) = \begin{cases} \psi_{M,\lambda}(Y + 1) & Y \geq 0 \\ -\psi_{M,2-\lambda}(-Y + 1) & Y < 0 \end{cases}$
- (2) Usage: Transform the response/regressors simultaneously, each variable might have its own parameter value  $\lambda$ . This can be regarded as a multivariate version of the Box-Cox method, which can provide a preparational modification of linearity before we try to fit a linear model. The importance of doing such a preparational transform is that when we doing variable selection we are actually dealing with a linear model instead of a curved model.
- (3) Attention
  - (a)  $Y$  could be negative or positive. Another method is dealing with possible negative value is to translate the data wholly by the same amount  $\gamma$ .
  - (b) The scaled power transformations preserves the direction of correlation.
  - (c) The scale power transformations are used for selecting the corresponding power transformations usually.
  - (d) If used for transformation for linearity,  $\lambda$  is chosen to minimize the residual sums of squares in the transformed model.

---

<sup>2</sup>Although Box-Cox is derived for fixing the non-normality, sometimes the linearity will be affected after transformation and thus L and N have both to be checked after transformation.



- (e) If used for transformation for normality,  $\lambda$  is chosen to make the residuals from the transformed model as close to normal as possible.

### 8.5 Miscellanea

(1) Tradeoffs

These transformations are usually used for transformation for homoscedasticity. However, the goal of finding transformation for linearity OR normality OR homoscedasticity may not be simultaneously achieved. So after each transformation we may want to diagnostic again to ensure that the transformation is working all very well.

(2) Tukey's ladder of transformations

(3) Regressor-response order

We usually transform the predictors/regressors first to yield a linear mean function and since then we can yet still modify the response by transformation to make it seem more appropriate modelled by a linear model.

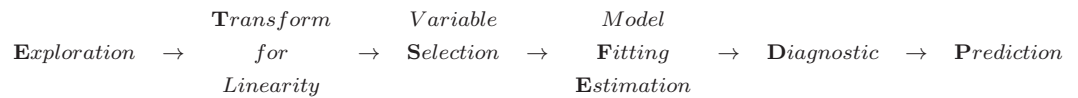
- (a) Inverse fitted value plot: Plot the  $Y$  against  $\hat{Y}$  (The fitted valued regressed on *the transformed response*  $\psi_\lambda(Y)$ ), if the regressors are approximately *linearly related*, the transformation with  $\lambda$  can be selected to minimize RSS or visually.

By calling a set of regressors linearly related, we mean that the graphs of any linear combination of these predictors versus any other linear combination of these predictors are linear.

- (b) The linearity is mostly concerned with the regressors while the normality is mostly concerned with the response, so Box-Cox is usually used for response transformations.

(4) General approach to linear regression analysis

- (a) Transform predictors in order to yield a set of linearly related regressors. i.e. The linear regressors all lie in the same linear space.
- (b) Do a formal variable selection procedure to establish a reasonable linear model to fit the observation dataset.
- (c) Use effect plot to show the observation response against the fitted mean function, and do a full diagnostic to see if we need to remedy any other issue.



## CHAPTER 9

# Regression Diagnostics

### 9.1 Linearity

#### (1) Graphical methods

- (a) Scatter plots: Nonlinear shape. Detecting the linearity of the trend.
- (b) Residual plots<sup>1 2</sup>: Nonlinear shape. Detecting the linearity of the residuals, we want to use RAW residuals here.<sup>3</sup>

(i) Pearson residual:  $p_i = \frac{\sqrt{w_i}(y_i - \hat{y}_{i(-)})}{\hat{\sigma}\sqrt{1 - x_i'(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})^{-1}x_i}}$ , used for having equality in SS.

(ii) Standardized residual:  $r_i = \frac{y_i - \hat{y}_{i(-)}}{\hat{\sigma}\sqrt{1 - x_i'(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})^{-1}x_i}}$ , used for having  $Var \equiv 1$ .

(iii) Studentized residual:  $t_i = \frac{y_i - \hat{y}_{i(-)}}{\hat{\sigma}(\hat{\sigma}_{(-i)})\sqrt{1 - x_i'(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})^{-1}x_i}}$ , used for independent numerator and denominator.<sup>4</sup>

#### (2) Analytic methods

- (a) Pearson's  $\chi^2$  fitness test for categorical data.

Suppose there are  $n$  observations in all,  $c$  is the number of categories and  $p$  is the degree of freedom or number of regressors+1.

$$H_0 : \mu_{Y|X} = \beta_0 + \beta_1 X \text{ v.s. } H_1 : \mu_{Y|X} \neq \beta_0 + \beta_1 X$$

$$\sum_{i=1}^c \frac{[Observed_i - Expected_i]^2}{Expected_i} = \sum_{i=1}^c \frac{\left[\frac{Observed_i}{n} - p_i\right]^2}{p_i} \overset{H_0}{\sim} \chi_p^2$$

- (b) Lack of fit test.

We need more than one observations at, at least one level, if there are some levels with a single replicate, that is also fine. Attention, we require a *valid replication* in each category. That is to say, if block effect or *pseudo-replication* structure is involved, then we might probably want to consider validness of LOF<sup>5</sup> test again.

Suppose there are  $n$  observations in all,  $c$  is the number of categories and  $p$  is the degree of freedom or number of regressors+1.

<sup>1</sup>One never plot residuals against  $Y_i$  since all of them were correlated.

<sup>2</sup>Although the null plot indicates that there is nothing wrong with the model, any other pattern of residual plot cannot be associated with a particular problem, it only indicates there is something wrong with the model but does not indicate what is exactly wrong.

<sup>3</sup>Residual plot is for SLR or marginal MLR; But in MLR the residual plots can be interpreted as in SLR if the predictors are correlated/linearly related.

<sup>4</sup>Since  $h_{ii}$  are NOT all the same, so the variance of residuals are NOT all the same even if the homoscedasticity is TRUE.

<sup>5</sup>Lack of fit can be understood as how good a regression model is compared to a mere mean valued model.

$$\begin{aligned}
H_0 : \mu_{Y|X} &= \beta_0 + \beta_1 X \text{ v.s. } H_1 : \mu_{Y|X} \neq \beta_0 + \beta_1 X \\
\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 &= \sum_{j=1}^c \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 + \sum_{j=1}^c n_j (\bar{Y}_{.j} - \hat{Y}_j)^2 \\
\Downarrow & \qquad \qquad \qquad \Downarrow \qquad \qquad \qquad \Downarrow \\
SSE &= SSPE + SSLF
\end{aligned}$$

(c) Lack of fit ANOVA: If the lack of fit test is significant, one use  $\hat{\sigma}^2 = MSPE = \frac{SSPE}{n-c}$ ; otherwise use  $\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$ .

Source	df	SS	F
Regression	$p - 1$	SSR	
Error	$n - p$	SSE=RSS	
Lack of Fit	$c - p$	SSLF	$\frac{SSLF/(c-p)}{SSPE/(n-c)}$
Pure Error	$n - c$	SSPE	
Total	$n - 1$	SStot	
$p = \text{rank}(\mathbf{X}); c = \text{number of categories}; n = \text{number of observations}$ $\text{rank}(\mathbf{X}) = p \stackrel{\text{model-fit}}{\leq} c \stackrel{\text{definition}}{\leq} n = \sum_{j=1}^c n_j$			

If there is a good amount of LOF, then all predictions from this model are not trustful. <sup>6 7</sup>

(d) Simple test for higher-power regressors: add new higher order regressor  $U$  and refit the model, see if the t-test for its coefficient is significant.

- (i) If  $U$  does not depend on estimated coefficients, then the test statistic should be compared with the standard normal to decide significant levels. (linearly independent of existing regressors)
- (ii) If  $U$  does depend on estimated coefficients, then use *Tukey's additive test* assuming the product interaction.

(3) Remedies

- (a) Including higher order trends: balance between explanatory power as well as complexity of the model; Pay attention to the principle of marginality.
- (b) Transformation for linearity
  - (i) Log transformation for regressors
  - (ii) Log transformation for response

<sup>6</sup>If the lack of fit test is *significant*, one must use MSPE as an estimate of  $\sigma^2$ ; Other wise we can still use MSE.

<sup>7</sup>One strategy for dealing with data that do not have repeated observations of the response variable at common sampling level of the explanatory variable is to pool **nearby** observations to permit estimation of pure error apart from the error from lack of fit. Further, we must check that whether we do such a pooling or not, the estimated coefficients of the model will not change much in order to validate our pooling.

## 9.2 Robustness

### Outlier

- (1) Graphical methods
  - (a) Scatter plots: Far-away point on the response scale.
  - (b) Residual plots: Far-away point on the residual scale. (Please use  $3\hat{\sigma}$ , ( $\hat{\sigma}$  is the estimated variance  $\frac{SSE}{n-p}$ ) principle to scan all possible outliers, which is the most reliable way even if the normality is broken in later checks. This is the strict rule, visual check is highly NOT recommended! )
- (2) Analytic methods
  - (a) Mean-shift outlier model
 

$H_0 : \delta = 0$  v.s.  $H_1 : \delta \neq 0$  assuming that  $\mathbb{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i + \delta$  while the general form of mean function is  $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$

To test this set of hypotheses, we can add an indicator regressor

$$U = \begin{cases} 1 & X = x_i \\ 0 & otherwise \end{cases}$$

into the model and test the significance of this regressor variable.

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + \mathbf{x}_i' (\mathbf{X}_{(i)}' \mathbf{X}_{(i)})^{-1} \mathbf{x}_i}} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}} \stackrel{H_0}{\sim} t(n - p - 1), \text{ exactly the same as studentized residual.}$$
  - (b) Residual distribution
 

Since residuals themselves  $\hat{e}_i = Y_i - \hat{Y}_i \sim N(0, \sigma^2)$ , if we use  $\hat{\sigma}^2 = MSE$  replaced  $\sigma^2$  and do a t-test, then we can find  $\alpha\%$  outliers by checking their residuals with  $t_{n-p, \alpha}$ . Bonferroni correction can be applied here, when there are  $m$  observations in total, we compare the  $\frac{Y_i - \hat{Y}_i}{\sqrt{MSE}}$  with  $t_{n-p, \frac{\alpha}{2m}}$ .
- (3) Remedies
  - (a) Throw one out and refit
  - (b) Sensitive consideration: Do not want a regression heavily depends on a few outliers.
  - (c) Possible cause of outliers:
    - (i) Lurking variable
    - (ii) Wrong collecting method of data
    - (iii) Wrong model and outliers should be included

### Leverage

- (1) Graphical methods:
  - (a) Scatter plots: Far-away point on the regress scale.
  - (b) Added-variable plots: Added-variable plot that do not match the general trend in the scatter plot might likely to be influential.
- (2) Analytic methods:
  - (a) Hat matrix: we directly calculate the leverage value of a data point by using the diagonal entries of the matrix.  $\frac{\partial \hat{\mathbf{Y}}}{\partial \mathbf{Y}} = \mathbf{H}$ ,  $h_{ii}$  represents

the rate of change of the fitted observation  $\hat{\mathbf{Y}}_i$  w.r.t. the actual data observed. Or it summarized how sensitive the fitted value will be to the observed data.

- (b) Cook's distance  $D_i = \frac{(\widehat{\beta_{\{i\}}} - \hat{\beta})' (\mathbf{X}'\mathbf{X}) (\widehat{\beta_{\{i\}}} - \hat{\beta})}{p\widehat{\sigma^2}} = \frac{\mathbf{Y}_{(-i)}' \mathbf{Y}_{(-i)}}{p\widehat{\sigma^2}} = \frac{1}{p} r_i^2 \frac{h_{ii}}{1-h_{ii}}$ , where  $r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$  is the standardized residual,  $p$  is the number of regressors.

The interpretation of Cook's distance is that if  $D_i$  were exactly equal to the  $F_{\alpha, p, n-p}$  cut-off point of an F-distribution, then deletion of the  $i$ -th case would move the estimate of  $\hat{\beta}$  to the edge of a  $1 - \alpha$  confidence region based on the complete data. So if  $D_i \ll 1$ , deletion of a case will NOT change the estimate of parameter much.

(3) Remedies

- (a) There could be error in collection of data OR on the contrary,
- (b) The regression fit is NOT appropriate except in the region defined without leverage/outlier.

### 9.3 Independence

(1) Graphical methods

- (a) Scatter plots: Non-band shape. Non-elliptic shape.
- (b) Residual plots: Unusual nonrandom pattern.

(2) Analytic methods

(a) Wald test

$H_0 : e'_i \text{ are independent v.s. } H_1 : e'_i \text{ are dependence}$

*Testing statistics:* Number of runs (consecutive strings)  $u \stackrel{H_0}{\sim} N\left(\frac{2n_-n_+}{n} + 1, \frac{2n_-n_+(2n_-n_+-n-n)}{n^2(n-1)}\right)$ ,  $Z = \frac{u - \mu_u + C}{\sigma_u} \sim N(0, 1)$

(b) Durbin-Watson test

$H_0 : \rho = 0$  v.s.  $H_1 : \rho \neq 0$  assuming that  $\epsilon_i = \rho\epsilon_{i-1} + u_i, u_i \stackrel{i.i.d}{\sim} N(0, \sigma^2), |\rho| < 1$

*Testing statistics:* Number of runs (consecutive strings)  $D = \frac{\sum_{i=2}^n (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^n \hat{e}_i^2} \approx$

$$2 \left( 1 - \frac{\sum_{i=2}^n \hat{e}_i \hat{e}_{i-1}}{\sum_{i=1}^n \hat{e}_i^2} \right)$$

If  $D < 2 (> 2)$  then we have evidence of positive(negative) correlation since it means the value of residuals in the numerator are often close to(far from) previous one, positively(negatively) correlated and hence the sum of differences will be small.

If  $D = 2$  then we have evidence that there is no relation between errors.

### 9.4 Homoscedasticity

(1) Graphical methods

- (a) Scatter plots: Non-band shape. Non-elliptic shape.

- (b) Residual plots: Megaphone shape. We want to use STD residuals here.

(2) Analytic methods

- (a) Brown-Forsythe test

$H_0$  : constant variance v.s.  $H_1$  : nonconstant variance

Testing statistics: Number of runs(consecutive strings)  $t_{BF}^* = \frac{\overline{d_1 - d_2}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim}$

$t(n-2)$  where  $\overline{d_1}, \overline{d_2}$  are mean of

$$\{d_{i1} := |e_{i1} - \text{median of group1}|\}, \{d_{i2} := |e_{i2} - \text{median of group2}|\}$$

$$\text{and } s^2 = \frac{\sum_{i=1}^{n_1} (d_{i1} - \overline{d_1})^2 + \sum_{i=1}^{n_2} (d_{i2} - \overline{d_2})^2}{n-2}$$

- (b) Breusch-Pagan OR Cook-Weisberg test

$H_0$  :  $\gamma_1 = 0$  v.s.  $H_1$  :  $\gamma_1 \neq 0$  assuming that  $\log(\sigma_i^2) = \gamma_0 + \gamma_1 X_i$

Testing statistics: Number of runs(consecutive strings)  $X_{CW}^2 = \frac{\frac{SSR^*}{2}}{\left(\frac{SSE}{n}\right)^2} \overset{H_0}{\sim}$

$\chi_1^2$  where  $SSR^*$  is the SSR when regressing  $\hat{\epsilon}^2$ 's on  $X$ .

If  $D < 2 (> 2)$  then we have evidence of positive correlation since it means the value of residuals in the numerator are often close to (far from) previous one, positively (negatively) correlated and hence the sum of differences will be small.

This test is not sensitive to the specific form of variance. We can also compare nested choices for  $X_i$  by taking the difference between each test and comparing the resulting value with  $\chi^2$  distribution of df equaling to the difference in their dfs.

(3) Remedies

- (a) Transformation for homoscedasticity (Variance-stabilizing transformations)

(i)  $\psi(Y) = \sqrt{Y}$  when  $Var(Y|X) \propto E(Y|X)$  like Poisson-distributed data.

(ii)  $\psi(Y) = \log(Y)$  when  $Var(Y|X) \propto [E(Y|X)]^2$  like multiplicative error.

(iii)  $\psi(Y) = \frac{1}{Y}$  when  $Var(Y|X) \propto [E(Y|X)]^4$  like most responses are close to zero.

(iv)  $\psi(Y) = \arcsin(\sqrt{Y})$  when  $Y \in [0, 1]$  as some sort of ratio.

- (b) Use WLS/GLS: The replications can be used for approximating weights.

- (c) Use Generalized linear models accounting for the non-homoscedasticity.

- (d) Estimate the variance of a non-linear parametric function

(i) Delta method:  $\hat{\theta} \sim N(\theta, \Sigma) \Rightarrow \mathbf{g}(\hat{\theta}) \sim N(\mathbf{g}(\theta_{true}), \nabla \mathbf{g}(\theta_{true})' \Sigma \nabla \mathbf{g}(\theta_{true}))$

(ii) Bootstrapping: Use the resampling with replacement to estimate the theoretical distributions. This can tell us the exact variances approximately. Bootstrapping can also be understood as new simulated datasets are created from re-weighting

the existing observations. This is exactly how Bayesians will draw a posterior sample.

- (iii) Use the correction of mis-specified variance at the cost of decreased efficiency of estimates.

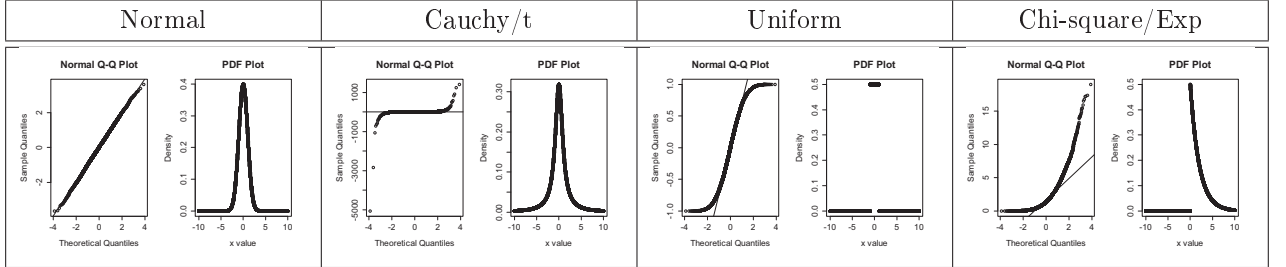
## 9.5 Normality

### (1) Graphical methods

- (a) Scatter plots: Non-elliptic shape.
- (b) Residual plots: Patterned shape.
- (c) Residual histogram: The correct normality should exhibits a bell-shape while a wrong histogram might deviate from bell-shape. Pay special attention to heavy tails.
- (d) Q-Q plots: Non-45 degree straight line. Detecting the normality assumption imposed on the error term. This also refer to *normality outlier problem* in another sense.

One thing to mention is that there is a kind of paradox that the outliers are usually near the tail of the q-q plots meanwhile the great uncertainty also arise around the tails.

$$\mathbb{E}(\widehat{e}_{(i)}) = \mu + \sigma z_{(i)}, z_{(i)} \sim N(0, 1) \text{ random ordered sample.}$$



### (2) Analytic methods

- (a) Kolmogorov-Smirnov test

$H_0$  : *normality* v.s.  $H_1$  : *nonnormality*

*Testing statistics*:  $D_n = \sup_y |F_n(y) - F(y)|$  where  $F_n(y)$  is the empirical c.d.f. of a set of observations of size  $n$ , the  $F(y)$  is the theoretical c.d.f. which is normal distribution in our case.

- (b) Shapiro-Wilks test

$H_0$  : *normality* v.s.  $H_1$  : *nonnormality*

*Testing statistics*:  $W = \frac{\sum_{i=1}^n a_i y_{(i)}}{\sum_{i=1}^n (y_i - \bar{y})^2}$  where  $(a_1, \dots, a_n) := \frac{m'V^{-1}}{\sqrt{m'V^{-1}V^{-1}m}}$ ,  $(m_1, \dots, m_n)$  are the expected values of the order statistics from  $N(0, 1)$  with a covariance matrix  $V$ .

### (3) Remedies

- (a) Transformation for normality: Box-Cox family. We usually use the inverse residual plot to check if there is a need to transform the response in order to reach normality.

- (b) Throw out some outliers and explain the reason why we should throw them out in a practical background.
- (c) Generalized linear models.

## 9.6 Lurking

- (1) Graphical methods
  - (a) Scatter plots: One-dimensional pattern shape. For example, parallel lines might indicate some lurking.
  - (b) Residual plots: Patterned shape. When we plot the residuals against some other potential regressors, the trend should be very clear. Another possibility is that if a covariate which is NOT included in current model plotted against residuals and this residual plot seems good then the covariate might serve as a lurking variable.
- (2) Remedies
  - (a) Transformation for normality: Box-Cox family.
  - (b) Check the high dimensional scatterplot.

## 9.7 Size

This is a very tricky diagnostic, when the sample size is large, any predictor having *some association* with the response will look statistically significant. This is because the confident intervals will have shorter length as the sample size grows. In a large sample, we generally trust our eyes and try to introduce some more predictors.



## CHAPTER 10

# Variable Selection

### 10.1 Selection Criteria

#### (1) Parameter assessment

- (a) The correlation coefficient  $R^2 := \frac{SS_{reg}}{SS_{total}} = 1 - \frac{SS_{resid}}{SS_{total}}$
- (b) The partial correlation coefficient  $R^2_{Y, X_1|X_2} := R^2_{Y, X_1, X_2} - R^2_{Y, X_2}$ .

- (i) Correction for the t-test: For testing the following pair of hypotheses:

The model  $Y = \beta_0 + \beta_1 X_1$  is nested within  $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2$

$H_0 : \gamma_1 = 0$  v.s.  $H_1 : \gamma_1 \neq 0$ , let  $t_\gamma$  be the t-statistic for this test.

$H_0 : \beta_1 = 0$  v.s.  $H_1 : \beta_1 \neq 0$ , let  $t_\beta$  be the t-statistic for this test.

$t_\gamma^2 = t_\beta^2 \frac{n-d_1-d_2-1}{n-d_2-1} \cdot \frac{R^2_{Y, X_1|X_2}}{R^2_{Y, X_1}(1-R^2_{Y, X_2|X_1})}$ , where  $d_i$  is the number of regressors in  $X_i$ .

- (ii) Interpretation along with colinearity.

$R^2_{X_1, X_2} \approx 0 \iff \frac{R^2_{Y, X_1|X_2}}{R^2_{Y, X_1}} \approx 1 \iff$  If  $q$  is not too large, the analysis with the model including  $X_1, X_2$  would be appropriate.

$R^2_{X_1, X_2} \approx 1 \iff \frac{R^2_{Y, X_1|X_2}}{R^2_{Y, X_1}} \approx 0 \iff$  If  $q$  is not too large, the t-statistic will be very unstable and  $t_\gamma$  will be not useful.

$R^2_{X_1, X_2} \in (0, 1) \iff \frac{R^2_{Y, X_1|X_2}}{R^2_{Y, X_1}} \in (0, 1) \iff$  If  $q$  is not too large, adding new predictors correlated with response  $Y$  will generally increase  $t_\gamma$ .

- (c) The adjusted correlation coefficient  $R^2 := \frac{SS_{reg}/df_{reg}}{SS_{total}/df_{total}}$ , this is equivalent to choose a model with minimum MSE yet does not consider the complexity.

#### (2) Discovery

General approach to finding the active predictors is to consider all possible choices for active predictors and select those predictors optimizing some selection criteria. <sup>1 2</sup>

<sup>1</sup>Diagnostic is always needed after we finished the variable selection, since the variable selection procedure may be affected greatly by the colinearity as well as other problems.

<sup>2</sup>One type of question is to ask you to choose a best model based on the output of model selection procedure. Usually we are looking for a model with the lowest . Do not consider the principle of parsimony when you tried to interpret all the information criteria because they all incorporate the complexity measure of this model.

All these criteria for discovery is constructed based on two concerns:

- (1) The lack of fit of a model, the  $RSS_p$ .  $RSS_p$  is a terminology used for a specific training set of data,  $SSE_p$  is more often used when the whole dataset is used for model building, they are often the same thing.
- (2) The complexity of a model, the number  $p$  of regressors included in the model.

- (a) AIC.  $AIC := n \log(\frac{RSS_p}{n}) + 2 \cdot p$ , where  $p$  is the number of predictors and  $n$  is the sample size. Smaller values are preferred.<sup>3</sup>
- (b) BIC.  $BIC := n \log(\frac{RSS_p}{n}) + \log n \cdot p$ , where  $p$  is the number of predictors and  $n$  is the sample size. Smaller values are preferred.
- (c) Mallows's  $C_p := \frac{\widehat{MSPE}}{\hat{\sigma}^2} = \frac{\sum_i (y_i - \hat{y}_i)^2 - \hat{\sigma}^2(n-2p)}{\hat{\sigma}^2} = \frac{SSE_p}{MSE_n} - n + 2p = \frac{SSE_{reduced}}{MSE_{full}} - (full\ para - 2 \cdot reduced\ para)$  is a measure of predictor bias.

The mean square of prediction error is defined to be  $MSPE := \sum_i \mathbf{E}[y_i - \hat{y}_i]^2 = \sum_i (\mathbf{E}[y_i - \hat{y}_i])^2 + \sum_i Var(\hat{y}_i)$

Compare the  $C_p$  with the number of predictors  $p$ .  $C_p > p$ , bias dominate;  $C_p < p$  variance dominate.

- (d) F-statistic. We basically choose the model with the greatest increase of  $SS_{reg}$ , which means the most significant increase in the F-statistic.

### (3) Prediction

- (a) Cross-validation
- (b) Outer cross-validation

The outer validation is to use a validation set of data with “correct” response which we do not use to fit our model.

$CV := \sum_{i=1}^{N_{validation}} (Y_{prediction,i} - Y_{validation,i})^2$  where  $Y_{prediction}$  is from the fitted model and  $Y_{validation}$  is from validation set.

If  $\frac{CV}{N_{validation}-p} \gg MSE$ , where  $p$  is the number of predictors in the fitted model, then the original fitted model is not good at predicting.

- (c) PRESS:

This is like the Cook's distance used for outliers detection. And  $PRESS$  is a measure of LOF of the model, we generally prefer it to be smaller OR close to  $SSE$ .

$$PRESS := \sum_{i=1}^n (Y_i - \hat{Y}_{-i}(x_i))^2$$

If  $PRESS \gg SSE$ , then the original fitted model is not good at predicting.

## 10.2 Algorithms

- (1) Exhaustive method

Calculate the selection criterion for each of  $2^n$  possible models and choose afterwards.

<sup>3</sup>Sometimes you can also calculate  $RSS_p$  using the output of AIC/BIC to deduce  $RSS_p$ .

## (2) Backward elimination

We choose  $X_i$  with the *smallest*  $F_i^* := \frac{MS_{reg}(X_i \in MODEL | X_1, \dots, \hat{X}_i, \dots, X_p)}{MSE(X_1, \dots, X_p)}$  for  $X_i$  in current model.  $X_1, \dots, X_p$  are predictors in MODEL.

## (3) Forward selection

We choose  $X_j$  with the *largest*  $F_j := \frac{MS_{reg}(X_j \notin MODEL | X_1, \dots, X_p)}{MSE(X_1, \dots, X_p)}$  for  $X_j$  NOT in current model.  $X_1, \dots, X_p$  are predictors in MODEL.

## (4) Stepwise regression

We choose  $X_i$  with the smallest  $F_i^*$  and  $X_j$  with the largest  $F_j$  in a single step. This procedure usually provides a model with the lowest information criteria.

There is no guarantee that all of these models will reach the same optimal model. The resulting model for BE/FS/SR might NOT be the model with the optimal selection criterion.

### 10.3 Miscellanea

Model	Criterion	Loss	Variable selection method
MLR/Polynomial	$S(\beta) := \sum_{i=1}^n \ \mathbf{Y}_i - \mathbf{X}_i\beta\ _2^2$	$L^2$	F-test
Spline	$S(\beta) := \sum_{i=1}^n \ \mathbf{Y}_i - \mathbf{X}_i\beta\ _2^2 + \sum_{j=1}^p \lambda_j \int \left[f_j''\right]^2 dx$	$L^2$ +Smoothness	$\lambda_j \begin{cases} V = \frac{1}{n} \sum_{i=1}^n [\hat{y}_{-i}(x_i) - y_i]^2 & OCV \\ V = \frac{n}{tr(\mathbf{I}-\mathbf{H})^2} \sum_{i=1}^n [\hat{y}_i - y_i]^2 & GCV \end{cases}$
Ridge regression	$S(\beta) := \sum_{i=1}^n \ \mathbf{Y}_i - \mathbf{X}_i\beta\ _2^2 + \lambda \ \beta\ _2^2$	$L^2$ +Shrinkage	
LASSO	$S(\beta) := \sum_{i=1}^n \ \mathbf{Y}_i - \mathbf{X}_i\beta\ _2^2 + \lambda \ \beta\ _1^2$		
Logistic/Nonlinear	$S(\beta) := \log L(\beta \mathbf{X})$	Log-likelihood	

## CHAPTER 11

# Nonlinear Regression

## CHAPTER 12

# Binomial and Poisson Regression

### 12.1 Logistic regression

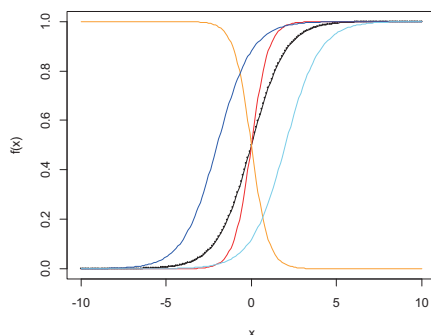
(1) Logistic link function

The logistic regression transformation function is  $\theta : (-\infty, \infty) \rightarrow (0, 1), x \mapsto$

$$\frac{e^x}{1+e^x};$$

Its inverse is known as *logit link function*  $x : (0, 1) \rightarrow (-\infty, \infty), \theta \mapsto$

$$\log\left(\frac{\theta}{1-\theta}\right)$$



Problems of using MLR to fit counted data.

- (a) Cannot take care of the natural range  $(0, 1)$ .
- (b) Cannot model the non-normal error term.
- (c) Cannot capture the nonconstant variance of each response because it depends on  $\beta$ .

(2) Logistic likelihood function

This is a conditional likelihood conditioned on the probability  $\theta$ .

$L(\theta) := \sum_{i=1}^N y_i \log p(x_i; \theta) + (1-y_i) \log(1-p(x_i; \theta))$  where  $p(x_i; \theta)$  is the probability of success in the  $i$ -th trial and  $y_i = 0, 1$  is the counting response of the data.

(3) Odds

The odds of success in a binomial model  $Binom(n, \theta)$  is  $\frac{\theta}{1-\theta}$ , which can be understood as:

- (a) If the probability of success is 0.25, then the odds are  $\frac{1}{3}$  which means 1 success to *each* 3 failures.
- (b) If the probability of success is 0.75, then the odds are 3 which means 3 success to *each* 1 failures.

## (4) Nonlinear model

As all other nonlinear model, the logistic model is fitted using MLE instead of OLE, so its inference is based on the asymptotic normality of MLE and there is no RSS associated with each fitted model.

So the coefficient estimate for the logistic model  $Y|X \sim \text{Binom}(n, \theta(\beta_0 + \beta_1 X))$  attain its coefficient estimate by minimizing

$$l(\beta) = \sum_{i=1}^n Y_i (\mathbf{X}\beta)' - \sum_{i=1}^n \log(1 + e^{\beta_0 + \beta_1})$$

- (a) For any estimate  $\hat{\beta}_1$ , a one-unit increase in  $X$  with all other regressors fixed, there is  $e^{\hat{\beta}_1}$  unit multiplicative factor increase in the *odd ratio*. The corresponding test is  $H_0 : \beta_1 = 0$  v.s.  $H_1 : \beta_1 \neq 0$  with confident interval  $\beta_1 \in \left[ \hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\beta}_1)} \right]$ , where  $\text{Var}(\hat{\beta}_1)$  is estimated by observed information number  $\frac{\partial^2 l}{\partial \beta_1^2} |_{\beta_1 = \hat{\beta}_1}$

- (b) For any estimate  $\hat{\beta}_1$ , a one-unit increase in  $X$  with all other regressors fixed, there is  $\hat{\beta}_1$  unit additive increase in the *log-odd ratio*. The corresponding test is  $H_0 : e^{\beta_1} = 1$  v.s.  $H_1 : e^{\beta_1} \neq 1$  with confident interval  $e^{\beta_1} \in \left[ e^{\hat{\beta}_1} \cdot e^{\pm Z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\beta}_1)}} \right]$ , where  $\text{Var}(\hat{\beta}_1)$  is estimated by observed information number  $\frac{\partial^2 l}{\partial \beta_1^2} |_{\beta_1 = \hat{\beta}_1}$

(5) Deviance:  $G_{M_1, M_2}^2 := -2 \log \left( \frac{L(\beta | \text{MODEL}_1)}{L(\beta | \text{MODEL}_2)} \right)$ 

When  $\text{MODEL}_2$  is a saturated model we get the saturated deviance we usually use,  $G_M^2 = 2 \sum_{i=1}^n \left[ Y_i \log\left(\frac{Y_i}{Y_i}\right) + (1 - Y_i) \log\left(\frac{1 - Y_i}{1 - Y_i}\right) \right] \geq 0$  and  $G_{M_1, M_2}^2 = G_{M_1}^2 - G_{M_2}^2$

The larger the null deviance is, the less LOF the model has.

The smaller the (saturated) deviance is, the less LOF the model has.

To test the following hypothesis, we can make use of deviance  $G_{M_1, M_2}^2$ .

$H_0 : M_1$  is significant v.s.  $H_1 : M_2$  is significant

$G_{M_1, M_2}^2 = G_{M_1}^2 - G_{M_2}^2 \stackrel{H_0}{\sim} \chi_{df_1 - df_2}^2$ . where  $df_i$  is the degree of freedom in the  $i$ -th model.

## 12.2 Poisson regression