



MATH

564/484

Simple Linear
Regression

Lulu Kang

MATH 564/484 Applied Statistics/Regression

Linear Regression

Lulu Kang

Department of Applied Mathematics
RM 234 in RE Building, Email: lkang2@iit.edu,
Thur 1:30–3:30 pm or by appointment.

Course Information on Blackboard



DATA111

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

Part I

Simple Linear Regression

- 2.1 Overview of Supervised Learning
- 2.2 Simple Linear Regression
- 2.3 Introduction to R



Overview of Supervised Learning

DATA

564 rows

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

Supervised Learning (directed data mining, learning with a teacher):

- The observed data is of the form of $(Y_i, X_{i,1}, \dots, X_{i,p})$ for $i = 1, \dots, n$, where the variables can be split into two groups:
 - **independent variables** (explanatory variables, inputs, predictors)
 $X = (X_1, \dots, X_p)$ and
 - One (or more) **dependent variable** (outputs, responses) Y .
- The objective is to predict Y given values of the input X .



Supervised Learning

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

- Observed Data (Training Data):

$$(Y_i, X_{i,1}, \dots, X_{i,p}) \text{ for } i = 1, \dots, n$$

- Objective: find a function $f(x_{new}) = f(x_1, \dots, x_p)$ that can predict Y well for any given input $x_{new} = (x_1, \dots, x_p)$.
- Deterministic relationship? (many classification tasks in machine learning)



The Additive Error Model

DATA

564/494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS
Inference
Summary

Introduction
to R

- Key Statistical Ideas: Observed Data = True Value + Noise
- For the observed training data,

$$Y_i = f(x_{i,1}, \dots, x_{i,p}) + \epsilon_i$$

for $i = 1, \dots, n$, where the error ϵ_i 's are iid with mean 0 and are independent of X 's.

- Find the function $f(x_1, \dots, x_p)$ or find its approximation!!! (Generative vs. Predictive models)
- The simplest case: when $p = 1$, $f(x) = \beta_0 + \beta_1 x$
Simple linear regression: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$



Empirical Models: Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

- Many engineering and scientific problems are concerned with determining a relationship between a set of variables.
- For example: $Y = \text{College GPA at 1st year}$; $X = \text{high school GPA}$
Or $Y = \text{Mortality rate}$; $X = \text{Immunization}$.
- Knowledge of such a relationship would enable us to predict the output for Y .
- **Regression analysis** is a statistical technique that is very useful for these types of problems, as it can be used to build a model to predict Y at a given X value.



Example: Immunized and Mortality

DATA

564 494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS
Inference
Summary

Introduction
to R

- Suppose one wants to investigate the relationship between the percentage of children who have been immunized against the infectious disease diphtheria, pertussis, and tetanus (DPT) in a given country and the corresponding mortality rate for children under five years of age in that country.
- The UN Children's Fund (UNICEF) considers the under-five mortality rate to be one of most important indicators of the level of well-being for children.



Data

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS
Inference
Summary

Introduction
to R

- X = Percentage of children immunized against DPT;
- Y = under-five mortality rate per 1000 live births, in 1992

Nation	X	Y	Nation	X	Y	Nation	X	Y
Bolivia	77	118	Ethiopia	13	208	Mexico	91	33
Brazil	69	65	Finland	95	7	Poland	98	16
Cambodia	32	184	France	95	9	Russian	73	32
Canada	85	8	Greece	54	9	Senegal	47	145
China	94	43	India	89	124	Turkey	76	87
Czech Republic	99	12	Italy	95	10	UK	90	9
Egypt	89	55	Japan	87	6			



Scatter plot

- DATA
- 564 observations
- Simple Linear Regression
- Lulu Kang
- Supervised Learning
- Simple Linear Regression
- Introduction via examples
- Linear Regression Model
- Least Square Estimation
- Properties of OLS
- Inference
- Summary
- Introduction to R

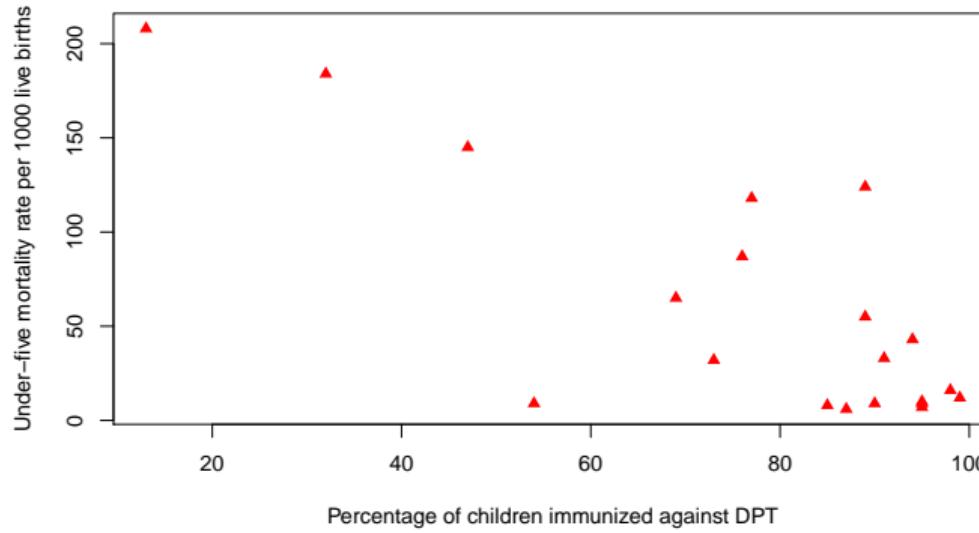


Figure: The plot shows that Mortality rate tends to decrease as the percentage of children immunization increases.



Question

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- X = Percentage of children immunized against DPT;
- Y = under-five mortality rate per 1000 live births, in 1992

Question:

- Are Y and X related (associated), and how?
- Does better immunization improve mortality rate?
- Can we use the data to develop a model for predicting under-five mortality rate from the percentage of children immunized against DPT?



Linear Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- It is interesting both theoretically because of the elegance of the underlying theory, and from an applied point view, because of the wide variety of uses.
- Fit a models for a dependent variable as a function of one or more independent variables.
- We will talk about
 - Building models
 - Assessing fit and reliability
 - Drawing conclusions



A Simple Linear Regression

DATA

564 rows

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

- We are interested in developing a linear equation that best summarizes the relationship in a sample between the response variable (Y) and the predictor variable (or independent variable) X

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the ϵ_i 's are independent with mean 0 and variance σ^2 .

- The equation is also used to predict Y from X .

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Interpretation

DATA

564 494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

$$y = \mu_{y|x} + \epsilon = \beta_0 + \beta_1 x + \epsilon$$

- $\mu_{y|x}$ is the conditional expectation of Y given $X = x$, i.e., $\mu_{y|x} = E(Y|X = x)$ or $E(Y|x)$.
- Linear regression assumes that $E(Y|x) = \beta_0 + \beta_1 x$, linear in x and β_0 and β_1 .
- The mismatch between the true possible value of Y and $E(Y|x) = \beta_0 + \beta_1 x$ is called *error*, denoted as ϵ .
- Linear regression assumes that ϵ_i from the data $\{y_i, x_i\}$ are iid, with $E(\epsilon_i) = 0$, and $\text{var}(\epsilon_i) = \sigma^2$, and ϵ_i are also independent of Y and X .



Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

A consultant for an electrical distributor is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week and the time required to prepare the bids. Suppose that regression model is applicable and is as follows:

$$Y_i = 9.5 + 2.1X_i + \epsilon_i$$

where X is the number of bids prepared in a week and Y is the number of hours required to prepare the bids.



Example (Cont.)

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

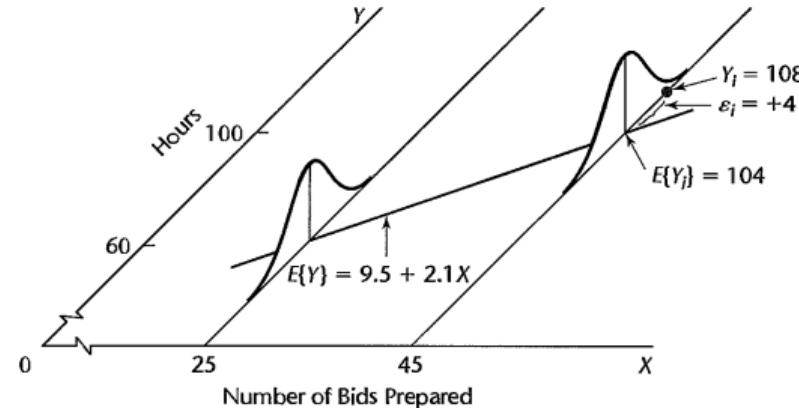


Figure: This figure shows the probability distribution of Y when $X = 25$. Note that the distribution exhibits the same variability as the probability distribution when $X = 45$.



Interpolation

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

- β_0 is the y -intercept. β_0 is the mean value of y when x is equal to 0. (Figure 3.2 (c))
- β_1 is the slope. β_1 is the change (amount of increase or decrease) in the mean value of y associated with a one-unit increase in x . If β_1 is positive, the mean value of y increases as x increases. If β_1 is negative, the mean value of y decreases as x increases.
- Both β_0 and β_1 are called *regression parameters* or *regression coefficients*.



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

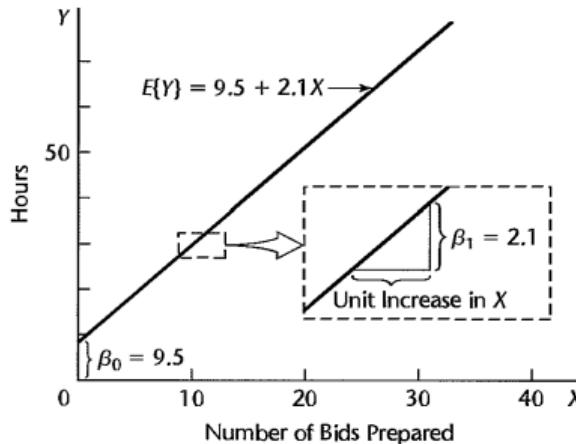
Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R



The slope $\beta_1 = 2.1$ indicates that the preparation of one additional bid in a week leads to an increase in the mean of the probability distribution of Y of 2.1 hours. The intercept $\beta_0 = 9.5$ indicates the value of regression function at $X = 0$. However, since the linear regression model was formulated to apply to weeks where the number of bids prepared ranges from 20 to 80, β_0 does not have any intrinsic meaning of its own here.



(a) How to estimate β 's

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- Observe n data (Y_i, x_i) , and assume

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the ϵ_i 's are independent with mean 0 and variance σ^2 .

- How to estimate β 's



Method of Least Squares

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- The (ordinary) least squares estimator: Choose β_0 and β_1 to minimize the residual of sum square (RSS)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2$$



Why Least Squares? [Board Derivation]

DATA

564/494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

- It is the Maximum Likelihood Estimators (MLE) of β_0 and β_1 when the errors ϵ_i 's are iid $N(0, \sigma^2)$.
- It leads to the best linear unbiased estimators (BLUE) of β_0 and β_1 , no matter whether the error ϵ_i 's are normally distributed or not.

[A linear estimator is of the form $\sum_{i=1}^n c_i Y_i$. The meaning of BLUE for β_1 :

$$\min \text{var}(\sum c_i Y_i) = \sigma^2 (\sum c_i^2)$$

$$\text{subject to } E(\sum c_i Y_i) = \sum c_i (\beta_0 + \beta_1 x_i) = \beta_1$$

for all β_0 and β_1 , i.e., subject to $\sum c_i \beta_0 = 0$ and $\sum c_i x_i = 1$.]



Method of Least Squares [Board Derivation]

DATA

564 494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

When minimizing the residual of sum of squares (RSS)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2$$

the solutions are:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}$$

and $\bar{x} = \sum_{i=1}^n x_i / n$ and $\bar{y} = \sum_{i=1}^n y_i / n$.



Example (Cont.)

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS
Inference
Summary

Introduction
to R

- X = Percentage of children immunized against DPT;
- Y = under-five mortality rate per 1000 live births, in 1992

Nation	X	Y	Nation	X	Y	Nation	X	Y
Bolivia	77	118	Ethiopia	13	208	Mexico	91	33
Brazil	69	65	Finland	95	7	Poland	98	16
Cambodia	32	184	France	95	9	Russian	73	32
Canada	85	8	Greece	54	9	Senegal	47	145
China	94	43	India	89	124	Turkey	76	87
Czech Republic	99	12	Italy	95	10	UK	90	9
Egypt	89	55	Japan	87	6			



Answer

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

■ For our data

$$n = 20, \bar{x} = 77.4, \bar{y} = 59, \sum_i x_i^2 = 130446, \sum_i x_i y_i = 68626$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2 = 10630.8$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} = -22706$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-22706}{10630.8} = -2.1359$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 59 + 2.1359 \times 77.4 = 224.3163$$

■ Thus, the fitted (simple linear regression) model is

$$\hat{Y} = 224.3163 - 2.1359x$$



(b) Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992 The fitted (simple linear regression) model is

$$\hat{Y} = 224.3163 - 2.1359x$$

- Estimate the mean under-five mortality rate per 1000 live births when $x=10$?
- Repeat the question when $x= 90$?

[202.9573; 32.0853]



(c) How to estimate σ^2 ?

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- Recall that the model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the ϵ_i 's are iid with mean 0 and variance σ^2 .

- We got the estimator $\hat{\beta}_0$, and $\hat{\beta}_1$ and how to estimate the third parameter σ^2 ?

Answer:

- It is natural to use the observed fitting error $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ and the residual sum of squares $RSS = \sum_{i=1}^n e_i^2$.
- The estimator of σ^2 is $\hat{\sigma}^2 = \frac{RSS}{n-2}$ [and $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$.]
- In practice, it is easier to compute RSS as follows:

$$RSS = \sum_{i=1}^n e_i^2 = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \quad [\text{Board Drivation}]$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

In our example, the fitted (simple linear regression) model is

$\hat{Y} = 224.3163 - 2.1359x$. Find an estimation of $\sigma^2 = \text{var}(\epsilon)$.

■ Two ways to calculate the residual sum of squares RSS:

- Calculate the observed fitting error (residual) $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ and then $RSS = \sum_{i=1}^n e_i^2 = 29000.95$.
- Use $S_{xx} = 10630.8$, $S_{xy} = -22706$, $S_{yy} = 77498$, and

$$RSS = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 77498 - (-22706)^2 / 10630.8 = 29000.95.$$

■ The estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = 1611.164 \quad (\text{or } \hat{\sigma} = \sqrt{1611.164} = 40.1393).$$



R code (calculator-type)

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

```
x <- c(77, 69, 32, 85, 94, 99, 89, 13, 95, 95, 54,  
89, 95, 87, 91, 98, 73, 47, 76, 90);  
y <- c(118, 65, 184, 8, 43, 12, 55, 208, 7, 9, 9,  
124, 10, 6, 33, 16, 32, 145, 87, 9);  
Sxx <- sum( x * x) - length(x) * (mean(x))^2  
Sxy <- sum(x *y ) - length(x) * mean(x) * mean(y)  
Syy <- sum( y * y) - length(y) * (mean(y))^2  
beta1hat <- Sxy / Sxx  
beta0hat <- mean(y) - beta1hat * mean(x)  
### Two ways to compute RSS  
error <- y - (beta0hat + beta1hat * x)  
RSS <- sum( error * error) ### Or RSS <- Syy { Sxy^2 / Sxx  
sigma2hat <- RSS / (length(x) - 2)  
c(beta0hat, beta1hat, sigma2hat)
```



(d) Property of OLS estimators

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- To derive the statistical inference of the (ordinary) least squares $\hat{\beta}_1$ and $\hat{\beta}_0$, we need to find
 - $E(\hat{\beta}_i)$
 - $\text{var}(\hat{\beta}_i)$

Then by the central limit theorem, asymptotically

$$\frac{\hat{\beta}_i - E(\hat{\beta}_i)}{\sqrt{\text{var}(\hat{\beta}_i)}} \approx N(0, 1)$$



Normal distribution assumption

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- Previously, we only assume that $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2$ and ϵ_i 's are iid.
- Now to obtain the inference on the parameters, we need to assume the exact distribution of ϵ . The typical assumption is $\epsilon \sim N(0, \sigma^2)$.
- The distribution of $Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2)$, and for the data $\{y_i, x_i\}$, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and y_i 's are independent of each other, i.e., $\text{cov}(y_i, y_j) = 0$ for $i \neq j$.



(d) Properties of OLS [Board Derivation]

DATA

564/494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

- Unbiased: $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$
- Variance:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad \text{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

where

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

- Note that they are correlated

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{S_{xx}}$$



Distributions on $\hat{\beta}_1$

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- Since $\hat{\beta}_1$ is a linear combination of Y_i 's, and Y_i 's are normally distributed. So $\hat{\beta}_1$ is also normally distributed (Why?). Thus

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / S_{xx}).$$

- Using $\hat{\sigma}^2$ (or s^2) to replace σ^2 in the distribution of $\hat{\beta}_1$, from the definition of t-distribution,

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} \sim t_{n-2},$$

$df = n - 2$ is the degrees of freedom of the t-distribution.



Distributions on $\hat{\beta}_0$

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- Similarly, the distribution of $\hat{\beta}_0$ is

$$\hat{\beta}_0 \sim N \left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \right).$$

- Using $\hat{\sigma}^2$ (or s^2) to replace σ^2 in the distribution of $\hat{\beta}_0$,

$$\frac{\hat{\beta}_0 - \beta_0}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2},$$

$df = n - 2$ is the degrees of freedom of the t-distribution.



CI and Tests

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- Since σ^2 is unknown, consider $\hat{\sigma}^2 = \frac{RSS}{n-2}$ and thus

$$s.e(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \quad s.e.(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

- Then

$$\frac{\hat{\beta}_0 - \beta_0}{s.e.(\hat{\beta}_0)} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)}$$

have t-distribution with $n - 2$ degree of freedom.



(d1) Inference on β_1

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- When testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ the test statistic is

$$T_{obs} = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}}$$

and we reject H_0 if $|T_{obs}| \geq t_{\alpha/2, n-2}$

- A $1 - \alpha$ confidence interval on β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

The fitted (simple linear regression) model is $\hat{Y} = 224.3163 - 2.1359x$.

- Test $H_0 : \beta_1 = 0$, versus $H_1 : \beta_1 \neq 0$ at $\alpha = 5\%$ level.

[Recall $S_{xx} = 10630.8$, $\hat{\sigma} = 40.1393$, $t_{\alpha/2, n-2} = t_{0.025, 18} = 2.101$,

$$T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{-2.1359}{40.1393/\sqrt{10630.8}} = -5.533$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

The fitted (simple linear regression) model is $\hat{Y} = 224.3163 - 2.1359x$.

- Find a 95% confidence interval on β_1 .

[Recall $S_{xx} = 10630.8$, $\hat{\sigma} = 40.1393$, $t_{\alpha/2, n-2} = t_{0.025, 18} = 2.101$,
 $\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = -2.1359 \pm 0.8179 = [-2.9538, -1.3180]$]



(d2) Inference on β_0

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- When testing $H_0 : \beta_0 = b_0$ versus $H_1 : \beta_0 \neq b_0$ the test statistic is

$$T_{obs} = \frac{\hat{\beta}_0 - b_0}{s.e(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - b_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

and we reject H_0 if $|T_{obs}| \geq t_{\alpha/2, n-2}$

- A $1 - \alpha$ confidence interval on β_0 is

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

The fitted (simple linear regression) model is $\hat{Y} = 224.3163 - 2.1359x$.

- Test $H_0: \beta_0 = 210$ versus $H_1: \beta_0 \neq 210$ at $\alpha = 5\%$ level.

[Recall $S_{xx} = 10630.8$, $\hat{\sigma} = 40.1393$, $t_{\alpha/2, n-2} = t_{0.025, 18} = 2.101$, So

$$T_{obs} = \frac{\hat{\beta}_0 - b_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} = 0.455.$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

The fitted (simple linear regression) model is $\hat{Y} = 224.3163 - 2.1359x$.

- Find a 95% confidence interval on β_0

[Recall $S_{xx} = 10630.8$, $\hat{\sigma} = 40.1393$, $t_{\alpha/2, n-2} = t_{0.025, 18} = 2.101$, So

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} = 224.3163 \pm 66.0562 = [158.26, 290.37].$$



(d3) Inference on $\beta_0 + \beta_1 x_{new}$

For the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

For a given x_{new} what is the confidence interval for the mean response

$$E(Y) = \beta_0 + \beta_1 x_{new}$$

Point estimator: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new} = \sum_{i=1}^n \left(\frac{1}{n} + k_i(x_{new} - \bar{x}) \right) Y_i$

- $E(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) = \beta_0 + \beta_1 x_{new}$, $\text{var}(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}} \right)$.
- $\hat{\beta}_0 + \hat{\beta}_1 x_{new} \sim N \left(\beta_0 + \beta_1 x_{new}, \sigma^2 \left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}} \right) \right)$.
- The $1 - \alpha$ confidence interval on the mean response is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{new} \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

The fitted (simple linear regression) model is $\hat{Y} = 224.3163 - 2.1359x$.

- Find a 95% confidence interval on the mean under-five mortality rate when $x = 10$

[Recall $S_{xx} = 10630.8$, $\hat{\sigma} = 40.1393$, $t_{\alpha/2, n-2} = t_{0.025, 18} = 2.101$.

$$\hat{Y} \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}} = 202.9573 \pm 58.2641 = [144.6932, 261.2214]$$



(e) Prediction on new observation

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

For the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. How to predict future observation Y corresponding to a given x_{new} ?

- Point estimator $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$
- How about a confidence interval on Y ? This is often called *prediction interval*.



Key Idea

DATA

564/494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

For the future response

$$Y = \beta_0 + \beta_1 x_{new} + \epsilon_{future}$$

Consider the estimator $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$.

- $E(Y - \hat{Y}) = 0$



$$\begin{aligned}\text{var}(Y - \hat{Y}) &= \text{var} \left(\beta_0 + \beta_1 x_{new} + \epsilon_{future} - (\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \right) \\ &= \text{var}(\epsilon_{future}) + \text{var}(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \\ &= \sigma^2 + \frac{\sigma^2}{n} + (x_{new} - \bar{x})^2 \frac{\sigma^2}{S_{xx}}\end{aligned}$$



Key Idea (Cont.)

DATA

564/494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

For the future response

$$y = \beta_0 + \beta_1 x_{new} + \epsilon$$

Consider the estimate $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$.

- $$\frac{y - \hat{Y}}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}} \sim N(0, 1)$$

- So

$$\frac{y - \hat{Y}}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}} \sim T_{n-2}$$



Prediction Interval

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

For the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. How to predict future observation \hat{Y} corresponding to a given x_{new} ?

- Point estimator $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$
- The $1 - \alpha$ prediction interval is

$$\hat{Y} \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

The fitted (simple linear regression) model is $\hat{Y} = 224.3163 - 2.1359 \cdot X$.

- Find a 95% prediction interval on Y when $x = 10$

[Recall $S_{xx} = 10630.8$, $\hat{\sigma} = 40.1393$, $t_{\alpha/2, n-2} = t_{0.025, 18} = 2.101$.

$$\hat{Y} \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}} = 202.9573 \pm 102.5022 = [100.4551, 305.4595]$$



Example (Cont.)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

X = Percentage of children immunized against DPT;

Y = under-five mortality rate per 1000 live births, in 1992

The fitted (simple linear regression) model is $\hat{Y} = 224.3163 - 2.1359 \cdot X$.

- Find a 95% prediction interval on Y when $x = 90$

[Recall $S_{xx} = 10630.8$, $\hat{\sigma} = 40.1393$, $t_{\alpha/2, n-2} = t_{0.025, 18} = 2.101$.

$$\hat{Y} \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}} = 32.0853 \pm 87.0276 = [-54.9423, 119.1429]$$



Summary (I): point estimation

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

Assume that we observe (x_i, y_i) for $i = 1, \dots, n$ and we consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where the ϵ_i 's are iid with mean 0 and variance σ^2 .

■ Define

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2,$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

■ The least squares estimators are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Summary (II): Estimation of σ^2 and Inference

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- The estimator of σ^2 is $\hat{\sigma}^2 = \frac{RSS}{n-2}$ where $RSS = \sum_{i=1}^n e_i^2$ and residuals $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. In practice, it is better to use

$$RSS = \sum_{i=1}^n e_i^2 = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

- $\frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)} \sim T_{n-2}$; $s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$
- $\frac{\hat{\beta}_0 - \beta_0}{s.e.(\hat{\beta}_0)} \sim T_{n-2}$; $s.e.(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$



Summary (III): Inference

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

At a given x_{new}

- the point estimator of Y is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$
- a $1 - \alpha$ confidence interval on the mean response Y is

$$\hat{Y} \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$



What is R

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- R is a system for statistical computation and graphics
- It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files
- Free software
- OS: Windows, Unix, Linux, Mac OS.
- Homepage: <http://www.r-project.org>



Data With R

DATA

564-494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

- Objects: vector, factor, array, matrix, data.frame, ts, list
- Mode (numerical, character, complex, and logical); Length
- Read data stored in text (ASCII) files `read.table()`, `scan()`, and `read.fwf()`
- Saving data: `write(x, file=\data.txt")`, `write.table()` write in a file a `data.frame`.
- Generating data



Linear Regression in R

DATA

564 494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference

Summary

Introduction
to R

```
x <- c(77, 69, 32, 85, 94, 99, 89, 13, 95, 95, 54, 89,  
95, 87, 91, 98, 73, 47, 76, 90);  
y <- c(118, 65, 184, 8, 43, 12, 55, 208, 7, 9, 9, 124,  
10, 6, 33, 16, 32, 145, 87, 9);  
  
fm1 <- lm( y ~ x)
```

```
fm1
```

Call:

```
lm(formula = y ~ x)
```

Coefficients: (Intercept) x

```
224.316 -2.136
```



summary(fm1)

DATA

564 494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

```
> summary(fm1)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-99.97934	-16.57854	0.06684	20.84946	89.77608

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	224.3163	31.4403	7.135	1.20e-06 ***
x	-2.1359	0.3893	-5.486	3.28e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.14 on 18 degrees of freedom

Multiple R-Squared: 0.6258, Adjusted R-squared: 0.605

F-statistic: 30.1 on 1 and 18 DF, p-value: 3.281e-05



Confidence Interval on coefficients

DATA

564_494

Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

```
> confint(fm1)
              2.5%      97.5%
(Intercept) 158.262579 290.369998
x           -2.953763  -1.317976
```

```
> confint(fm1, level = 0.99)
              0.5%      99.5%
(Intercept) 133.817133 314.815444
x           -3.256453  -1.015286
```



Intervals for x_{new}

DATA
564-494
Simple Linear
Regression

Lulu Kang

Supervised
Learning

Simple Linear
Regression

Introduction via
examples

Linear Regression
Model

Least Square
Estimation

Properties of OLS

Inference
Summary

Introduction
to R

```
> xnew <- data.frame(x = c(10, 90))
## Confidence intervals on the mean response
> predict(fm1, xnew, interval="confidence", level=0.95)
      fit        lwr        upr
1 202.95759  144.69566  261.21953
2 32.08805   10.59907  53.57702

## Prediction intervals for future observations
> predict(fm1, xnew, interval="prediction", level=0.95)
      fit        lwr        upr
1 202.95759  100.45917  305.4560
2 32.08805   -54.93637 119.1125
```



DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

Part II

R-square, Correlation Analysis and Variable Transformation

- 2. R-square**
- 2. Correlation Analysis**
- 2. Variable Transformation**



Simple Linear Regression: R-square

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

Simple Linear Regression: assume we observe n pairs of data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. The model is $y_i = \beta_0 + \beta_1 x + \epsilon_i$ where the ϵ_i 's are iid $N(0, \sigma^2)$.

- R-square: a statistic that effectively summarizes how well the X can be used to predict Y .

$$\begin{aligned} R^2 &= 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= 1 - \frac{S_{yy} - S_{xy}^2/S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} \end{aligned}$$



R-square

DATA

564/494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- Here SS_{err} is the sum of squares of errors, which is the same as residual sum of squares RSS .
- SS_{tot} stands for sum of squares of total variation, meaning when the assumption $\beta_1 = 0$ is true, $\hat{\beta}_0 = \bar{y}$ and $\sum_{i=1}^n (y_i - \bar{y})^2 = (n-1) \times \text{sample variance of } y_1, \dots, y_n$.
- R^2 is also called *coefficient of variation* or *coefficient of determination*.
- The larger R^2 is, the more the predictor X explains the variability in the response Y .



R-square

DATA

564/494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- When all observations fall on the fitted line, $R^2 = 1$.
- When the fitted slope $\hat{\beta}_1 = 0$, we have

$$\hat{\beta}_0 = \bar{Y} \text{ and } \hat{Y}_i = \bar{Y}. \text{ So } R^2 = 0.$$

- The closer R^2 is to 1, the greater the degree of **linear association** between X and Y .
- **But** a high R^2 does not indicate that useful prediction can be made (CI can be wide). A small R^2 does not mean that X and Y are uncorrelated.



$$SS_{tot} = SS_{reg} + SS_{err}$$

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation

Important result on the Decomposition of Variation: $SS_{tot} = SS_{reg} + SS_{err}$.
[\[Board Derivation\]](#)

- $SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$ represents the total variation of Y .
- $SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ represents the variation that can be explained by the regression model. [\bar{y} is also the mean of $\hat{y}_1, \dots, \hat{y}_n$.]
[\[Board Derivation\]](#)
- $SS_{err} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ represents the variation in the residuals that cannot be explained by the model.



Warning: when there is no intercept

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

If the model does not contain an intercept by assumption, i.e., $Y_i = \beta_1 x_i + \epsilon_i$, the R^2 is

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^n Y_i^2}.$$

If under the null hypothesis assumption $\beta_0 = 0$, $\bar{Y} = E(Y) = 0$. So $SS_{tot} = \sum_{i=1}^n (Y_i - 0)^2 = \sum_{i=1}^n Y_i^2$.



Pearson's Correlation Coefficient

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- Correlation between any two random variables:

$$\rho = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}Y}}.$$

According to Chauchy-Schwartz inequality, $|\text{cov}(X, Y)|^2 \leq \text{var}(X) \text{ var}(Y)$ and thus $|\text{cor}(X, Y)| \leq 1$.

- Sample correlation between two variables with data:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}.$$

- Note that for simple linear regression R^2 is the same as r^2 , since $R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$.



Properties of Pearson's Correlation

DATA

564/494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- $-1 \leq r \leq 1$
- If Y_i increases (decreases) linearly with X_i , then $r > 0$ ($r < 0$).
- The closer the points (x_i, y_i) come to forming a straight line, the closer r is to ± 1 .
- The magnitude of r is unchanged if either the X or Y sample is transformed linearly. [Why?](#)
- The correlation does not depend on which variable is called Y and which is called X .



Properties of r

DATA

564/494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- If r is near ± 1 , then there is a strong linear relationship between Y and X .
Might predict Y from X via linear regression.
- If r is near 0, there is a weak linear relationship between Y and X .
- Note that $r = 0$ does not imply that Y and X are not related. For example,
 $Y_i = x_i^2$ where $X_i = -2, -1.9, \dots, 0, \dots, 2$.



Example: Data

DATA
564-494
Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- X = Percentage of children immunized against DPT;
- Y = under-five mortality rate per 1000 live births, in 1992

Nation	X	Y	Nation	X	Y	Nation	X	Y
Bolivia	77	118	Ethiopia	13	208	Mexico	91	33
Brazil	69	65	Finland	95	7	Poland	98	16
Cambodia	32	184	France	95	9	Russian	73	32
Canada	85	8	Greece	54	9	Senegal	47	145
China	94	43	India	89	124	Turkey	76	87
Czech Republic	99	12	Italy	95	10	UK	90	9
Egypt	89	55	Japan	87	6			



Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- X = Percentage of children immunized against DPT;
- Y = under-five mortality rate per 1000 live births, in 1992

Question

- Are Y and X related (associated), and how?

```
x <- c(77, 69, 32, 85, 94, 99, 89, 13, 95, 95,  
54, 89, 95, 87, 91, 98, 73, 47, 76, 90);  
y <- c(118, 65, 184, 8, 43, 12, 55, 208, 7, 9,  
9, 124, 10, 6, 33, 16, 32, 145, 87, 9);  
cor(x,y) # same as cor(x,y, method = "pearson")  
[1] -0.7910654
```

Want to test $H_0 : \rho = 0$ based on magnitude for r .



Test: $H_0 : \rho = 0$

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- One estimator for the correlation $\rho = \text{cor}(X, Y)$ between random variable X and Y can be the Pearson's correlation coefficient.
- How to use r to test the hypothesis that there is no correlation between X and Y ($H_0 : \rho = 0$)?
- We need to make some assumption: the (X, Y) are from a bivariate normal distribution.



The relationship of r in linear regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- In the simple linear regression: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r \sqrt{\frac{S_{yy}}{S_{xx}}}, \text{ since } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

so $\hat{\beta}_1 = 0$ if and only $r = 0$.

- Thus, testing $H_0 : \rho = 0$ is equivalent to testing $H_0 : \beta_1 = 0$.
(when the (X, Y) are from a bivariate normal distribution and (x_i, ϵ_i) are independent, then $\rho = \beta_1 \sigma_x / \sigma_y$ or $\beta_1 = \rho \sigma_y / \sigma_x$.)



Test Statistic

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- When testing $H_0 : \beta_1 = 0$, the test statistic is

$$T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}}$$

- Since $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}$, $\hat{\sigma}^2 = \frac{SS_{err}}{n-2} = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$, we have

$$T_{obs} = r \sqrt{\frac{S_{yy}}{S_{xx}} \frac{n-2}{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}} = r \sqrt{\frac{n-2}{1-r^2}}$$

- Under $H_0 : \rho = 0$, T_{obs} has a t-distribution with $df = n - 2$.



Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

X = Percentage of children immunized against DPT;
 Y = under-five mortality rate per 1000 live births, in 1992

- Observe $r = -0.7910654$ and $n = 20$.
- To decide whether Y and X are associated, we can formulate it as the problem of testing the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis $H_1 : \rho \neq 0$.
- Decision Rule: The test statistic is $T_{obs} = r \sqrt{\frac{n-2}{1-r^2}}$ we reject H_0 at the α level if and only if $|T_{obs}| \geq t_{\alpha/2, n-2}$.



Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

X = Percentage of children immunized against DPT;
 Y = under-five mortality rate per 1000 live births, in 1992

- Now we observe $r = -0.7910654$ and $n = 20$, so we have

$$T_{obs} = r \sqrt{\frac{n-2}{1-r^2}} = -5.49$$

Since $t_{0.025,18} = 2.101$, $|T_{obs}| > t_{\alpha/2,n-2}$, and thus we reject $H_0 : \rho = 0$ at the 5% level.

- Now suppose one claims that $\rho = -0.7$. Does the observed value differ significantly from this value?



Key Fact

DATA

564/494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- R. A. Fisher showed that

$$Z_r = \frac{1}{2} \log \frac{1+r}{1-r}$$

has approximate normal distribution

$$Z_r \sim N \left(\frac{1}{2} \log \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

when $n = \#$ of observations (here and below $\log = \log_e$, not \log_{10}).

- The results is not sensitive to the Bivariate normal assumption, and is useful quite broadly.



Decision Rule

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

$$Z_r \sim N \left(\frac{1}{2} \log \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

- In the problem of testing $H_0 : \rho = \rho_0$ vs. $H_1 : \rho \neq \rho_0$. The test statistic is

$$Z_{obs} = \frac{\frac{1}{2} \log \frac{1+r}{1-r} - \frac{1}{2} \log \frac{1+\rho_0}{1-\rho_0}}{\sqrt{\frac{1}{n-3}}}$$

and we reject H_0 at the α level if $|Z_{obs}| \geq Z_{\alpha/2}$.

[Some useful critical value for standard normal: $z_{0.01} = 2.326$, $z_{0.025} = 1.960$,
 $z_{0.05} = 1.645$, $z_{0.10} = 1.282$.]



Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- Observe $r = -0.7910654$ and $n = 20$. Want to test $H_0 : \rho = -0.7$ vs $H_1 : \rho \neq -0.7$ at 10% level.
- The observed $Z_r = \frac{1}{2} \log \frac{1+r}{1-r} = -1.0743$
- Under $H_0 : \rho = -0.7$,

$$Z_r \sim N\left(\frac{1}{2} \log \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right) = N(-0.8673, 0.05882)$$

- The corresponding (normalized) test statistic is

$$Z_{obs} = \frac{Z_r - \mu_0}{s.e.(Z_r)} = \frac{(-1.0743) - (-0.8673)}{\sqrt{0.05882}} = -0.8535$$

- This value does not exceed $Z_{0.05} = 1.645$, and thus there is no evidence to reject $H_0 : \rho = -0.7$ at the 10% level.



Confidence Interval for ρ

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

$$Z_r \sim N\left(z_\rho, \frac{1}{n-3}\right), \quad z_\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$$

How to find confidence interval for ρ

- First, $\frac{Z_r - z_\rho}{\sqrt{1/(n-3)}} \sim N(0, 1)$. Thus for the observed Z_r , $100(1 - \alpha)\%$ CI for z_ρ is $Z_r \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}} = [Z_L, Z_U]$.
- Second, if $Z = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$ then $\rho = \frac{e^{2Z}-1}{e^{2Z}+1}$. So transform back to find $100(1 - \alpha)\%$ CI for ρ

$$\left[\frac{e^{2Z_L} - 1}{e^{2Z_L} + 1}, \frac{e^{2Z_U} - 1}{e^{2Z_U} + 1} \right]$$



Example: confidence interval for ρ

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- The observed $Z_r = \frac{1}{2} \log \frac{1+r}{1-r} = -1.0743$.
- The 90% CI for $Z_\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$ is given by

$$Z_r \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}}$$

$$\left(-1.0743 - \frac{1.645}{\sqrt{20-3}}, -1.0743 + \frac{1.645}{\sqrt{20-3}} \right) = (-1.4733, -0.6753)$$

- Thus the 90% CI for ρ is

$$\left(\frac{e^{2*(-1.4733)} - 1}{e^{2*(-1.4733)} + 1}, \frac{e^{2*(-0.6753)} - 1}{e^{2*(-0.6753)} + 1} \right) = (-0.9002, -0.5885).$$



Pearson Correlation r

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation

- The Pearson correlation r can be highly influenced by outliers in one or both samples.
- If we delete the one extreme case with the largest X and smallest Y , then r can change from -1 to $0!!!$
- To avoid the conclusion depending heavily on a single observation, use nonparametric approach.



Spearman's rank correlation r_s

DATA

564/494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- 1 Order the X_i 's and assign them ranks.
- 2 Do the same for the Y_i 's, and replace the original data pairs by the pairs of ranks values. (Ties are treated by the mid-ranks.)
- 3 The Spearman rank correlation is the Pearson correlation computed from the pairs of ranks.

[Then we can use $T_{obs} = r_s \sqrt{\frac{n-2}{1-r_s^2}}$ for testing]



Spearman's rank correlation r_s

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

	X	Rank(X)	Y	Rank(Y)	$d = \text{Rank}(X) - \text{Rank}(Y)$
1	X_1	R_{X_1}	Y_1	R_{Y_1}	$d_1 = R_{X_1} - R_{Y_1}$
2	X_2	R_{X_2}	Y_2	R_{Y_2}	$d_2 = R_{X_2} - R_{Y_2}$
...
n	X_n	R_{X_n}	Y_n	R_{Y_n}	$d_n = R_{X_n} - R_{Y_n}$

- If there are no ties

$$r_s = r_{R_X} r_{R_Y} = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$



Example: Data (rank)

DATA
564-494
Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- X = Percentage of children immunized against DPT;

- Y = under-five mortality rate per 1000 live births, in 1992

Nation	X	Y	Nation	X	Y	Nation	X	Y
Bolivia	77 (8)	118 (16)	Ethiopia	13 (1)	208 (20)	Mexico	91 (14)	33 (11)
Brazil	69 (5)	65 (14)	Finland	95 (17)	7 (2)	Poland	98 (19)	16 (9)
Cambodia	32 (2)	184 (19)	France	95 (17)	9 (5)	Russian	73 (6)	32 (10)
Canada	85 (9)	8 (3)	Greece	54 (4)	9 (5)	Senegal	47 (3)	145 (18)
China	94 (15)	43 (12)	India	89 (11.5)	124 (17)	Turkey	76 (7)	87 (15)
Czech Republic	99 (20)	12 (8)	Italy	95 (17)	10 (7)	UK	90 (13)	9 (5)
Egypt	89 (11.5)	55 (13)	Japan	87(10)	6 (1)			



Spearman's rank correlation r_s

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

```
x <- c(77, 69, 32, 85, 94, 99, 89, 13, 95, 95, 54,  
89, 95, 87, 91, 98, 73, 47, 76, 90);  
y <- c(118, 65, 184, 8, 43, 12, 55, 208, 7, 9, 9,  
124, 10, 6, 33, 16, 32, 145, 87, 9);
```

```
cor(x,y); # Pearson's correlation r  
[1] -0.7910654
```

```
cor(x,y, method = "spearman")  
[1] -0.5431913
```

```
# Alternative method to compute Spearman's rank correlation  
a <- rank(x); b <- rank(y); cor(a,b) [1] -0.5431913
```



Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

X = Percentage of children immunized against DPT; Y = under-five mortality rate per 1000 live births, in 1992

- Spearman's rank correlation $r_s = -0.5431913$ and $n = 20$.
- In the problem of testing $H_0 : \rho = 0$ (no association) vs $H_1 : \rho \neq 0$, we can use the test statistics

$$T_{obs} = r_s \sqrt{\frac{n-2}{1-r_s^2}} = -5.4864$$

and a significant level α test to reject H_0 if and only if $|T_{obs}| \geq t_{\alpha/2, n-2}$. Since $t_{18, 0.025} = 3.20$, we reject $H_0 : \rho = 0$ at the 5% level.



Spearman's rank correlation r_s

DATA

564/494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- It is not sensitive to outliers.
- In samples without unusual observations and a linear trend, we often have $r_s = r$.
- The magnitude of the Spearman correlation does not change if either X or Y or both are monotonically transformed.
- If r_s is noticeably greater than r , then a transformation of the data might provide a stronger linear relationship.



3. Variable Transformation

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- If the model fit is inadequate, it does not mean that a regression is not useful, it just means that the linear regression you proposed is not useful.
- One problem might be that the relationship between X and Y is not exactly linear.
- To model the nonlinear relationship, we can transform X and Y (or both) by some nonlinear function, e.g., $f(t) = t^a$ or $\log(t)$.
- Example: assume (X, Y) are related through $y = \gamma e^{\theta x}$. We can transformation Y to $y^* = \log(y)$ and then fit the new model as

$$y^* = \log y = \log \gamma + \theta x = \beta_0 + \beta_1 x.$$



Box-Cox Transformation

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- Does the response Y need transformation?
- A useful tool is from the “MASS” library of R to perform Box-Cox Transformation.
- It considers the power transformation of the forms Y^λ or $\log Y$ ($\lambda = 0$) and find the maximum likelihood estimate when fitting data to model $(Y_i)^\lambda = \beta_0 + \beta_1 x_i + \epsilon_i$ when ϵ_i are iid $N(0, \sigma^2)$.
- Equivalently, the transformation is

$$Y' = \frac{Y^\lambda - 1}{\lambda} = \begin{cases} (Y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0 \\ \log Y, & \text{if } \lambda = 0 \end{cases}$$

- When $\lambda \approx 1$, no need to transform Y .



R Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

Ex. 1.19. The director of admissions of a small college selected $n = 10$ students at random from the new freshman class in a preliminary study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (x). The results of the study follows.

```
GPAdata <- read.table('GPAdat.txt')
lm1 <- lm(V1~V2, data=GPAdat)
library(MASS)
boxcox(lm1)
### Or alternatively
boxcox(V1~V2,data=GPAdat)
```



boxcox(lm1)

DATA

564 rows

Simple Linear

Regression

Lulu Kang

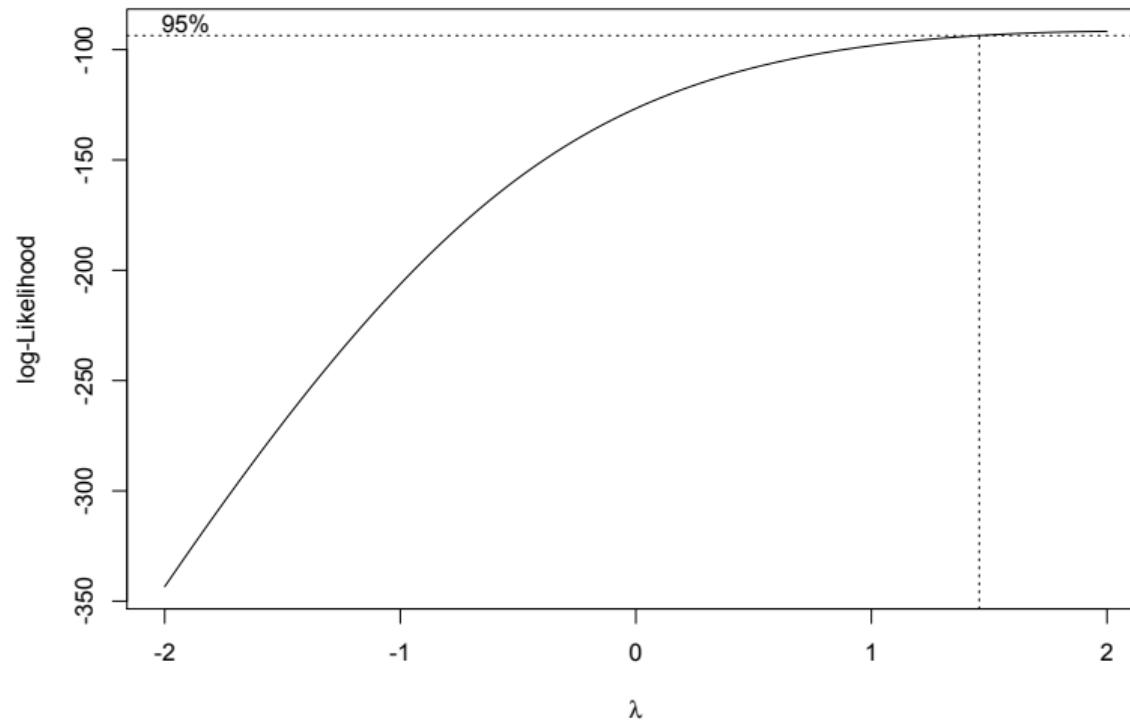
R-square

Correlation analysis

Pearson's correlation

Spearman's correlation

Variable transformation





boxcox(lm1, lambda=seq(-10,10))

DATA
564 rows
Simple Linear
Regression

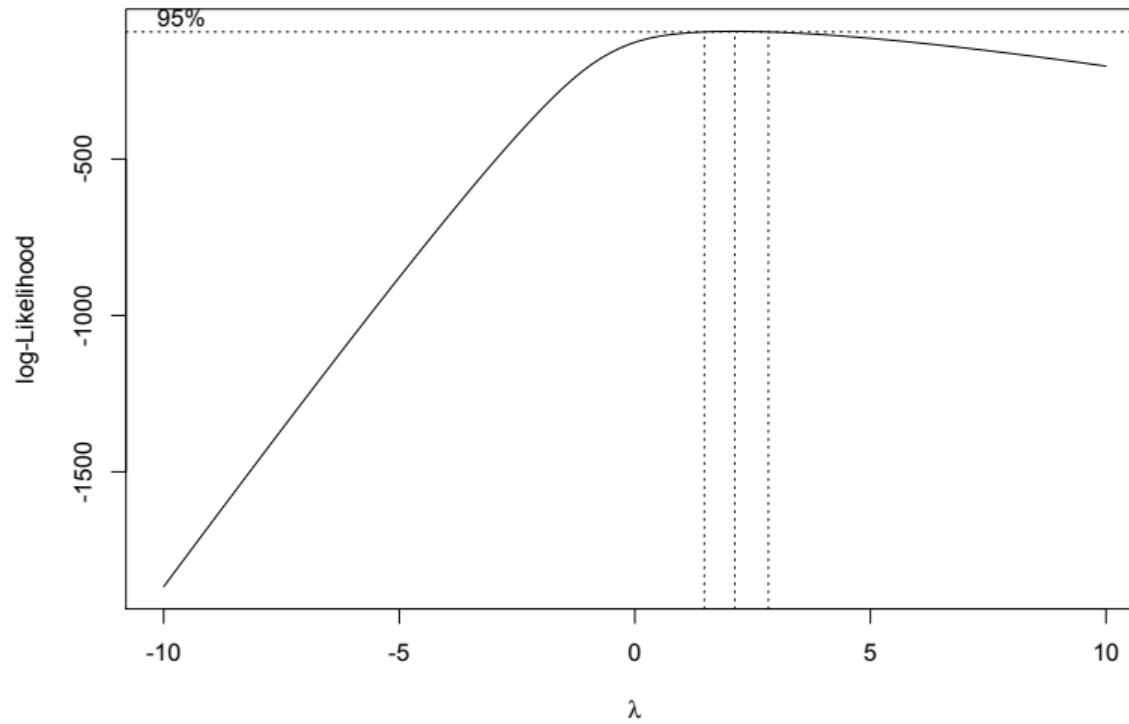
Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation





boxcox(lm1, lambda=seq(0,5,0.1))

DATA
564 494
Simple Linear
Regression

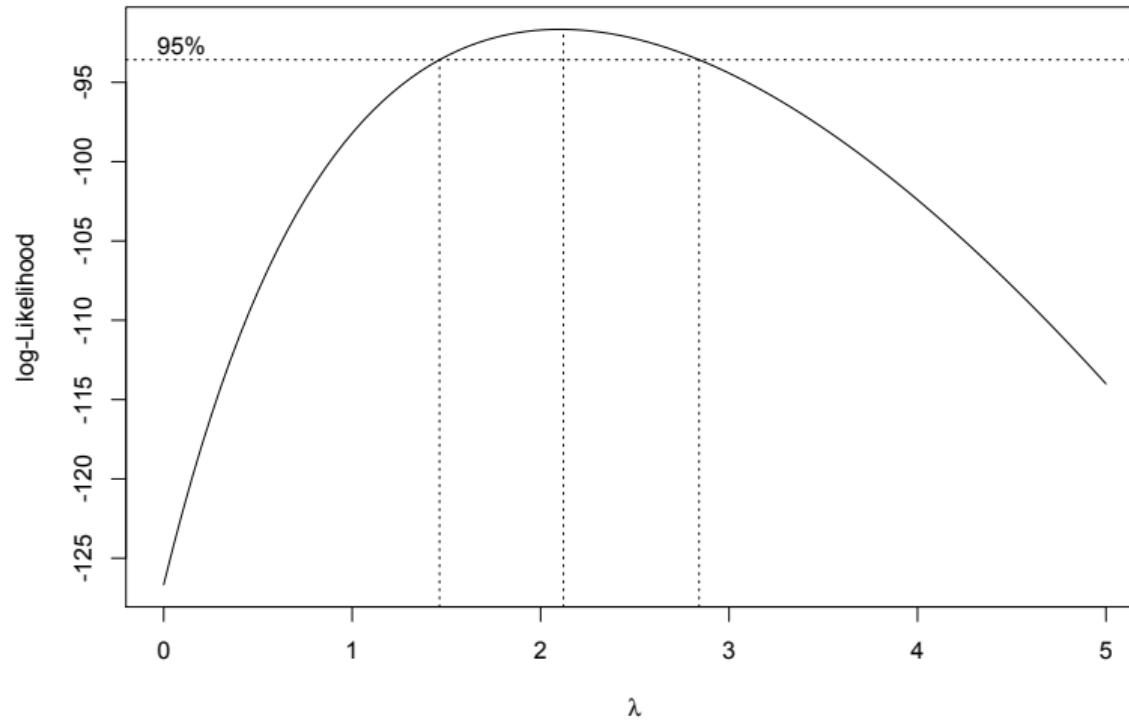
Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation





boxcox(lm1, lambda=seq(1,3,0.1))

DATA
564 494
Simple Linear
Regression

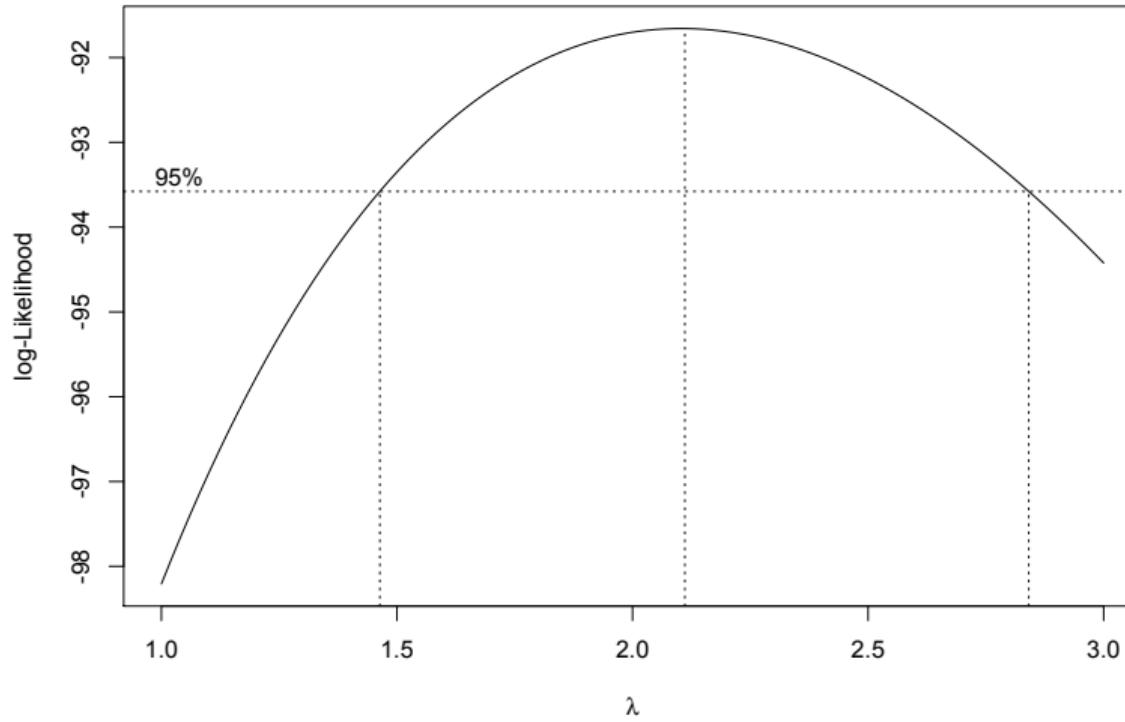
Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation





boxcox(lm1, lambda=seq(1,3,0.1))

DATA

564 494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation

- The confidence interval for λ is about [1.5, 2.9].
- We see perhaps $Y' = Y^2$ might be best here.
- $Y^{1.5}$ or $Y^{2.5}$ are also possible.
- Transformation can be useful here.



Try the following R code

DATA
564 494
Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

```
GPAdat<- read.table('GPAdat.txt')  
GPAdat$V3 <- GPAdat$V1^2  
  
plot(GPAdat$V2, GPAdat$V1, xlab="ACT", ylab="GPA")  
plot(GPAdat$V2, GPAdat$V3, xlab="ACT", ylab="GPA Square")  
  
model1 <- lm(V1 ~ V2, data= GPAdat)  
model2 <- lm(V3 ~ V2, data= GPAdat)  
summary(model1);  
summary(model2);
```



Scatter plot

Data

564 rows

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

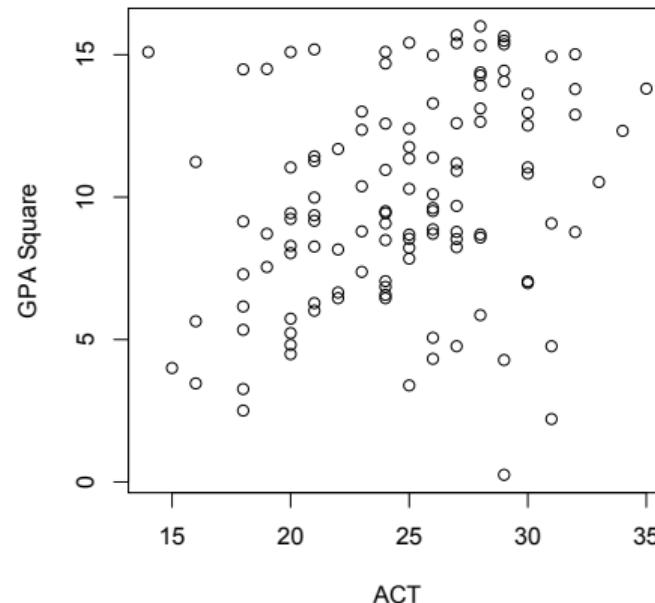
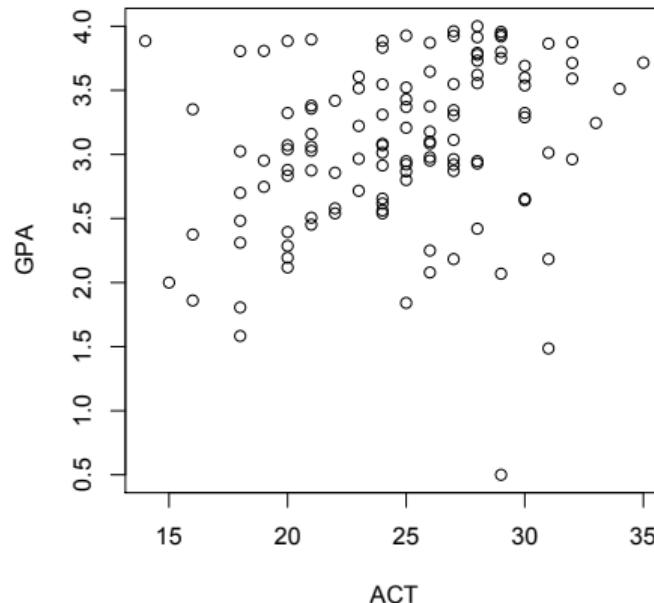


Figure: Left: x vs. original Y . Right: x vs. Y^2



Plot the data and fitted lines

DATA
564 rows
Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation

Add fitted regression line to the scatter plots.

```
##If you want to plot two plots together (2 columns)
par(mfcol=c(1,2))
plot(GPAdatadata$V2, GPAdatadata$V1, xlab="ACT", ylab="GPA")
abline(model1);
plot(GPAdatadata$V2, GPAdatadata$V3, xlab="ACT", ylab="GPA Square")
abline(model2)
```



Scatter plot and fitted lines

DATA

564 rows

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

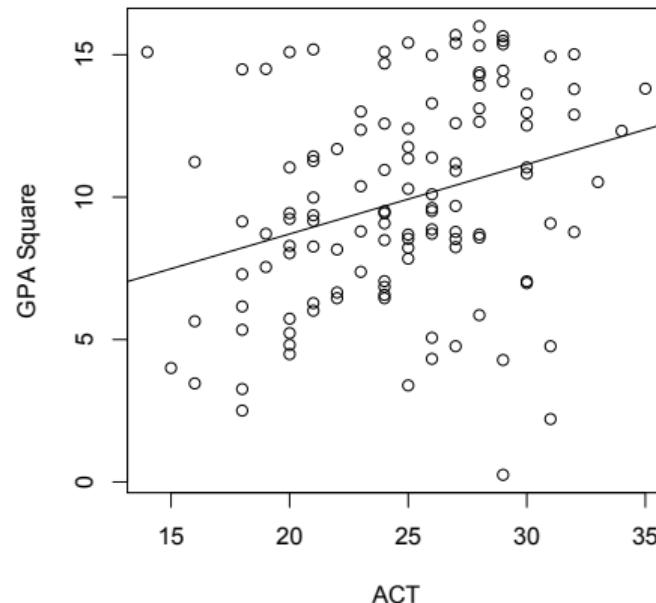
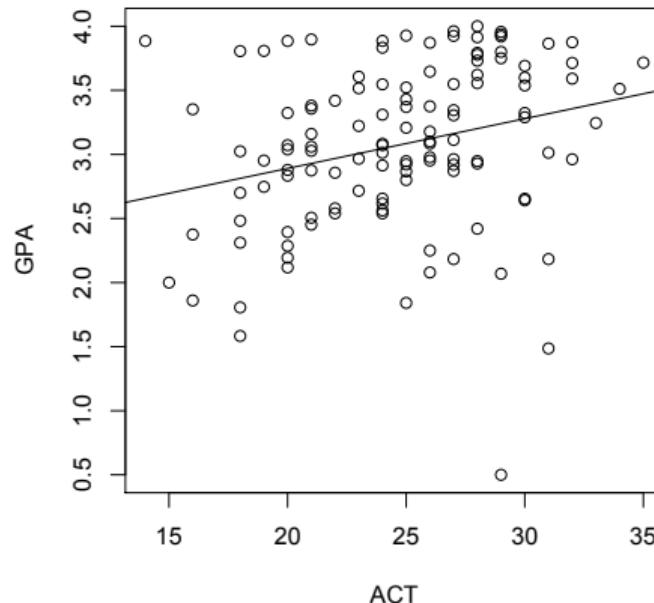


Figure: Left: x vs. original Y . Right: x vs. Y^2



Residual vs Fitted

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

- Plot the residuals against fitted values.

```
par(mfcol=c(1,2))
plot(fitted(model1), resid(model1))
abline(0,0)
plot(fitted(model2), resid(model2))
abline(0,0)
```



Scatter plot of the residuals

DATA
564/494
Simple Linear
Regression

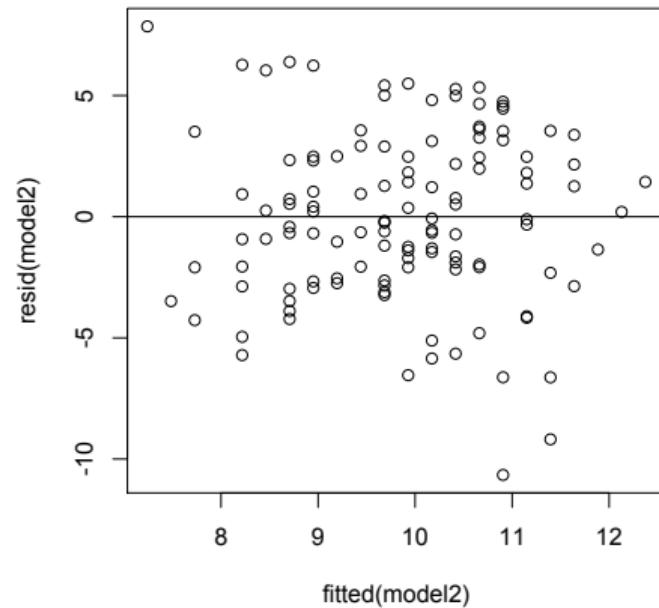
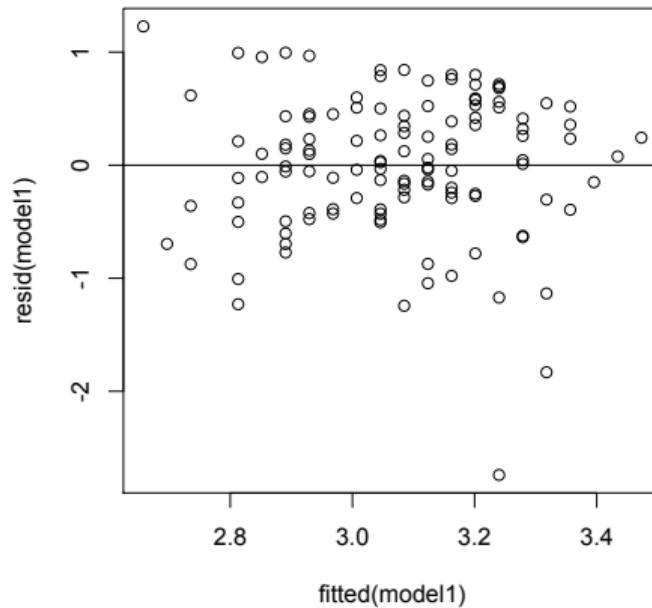
Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation





qqnorm for residuals

DATA

564-494

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation

Spearman's
correlation

Variable
transformation

Try the following R codes

```
par(mfcol=c(1,2))
qqnorm(residuals(model1))
qqline(residuals(model1))
qqnorm(residuals(model2))
qqline(residuals(model2))
```



QQ plot

DATA
564-494
Simple Linear
Regression

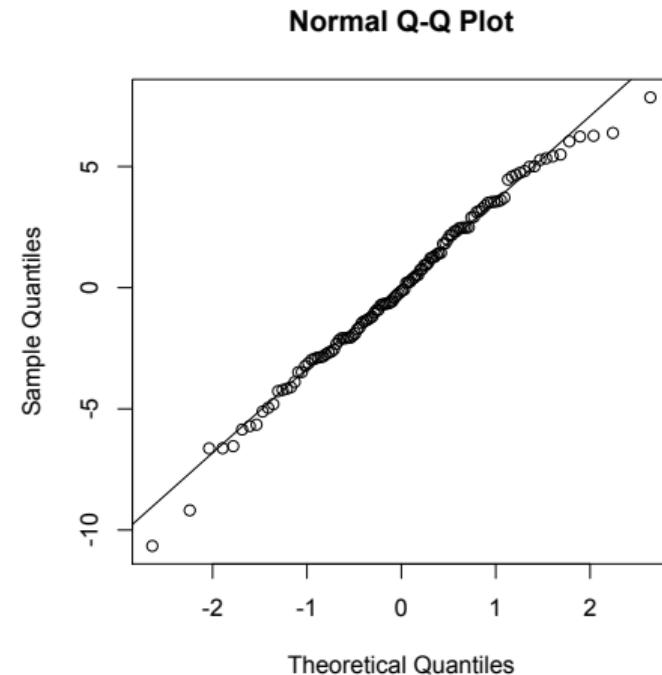
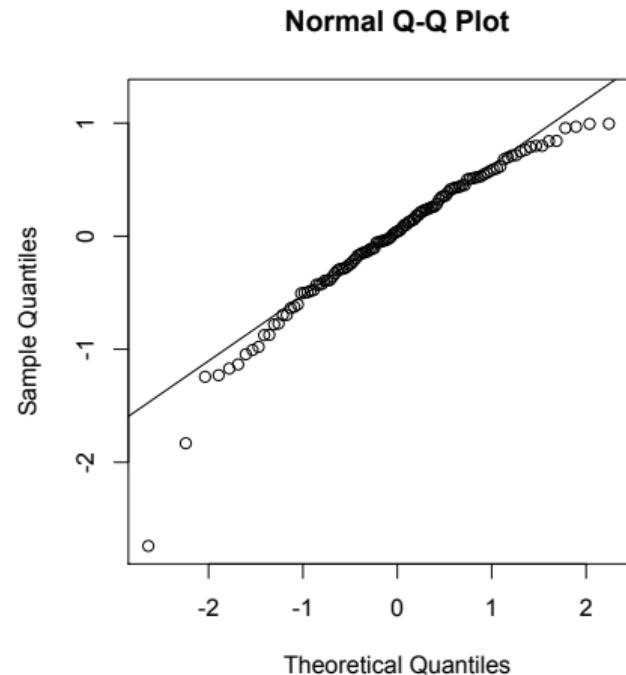
Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation





Transformation predictors X ?

GPAdata

564 rows

Simple Linear
Regression

Lulu Kang

R-square

Correlation
analysis

Pearson's correlation
Spearman's
correlation

Variable
transformation

```
> update(lm1, .~ . +I(V2^2))
Call:
lm(formula = V1 ~ V2 + I(V2^2), data = GPAdata) Coefficients:
(Intercept)      V2          I(V2^2)
1.516750       0.089425     -0.001036
```

Multiple linear regression: examine the coefficient of X' in regression.



DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Part III

Multiple Linear Regression

- 2. Multiple Linear Regression
- 2. Least Square Estimation
- 2. ANOVA, R^2 , and adjusted- R^2
- 2. Inference of $\hat{\beta}$
- 2. Estimation and Prediction
- 2. Diagnostics and Remedies



Multiple Regression Models

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Why we need to do multiple regression models? Most systems/processes are more complicated and are affected by more than just one input variable, or affected by one variable nonlinearly.
- First-Order model with $p - 1$ predictor variables.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad (1)$$

for the data $\{Y_i, X_{i1}, \dots, X_{i,p-1}\}_{i=1}^n$.

- When $p = 2$, we have the simple linear regression.
- When $p = 3$, $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is a *plane* in the 3-D space of (Y, X_1, X_2) .
- When $p > 3$, $E(Y|X)$ is a *hyperplane* in the p -dimension space of (Y, X_1, \dots, X_{p-1}) .



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- $Y = \text{sales}$; $X_1 = \text{number of persons aged 16 or younger}$; $X_2 = \text{per capita disposable personal income}$

```
dwaine<- read.table('CH06FI05.txt')
colnames(dwaine)<-c('X1','X2','Y')
library('scatterplot3d')
par(mfrow=c(2,2))
plot(dwaine$X1,dwaine$Y,xlab='X1',ylab='Y')
plot(dwaine$X2,dwaine$Y,xlab='X2',ylab='Y')
plot(dwaine$X1,dwaine$X2,xlab='X1',ylab='X2')
scatterplot3d(dwaine, angle = 55)
```



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

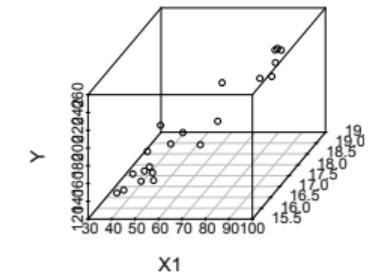
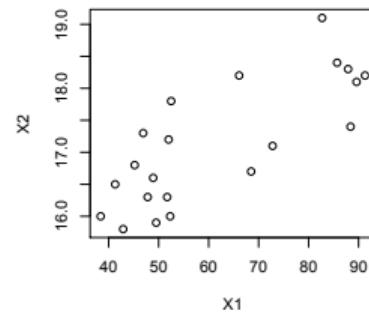
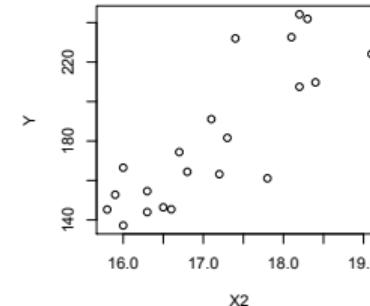
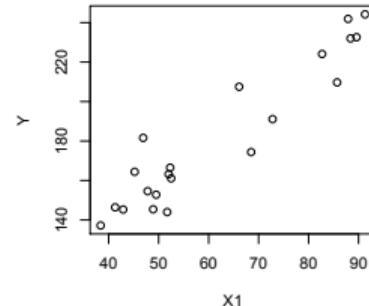
Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies





Example: Dwaine Studio

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Two the ways to plot interactive 3d plot.

```
library(rgl)
plot3d(dwaine$X1,dwaine$X2,dwaine$Y,xlab='Targtpop',
       ylab='Dispoinc',zlab='Sales',size=0.75,type='s',lit=FALSE)

interleave <- function(v1, v2) as.vector(rbind(v1,v2))

segments3d(interleave(dwaine$X1, dwaine$X1),
           interleave(dwaine$X2,dwaine$X2),interleave(dwaine$Y, min(dwaine$Y)),
           alpha=0.4, col="blue")

library(car)
scatter3d(Y~X1+X2,data=dwaine)
```



General linear regression model

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Response Y .
- Input variables x_1, \dots, x_{p-1} , and $\mathbf{x} = (x_1, \dots, x_{p-1})'$ is a column vector of the input variables.
- Model:

$$Y = \beta_0 + \beta_1 f_1(\mathbf{x}) + \beta_2 f_2(\mathbf{x}) + \dots + \beta_k f_k(\mathbf{x}) + \epsilon. \quad (2)$$

- $\epsilon \sim N(0, \sigma^2)$ and if we have data $\{Y_i, \mathbf{x}_i\}$, then ϵ_i 's are iid and also independent of Y and \mathbf{x} .
- $\beta_0, \beta_1, \dots, \beta_k$ are parameters of the regression, or *coefficient parameters*.
- f_1, f_2, \dots, f_k are known functions of the input variables \mathbf{x} .



Polynomial Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- f_i 's are polynomial terms of the input variables, e.g., $f_1(\mathbf{x}) = x_1$, $f_2(\mathbf{x}) = x_1^2$, etc.
- Example: 2nd order polynomial regression of one predictor variable

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon.$$

- Example: 2nd order polynomial regression of two predictor variables

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

Here x_i is the linear terms, x_i^2 is the quadratic term, and $x_1 x_2$ is called *interaction* term between x_1 and x_2 .



Qualitative predictor variables

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

When the predictor variable is k -level qualitative (categorical) variable, i.e., the variable is not a continuous variable but rather has limited number of settings (levels) to be.

- two-level: Y —hospital stay, X_1 —patient age, X_2 —patient gender; we can use dummy variable

$$X_2 = \begin{cases} 1 & \text{if patient female} \\ 0 & \text{if patient male.} \end{cases}$$

- Model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
- $E(Y|X_2 = 1) = \beta_0 + \beta_1 X_1 + \beta_2$ vs $E(Y|X_2 = 0) = \beta_0 + \beta_1 X_1$. So $\beta_2 = E(Y|X_2 = 1, X_1) - E(Y|X_2 = 0, X_1)$ is the mean difference between the hospital stay of the female and male patients of the same age.



Qualitative predictor variables

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- three-level: Y —hospital stay, X_1 —patient age, X_2 —patient gender; we can use dummy variable; and the third variable about seriousness of their illness that has three levels { severe, medium, minor }.

$$X_3 = \begin{cases} 1 & \text{if illness is severe} \\ 0 & \text{otherwise.} \end{cases} \quad X_4 = \begin{cases} 1 & \text{if illness is medium} \\ 0 & \text{otherwise.} \end{cases}$$

- Model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$.
- minor illness (X_3, X_4) = (0, 0), medium (X_3, X_4) = (0, 1), severe (X_3, X_4) = (1, 0). Minor level is the baseline.
- $E(Y|\text{minor}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ vs $E(Y|\text{medium}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4$ vs $E(Y|\text{severe}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3$. So $\beta_3 = E(Y|\text{severe}) - E(Y|\text{minor})$ is the mean difference between the hospital stay of patients with severe illness and the ones with minor illness of the same gender and age, and $\beta_4 = E(Y|\text{medium}) - E(Y|\text{minor})$ is the mean difference between the hospital stay of patients with medium illness and the ones with minor illness of the same gender and age.



Matrix Form

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Use matrix notation:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1, & x_{11}, & \dots, & x_{1,p-1} \\ 1, & x_{21}, & \dots, & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1, & x_{n1}, & \dots, & x_{n,p-1} \end{bmatrix}_{n \times p}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The linear regression model for the data $\{x_{i1}, \dots, x_{i,p-1}, y\}_{i=1}^n$ can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{and } E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \text{ and } \text{var}(\mathbf{y}) = \sigma^2 \mathbf{I}_n.$$



Least Square Estimation

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

The least square estimation can be derived as follows.

$$Q = RSS = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \quad \Rightarrow \quad (\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}.$$

If $\mathbf{X}'\mathbf{X}$ is invertible, then the least square estimation $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
(Discussion: under what condition $\mathbf{X}'\mathbf{X}$ is nonsingular?)



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Using R as a calculator.

```
X <- cbind(rep(1,nrow(dwaine)),dwaine$X1,dwaine$X2)
X2 <- t(X) %*% X
XY <- t(X) %*% dwaine$Y
X2.inv <- solve(X2)
beta_hat <- X2.inv %*% XY
```



Example: Dwaine Studio

Use the lm function.

```
> fit <- lm(Y~X1+X2,data=dwaine)
> summary(fit)
Call:
lm(formula = Y ~ X1 + X2, data = dwaine)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.4239	-6.2161	0.7449	9.4356	20.2151

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-68.8571	60.0170	-1.147	0.2663
X1	1.4546	0.2118	6.868	2e-06 ***
X2	9.3655	4.0640	2.305	0.0333 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.01 on 18 degrees of freedom

Multiple R-squared: 0.9167, Adjusted R-squared: 0.9075

F-statistic: 99.1 on 2 and 18 DF, p-value: 1.921e-10

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies



Fitted Values and Residuals

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Both point estimation and prediction are \hat{y}

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_{p-1} x_{p-1}.$$

or $\hat{y} = \mathbf{x}' \hat{\beta}$, where $\mathbf{x} = (1, x_1, \dots, x_{p-1})'$.

- Fitted values for the training data set $\hat{\mathbf{y}} = \mathbf{X} \hat{\beta} = \mathbf{H} \mathbf{y}$.
- Residual $e_i = y_i - \hat{y}_i$ and $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H}) \mathbf{y}$.



Example:Dwaine Studio

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

If using R as a calculator

```
X <- cbind(rep(1,nrow(dwaine)),dwaine$X1,dwaine$X2)
X2 <- t(X) %*% X; XY <- t(X) %*% dwaine$Y; X2.inv <- solve(X2)
beta_hat <- X2.inv %*% XY; Y_hat <- X %*% beta_hat
error_hat <- dwaine$Y-Y_hat
```

If using the lm function

```
fit <- lm(Y~X1+X2,data=dwaine); fit$fitted.values; fit$residuals
par(mfrow=c(1,2))
plot(fit$fitted.values,fit$residuals,type='p',xlab='fitted values',ylab='residuals')
plot(fit$fitted.values,dwaine$Y,type='p',xlab='fitted values',ylab='Y')
```



Example:Dwaine Studio

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

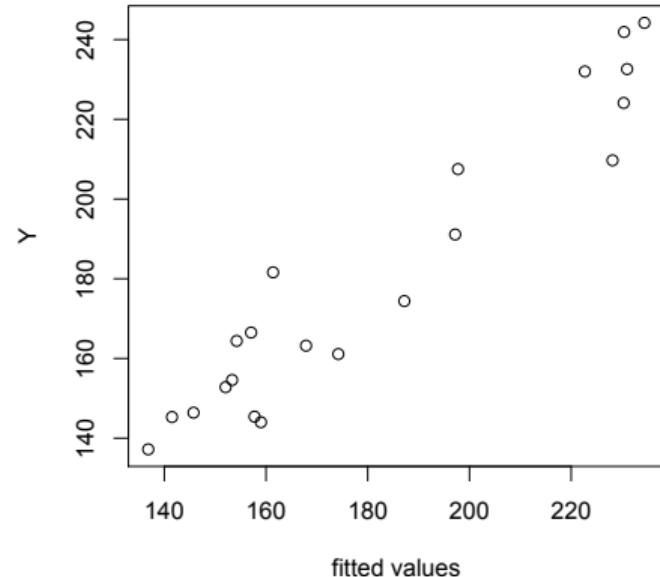
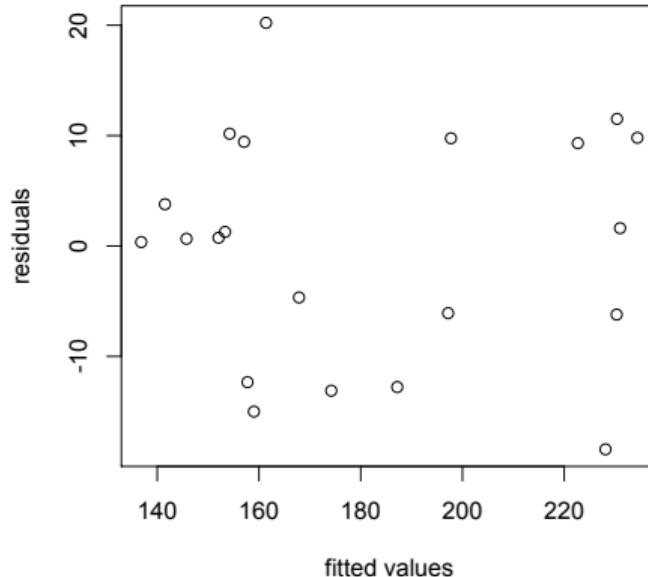
Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies





Hat matrix

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$.
- \mathbf{H} is a projection matrix: $\mathbf{H}' = \mathbf{H}$ (symmetric) and $\mathbf{H}^2 = \mathbf{H}$ (idempotent).
- For any projection matrix, its eigenvalues are either 1 or 0. [\[Derivation\]](#)
- For the projection matrix \mathbf{H} , $\text{rank}(\mathbf{H}) = \text{trace}(\mathbf{H})$. For least square estimation, $\text{trace}(\mathbf{H}) = \text{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \text{trace}((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})) = \text{trace}(\mathbf{I}_p) = p$.
- $\mathbf{I}_n - \mathbf{H}$ is also a projection matrix. $(\mathbf{I}_n - \mathbf{H})' = (\mathbf{I}_n - \mathbf{H})$, $(\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$ and $\text{rank}(\mathbf{I}_n - \mathbf{H}) = \text{trace}(\mathbf{I}_n - \mathbf{H}) = n - p$. $\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \mathbf{0}$. [\[Derivation\]](#)



Property of the residual

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- $e = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}.$
- Projection property: $e \perp \text{span}\{\mathbf{X}'\text{s columns including } \mathbf{1}\}$

$$\mathbf{e}'\mathbf{X} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{y}'(\mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = \mathbf{0}$$

So $e \perp$ every column of \mathbf{X} or any linear combinations of them.

- $e \perp \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$



Linear regression fitting=Projection of y in the space of X

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

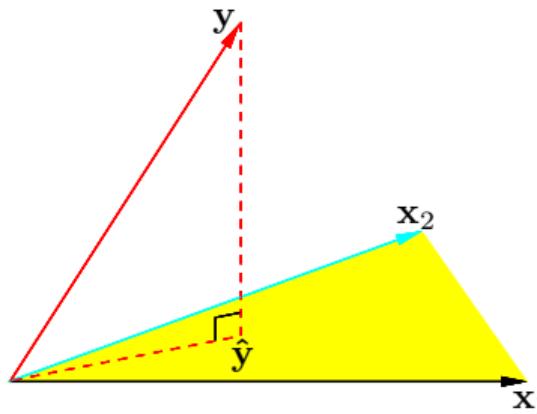


FIGURE 3.2. The N -dimensional geometry of least squares regression with two predictors. The outcome vector \mathbf{y} is orthogonally projected onto the hyperplane spanned by the input vectors \mathbf{x}_1 and \mathbf{x}_2 . The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions



ANOVA

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- The total variation

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2 = (\mathbf{y} - \frac{1}{n}\mathbf{1}'\mathbf{y})'(\mathbf{y} - \frac{1}{n}\mathbf{1}'\mathbf{y}) = \mathbf{y}'(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{y}.$$

- The sum of squares of error $SS_{err} = SSE = RSS = \sum_{i=1}^n e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}.$

- The sum of squares of regression $SS_{reg} = SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (\mathbf{H}\mathbf{y} - \frac{1}{n}\mathbf{1}'\mathbf{y})'(\mathbf{H}\mathbf{y} - \frac{1}{n}\mathbf{1}'\mathbf{y}) = \mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{y}.$

- Decomposition of variation still holds: $SS_{tot} = SSR + SSE.$



Distribution of MSE

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Sum of squares of errors $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, and SSE/σ^2 follows χ_{n-p}^2 distribution, where p is the regression parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$. [\[Derivation\]](#)
- Sum of squares of regression $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{y}$ and SSR/σ^2 follows χ_{p-1}^2 distribution. [Why?](#)
- SSR and SSE are statistically independent. [Why?](#)
- $\hat{\sigma}^2 = MSE = \frac{SSE}{n-p}$ and standard error $s = \sqrt{MSE}$.



ANOVA

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Source of Variation	SS	df	MS
Regression	$SSR = \mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{y}$	$p - 1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = \mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}$	$n - p$	$MSE = \frac{SSE}{n-p} = \hat{\sigma}^2$
Total	$SS_{tot} = \mathbf{y}'(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{y}$	$n - 1$	

- F-ratio = $\frac{MSR}{MSE} = \frac{SSR/(p-1)}{\hat{\sigma}^2}$.
- F-ratio $\sim F_{p-1, n-p}$. [Recall the distribution of SSR and SSE.]
- Hypothesis $H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ vs $H_1 : \text{at least one of } \beta_1, \dots, \beta_{p-1} \text{ is not zero.}$
- Reject H_0 if F-ratio $> F_{\alpha, p-1, n-p}$.



Example: Dwaine Studio

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Source of Variation	SS	df	MS
Regression	$SSR = 24015.3$	$3 - 1 = 2$	$MSR = 12007.65$
Error	$SSE = 2180.9$	$21 - 3 = 18$	$MSE = 121.2$
Total	$SS_{tot} = 26196.2$		

- F-ratio = $\frac{MSR}{MSE} = 99.07$.
- F-ratio $\sim F_{2,18}$.
- Hypothesis $H_0 : \beta_1 = \beta_2 = 0$ vs $H_1 : \text{at least one of } \beta_1, \beta_2 \text{ is not zero.}$
- Reject H_0 because F-ratio $> F_{0.05,2,18} = 3.55$.

[Check out R function `anova`. To find $F_{0.05,2,18}$, call function `qf(0.95, 2, 18)`.]



R-square

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- R-square is still defined as $R^2 = \frac{SSR}{SS_{tot}} = 1 - \frac{SSE}{SS_{tot}}$.
- R-square has a problem. R^2 is always increasing as the number of input variables in the model increases. In other words, R^2 increases as the linear model becomes larger. [Why?](#)
- R^2 only has small increase if we add more input variables that are not significant. Thus R^2 is not a perfect criterion for comparing different linear regression models.



Adjusted R-square

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

■ Definition:

$$R_a^2 = 1 - \frac{SSE/n - p}{SST/(n - 1)} = 1 - (1 - R^2) \frac{n - 1}{n - p} = (R^2 - \frac{p - 1}{n - 1}) \frac{n - 1}{n - p}.$$

- Adjusted R-square adjusts for the number of explanatory terms in a model relative to the number of data points. It can be understood as the percentage of explained variation per df by the model.
- R_a^2 can be negative. If this happens, any statistical software will set $R_a^2 = 0$.
- $R_a^2 \leq R^2$ for the same linear regression model.
- R_a^2 increases when a new predictor variable is included only if it improves R^2 more than it would be expected by chance.



Inference of $\hat{\beta}$

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Least square estimation $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.
- Each individual $\hat{\beta}_i \sim N(\beta_i, \sigma^2 L_i)$ $i = 0, 1, \dots, p - 1$, where L_i is the $(i + 1)$ th element of the diagonal entries of $(\mathbf{X}'\mathbf{X})^{-1}$, starting the index of the matrix row and column from $1, \dots, p$.
- So $\hat{\beta}_i$ and $\hat{\beta}_j$ are correlated for any pair of different i and j , if the off-diagonal entries $(\mathbf{X}'\mathbf{X})^{-1}$ are not zero.
- If replace σ^2 by $\hat{\sigma}^2$, then $t = \frac{\hat{\beta}_i - \beta}{s_{\hat{\beta}_i}} \sim t_{n-p}$, where $s_{\hat{\beta}_i} = \hat{\sigma}\sqrt{L_i}$.



Inference of $\hat{\beta}$

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Hypothesis Test for β_i for $i = 0, 1, \dots, p - 1$.

Null Hypothesis	Alternative Hypothesis	Rejection Condition	p-value
$H_0: \beta_i = 0$	$H_a: \beta_i \neq 0$	$ t > t_{\alpha/2, n-p}$	Twice the area under the t-curve to the right of $ t $.
$H_0: \beta_i \leq 0$	$H_a: \beta_i > 0$	$t > t_{\alpha, n-p}$	The area under the t-curve to the right of t .
$H_0: \beta_i \geq 0$	$H_a: \beta_i < 0$	$t < -t_{\alpha, n-p}$	The area under the t-curve to the left of t .

Or reject H_0 if p-value $< \alpha$.

- Two-sided C.I.

Two-sided confidence interval of β_i is $\hat{\beta}_i \pm t_{\alpha/2, n-p} s_{\beta_i}$ for $i = 0, 1, \dots, p - 1$.



Example: Dwaine Studio

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-68.8571	60.0170	-1.147	0.2663
X1	1.4546	0.2118	6.868	2e-06 ***
X2	9.3655	4.0640	2.305	0.0333 *

- $\hat{\beta}_1 = 1.4546$, 95% CI is $1.4546 \pm t_{0.025, 18} 0.2118 = [1.0098, 1.8994]$.
- $\hat{\beta}_2 = 9.3655$, 95% CI is $9.3655 \pm t_{0.025, 18} 4.0640 = [0.8311, 17.8999]$.

$[t_{0.025, 18} = 2.10$. To find this value, call R function `qt(0.975, 18)`.]



Joint inference: Bonferroni method

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Like F -test, we can use multiple comparison (or joint inference) to test if a subset of $\beta_1, \dots, \beta_{p-1}$ meets $\beta_{i_1} = \beta_{i_2} = \dots = \beta_{i_k} = 0$, and the alternative is at least one of the parameter in the subset is not equal to zero.
- To perform statistical inference (hypothesis testing or confidence interval) on multiple parameters (g parameters) simultaneously, *Bonferroni* method basically adjusts the significant level (or Type I error) α for the individual test (CI) of each parameter to be α/g , so that the overall significant level (or Type I error) of the g tests (CI's) is $\leq \alpha$.
- Dwaine Studio Example: joint test β_1 and β_2 . Given $\alpha = 0.05$, we need to find $t_{\alpha/(2\times g), 18}$, where $g = 2$, and $t_{0.05/4, 18} = 2.45$. Only the t-ratio for β_1 is larger than $t_{0.05/4, 18}$, so we reject the null hypothesis. Same conclusion as the F-test from ANOVA.



Estimation of Mean and Inference

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Estimation of $E(Y_h)$ for $\mathbf{x}_h = (1, x_{h,1}, \dots, x_{h,p-1})'$, $\hat{Y}_h = \mathbf{x}'_h \hat{\beta}$.
- Distribution of \hat{Y}_h : $\hat{Y}_h \sim N(\mathbf{x}'_h \beta, \sigma^2 \mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h)$.
- If replace σ^2 with $\hat{\sigma}^2$, and under $H_0 : E(Y_h) = \mu_h$

$$\frac{\hat{Y}_h - \mu_h}{\hat{\sigma} \sqrt{\mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h}} \sim t_{n-p}.$$

- Reject H_0 if the absolute value of the t-ratio is larger than $t_{\alpha/2, n-p}$, or the p-value is smaller than α .
- $100(1 - \alpha)\%$ confidence interval $\hat{Y}_h \pm t_{\alpha/2, n-p} \hat{\sigma} \sqrt{\mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h}$.



Confidence Region

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- The $100(1 - \alpha)\%$ confidence region for the entire regression surface is obtained by

$$\hat{Y}_h \pm W\hat{\sigma}\sqrt{\mathbf{x}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h},$$

where $W^2 = p \times F_{\alpha, p, n-p}$.

- Why cannot we use $\hat{Y}_h \pm t_{\alpha/2, n-p}\hat{\sigma}\sqrt{\mathbf{x}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h}$ for any \mathbf{x}_h ? It is because we make the confidence region that must encompass the entire regression surface, whereas the confidence interval for $E(Y_h)$ at \mathbf{x}_h applies only at the single point \mathbf{x}_h . Same reasoning as in joint inference. This confidence region is wider than the confidence interval.
- It is called confidence band for simple linear regression.



Simultaneous Confidence Intervals for Several Mean Responses

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Using the Working-Hotelling confidence region, for each x_h for $h = 1, \dots, g$, return $\hat{Y}_h \pm W s.e(\hat{Y}_h)$.
- Using Bonferroni simultaneous confidence interval, replace α to α/g in the t-quantile, $t_{\alpha/2g, n-p}$.



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Estimate the mean response for $(x_{h1}, x_{h2}) = (65.4, 17.6)$.

$$\blacksquare \hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_{h1} + \hat{\beta}_2 x_{h2} = [1, 65.4, 17.6] \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix} = 191.10.$$

$$\blacksquare \hat{\sigma}^2 \mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h = 121.1626 [1, 65.4, 17.6] \begin{bmatrix} 29.7289 & 0.0722 & -1.9926 \\ 0.0722 & 0.00037 & -0.0056 \\ -1.9926 & -0.0056 & 0.1363 \end{bmatrix} \begin{bmatrix} 1 \\ 65.4 \\ 17.6 \end{bmatrix} = 7.656.$$

$$\blacksquare \hat{Y}_h \pm t_{\alpha/2, n-p} s.e.(\hat{Y}_h) = 191.10 \pm 2.10 \times 2.77 = [185.3, 196.9].$$



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

```
dwaine<- read.table("CH06FI05.txt")
colnames(dwaine)<-c("X1", "X2", "Y")
fit <- lm(Y~X1+X2,data=dwaine)
X1_seq<-seq(from=min(dwaine$X1),to=max(dwaine$X1),length=20)
X2_seq<-seq(from=min(dwaine$X2),to=max(dwaine$X2),length=20)
pred <- predict(fit,newdata=data.frame(X1=rep(X1_seq,each=20),
X2=rep(X2_seq,length.out=20*20)),se.fit=TRUE)
W <-sqrt(3*qf(1-0.05,3,21-3))
CI_band_upper <- pred$fit+W*pred$se.fit
CI_band_lower <- pred$fit-W*pred$se.fit

library(rgl)
persp3d(x=X1_seq,y=X2_seq,z=matrix(pred$fit,ncol=20,byrow=TRUE),col='lightblue',
xlab='X1',ylab='X2',zlab='Y',alpha=0.7)
persp3d(x=X1_seq,y=X2_seq,z=matrix(CI_band_upper,ncol=20,byrow=TRUE),col='orange',
xlab='X1',ylab='X2',zlab='Y',alpha=0.7,add=TRUE)
persp3d(x=X1_seq,y=X2_seq,z=matrix(CI_band_lower,ncol=20,byrow=TRUE),col='orange',
xlab='X1',ylab='X2',zlab='Y',alpha=0.7,add=TRUE)
points3d(x=dwaine$X1,y=dwaine$X2,z=dwaine$Y,col='green')
```



Prediction of New Observations

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Prediction of Y_h at $\mathbf{x}_h = (1, x_{h,1}, \dots, x_{h,p-1})'$ is still $\hat{Y}_h + \hat{\epsilon} = \mathbf{x}'_h \hat{\beta}$.
- But $\text{var}(\hat{Y}_h + \hat{\epsilon}) = \sigma^2(1 + \mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h)$, and $\hat{Y}_h + \hat{\epsilon} \sim N(\mathbf{x}'_h \hat{\beta}, \sigma^2(1 + \mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h))$.
- If replace σ^2 with $\hat{\sigma}^2$, and under $H_0 : Y_h = \mu_h$

$$\frac{\hat{Y}_h - \mu_h}{\hat{\sigma} \sqrt{1 + \mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h}} \sim t_{n-p}.$$

- Reject H_0 if the absolute value of the t-ratio is larger than $t_{\alpha/2, n-p}$, or the p-value is smaller than α .
- $100(1 - \alpha)\%$ prediction interval $\hat{Y}_h \pm t_{\alpha/2, n-p} \hat{\sigma} \sqrt{1 + \mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h}$.



Prediction of g new observations

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Using the simultaneous Scheffé prediction limits, for each x_h for $h = 1, \dots, g$, return $\hat{Y}_h \pm S s.e(\hat{Y}_h + \hat{\epsilon})$, where $S^2 = g \times F_{\alpha, g, n-p}$.
- Using Bonferroni simultaneous confidence interval, replace α to α/g in the t-quantile, $t_{\alpha/2g, n-p}$



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Predict the response for $(x_{h1}, x_{h2}) = (65.4, 17.6)$.

$$\blacksquare \hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_{h1} + \hat{\beta}_2 x_{h2} = [1, 65.4, 17.6] \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix} = 191.10.$$

$$\blacksquare \hat{\sigma}^2 \mathbf{x}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_h = 7.656.$$

■ Prediction interval

$$\hat{Y}_h \pm t_{\alpha/2, n-p} s.e.(\hat{Y}_h + \hat{\epsilon}) = 191.10 \pm 2.10 \times \sqrt{7.656 + 1} = [184.922, 197.278].$$



Scatter plot and correlation matrix

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Plot Y against each X_1, \dots, X_{p-1} to detect the general pattern of relationship between Y and the input variables.
- Correlations between Y and X_1, \dots, X_{p-1} and among X_1, \dots, X_{p-1} themselves can show how strong the linear relationship between Y and the input variables. The collinearity between X_1, \dots, X_{p-1} is also a problem. Strong collinearity will lead to singularity of $\mathbf{X}'\mathbf{X}$.



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

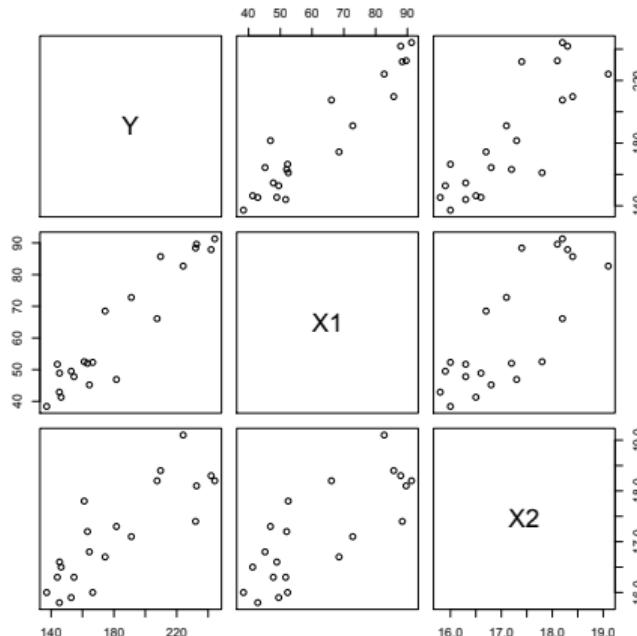
Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

```
dwaine<- read.table("CH06FI05.txt")
colnames(dwaine)<-c("X1", "X2", "Y")
pairs(~Y+X1+X2, data=dwaine)
cor(dwaine)
```

	X1	X2	Y
X1	1.0000000	0.7812993	0.9445543
X2	0.7812993	1.0000000	0.8358025
Y	0.9445543	0.8358025	1.0000000





Check the normal assumption

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

We need to check the assumption on $\epsilon \sim N(0, \sigma^2)$ in the following way.

- Residual plot: e_i v.s. \hat{Y}_i , e_i v.s. X_j for $j = 1, \dots, p - 1$. If there is no systematic pattern, the residuals should appear randomly around zero with constant variance.
- Quantile-Quantile (QQ) plot: check the normal distribution assumption.
- Brown-Forsythe test and Breusch-Pagan test from Section 3.6 in the textbook.
(skip in lecture).
- Plot residuals against the index to see if auto-correlation exists and check independence.



QQ plot

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Steps to draw QQ plot:

- 1 Sort residuals increasingly, $e_{(1)} \leq \dots \leq e_{(n)}$.
- 2 Compute the corresponding normal quantiles via transformation,
$$z_{(i)} = \Phi^{-1} \left(\frac{i-0.5}{n} \right)$$
, or
$$z_{(i)} = \Phi^{-1} \left(\frac{i}{n+1} \right)$$
, or
$$z_{(i)} = \Phi^{-1} \left(\frac{3i-1}{3n+1} \right)$$
, where Φ^{-1} is the inverse CDF for the standard normal distribution.
- 3 Plot $e_{(i)}$ v.s. $z_{(i)}$. If the points are distributed along the line, then the normality assumption holds. Otherwise, it does not hold. Typically, the tail parts of the points will diverge from the straight line if the underlying distribution is t -distribution.



QQ plot

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Logic behind QQ plot:

- The nonparametric estimation of the probability $P(\text{residuals} \leq e_{(i)})$ is i/n .
- If normal assumption holds, $e_{(i)}$'s as the estimate of $\epsilon_{(i)}$, should approximately have the probability as $P(\text{residuals} \leq e_{(i)}) = i/n = \Phi(z_{(i)})$ where $z_{(i)} = \Phi^{-1}(i/n)$. Thus $e_{(i)}$ is proportional to $z_{(i)}$, thus $e_{(i)}$ vs $z_{(i)}$ plot should have a straight line appearance.
- To avoid the probably $z_{(n)} = \Phi^{-1}(n/n) = \Phi^{-1}(1) = \infty$, we adjust the ratio i/n to, $(i - 0.5)/n$, or $(3i - 1)/(3n + 1)$, or $i/(n + 1)$.



Example: Dwaine Studio

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

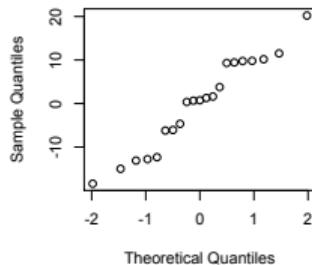
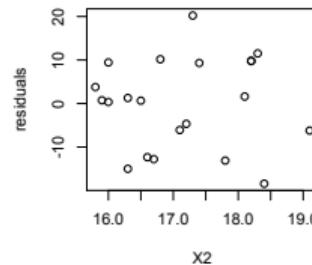
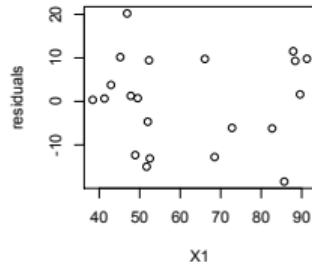
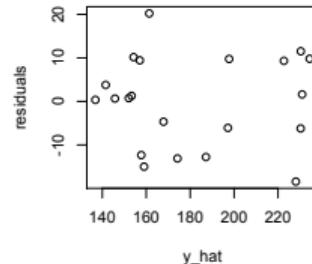
Diagnostics
and Remedies

```
fit<-lm(Y~X1+X2,data=dwaine)
par(mfrow=c(2,2))
plot(fit$fitted,fit$resid,
xlab="y_hat",ylab="residuals")

plot(dwaine$X1,fit$resid,
xlab="X1",ylab="residuals")

plot(dwaine$X2,fit$resid,
xlab="X2",ylab="residuals")

qqnorm(fit$resid)
```





Lack-of-fit Test

DATA
564-494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Assume the data has some replications for different values of the input variables.

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip-1} + \epsilon_{ij} = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_{ij}$$

for $j = 1, \dots, n_i$ and $i = 1, \dots, m$. The total number of observations is
 $n_1 + \dots + n_m = n$

- There are m different values of input variables, x_1, \dots, x_m .
- At each x_i , there are n_i repeated measures of Y , i.e., replications.



Decomposition of Variation

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- $SST = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$ represents the total variation from all data.
- $SSR = \sum_{i=1}^m \sum_{j=1}^{n_i} (\hat{y}_{ij} - \bar{y})^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^m n_i (\hat{y}_i - \bar{y})^2$ represents the error from the regression.
- $SSE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$.
- $SSPE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ represents the pure error that are affected by possible measurement error. Here $\bar{y}_j = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$.
- $SSLF = \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_{ij})^2 = \sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2$

The error can be further decomposed into lack-of-fit (inadequacy of the model) and pure error (measurement noise).

$$SST = SSR + SSE$$

$$SSE = SSLF + SSPE$$



ANOVA with replications

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Source of Variation	SS	df	MS
Regression	SSR	$p - 1$	$MSR = \frac{SSR}{p-1}$
Error	SSE	$n - p$	$MSE = \frac{SSE}{n-p}$
Lack-of-fit	SSLF	$m - p$	$MSLF = \frac{SSLF}{m-p}$
Pure Error	SSPE	$n - m$	$MSPE = \frac{SSPE}{n-m}$
Total	SST	$n - 1$	

$$E(MSPE) = \sigma^2$$

$$E(MSLF) = \sigma^2 + \frac{\sum_{i=1}^m n_i(E(Y|\mathbf{x}_i) - \mathbf{x}'_i \boldsymbol{\beta})^2}{m - p}$$

ANOVA with replications and lack-of-fit

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

Source of Variation	SS	df	MS
Regression	SSR	$p - 1$	$MSR = \frac{SSR}{p-1}$
Error	SSE	$n - p$	$MSE = \frac{SSE}{n-p}$
Lack-of-fit	SSLF	$m - p$	$MSLF = \frac{SSLF}{m-p}$
Pure Error	SSPE	$n - m$	$MSPE = \frac{SSPE}{n-m}$
Total	SST	$n - 1$	

- F-ratio₂ = $\frac{MSLF}{MSPE} \sim F_{m-p, n-m}$
- $H_0 : E(Y|X) = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1}$ v.s.
 $H_a : E(Y|X) \neq \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1}$.
- Reject H_0 if F-ratio₂ > $F_{\alpha, m-p, n-m}$.



Example: Bank Data

MATH
564:494
Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

```
Y <- c(160,112,124,28,152,156,42,124,150,104,136)
X <- c(125,100,200,75,150,175,75,175,125,200,100)
fit<-lm(Y~X)
library(alr3);pureErrorAnova(fit)
Analysis of Variance Table

Response: Y
            Df  Sum Sq Mean Sq F value    Pr(>F)
X             1  5141.3  5141.3  22.393 0.005186 ***
Residuals     9 14741.6   1638.0
Lack of fit   4 13593.6   3398.4  14.801 0.005594 ***
Pure Error    5   1148.0    229.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternatively,

```
X.f<- as.factor(X)
fit2 <- lm(Y~X.f)
anova(fit,fit2)
```



Remedies

DATA

564-494

Simple Linear
Regression

Lulu Kang

Multiple
Regression

Least square
estimation

ANOVA, R^2 ,
 R_a^2

Inference of $\hat{\beta}$

Estimation
and Prediction

Diagnostics
and Remedies

- Add missing terms if residual plot show any curvature pattern with respect to \hat{Y} or any of X_j 's.
- Apply box-cox transformation to the response if the variance of the residuals is not constant.
- Look for new input variables if the lack-of-fit test reject the null hypothesis.



DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Part IV

Multiple Linear Regression II

- 2. Extra Sums of Squares**
- 2.1 Tests for Regression Coefficients**
- 2. Standardized Multiple Regression Model**
- 2.4 Multicollinearity and Its Effects**
- 2.5 Polynomial Regression of Quantitative Predictors**
- 2.6 Qualitative Predictors**



Extra Sums of Squares: basic ideas

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model.
- One can view an extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model

Example (Body Fat)

- $n = 20$
- Y amount of body fat of healthy females of 25-34 years old.
- X_1 triceps skinfold thickness, X_2 thigh circumference, X_3 midarm circumference



Example: Body Fat

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Compare four models:

- (a) Regression of Y on X_1 : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$
- (b) Regression of Y on X_2 : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_2$
- (c) Regression of Y on both X_1 and X_2 : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$
- (d) Regression of Y on X_1, X_2, X_3 : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$

Question: which one should be the final model we use to make interpretation, prediction and conclusion?



Example: Body Fat

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

```
bodyfat<-read.table("BodyFat.txt")
colnames(bodyfat)<- c("X1","X2","X3","Y")
fit1 <- lm(Y~X1,data=bodyfat)
summary(fit1)
anova(fit1)
```

Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	

Variable	Estimation	Standard Deviation	t-value
X_1	$\hat{\beta}_1 = 0.8572$	$s(\hat{\beta}_1) = 0.1288$	6.66



Example: Body Fat

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

```
fit2 <- lm(Y~X2,data=bodyfat)
summary(fit2)
anova(fit2)
```

Source of Variation	SS	df	MS
Regression	382.97	1	382.97
Error	113.42	18	6.30
Total	495.39	19	

Variable	Estimation	Standard Deviation	t-value
X_2	$\hat{\beta}_1 = 0.8565$	$s(\hat{\beta}_1) = 0.1100$	7.79



Example: Body Fat

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

```
fit3 <- update(fit1,.~.+X2)
summary(fit3)
```

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.55	17	6.47
Total	495.39	19	
Variable	Estimation	Standard Deviation	t-value
X_1	$\hat{\beta}_1 = 0.2224$	$s(\hat{\beta}_1) = 0.3034$	0.73
X_2	$\hat{\beta}_2 = 0.6594$	$s(\hat{\beta}_2) = 0.2912$	2.26

$$\begin{aligned}SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) = 385.44 - 352.27 = 33.17 \\&= SSE(X_1) - SSE(X_1, X_2) = 143.12 - 109.95 = 33.17\end{aligned}$$



Example: Body Fat

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
X_1	$SSR(X_1) = 352.27$	1	352.27
X_2	$SSR(X_2 X_1) = 33.17$	1	33.17
Error	109.55	17	6.47
Total	495.39	19	
Variable	Estimation	Standard Deviation	t-value
X_1	$\hat{\beta}_1 = 0.2224$	$s(\hat{\beta}_1) = 0.3034$	0.73
X_2	$\hat{\beta}_2 = 0.6594$	$s(\hat{\beta}_2) = 0.2912$	2.26

$$\begin{aligned} SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) = 385.44 - 352.27 = 33.17 \\ &= SSE(X_1) - SSE(X_1, X_2) = 143.12 - 109.95 = 33.17 \end{aligned}$$



Example: Body Fat

If we change the order of X_1 and X_2 entering the model, and call
`anova(update(fit2,.~.+X1))...`

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
X_2	$SSR(X_2) = 381.97$	1	381.97
X_1	$SSR(X_1 X_2) = 3.47$	1	3.4711
Error	109.55	17	6.47
Total	495.39	19	
Variable	Estimation	Standard Deviation	t-value
X_1	$\hat{\beta}_1 = 0.2224$	$s(\hat{\beta}_1) = 0.3034$	0.73
X_2	$\hat{\beta}_2 = 0.6594$	$s(\hat{\beta}_2) = 0.2912$	2.26

$$\begin{aligned} SSR(X_1|X_2) &= SSR(X_1, X_2) - SSR(X_2) = 385.44 - 381.97 = 3.47 \\ &= SSE(X_2) - SSE(X_1, X_2) = 113.42 - 109.95 = 3.47 \end{aligned}$$

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors



Example: Body Fat

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

```
fit4 <- fit4 <- update(fit3, .~.+X3)
summary(fit4)
```

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	

Variable	Estimation	Standard Deviation	t-value
X_1	$\hat{\beta}_1 = 4.334$	$s(\hat{\beta}_1) = 3.016$	1.44
X_2	$\hat{\beta}_2 = -2.857$	$s(\hat{\beta}_2) = 2.582$	-1.11
X_3	$\hat{\beta}_3 = -2.186$	$s(\hat{\beta}_3) = 1.596$	-1.37



Example: Body Fat

MATH
564:494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
X_1	$SSR(X_1) = 352.27$	1	352.27
X_2	$SSR(X_2 X_1) = 33.17$	1	33.17
X_3	$SSR(X_3 X_1, X_2) = 11.54$	1	11.55
Error	98.41	16	6.15
Total	495.39	19	
Variable	Estimation	Standard Deviation	t-value
X_1	$\hat{\beta}_1 = 4.334$	$s(\hat{\beta}_1) = 3.016$	1.44
X_2	$\hat{\beta}_2 = -2.857$	$s(\hat{\beta}_2) = 2.582$	-1.11
X_3	$\hat{\beta}_3 = -2.186$	$s(\hat{\beta}_3) = 1.596$	-1.37

$$\begin{aligned} SSR(X_3|X_1, X_2) &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) = 396.98 - 385.44 = 11.54 \\ &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = 109.95 - 98.41 = 11.54 \end{aligned}$$



Definitions

MATH
564:464
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- The extra sum of squares of adding one predictor X_k to a regression model with existing predictors X_1, \dots, X_{k-1} is

$$\begin{aligned}SSR(X_k|X_1, \dots, X_{k-1}) &= SSR(X_1, \dots, X_{k-1}, X_k) - SSR(X_1, \dots, X_{k-1}) \\&= SSE(X_1, \dots, X_{k-1}) - SSE(X_1, \dots, X_{k-1}, X_k)\end{aligned}$$

The corresponding df or $SSR(X_k|X_1, \dots, X_{k-1})$ is 1 if X_k is a single quantitative term.

- The extra sum of squares of adding another p predictor X_{k+1}, \dots, X_{k+p} to a regression model with existing predictors X_1, \dots, X_k is

$$\begin{aligned}SSR(X_{k+1}, \dots, X_{k+p}|X_1, \dots, X_k) \\&= SSR(X_1, \dots, X_k, X_{k+1}, \dots, X_{k+p}) - SSR(X_1, \dots, X_k) \\&= SSE(X_1, \dots, X_k) - SSE(X_1, \dots, X_k, X_{k+1}, \dots, X_{k+p})\end{aligned}$$

The corresponding df or $SSR(X_{k+1}, \dots, X_{k+p}|X_1, \dots, X_k)$ is p if X_{k+1}, \dots, X_{k+p} are all quantitative terms.



Decomposition of SSR into Extra SS

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Consider the regression model for the Body Fat example, the regression model contain all predictors is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$.

- $SST = SSR(X_1, X_2, X_3) + SSE(X_1, X_2, X_3)$
- $SSR(X_1, X_2, X_3) = SSR(X_1, X_2) + SSR(X_3|X_1, X_2)$
- $SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$
- $SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2, X_3|X_1)$



Test whether a single $\beta_k = 0$

MATH

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- We need to recognize a full model:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \dots + \beta_{p-1} X_{p-1} + \epsilon.$$

- Question, whether a particular X_k should be dropped from the model, i.e., $\beta_k = 0$?

- Compare the full model with the reduced model

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1} + \beta_{k+1} X_{k+1} + \dots + \beta_{p-1} X_{p-1} + \epsilon.$$

- Use the partial F test:

$$\begin{aligned} F - ratio &= \frac{SSR(X_k | X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1}) / 1}{MSE(X_1, \dots, X_{p-1})} \\ &= \frac{(SSR_F - SSR_R) / (df_F - df_r)}{MSE_F} \end{aligned}$$

which follows $F_{1,n-p}$ distribution.



Test whether a single $\beta_k = 0$

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

■ Partial F test:

$$\begin{aligned} F - ratio &= \frac{SSR(X_k | X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1}) / 1}{MSE(X_1, \dots, X_{p-1})} \\ &= \frac{(SSR_F - SSR_R) / (df_F - df_r)}{MSE_F} \end{aligned}$$

■ Equivalent test: $H_0: \beta_k = 0$ v.s. $H_a: \beta_k \neq 0$

$$t - ratio = \frac{\hat{\beta}_k}{s.e.(\hat{\beta}_k)},$$

under H_0 , t-ratio following t_{n-p} distribution.



Test whether some $\beta_k = 0$

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

■ Hypotheses:

$$H_0: \beta_q = \beta_{q+1} = \dots = \beta_{p-1} = 0,$$

H_a : not all of the β_k in H_0 equal zero.

■ Partial F-test:

$$\begin{aligned} F - ratio &= \frac{SSR(X_q, \dots, X_{p-1} | X_1, \dots, X_{q-1}) / (p - q)}{MSE(X_1, \dots, X_{p-1})} \\ &= \frac{(SSR_F - SSR_R) / (df_F - df_r)}{MSE_F}, \end{aligned}$$

which follows $F_{p-q, n-p}$ distribution.



Other Tests

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- Full model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Hypotheses: $H_0: \beta_1 = \beta_2$ vs $H_a: \beta_1 \neq \beta_2$
- Reduced model $Y = \beta_0 + \beta_c(X_1 + X_2) + \beta_3 X_3 + \epsilon$
- F-ratio: $\frac{(SSR_F - SSR_R)/(df_F - df_r)}{MSE_F}$, $df_F = 3$, $df_r = 2$, which follows $F_{df_F - df_r, n-p}$ distribution.



Other Tests

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- Full model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Hypotheses: $H_0: \beta_1 = 3, \beta_3 = 5$ vs $H_a: \beta_1 \neq 3$ or $\beta_3 \neq 5$.
- Reduced model $Y^* = \beta_0 + \beta_2 X_2 + \epsilon$, where $Y^* = Y - 3X_1 - 5X_3$
- F-ratio: $\frac{(SSRF - SSR_R)/(df_F - df_r)}{MSE_F}$, $df_F = 3$, $df_r = 1$, which follows $F_{df_F - df_r, n-p}$ distribution.



Example: Body fat

DATA

564-490

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Example (Body Fat)

Consider the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$. Which two of the predictors are redundant? We need to find the two predictors that would conclude with the null hypothesis $H_0: \beta_i = \beta_j = 0$.

```
bodyfat<-read.table("BodyFat.txt")
colnames(bodyfat)<- c("X1", "X2", "X3", "Y")
fit1<-lm(Y~X1,data=bodyfat); fit2<-lm(Y~X2,data=bodyfat)
fit3<-lm(Y~X3,data=bodyfat); fitf<-lm(Y~X1+X2+X3,data=bodyfat)
anova(fit1,fitf); anova(fit2,fitf); anova(fit3,fitf)
```



Roundoff Errors

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- a is a real number, and might have infinite digits, but it is stored in computer with a finite number of digits as \tilde{a} . The difference between a and \tilde{a} is roundoff error.
- Roundoff errors are mostly likely to be involved in calculating the inverse of $\mathbf{X}'\mathbf{X}$. (What kind of arithmetic operation would like to produce infinite number of digits of a number? Division.)
- The danger of the roundoff error is that the roundoff error is to be magnified by the following calculation for $\hat{\boldsymbol{\beta}}$, which is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.



What can cause the most serious roundoff error?

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- $\mathbf{X}'\mathbf{X}$ has a determinant that is close to zero. Multi-collinearity can lead to this.
- If the elements of $\mathbf{X}'\mathbf{X}$ differ substantially in order of magnitude. This can easily happen when different predictors have different ranges of values, naturally based on the practical problem.



Lack of Comparability in Regression Coefficients

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- The size of the coefficients β_i depends on many factors, particularly the scale of the predictors corresponding to β_i . So we cannot judge the statistical significance of β_i just based on the size of $\hat{\beta}_i$. Instead, we have to use the t-ratio, $\hat{\beta}_i / s.e.(\hat{\beta}_i)$, which is scale-free.
- We certainly should not compare the significance of different predictors based on the values of $\hat{\beta}_i$'s when the scale of the predictors are very different.



Correlation Transformation

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

How to do it:

- Perform the standardization for each of Y, X_1, \dots, X_{p-1} as follows:

$$\frac{Y_i - \bar{Y}}{s_Y}, \quad \frac{X_{ik} - \bar{X}_k}{s_k}, \text{ for } k = 1, \dots, p-1, \text{ and } i = 1, \dots, n.$$

- \bar{Y} is the sample mean of Y_1, \dots, Y_n and \bar{X}_k is the sample mean for each $X_{1,k}, \dots, X_{n,k}$, for $k = 1, \dots, p-1$.
- s_Y is the sample standard deviation for Y , and s_i is the sample standard deviation for predictor X_i .
- *Correlation Transformation* is

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{s_Y} \right),$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{s_k} \right), \text{ for } k = 1, \dots, p-1, \text{ and } i = 1, \dots, n.$$



Standardized Regression Model

MATH

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Perform regression on the correlation transformed data.

$$Y_i^* = \beta_1^* X_{i1}^* + \dots + \beta_{p-1}^* X_{i,p-1}^* + \epsilon_i^*.$$

- No intercept, because if there is one, it will be zero. For the original model least square estimated model

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

- Based on the least estimation, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \dots - \hat{\beta}_{p-1} \bar{X}_{p-1}$.
- Take the above representation back to the fitted regression model:

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_1 (X_{i1} - \bar{X}_1) + \dots + \hat{\beta}_{p-1} (X_{i,p-1} - \bar{X}_{p-1})$$

- Let $\hat{\beta}_i^* = \frac{s_k}{s_Y} \hat{\beta}_i$ for $i = 1, \dots, p-1$. Then replace $\hat{\beta}_i$ by $\hat{\beta}_i^*$ in the last equation, we obtain

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



Estimated the Standardized Regression Coefficient

- $\mathbf{X}^{*'} \mathbf{X}^*$ matrix for the correlation transformed data is following

$$\mathbf{r}_{XX} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1,p-1} \\ r_{21} & 1 & \dots & r_{2,p-1} \\ \vdots & \vdots & & \vdots \\ r_{p-1,1} & r_{p-1,2} & \dots & 1 \end{bmatrix},$$

which is the correlation matrix for the transformed predictors X_1^*, \dots, X_{p-1}^* of size $(p - 1) \times (p - 1)$. [Derivation omitted. See Page 275 of textbook.]

- Define another correlation vector between Y and X_1, \dots, X_{p-1} .

$$\mathbf{r}_{XY} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{bmatrix}.$$

It can be shown that the correlation transformed data has $\mathbf{X}^{*'} \mathbf{Y}^* = \mathbf{r}_{XY}$.

MATH
564:494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors



Estimated Standardized Regression Coefficients

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- For regular regression, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
- For the correlation transformed data, $\hat{\beta}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^* = \mathbf{r}_{XX}^{-1}\mathbf{r}_{XY}$.
- The original regression coefficient estimates are

$$\hat{\beta}_k = \left(\frac{s_Y}{s_k} \right) \hat{\beta}_k^*, \text{ for } k = 1, \dots, p-1$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \dots - \hat{\beta}_{p-1} \bar{X}_{p-1},$$



Example: Body Fat

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

```
bodyfat<-read.table("BodyFat.txt")
colnames(bodyfat)<- c("X1","X2","X3","Y")
r_XX <- cor(bodyfat[,-4])
r_XY<-cor(bodyfat[,-4],bodyfat[,4])
beta2 <- solve(r_XX,r_XY)
M <- colMeans(bodyfat)
S <- apply(bodyfat, 2, sd)
beta <- (S[4]/S[-4])*beta2
beta0 <- M[4]-M[-4]%^%beta
beta <- c(beta0,beta)
fit <-lm(Y~.,data=bodyfat)
summary(fit)
```



Uncorrelated predictor variables

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- If all the predictors are not correlated, i.e., $\text{cor}(X_i, X_j) = 0$ for all $i \neq j$, then $(\mathbf{X}'\mathbf{X})$ is a diagonal matrix, and thus $(\mathbf{X}'\mathbf{X})^{-1}$ is very easy to compute.
- The extra sum of squares would be equal to the marginal ones.

$$SSR(X_1|X_2) = SSR(X_1), \quad SSR(X_2|X_1) = SSR(X_2).$$

- Adding new predictors would affect the coefficients estimates of the existing predictors. [Page 281 of textbook.]
- In practice, we seldom find predictor variables that are perfectly uncorrelated in the non-experimental data. Some or all predictor variables are correlated among themselves.



Effects of Multicollinearity

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- $\mathbf{X}'\mathbf{X}$ will have large condition number, which makes the inverse of $\mathbf{X}'\mathbf{X}$ numerically unstable. It will lead to large roundoff error. To make things worse, if one column of \mathbf{X} is perfectly collinear with other columns of \mathbf{X} , \mathbf{X} does not have full rank, and thus $\mathbf{X}'\mathbf{X}$ is singular and not invertible.
- Even if $\mathbf{X}'\mathbf{X}$ is still invertible and only with large condition number, the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ would have large diagonal entries, which makes the variance of $\hat{\beta}$ very large.
- Multicollinearity can cause large condition number of $\mathbf{X}'\mathbf{X}$. Small changes in \mathbf{y} will cause large changes in $\hat{\beta}$. But the underlying unknown model should not be so sensitive to small change in the response.
- If X_j is collinear with some or all of the other predictor variables, then the size of the t-statistic and the related p-value measure the additional importance of the predictor variable X_j over the combined importance of the other predictor variables. This is called effects confounding.



Variance Inflation Factor (Section 10.5)

MATH

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- To detect if the model term X_k has strong collinearity with other model terms, use variance inflation factor (VIF) for this model term.

$$VIF(X_k) = \frac{1}{1 - R_k^2},$$

where R_k^2 is the coefficient of determination of regressing X_k with respect to all the other input model terms.

- If $R_k^2 = 0$, then X_k is not related to any other model terms. So $VIF(X_k) = 1$.
- If $R_k^2 \rightarrow 1$, then X_k is perfectly linearly related to some or all of the other model terms, and $VIF(X_k) \rightarrow \infty$.
- So the larger the $VIF(X_k)$ is, the worse the collinearity of X_k is with other input model terms.



Variance Inflation Factor (Section 10.5)

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

The maximum VIF (denote maxVIF in my notes) and the averaged VIF (denoted as \bar{VIF}) of all the input model terms are used to evaluate the overall multicollinearity of the regression model. If $\text{max } VIF > 10$ (equivalently $R_j^2 > 0.9$) or $\bar{VIF} \gg 1$, then we would say that there exists multicollinearity in the model.



Example: Body Fat (Section 10.5)

MATH

564.49

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Example

The VIF values for the three predictors are $VIF_1 = 708.84$, $VIF_2 = 564.34$, $VIF_3 = 104.61$, so $\bar{VIF} = 459.26$ and $\max VIF = VIF_1 = 708.84$. There are some serious multicollinearity between the predictors.

```
bodyfat<-read.table("BodyFat.txt")
colnames(bodyfat)<- c("X1", "X2", "X3", "Y")
library(car)
fit <- lm(Y~., data=bodyfat)
vif(fit)
```



General Polynomial Regression Models

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

If X_1, X_2, \dots, X_k are all quantitative predictors, i.e., the possible values for the predictors are not discrete, then the polynomial regression model of order q is

$$E(Y) = \beta_0 + \sum_{i=1}^k \beta_i X_i + \sum_{i>j} \beta_{ij} X_i X_j + \sum_{j=1}^k \beta_{jj} X_j^2 + \dots + \dots$$

The omitted terms can be denoted in general as $\beta_l X_1^{i_1} X_2^{i_2} \dots X_k^{i_k}$ and $i_1 + i_2 + \dots + i_k \leq q$. A quadratic (second-order) polynomial model is

$$E(Y) = \beta_0 + \sum_{i=1}^k \beta_i X_i + \sum_{i>j} \beta_{ij} X_i X_j + \sum_{j=1}^k \beta_{jj} X_j^2.$$



Interaction Terms

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- The terms involving multiple predictors are called *interaction* terms.
- Interpretation of the interaction term.

Example

Consider the model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$. Whether β_3 is zero (or statistically significant) would lead to different patterns of the *interaction plot* or *conditional effects plot*.



Interaction Terms

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

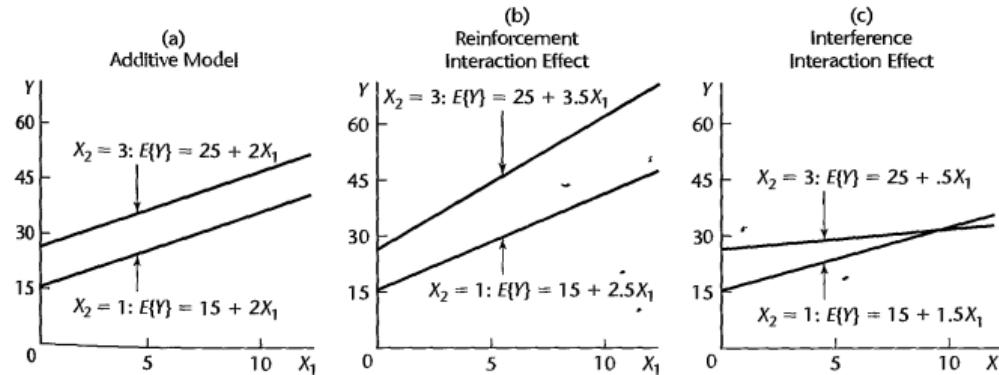
Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

FIGURE 8.7 Illustration of Reinforcement and Interference Interaction Effects—Sales Promotion Example.



- $E(Y) = 10 + 2X_1 + 5X_2$
- $E(Y) = 10 + 2X_1 + 5X_2 + 0.5X_1X_2$
- $E(Y) = 10 + 2X_1 + 5X_2 - 0.5X_1X_2$



Interaction Terms

MATH
564:494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

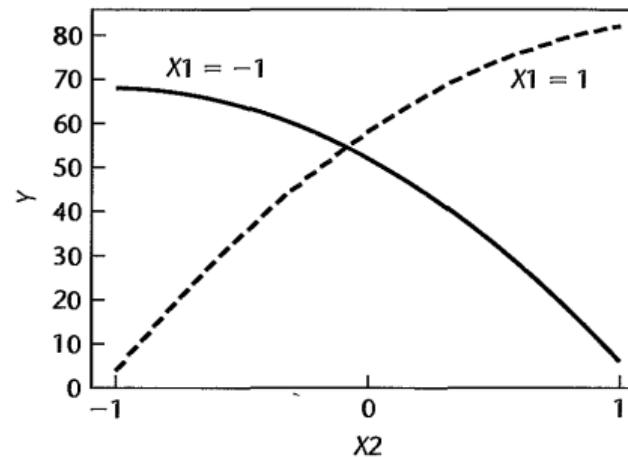
Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

When quadratic terms of X_1 and X_2 exist in the model, the interaction plot are curves (whether parallel or intersect) not straight lines.



$$E(Y) = 65 + 3X_1 + 4X_2 - 10X_1^2 - 15X_2^2 + 35X_1X_2$$



Orthogonal polynomials

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- Strong interaction can occur between the polynomial terms which lead to severe collinearity problem.
- Correlation transformation can only remove the correlation between the linear terms.
- Orthogonal polynomials are recommended. How to do it?
 - Optional but suggested: standardize the predictor variables.
 - Apply Hermite polynomials, if the predictors can be roughly seen as normally distributed.
 - Apply Legendre polynomials, if the predictors can be roughly seen as uniform distributed.



Qualitative predictor: Chain Store Example

DATA

564-400

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Example

- X_1 : number of households in the area;
- X_2 : locations of the stores can be “street”, “mall”, and “downtown”;
- Y : sales.

X_2 has three levels: mall, street and downtown. Use two dummy variables D_M and D_D , such that,

$$D_M = \begin{cases} 1, & \text{if } x_2 \text{ is mall,} \\ 0, & \text{otherwise.} \end{cases} \quad D_D = \begin{cases} 1, & \text{if } x_2 \text{ is downtown,} \\ 0, & \text{otherwise.} \end{cases}$$



Qualitative predictor: Chain Store Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

Linear regression model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 D_M + \beta_3 D_D + \epsilon$.

- If X_2 is street, $Y = \beta_0 + \beta_1 X_1 + \epsilon$,
- If X_2 is mall, $Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$,
- If X_2 is downtown, $Y = \beta_0 + \beta_1 X_1 + \beta_3 + \epsilon$.

So $\beta_2 = E(Y|X_1, X_2 = \text{mall}) - E(Y|X_1, X_2 = \text{street})$, and can be interpreted as the difference between the mean sales between mall stores and downtown stores, given the same X_1 . So $\beta_3 = E(Y|X_1, X_2 = \text{downtown}) - E(Y|X_1, X_2 = \text{street})$, and can be interpreted as the difference between the mean sales between downtown stores and street stores, given the same X_1 . Street is the baseline level.



Qualitative predictor: Chain Store Example

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

ANOVA Table:

Source	SS	df	MS
X_1	$SSR(X_1)$	1	$SSR(X_1)/1$
X_2	$SSR(D_M, D_D X_1)$	2	$SSR(D_M, D_D X_1)/2$
Error	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - 4$	$MSE = SSE/(n - 4)$
Total	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$	

```
store<-read.table("store_location.txt",header=T,sep="\t")
store$X2 <- as.factor(store$X2)
fit<-lm(Y~.,data=store)
summary(fit)
anova(fit)
```



Qualitative predictor

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

In general if a qualitative variable X has q levels, we can use $q - 1$ dummy variables to model it.

$$D_i = \begin{cases} 1, & \text{if } X \text{ is equal to the } i\text{th level,} \\ 0, & \text{otherwise.} \end{cases} \quad \text{for } i = 1, 2, \dots, q - 1.$$

So the q level is used as the baseline level. For every β_i , the coefficient before D_i in the linear regression model, $\beta_i = E(Y|X = i) - E(Y|X = q)$ the difference between the mean responses with the qualitative variable equal to the i th level and the q th level, given the other variables fixed at the same value. The hypothesis testing and confidence interval on β_i can be performed in the same way as the quantitative variables.



Qualitative predictor

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

The difference $\beta_i - \beta_j$ is the difference between the mean responses with the qualitative variable equal to the i th level and the j th level, given the other variables fixed at the same value. The estimate of $\beta_i - \beta_j$ is $\hat{\beta}_i - \hat{\beta}_j$. So

$$E(\hat{\beta}_i - \hat{\beta}_j) = \beta_i - \beta_j$$

$$\text{var}(\hat{\beta}_i - \hat{\beta}_j) = \text{var}(\mathbf{a}'\hat{\beta}) = \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a},$$

where \mathbf{a} is a vector with the elements corresponding to β_i and β_j equal to 1 and -1, and other elements equal to 0.



Qualitative predictor

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

The distribution of $\hat{\beta}_i - \hat{\beta}_j$ is

$$\frac{\hat{\beta}_i - \hat{\beta}_j}{\hat{\sigma} \sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} \sim t_{n-k-1}.$$

The $100(1 - \alpha)\%$ C. I. is $\hat{\beta}_i - \hat{\beta}_j \pm t_{\alpha/2}^{n-k-1} \hat{\sigma} \sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$.



Chain store example: Interaction between X_1 and X_2

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

The regression model containing the interaction between quantitative and qualitative predictors is

$$E(Y) = Y = \beta_0 + \beta_1 X_1 + \beta_2 D_M + \beta_3 D_D + \beta_4 X_1 D_M + \beta_5 X_1 D_D.$$

The interaction between X_1 and X_2 can be shown through the interaction plot.



Chain store example: Interaction between X_1 and X_2

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

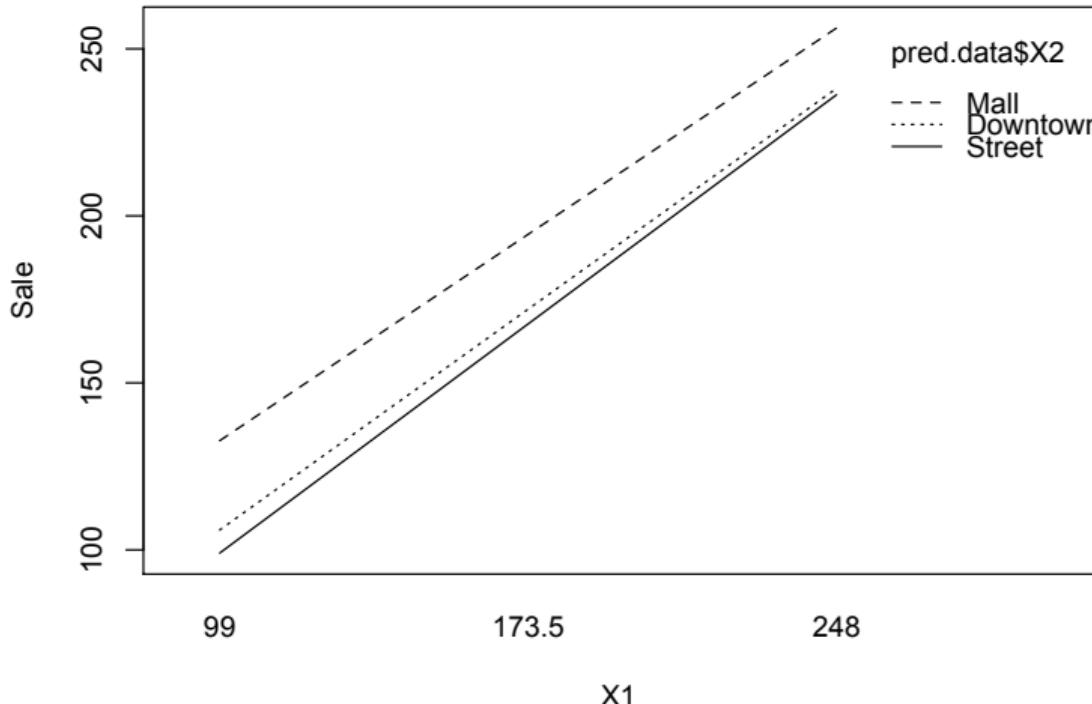
Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors





Chain store example: Interaction between X_1 and X_2

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

```
fit2 <- lm(Y~X1*X2,data=store)
summary(fit2)
pred.data <- data.frame(X1=rep(c(99,(99+248)/2,248),length=9),
X2=rep(levels(store$X2),each=3))
mean.y <- predict(fit2,newdata=pred.data)
interaction.plot(pred.data$X1,pred.data$X2,mean.y,type="l",
xlab='X1',ylab='Sale')
```



Chain store example: Interaction between X_1 and X_2

DATA
564-494
Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

ANOVA Table:

Source	SS	df	MS
X_1	$SSR(X_1)$	1	$SSR(X_1)/1$
X_2	$SSR(D_M, D_D X_1)$	2	$SSR(D_M, D_D X_1)/2$
$X_1 : X_2$	$SSR(X_1D_M, X_1D_D X_1, D_M, D_D)$	2	$SSR(X_1D_M, X_1D_D X_1, D_M, D_D)/2$
Error	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - 6$	$MSE = SSE/(n - 6)$
Total	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$	

`anova(fit2)`



More general case

DATA

564-494

Simple Linear
Regression

Lulu Kang

Extra Sums of
Squares

Tests for
Regression
Coefficients

Standardized
Multiple
Regression
Model

Multicollinearity
and Its Effects

Polynomial
Regression of
Quantitative
Predictors

Qualitative
Predictors

- The regression model can contain several quantitative predictors represented by their different polynomial terms, dummy variables of the qualitative variables, the interactions between dummy variables and polynomial terms, and interactions among the dummy variables.
- Comparing different models, the current tool we have learned is partial F-test. The key is to identify the bigger model of the two under comparison. So the limitation of using this approach is that the bigger model must contains all the terms of the reduced model.



DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

Part V

Building the Regression Model

- 2. Overview of Building Regression Model**
 - 2.1 Model Comparison Criteria**
 - 2.2 Automatic Search Procedures**
 - 2.3 Model Validation**
 - 2.4 Added-Predictor Plot**
 - 2.5 Outliers and Influential Observations**



Overview of Model Building Process

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

Study the flow chart of Figure 9.1 from the textbook. The entire model-building process starts with data collection and ends with model validation. It is a iterative multistage process.



Data collection

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Controlled experiments.
- Controlled experiments with covariates.
- Confirmatory observational studies: data are collected for explanatory variables that previous studies have shown to affect the response variable, as well as for the new variables involved in the hypothesis. Primary variables vs control variables.
- Exploratory observational studies: the opposite of controlled experiments.



Data preparation and preliminary model investigation

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

Data preparation:

- Prepare the raw data into the proper format for later analysis.
- Checks for gross data errors and extreme outliers.

Preliminary model investigation:

- the functional form in which the explanatory variables should enter the regression model
- important interactions that should be included in the model.



Reduction of explanatory variable

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Controlled Experiments: not a problem, as all the controllable variables are already expected to be influential. Previously learned methods can be used to determine whether the explanatory variable have effects on the response variable.
- Controlled Experiments with Covariates: not sure in advance that the selected covariates would be helpful in reducing the error variance. The number of covariates involved in controlled experiments are likely to be small.
- Confirmatory Observational Studies: no reduction of explanatory variables should take place in confirmatory observational studies, because the control variables were chosen one the basis of prior knowledge and should be retained for comparison with earlier studies even if some of the control variables turn out not to lead to any error variance reduction.
- Explanatory Observational Studies: ...



Reduction of explanatory variable

DATA

564 observations

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Explanatory Observational Studies: the number of explanatory variables is large, and the variables are frequently highly correlated.
- We wish to reduce the explanatory variables: (1) smaller model is easy to work with and understand (2) elevate the multicollinearity problem.
- Need to have a subset that contain all the potentially useful explanatory variables. It is a difficult task.
- Need to identify the right size of the model, too small would lead to under-fitting, too large would lead to over fitting.



Model refinement and selection, validation

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Build an initial model.
- Perform the model diagnostics.
- Remedies if necessary.
- Model validation: the stability and the reasonableness of the regression coefficients, the plausibility and usability of the regression function, and the ability to generalize inference.



Surgical unit example

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- X_1 : blood clotting score
- X_2 : prognostic index
- X_3 : enzyme function test score
- X_4 : liver function test score
- X_5 : age, in years
- X_6 : indicator variable for gender (0=male, 1=female)
- X_7 and X_8 : indicator variable of history of alcohol use. (Two dummy variables)

$$X_7 = \begin{cases} 1, & \text{if Moderate} \\ 0, & \text{otherwise} \end{cases} \quad X_8 = \begin{cases} 1, & \text{if Severe} \\ 0, & \text{otherwise} \end{cases}$$

- Y : survival time of the patient.



Surgical Example

DATA

564 observations

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
surgical<-read.table("surgical.txt",header=F)
colnames(surgical)<- c("X1","X2","X3","X4","X5",
 "X6","X7","X8","Y","lnY")

cor(surgical[,c(10,1:4)])
pairs(surgical[,c(10,1:4)])
fit1 <- lm(Y~., data=surgical[,-10])
fit2 <- lm(lnY~., data=surgical[,-9])
plot(fit1)
plot(fit2)
```



Criteria for model selection

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

We need to use certain criterion to compare different models. If there are $p - 1$ predictors, there 2^{p-1} possible models.

- R_p^2 and SSE_p
- $R_{a,p}^2$ and MSE_p
- Mallow's C_p
- AIC and BIC
- $PRESS_p$



R_p^2 and SSE_p

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- The subscript “p” is only to emphasize that these two varies with respect to the size of the model.
- $R_p^2 = 1 - SSE_p/SSTO$: percentage of explained variation by the linear regression models.
- R_p^2 (SSE_p) increases (decreases) as long as the size of the model p increases.
- For the models with the same size, we can use R_p^2 (SSE_p) to select the better model. But not for models with different size.
- R_p^2 and SSE_p are equivalent measure.



$R^2_{a,p}$ and MSE_p

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- $R^2_{a,p} = 1 - \frac{MSE_p}{SSTO/(n-1)}$ takes into account of the degrees of freedom, and $R^2_{a,p}$ to counter the inflation due to model size increase. Can be used to compare models of different sizes.
- $MSE_p = \frac{SSE}{n-p}$ measures how precise the linear regression model is. If MSE_p is small, all the variance of $\hat{\beta}$ and \hat{y} will be small as well.
- $R^2_{a,p}$ and MSE_p are equivalent measure.



■ Mallow's C_p

$$C_p = \frac{SSE}{s_p^2} - [n - 2k]$$

where s_p^2 is the MSE of the complete model (the model that contains all possible candidate model terms). s_p^2 can be seen as very close to σ^2 . The intuition of C_p :

$$E(C_p) = \frac{E(SSE_k)}{E(s_p^2)} - [n - 2k] \approx \frac{[n - k]\sigma^2}{\sigma^2} - [n - 2k] = k$$

Here k is the size of the current model being compared against the complete model (including intercept), and SSE_k is of the current model. So the mean value of C_p is close to the size of the model (including the intercept). To use C_p for model comparison, the smaller C_p is better.



Mallow's C_p

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

Example: x_1 , x_2 , and x_3 are three candidate model terms. What the candidate linear regression models?

\emptyset , $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, $\{x_1, x_2, x_3\}$. For each linear model, calculate the SSE and C_p . Here s^2_4 ($p = 4$) is the MSE of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$. Choose the model that has the smallest C_p . This approach is called *best subset model*.



AIC and BIC

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- AIC: Akaike Information Criterion

$$AIC = -2 \log(\hat{L}) + 2p$$

- BIC (Bayesian Information Criterion) is called *SBC* in the textbook.

$$BIC = -2 \log(\hat{L}) + \log(n)p.$$

- Here \hat{L} is the maximized likelihood of the observations. p is the number of unknown parameters.



AIC and BIC

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- For the linear regression model, the likelihood of the observations $\{y_i\}_{i=1}^n$ is

$$\begin{aligned}L &= \prod_{i=1}^n p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n \exp\left(-\frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma} \\&= \frac{1}{(\sqrt{2\pi})^n (\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)\end{aligned}$$

- The MLE is $\hat{\boldsymbol{\beta}}_{LS}$ and $\hat{\sigma}_{MLE}^2 = SSE/n$. Take the MLE into L and obtain \hat{L} .

$$\begin{aligned}AIC(\hat{\boldsymbol{\beta}}) &= n \log \hat{\sigma}_{MLE}^2 + n \log 2\pi + \frac{1}{\hat{\sigma}_{MLE}^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + 2p \\&= n \log(SSE_p) - n \log n + n \log 2\pi + n + 2p \\&= n \log(SSE_p) - n \log n + 2p + \text{constant}.\end{aligned}$$

- BIC

$$BIC = n \log(SSE_p) - n \log n + (\log n)p + \text{constant}$$



PRESS_p

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- PRESS: prediction sum of squares
- $\hat{Y}_{i(i)}$: using all the data but (X_i, Y_i) to fit the regression and then using the fitted model to predict the Y value at X_i .
- $PRESS_p = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2$
- It can be shown that $Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_i}$, where e_i is the i residual for Y_i from the original regression and h_i is the i th diagonal entry of the hat matrix of the original regression.
- $PRESS_p = \sum_{i=1}^n \frac{e_i^2}{(1-h_i)^2}$.



Best subset

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- For a given model size, compute the model criterion (such as Mallow's C_p) for all possible models of the same size.
- Do so for $k = 1, \dots, p - 1$.
- If there are $p - 1$ potential predictors, the best subset would do the exhaust search by fitting all 2^{p-1} models and pick the best one according to the chosen criterion.



Surgical unit example

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
best1<-regsubsets(x=surgical[,1:8], y=surgical$Y,nbest=1,  
nvmax=8,method="exhaustive")  
best8<-regsubsets(x=surgical[,1:8], y=surgical$Y,nbest=8,  
nvmax=8,method="exhaustive")  
report1<-summary(best1)  
report8<-summary(best8)  
k <- rowSum(report8$which)  
plot(k,report8$rsq,xlab="model size", ylab="R-square")  
lines(1:8,report1$rsq)  
plot(k,report8$adjr2,xlab="model size", ylab="Adjusted R-square")  
lines(1:8,report1$adjr2)  
plot(k,report8$cp,xlab="model size", ylab="Mallow's Cp")  
lines(1:8,report1$cp)  
plot(k,report8$bic,xlab="model size", ylab="BIC")  
lines(1:8,report1$bic)
```



Stepwise Regression

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

Three directions:

- **Forward selection**, which involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model.
- **Backward elimination**, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model the most by being deleted, and repeating this process until no further improvement is possible.
- **Bidirectional elimination**, a combination of the above, testing at each step for variables to be included or excluded.

AIC and BIC are the common criteria used in stepwise regression.



Surgical unit example

DATA

564 rows

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
fit0<-lm(Y~1,data=surgical)
fit.forward<-step(fit0,scope=list(lower=~1,
upper=~X1+X2+X3+X4+X5+X6+X7+X8),direction="forward")
fitf<-lm(Y~X1+X2+X3+X4+X5+X6+X7+X8,data=surgical)
fit.backward<-step(fitf, scope=list(lower=~1,
upper=~X1+X2+X3+X4+X5+X6+X7+X8),
direction="backward")

fit.both<-step(fitf, scope=list(lower=~1,
upper=~X1+X2+X3+X4+X5+X6+X7+X8), direction="both")
fit.both2<-step(fit0, scope=list(lower=~1,
upper=~X1+X2+X3+X4+X5+X6+X7+X8),direction="both")
```



Model Validation

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Collection of new data to check the model and its predictive ability. Compute Mean Square Prediction Error (MSPR)

$$MSPR = \frac{\sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2}{n^*} \text{ should be comparable to MSE}$$

- Comparison of results with theoretical expectations, earlier empirical results, and simulation results.
- Use of a holdout sample to check the model and its predictive ability.



Model Adequacy for a Predictor Variable

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

Added-variable plot.

Example

Consider a first-order multiple regression model with two predictor variables X_1 and X_2 . We question the nature of the regression effect for X_1 , given that X_2 is already in the model. Obtain:

$$\hat{Y}_i(X_2) = \hat{\beta}_0 + \hat{\beta}_2 X_{i2}$$

$$e_i(Y|X_2) = Y_i - \hat{Y}_i(X_2)$$

$$\hat{X}_{i1} = \hat{\beta}_0^* + \hat{\beta}_2^* X_{i2}$$

$$e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

The added-variable plot is to plot $e(Y|X_2)$ against $e(X_1|X_2)$.



Added-Variable Plot

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

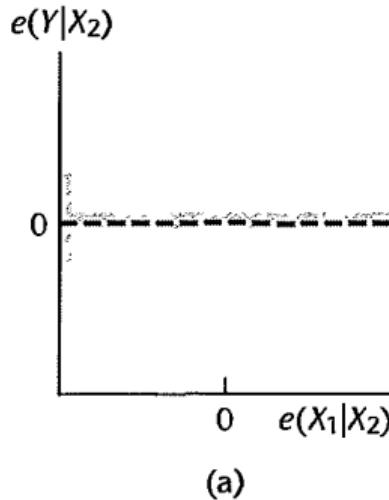
Model
Comparison
Criteria

Automatic
Search
Procedures

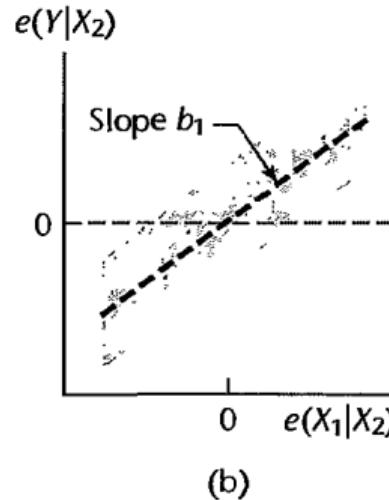
Model
Validation

Added-
Predictor Plot

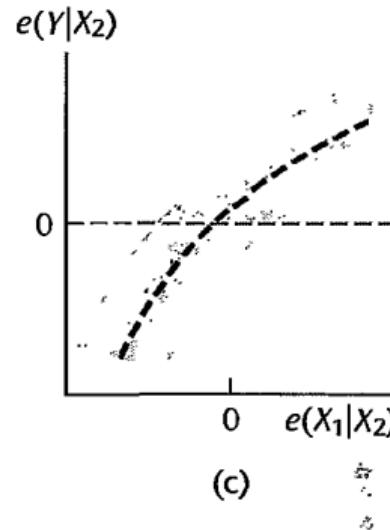
Outliers and
Influential
Observations



(a)



(b)



(c)



Example: Average Income

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
income<-read.table("annual_income.txt",header=FALSE)
colnames(income)<-c("X1","X2","Y")
fit1<-lm(Y~.,data=income)
plot(income$X1,fit1$residuals,xlab="X1",ylab="Residuals")
fit2<-lm(Y~X2,data=income)
fit_X1<-lm(X1~X2,data=income)
plot(fit_X1$residuals,fit2$residuals,xlab="e(X1|X2)",ylab="e(Y|X2)")
fit3<-lm(Y~X1+X2+I(X1^2),data=income)
plot(income$X1,fit3$residuals,xlab="X1",ylab="Residuals")
```



Outliers

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

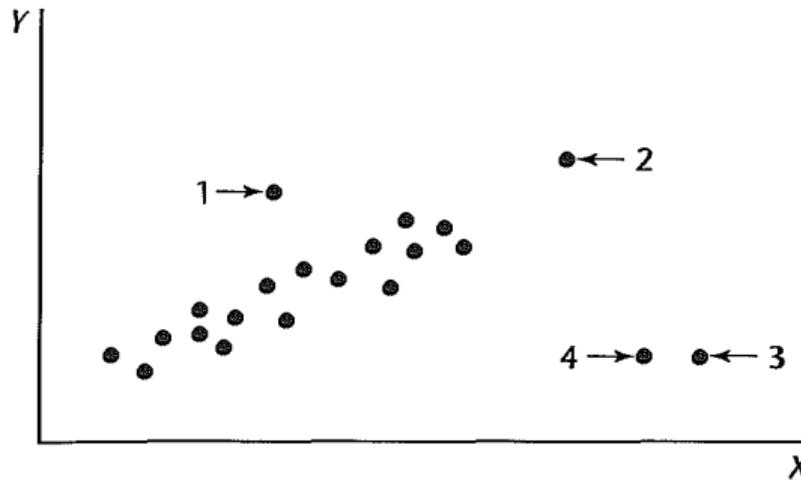
Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations



Case 1 Outlying in terms of Y .

Case 2 Outlying in terms of X .

Case 3 & 4 Outlying in terms of both X and Y , thus *influential*.



Studentized Residuals

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Residuals $e = (\mathbf{I} - \mathbf{H})\mathbf{y}$.
- Distribution of the residuals $e \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$.
- The standard error of e_i is $s.d.(e_i) = \sqrt{MSE(1 - h_{ii})}$.
- The studentized residual

$$r_i = \frac{e_i}{s.d.(e_i)} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}} \sim t(n - p).$$



Deleted Residuals

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Recall in PRESS, we define $d_i = Y_i - \hat{Y}_{i(i)}$ and its shortcut formula is

$$d_i = \frac{e_i}{1 - h_{ii}}.$$

- Recall if we have the rest of the $(n - 1)$ data with the i data point removed, then the standard error of the prediction at X_i is

$$s.d.(d_i) = \sqrt{MSE_{(i)}(1 + \mathbf{x}'_i(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{x}_i)}.$$

It can be shown that $(1 + \mathbf{x}'_i(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{x}_i) = (1 - h_{ii})^{-1}$ and

$$s.d.(d_i) = \sqrt{\frac{MSE_{(i)}}{1 - h_{ii}}}.$$



Studentized Deleted Residuals

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

■ Studentized deleted residuals

$$t_i = \frac{d_i}{s.d(d_i)} \sim t(n - p - 1).$$

- $(n - p)MSE = (n - p - 1)MSE_{(i)} + \frac{e_i^2}{1 - h_{ii}}$ Why? Try.
- Take $MSE_{(i)}$ obtained from the above equation to t_i and we have

$$t_i = e_i \left[\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2} \right]^{1/2}.$$

- Test for outliers in terms of Y : Bonferroni test procedure to test n studentized deleted residuals $|t_i|$ for $i = 1, \dots, n$, with critical value as $t(1 - \alpha/2n, n - p - 1)$.



Identifying Outlying X Observations

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

- Hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- Leverage h_{ii} for $i = 1, \dots, n$.
- We already knew that $\sum_{i=1}^n h_{ii} = \text{trace}(\mathbf{H}) = p$.
- h_{ii} is a distance measure of point \mathbf{x}_i to the centroid of all the inputs of the predictors $\bar{\mathbf{x}}$. It has nothing to do with the Y value.
- Rule of thumb: if h_{ii} is larger than twice of $\bar{h} = p/n$, then \mathbf{x}_i is outlying in terms of certain predictor variables.
- Interpret h_{ii} : a weight on Y_i . What happens if $h_{ii} \rightarrow 1$?



Example: Body Fat

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
body<-read.table("BodyFat.txt",header=FALSE)
colnames(body)<-c("X1","X2","X3","Y")
n<-nrow(body)
fit<-lm(Y~X1+X2,data=body)
summary(fit)
M <- model.matrix(fit)
H <- M%*%solve(t(M)%*%M)%*%t(M)
leverage<-diag(H)
t_value<-fit$residuals*sqrt((n-3-1)/
(sum(fit$residuals^2)*(1-leverage)-fit$residuals^2))

qt(1-0.1/2/20,df=16)
```



Influence on Single Fitted Value

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

■ DFFITS

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

■ A shortcut formula

$$(DFFITS)_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2} \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} = t_i \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2}.$$

■ Rule of thumb: influential if the $|DFFITS_i|$ exceeds 1 for small to medium data sets and $2\sqrt{p/n}$ for large data sets.



Influence on all fitted values

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

■ Cook's Distance

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \times MSE} = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p \times MSE}.$$

■ A shortcut formula

$$D_i = \frac{e_i^2}{p \times MSE} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

- D_i is related to distribution $F(p, n - p)$. If the percentile value of the corresponding D_i from the $F(p, n - p)$ distribution is near 50% or more, then the i case has a major influence on the fit of the regression function.



Influence on regression coefficients

DATA

564-494

Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

■ DFBETAS

$$(DFBETAS)_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}, \quad k = 0, 1, \dots, p - 1$$

where c_{kk} is the k th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

- Rule of thumb: a case is influential if the absolute value of DFBETAS exceeds 1 for small to medium data sets and $2/\sqrt{n}$ for large data sets.



Example: Body Fat Continued

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
DFFITS <- t_value*sqrt(leverage/(1-leverage))  
sort(abs(DFFITS))
```

```
CookD <- fit$residuals^2/(3*sum(fit$residuals^2)/fit$df)  
*(leverage)/(1-leverage)^2  
influence.measures(fit)
```



Example:Surgical Unit

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
surgical <- read.table("surgical.txt",header=F)
colnames(surgical) <- c("X1","X2","X3","X4","X5",
"X6","X7","X8","Y","lnY")
fit1<-lm(lnY~X1+X2+X3+X8,data=surgical)
plot(fit1$fitted.values,fit1$residuals,
xlab="Predicted Value",ylab="Residual")
plot(surgical$X5,fit1$residuals,xlab="X5",ylab="Residual")
fit_X5<-lm(X5~X1+X2+X3+X8,data=surgical)
plot(fit_X5$residuals,fit1$residuals,
xlab="e(X5|X1,X2,X3,X8)",ylab="e(lnY|X1,X2,X3,X8)")
abline(lm(fit1$residuals~fit_X5$residuals))
plot(fit1)
diagnostics<-influence.measures(fit1)
```



Example:Surgical Unit–Influential Plot

DATA
564-494
Simple Linear
Regression

Lulu Kang

Overview

Model
Comparison
Criteria

Automatic
Search
Procedures

Model
Validation

Added-
Predictor Plot

Outliers and
Influential
Observations

```
stud_resid<-fit1$residuals/sqrt(sum(fit1$residuals^2))
/fit1$df*(1-diagnostics$infmat[, "hat"])
plot(1:54,stud_resid,type='b')
plot(hatvalues(fit1),rstudent(fit1),ylim=c(-3,4),
type="n",main="Influential Plot for Surgical Unit Data",
xlab="Leverage",ylab="Studentized Residuals")
cook<-sqrt(cooks.distance(fit1))
points(hatvalues(fit1),rstudent(fit1),cex=10*cook/max(cook))
abline(v=mean(hatvalues(fit1))*2,lty=2)
abline(h=c(-2,0,2),lty=2)
identify(hatvalues(fit1),rstudent(fit1),row.names(surgical))
```



DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition

Splitting Algorithm
Node Impurities

Part VI

Advanced Topics in Regression

- 2.1 Weighted Least Squares
- 2.2 Ridge Regression
- 2.3 Robust Regression
- 2.4 Tree Regression



Weighted Least Squares

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm
Node Impurities

Recall the linear regression equation:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1}.$$

We have estimated the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1}$ by minimizing the sum of squared residuals.

$$\begin{aligned} SSE &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1})]^2. \end{aligned}$$



Weighted Least Squares

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition
Splitting Algorithm
Node Impurities

Sometimes we want to give some observations more weight than others. We achieve this by minimizing a *weighted* sum of squares:

$$\begin{aligned} WSSE &= \sum_{i=1}^n w_i (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n w_i [y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1})]^2. \end{aligned}$$

The resulting $\hat{\beta}$ are called *weighted least squares* (WLS) estimates, and the WLS residuals are

$$\sqrt{w_i} (y_i - \hat{y}_i).$$



Why use weights?

DATA
564-494
Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression
CART
Partition
Splitting Algorithm
Node Impurities

Suppose that the variance is not constant:

$$\text{var}(Y_i) = \sigma_i^2.$$

If we use weights

$$w_i \propto \frac{1}{\sigma_i^2},$$

the WLS estimates have smaller standard errors than the ordinary least squares (OLS) estimates.

That is, the OLS estimates are *inefficient*, relative to the WLS estimates.



Why use weights?

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition
Splitting Algorithm
Node Impurities

- In fact, using weights proportional to $1/\sigma_i^2$ is optimal: no other weights give smaller standard errors.
- When you specify weights, regression software calculates standard errors on the assumption that they are proportional to $1/\sigma_i^2$.



Weighted Least Square Estimates

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm
Node Impurities

It is straight forward to write the WLS version of the SSE in matrix format.

$$WSSE = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$ is a diagonal matrix of the weights. The WLS estimates of $\boldsymbol{\beta}$ is the optimal solution of minimizing WSSE, i.e.,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y}.$$

The prediction is

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y}.$$



How to choose the weights

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm

Node Impurities

- If you have many replicates for each unique combination of \mathbf{x}_i' , use s_i^2 to estimate $\sigma_i^2 = \text{var}(Y|\mathbf{x}_i)$.
- Often you will not have enough replicates to give good variance estimates.
 - The textbook suggests grouping observations that are “nearest neighbors”.
 - Alternatively you can use the regression diagnostic plots.



Example: diastolic blood pressure

DATA

564 observations

Simple Linear Regression

Lulu Kang

Weighted Least Squares

Ridge Regression

Robust Regression

Tree Regression

CART

Partition

Splitting Algorithm

Node Impurities

```
bpdata<-read.table("diastolic.txt",header=FALSE)
colnames(bpdata)<-c("age","bp")
fit<-lm(bp~age,data=bpdata)
summary(fit)
plot(fit)

absresid <- abs(fit$residuals)
fit.s<-lm(absresid~bpdata$age)
summary(fit.s)

w<-1/(fit.s$fitted.values)^2
fit.weighted <- lm(bp~age,data=bpdata,weights=w)
summary(fit.weighted)
plot(fit.weighted)
```



Note

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition
Splitting Algorithm
Node Impurities

- When you specify weights w_i , `lm` function fits the model

$$\sigma_i^2 = \sigma^2 / w_i$$

and the “residual standard error” s^2 is an estimate of σ^2 :

$$s^2 = \frac{\sum_{i=1}^n w_i(y_i - \hat{y}_i)^2}{n - p}.$$

- If you change the weights, the meaning of σ (and s) changes.
- You cannot compare the residual standard errors for different weighting schemes.



Definition

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm
Node Impurities

Ridge regression shrinks the regression coefficients by imposing a penalty on their size.

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^{p-1} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}.$$

equivalent to:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^{p-1} x_{ij} \beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p-1} \beta_j^2 \leq t$$



Computation

DATA

564,494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition

Splitting Algorithm
Node Impurities

- $\hat{\beta}^{\text{ridge}}$ are sensitive to the scales.
- After centering y and \mathbf{X} :

$$\text{SSE}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\beta$$

\mathbf{X} doesn't have the **1** column.

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

- If \mathbf{X} has orthogonal columns, $\hat{\beta}^{\text{ridge}} = \hat{\beta}^{ls}/(1 + \lambda)$.



Explanation: SVD of X

MATH

564-484

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

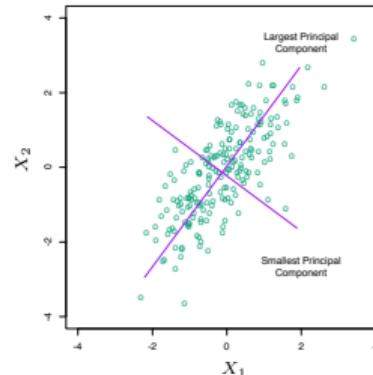
CART

Partition

Splitting Algorithm

Node Impurities

- 1 Singular value decomposition: $\mathbf{X} = \mathbf{UDV}'$. \mathbf{U} and \mathbf{V} are $N \times (p - 1)$ and $(p - 1) \times (p - 1)$ orthogonal matrices. The columns of \mathbf{U} span the column space of \mathbf{X} , and the columns of \mathbf{V} span the row space.
 $\mathbf{D} = \text{diag}\{d_1, d_2, \dots, d_{p-1}\}$, $d_1 \geq d_2 \geq \dots \geq d_{p-1} \geq 0$.
- 2 Relationship with Principle components: $\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$. Meaning, project \mathbf{X} into the equivalent space with orthogonal basis whose variations are decreasing.





Explanation: shrinkage coefficients

DATA
564-494
Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression
CART
Partition
Splitting Algorithm
Node Impurities

$$\begin{aligned}\mathbf{X} \hat{\boldsymbol{\beta}}^{ls} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{U}\mathbf{U}'\mathbf{y} \\ &= \sum_{j=1}^{p-1} \mathbf{u}_j \mathbf{u}_j' \mathbf{y}\end{aligned}$$

$$\begin{aligned}\mathbf{X} \hat{\boldsymbol{\beta}}^{\text{ridge}} &= \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda \mathbf{I})^{-1}\mathbf{D}\mathbf{U}'\mathbf{y} \\ &= \sum_{j=1}^{p-1} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j' \mathbf{y}\end{aligned}$$

Shrinkage $\frac{d_j^2}{d_j^2 + \lambda}$. The smaller the d_j is, the more the shrinkage is. d_j indicates the variation of the direction \mathbf{u}_j . Thus \mathbf{u}_j with smaller variation will have smaller β_j^{ridge} , and less significant. Underlying assumption: the response tend to vary most in the directions of high variance of the inputs.



Effective degree of freedom

DATA

564/494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm

Node Impurities

- Least square regression:

$$\text{df} = \text{tr}(\mathbf{H}) = p - 1.$$

- Ridge regression:

$$\text{df}(\lambda) = \text{tr}(\mathbf{X}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}')$$

$$= \text{tr}(\mathbf{H}_\lambda) = \sum_{j=1}^{p-1} \frac{d_j^2}{d_j^2 + \lambda}$$

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1} = (\mathbf{V}\mathbf{D}^2\mathbf{V}' + \lambda\mathbf{I})^{-1} = (\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})\mathbf{V}')^{-1} = \mathbf{V}'(\mathbf{D}^{-2} + \lambda^{-1}\mathbf{I}^{-1})\mathbf{V}.$$



Example

DATA
564-494
Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

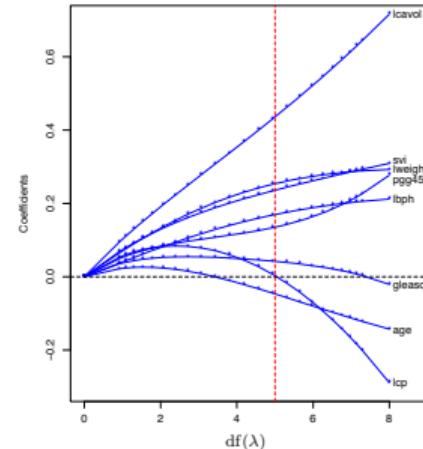
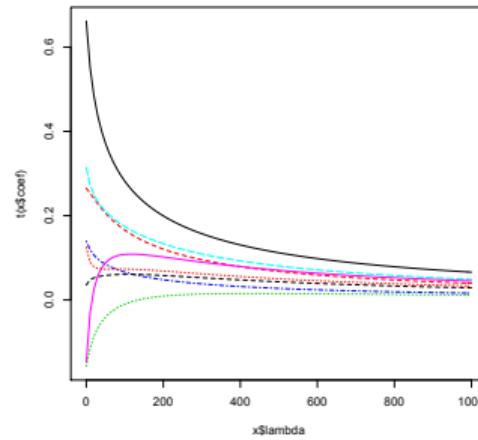
CART

Partition

Splitting Algorithm

Node Impurities

```
library(MASS)
prostate<-read.table('prostate.txt',header=TRUE,sep=',')
fit.ridge<-lm.ridge(lpsa~lcavol+lweight+age+lbph+svi
+lcp+gleason+pgg45, data=prostate, lambda=seq(0,1000,10))
plot(fit.ridge)
```





Robust Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition

Splitting Algorithm
Node Impurities

The goal of doing robust regression is trying to reduce the influence of the potential outliers to the regression. There are several classic robust regression approaches.

- Least absolute residuals (LAR) or least absolute deviations (LAD).

$$\min_{\beta} \sum_{i=1}^n |Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_{p-1} X_{i,p-1}|.$$

- Iteratively reweighted least squares (IRLS)
- Least median of squares (LMS)

$$\min_{\beta} \text{median}_{i=1,\dots,n} \{(Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_{p-1} X_{i,p-1})^2\}.$$



CART-Classification and Regression Trees

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression
CART

Partition
Splitting Algorithm
Node Impurities

Trees can be viewed as basis expansions of simple functions

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m)$$

with $R_1, R_2, \dots, R_m \subset \mathbb{R}^p$ disjoint. The CART algorithm is a heuristic, adaptive algorithm for basis function selection. A recursive, binary partition (a tree) is given by a list of splits

$$\{(t_{01}), (t_{11}, t_{12}), (t_{21}, t_{22}, t_{23}, t_{24}), \dots, (t_{n1}, \dots, t_{n2^n})\}$$

and corresponding split variable indices

$$\{(i_{01}), (i_{11}, i_{12}), (i_{21}, i_{22}, i_{23}, i_{24}), \dots, (i_{n1}, \dots, i_{n2^n})\}$$

$$R_1 = (x_{i_{01}} < t_{01}) \cap (x_{i_{11}} < t_{11}) \cap \dots \cap (x_{i_{n1}} < t_{n1})$$

and we can determine if $x \in R_1$ in n steps $\ll M = 2^n$.



Recursive Binary Partition

DATA

564,494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

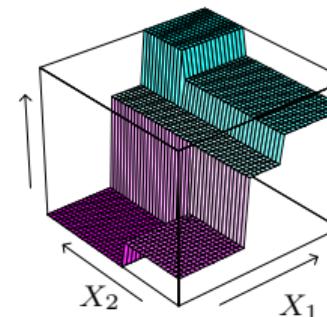
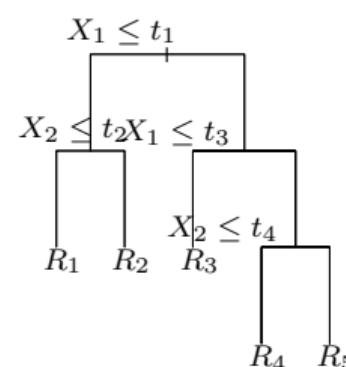
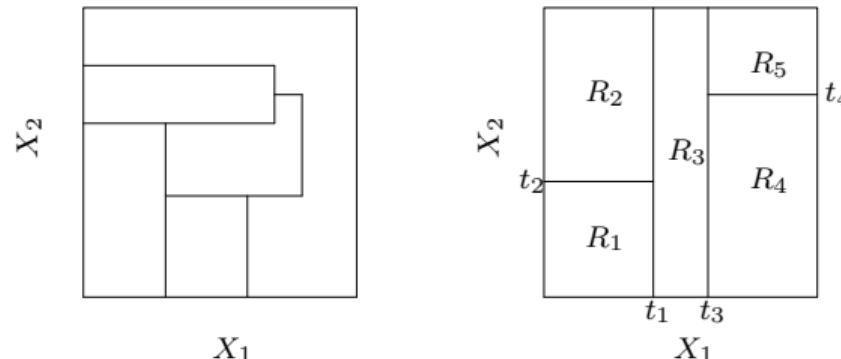
Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm
Node Impurities





Binary Partition

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm
Node Impurities

Recursive Binary Partition

- The recursive partition of $[0, 1]^2$ above and the representation of the partition by tree.
- A binary tree of depth n can represent up to 2^n partitions/basis functions.
- We can determine which R_j and x belongs to by n recursive yes/no questions.

General Partitions

- A general partition can not be represented as binary splits.
- With M sets in a general partition we would in general need of the order M yes/no questions to determine which of the sets an x belongs to.



Recursive Binary Partitions

DATA

564,494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition

Splitting Algorithm
Node Impurities

For a fixed partition R_1, \dots, R_M the least squares estimates are

$$\hat{c}_m = \bar{y}(R_m) = \frac{1}{N_m} \sum_{i:x_i \in R_m} y_i$$

$N_m = |\{i|x_i \in R_m\}|$. The recursive partition allows for rapid computation of the estimates and rapid prediction of new observations.



Greedy Splitting Algorithm

DATA

564,494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm
Node Impurities

With squared error loss and an unknown partition R_1, \dots, R_m we would seek to minimize

$$\sum_{i=1}^N (y_i - \bar{y}(R_{m(i)}))^2$$

over the possible binary, recursive partitions. But this is computationally difficult.

An optimal single split on a region R is determined by

$$\min_j \min_s \left(\sum_{i:x_i \in R_{(j,s)}} (y_i - \bar{y}(R(j,s)))^2 + \sum_{i:x_i \in R(j,s)^C} (y_i - \bar{y}(R(j,s)^C))^2 \right)$$

with $R(j,s) = \{x \in R | x_j < s\}$. The tree growing algorithm recursively does single, optimal splits on each of the partitions obtained in the previous step.



Tree Pruning

DATA

564 / 494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm
Node Impurities

The full binary tree, T_0 , representing the partitions R_1, \dots, R_M with $M = 2^n$ may be too large. We prune it by snipping of leafs or subtrees.

For any subtree T of T_0 with $|T|$ leafs and partition $R_1(T), \dots, R_{|T|}(T)$ the cost-complexity of T is

$$C_\alpha = \sum_{i=1}^N (y_i - \bar{y}(R_{m(i)}(T)))^2 + \alpha|T|$$

Theorem

There is a finite set of subtrees $T_0 \supseteq T_{\alpha_1} \supset T_{\alpha_2} \dots \supset T_{\alpha_r}$ with $0 \leq \alpha_1 < \alpha_2 < \dots < \alpha_r$ such that T_{α_i} minimizes $C_\alpha(T)$ for $\alpha \in [\alpha_i, \alpha_{i+1})$.



Node Impurities and Classification Trees

DATA

564-494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition

Splitting Algorithm
Node Impurities

Define the node impurity as the average loss for the node R

$$Q(R) = \frac{1}{N(R)} \sum_{i:x_i \in R} (y_i - \bar{y}(R))^2.$$

The greedy split of R is found by

$$\min_j \min_s (N(R(j, s))Q(R(j, s)) + N(R(j, s)^C)Q(R(j, s)^C))$$

with $R(j, s) = \{x \in R | x_j < s\}$ and we have

$$C_\alpha(T) = \sum_{m=1}^{|T|} N(R_m(T))Q(R_m(T)) + \alpha|T|.$$

If Y takes K discrete values we focus on the node estimate for $R_m(T)$ in tree T as being

$$\hat{p}_m(T)(k) = \frac{1}{N_m} \sum_{i:x_i \in R_m(T)} I(y_i = k).$$



Node Impurities and Classification Trees

DATA
564-494
Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART
Partition
Splitting Algorithm
Node Impurities

The loss functions for classification enter in the specification of the node impurities used for splitting and cost-complexity computations.

Examples of

- 0-1 loss gives misclassification error impurity:

$$Q(R_m(T)) = 1 - \max\{\hat{p}(R_m(T))(1), \dots, \hat{p}(R_m(T))(K)\}$$

- likelihood loss gives entropy impurity:

$$Q(R_m(T)) = - \sum_{k=1}^K \hat{p}(R_m(T))(k) \log \hat{p}(R_m(T))(k)$$

- The Gini index impurity:

$$Q(R_m(T)) = \sum_{k=1}^K \hat{p}(R_m(T))(k)(1 - \hat{p}(R_m(T))(k))$$



Node Impurities

DATA
564,494
Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

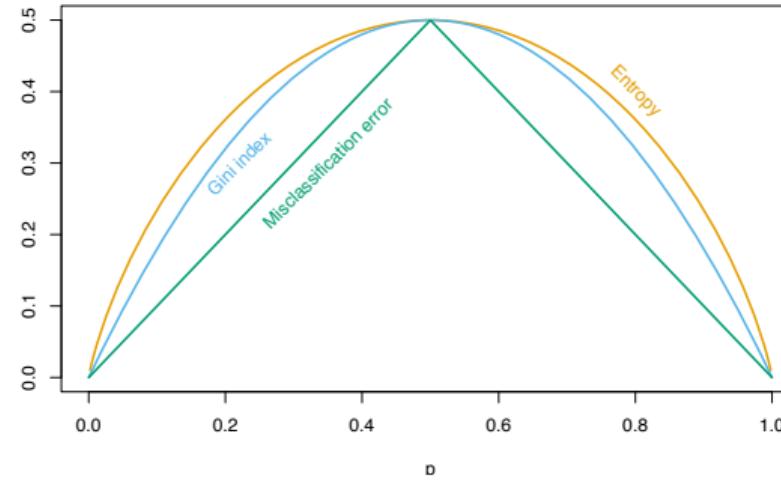
Robust
Regression

Tree
Regression

CART
Partition

Splitting Algorithm

Node Impurities





R code

DATA

564 494

Simple Linear
Regression

Lulu Kang

Weighted
Least Squares

Ridge
Regression

Robust
Regression

Tree
Regression

CART

Partition

Splitting Algorithm

Node Impurities

```
require(rpart) #the function "rpart" implements the estimation of trees
require(MASS) #For the iris data example
require(mboost) #The implementation chosen for boosting

data(iris)

##Setting cp, the complexity parameter,
irisTree <- rpart(Species~.,data=iris,cp=0.001)
plot(irisTree,margin=0.1)
text(irisTree,use.n=TRUE)
printcp(irisTree)

###Note that for classification, mboost relies on the encoding of
y as +1 and -1
iris.tree <- rpart(Species~.,data=iris)
```



DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Part VII

Generalized Linear Models

2. Logistic regression: model: model, estimation, interpretation, and the diagnostics.
2. Polytomous logistic regression.xs
2. Poisson regression.
2. Generalized linear model.



Regression Models with Binary Response

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous

Logistic

Regression

Poisson

Regression

GLM

- Consider the simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{where } Y_i = 0, \text{ or } 1.$$

The expected response $E(Y_i)$ has a special meaning in this case. Since $E(\epsilon_i) = 0$, we have

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

- Consider Y_i to be a Bernoulli random variable for which we can state the probability distribution as follows.

$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i.$$



Regression Models with Binary Response

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic

Regression

Poisson
Regression

GLM

- By the definition of expected value

$$E(Y_i) = 1 \times \pi_i + 0 \times (1 - \pi_i) = P(Y_i = 1).$$

- Following the regression model, we obtain,

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

- Non-normal error term: the errors for binary response is not normal

$$\text{When } Y_i = 1 : \epsilon_i = 1 - \beta_0 - \beta_1 X_i$$

$$\text{When } Y_i = 0 : \epsilon_i = 0 - \beta_0 - \beta_1 X_i.$$



Regression Models with Binary Response

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Non-constant Error Variance:

$$\sigma^2(Y_i) = E[(Y_i - E(Y_i))^2] = (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i) = \pi_i(1 - \pi_i).$$

Under the linear model,

$$\sigma^2(\epsilon_i) = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i).$$

- Constraints on response function, $0 \leq E(Y_i) \leq 1$.
- So the linear model is not applicable for binary response.



Sigmoidal Response Functions for Binary Responses: Probit Mean Response Function

DATA

564-490

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- For the binary response Y_i , and it is ordinal, we can imagine there is another hidden continuous variable, Y_i^c and

$$Y_i^c = \beta_0^c + \beta_1^c X_i + \epsilon_i^c,$$

where ϵ_i^c is normally distributed with mean 0 and variance σ_c^2 .

- Since Y_i is ordinal, we can consider it is changed from 0 to 1 when Y_i^c hits a threshold value a .

$$Y_i = \begin{cases} 1, & \text{if } Y_i^c \leq a \\ 0, & \text{if } Y_i^c > a. \end{cases}$$



Sigmoidal Response Functions for Binary Responses: Probit Mean Response Function

DATA

564-490

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous

Logistic

Regression

Poisson

Regression

GLM

$$\begin{aligned} P(Y_i = 1) &= \pi_i = P(Y_i^c \leq a) \\ &= P(\beta_0^c + \beta_1^c X_i + \epsilon_i^c \leq a) \\ &= P(\epsilon_i^c \leq a - \beta_0^c - \beta_1^c X_i) \\ &= P\left(\frac{\epsilon_i^c}{\sigma_c} \leq \frac{a - \beta_0^c}{\sigma_c} - \frac{\beta_1^c}{\sigma_c}\right) \\ &= P(Z \leq \beta_0^* + \beta_1^* X_i) \end{aligned}$$

So we define the *probit mean response function*:

$$E(Y_i) = \pi_i = \Phi(\beta_0^* + \beta_1^* X_i).$$



Sigmoidal Response Functions for Binary Responses: Probit Mean Response Function

DATA

564-490

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Applying the probit transformation, Φ^{-1} inverse of the standard normal CDF, we obtain the linear predictor

$$\Phi^{-1}(\pi_i) = \beta_0^* + \beta_1^* X_i.$$

- Symmetry property: for $Y'_i = 1 - Y_i$,

$$P(Y'_i = 1) = P(Y_i = 0) = 1 - \Phi(\beta_0^* + \beta_1^* X_i) = \Phi(-\beta_0^* - \beta_1^* X_i).$$



Sigmoidal Response Functions for Binary Responses: Logistic Mean Response Function

DATA

564-490

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Define the logistic distribution for a random variable ϵ_L with pdf and cdf

$$f_L(\epsilon_L) = \frac{\exp(\epsilon_L)}{[1 + \exp(\epsilon_L)]^2}, \quad F_L(\epsilon_L) = \frac{\exp(\epsilon_L)}{1 + \exp(\epsilon_L)}.$$

The mean $E(\epsilon_L) = 0$ and $s.d(\epsilon_L) = \pi/\sqrt{3}$.

- Now if ϵ_i^c in the model $Y_i^c = \beta_0^c + \beta_1^c X_i + \epsilon_i^c$ has a mean 0 and standard deviation σ_c , then

$$P(Y_i = 1) = P\left(\frac{\epsilon_i^c}{\sigma_c} \leq \beta_0^* + \beta_1^* X_i\right)$$



Sigmoidal Response Functions for Binary Responses: Logistic Mean Response Function

DATA

564-490

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- If instead of a normal distribution, we assume ϵ_i^c follows the logistic distribution with mean 0 and deviation $\pi/\sqrt{3}$ (like the random variable ϵ_L), then we need to scale ϵ_i^c ,

$$\begin{aligned} P(Y_i = 1) &= \pi_i = P\left(\frac{\pi}{\sqrt{3}}\frac{\epsilon_i^c}{\sigma_c} \leq \frac{\pi}{\sqrt{3}}\beta_0^* + \frac{\pi}{\sqrt{3}}\beta_1^*X_i\right) \\ &= P(\epsilon_L \leq \beta_0 + \beta_1 X_i) \\ &= FL(\beta_0 + \beta_1 X_i) \\ &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}. \end{aligned}$$



Sigmoidal Response Functions for Binary Responses: Logistic Mean Response Function

DATA

564-490

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- The Logistic mean response function is

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)},$$

which is equal to $\pi = (1 + \exp(-\beta_0 - \beta_1 X_i))^{-1}$.

- Apply for the inverse of F_L^{-1} , or the logit transformation of the probability π_i , and obtain the linear predictor of the logit response function

$$F_L^{-1}(\pi_i) = \log_e \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i.$$



Simple Logistic Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Simple logistic regression assumption:

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)},$$

where Y_i follows a Bernoulli distribution with $E(Y_i) = P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$.



The likelihood function

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

For the data (X_i, Y_i) for $i = 1, \dots, n$, we can write the likelihood function

$$l(Y_1, \dots, Y_n; \beta_0, \beta_1) = \prod_{i=1}^n f_i(Y_i) = \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i}.$$

The log transformation of l is

$$\begin{aligned}\log l(Y_1, \dots, Y_n; \beta_0, \beta_1) &= \log \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \\ &= \sum_{i=1}^n [Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i)] \\ &= \sum_{i=1}^n \left[Y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) \right] + \sum_{i=1}^n \log(1 - \pi_i)\end{aligned}$$



Maximum likelihood estimation

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic

Regression

Poisson
Regression

GLM

Take $\log \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1 X_i$ in the log-likelihood (Y_1, \dots, Y_n are observed data, considered as constant in log-likelihood. So $\log l$ is only a function of β_0, β_1).

$$\log l(\beta_0, \beta_1) = \sum_{i=1}^n Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^n \log [1 + \exp(\beta_0 + \beta_1 X_i)].$$

Maximize the $\log l(\beta_0, \beta_1)$ to estimate β_0 and β_1 .

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \max_{\beta_0, \beta_1} \sum_{i=1}^n Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^n \log [1 + \exp(\beta_0 + \beta_1 X_i)]$$



Optimization Method

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The maximum likelihood estimates are obtained by solving the score equations:

$$s(\beta_0) = \frac{\partial \log l}{\partial \beta_0} = \sum_{i=1}^n \left\{ Y_i - \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \right\} = \sum_{i=1}^n (Y_i - \pi_i) = 0,$$

$$s(\beta_1) = \frac{\partial \log l}{\partial \beta_1} = \sum_{i=1}^n \left\{ Y_i X_i - \frac{\exp(\beta_0 + \beta_1 X_i) X_i}{1 + \exp(\beta_0 + \beta_1 X_i)} \right\} = \sum_{i=1}^n (Y_i X_i - \pi_i X_i) = 0.$$



Optimization Method

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

A general method of solving score equations is the iterative algorithm *Fisher's Method of Scoring* (derived from a Taylor's expansion of $s(\beta)$). It is a version of Newton's method used in statistics to solve maximum likelihood equations.

In the r -th iteration, the new estimate $\beta^{(r+1)}$ is obtained from the previous estimate $\beta^{(r)}$ by

$$\beta^{(r+1)} = \beta^{(r)} + E \left(H(\beta^{(r)}) \right)^{-1} s(\beta^{(r)}),$$

where H is the *Hessian matrix*: the matrix of the second derivatives of the log-likelihood.



Optimization Method

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression
Probit and Logit
Transformation
Logistic Regression
Binomial Distribution
Inference on MLE
Polynomial Logistic
Regression
Variable Selection
Goodness-of-fit
Logistic Regression
Diagnostics
Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

For the log-likelihood of the logistic model,

$$\begin{aligned} H &= \begin{bmatrix} \frac{\partial^2 \log l}{\partial^2 \beta_0}, & \frac{\partial^2 \log l}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \log l}{\partial \beta_1 \partial \beta_0}, & \frac{\partial^2 \log l}{\partial^2 \beta_1} \end{bmatrix} = \begin{bmatrix} \frac{\partial s(\beta_0)}{\partial \beta_0}, & \frac{\partial s(\beta_0)}{\partial \beta_1} \\ \frac{\partial s(\beta_1)}{\partial \beta_0}, & \frac{\partial s(\beta_1)}{\partial \beta_1} \end{bmatrix} \\ &= \begin{bmatrix} -\sum_{i=1}^n \pi_i(1-\pi_i), & -\sum_{i=1}^n \pi_i(1-\pi_i)X_i \\ -\sum_{i=1}^n \pi_i(1-\pi_i)X_i, & -\sum_{i=1}^n \pi_i(1-\pi_i)X_i^2 \end{bmatrix} = -\mathbf{X}' \mathbf{W} \mathbf{X}. \end{aligned}$$

Here $\mathbf{W} = \text{diag}\{\pi_1(1-\pi_1), \dots, \pi_n(1-\pi_n)\}$ and \mathbf{X} is the model matrix $\mathbf{X} = [\mathbf{1}, \mathbf{X}_1]$.



Optimization Method

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

It turns out that the updates can be written as

$$\boldsymbol{\beta}^{(r+1)} = (\mathbf{X}' \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{(r)} \mathbf{z}^{(r)},$$

i.e., the score equations for a weighted least squares regression of $\mathbf{z}^{(r)}$ on \mathbf{X} with weights $\mathbf{W}^{(r)}$, where

$$z_i^{(r)} = \mathbf{x}_i' \boldsymbol{\beta}^{(r)} + \frac{(Y_i - \pi_i^{(r)})}{\pi_i(1 - \pi_i)},$$

for $i = 1, \dots, n$.



Optimization Method

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Hence the estimates can be found using an *iterative re-weighted least squares* algorithm:

- 1 Start with initial estimates of $\beta^{(0)}$.
- 2 Calculate *working responses* $z^{(r)}$ and working weights $W^{(r)}$.
- 3 Calculate $\beta^{(r+1)}$ by weighted least squares.
- 4 Repeat Step 2 and 3 until convergence.



The `glm` function

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Generalized linear models can be fitted in R using the `glm` function, which is similar to the `lm` function for fitting linear models. The arguments to a `glm` call are as follows.

```
glm(formula, family = gaussian, data, weights, subset,  
na.action, start = NULL, etastart, mustart, offset,  
control = glm.control(...), model = TRUE,  
method = "glm.fit", x = FALSE, y = TRUE,  
contrasts = NULL, ...)
```



Formula Argument

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The formula is specified to `glm` as, e.g., `y~x1+x2`

where `x1,x2` are the names of

- numeric vectors (continuous variables)
- factors (categorical variables)

All specified variables must be in the workspace or in the data frame passed to the `data` argument.



Formula Argument

DATA

564-494

Simple Linear

Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Other symbols that can be used in the formula include

- $a:b$ for an interaction between a and b
- $a*b$ which expands to $a+b+a:b$
- $.$ for first order terms of all variables in data
- $-$ to exclude a term or terms
- 1 to include an intercept (include by default)
- 0 to exclude an intercept



Family Argument

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The family argument takes (the name of) a family function which specifies

- the link function
- the variance function
- various related objects used by `glm`, e.g. `linkinv`

The exponential family function available in R are

- `binomial(link="logit")` or `binomial(link="probit")`
- `gaussian(link="identity")`
- `Gamma(link="inverse")`
- `inverse.gaussian(link="1/mu^2")`
- `poisson(link="log")`



Extractor Functions

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The `glm` function returns an object of class `c("glm", "lm")`. There are several `glm` or `lm` methods available for accessing/display components of the `glm` object, including:

- `residuals()`
- `fitted()`
- `predict()`
- `coef()`
- `deviance()`
- `formula()`
- `summary()`



Example: Programming Test

DATA

564-490

Simple Linear

Regression

Lulu Kang

Logistic

Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous

Logistic

Regression

Poisson

Regression

GLM

Example

- Y : whether an analyst can finish complex programming task in allocated time ($Y = 1$ yes, $Y = 0$ no).
- X : months of experiences.

```
program<-read.table("program.txt",header=FALSE)
colnames(program)<-c("X","Y","fitted")
fit<-glm(Y~X,data=program,family=binomial(link="logit"))
summary(fit)
```



Estimation and Inference

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

■ Estimated probability

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)}$$

- Interpretation of $\hat{\beta}_1$: log of the odds ratio.
If increasing X by 1 unit,

$$\text{logit}(\hat{\pi}(X)) = \log(\text{odds}_1) = \log \frac{\hat{\pi}(X)}{1 - \hat{\pi}(X)} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\text{logit}(\hat{\pi}(X + 1)) = \log(\text{odds}_2) = \log \frac{\hat{\pi}(X + 1)}{1 - \hat{\pi}(X + 1)} = \hat{\beta}_0 + \hat{\beta}_1(X + 1)$$

$$\hat{\beta}_1 = \text{logit}(\hat{\pi}(X + 1)) - \text{logit}(\hat{\pi}(X)) = \log \frac{\text{odds}_2}{\text{odds}_1}$$

$$\exp(\hat{\beta}_1) = \frac{\text{odds}_2}{\text{odds}_1}$$



Example: Programming Test

DATA

564 494

Simple Linear

Regression

Lulu Kang

Logistic

Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous

Logistic

Regression

Poisson

Regression

GLM

```
pred<-predict(fit,newdata=data.frame(X=4:32),type='response')
plot(4:32,pred,xlim=c(3,33),ylim=c(0,1),xlab="Months of Experiences",
ylab="Probability",type="l")
points(program$X,program$Y)
points(program$X,fit$fitted.values,pch=16)
```



Multiple Logistic Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Just like linear regression, we can extend the simple logistic regression to multiple logistic regression to include more explanatory variables in the model.

$$\mathbf{x}_i' \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1}$$

Define the matrix notation:

$$\boldsymbol{\beta}_{p \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \mathbf{X}_{p \times n} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ 1 & x_{21} & \dots & x_{2,p-1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{bmatrix} \quad \mathbf{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{i,p-1} \end{bmatrix}$$



Multiple Logistic Regression

DATA

564 / 494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Logistic model assumption

$$E(Y_i) = \pi_i = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})}$$

or equivalently

$$\mathbf{x}'_i \boldsymbol{\beta} = \log \frac{\pi_i}{1 - \pi_i}.$$

The log-likelihood is

$$\log l(Y_1, \dots, Y_n) = \sum_{i=1}^n Y_i \mathbf{x}'_i \boldsymbol{\beta} - \sum_{i=1}^n \log[1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})]$$



Multiple Logistic Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The score function

$$\mathbf{s}(\boldsymbol{\beta}) = \frac{\partial \log l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n (Y_i - \pi_i) \mathbf{x}_i.$$

The Hessian matrix

$$\mathbf{H} = \frac{\partial^2 \log l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \frac{\partial \mathbf{s}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = -\mathbf{X}^\top \mathbf{W} \mathbf{X},$$

where \mathbf{X} is defined in the same way as before. Here the weight matrix $\mathbf{W} = \text{diag}\{\pi_1(1 - \pi_1), \dots, \pi_n(1 - \pi_n)\}$.



Logistic distribution for binomial outcomes

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Repeated measures: Y_{ij} for $i = 1, \dots, n_j$ for $j = 1, \dots, c$. For each $\mathbf{x} = \mathbf{x}_j$, there were n_j repeated binary observations of Y_{ij} .
- $Y_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}$ follows a binomial distribution $\text{Binomial}(n_j, \pi(\mathbf{x}_j))$.
- The p.m.f. for the binomial random variable is

$$f(Y_{\cdot j}) = \frac{n_j!}{(Y_{\cdot j})!(n_j - Y_{\cdot j})!} \pi(\mathbf{x}_j)^{Y_{\cdot j}} (1 - \pi(\mathbf{x}_j))^{n_j - Y_{\cdot j}}.$$

- Consider the logistic model

$$\frac{E(Y_{\cdot j})}{n_j} = \pi(\mathbf{x}_j) = \frac{\exp(\mathbf{x}'_j \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_j \boldsymbol{\beta})}, \quad \text{or equivalently } \mathbf{x}'_j \boldsymbol{\beta} = \log \frac{\pi_j}{1 - \pi_j}$$



Logistic distribution for binomial outcomes

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The joint log likelihood for all the data $(\mathbf{x}_j, Y_{.j})$ for $j = 1, \dots, c$ is

$$\log l(\boldsymbol{\beta}) = \sum_{j=1}^c \left\{ \log \binom{n_j}{Y_{.j}} + Y_{.j} \mathbf{x}'_j \boldsymbol{\beta} - n_j \log [1 + \exp(\mathbf{x}'_j \boldsymbol{\beta})] \right\}$$

So the Bernoulli distribution is a special case of the binomial distribution with $n_j = 1$ for all j .

■ Score function:

$$\mathbf{s}(\boldsymbol{\beta}) = \sum_{j=1}^c (Y_{.j} - n_j \pi_j) \mathbf{x}_j.$$

■ Hessian matrix

$$\mathbf{H} = -\mathbf{X}' \mathbf{W} \mathbf{X}.$$

Here the weight matrix is

$$\mathbf{W} = \text{diag}\{n_1 \pi_1 (1 - \pi_1), n_2 \pi_2 (1 - \pi_2), \dots, n_c \pi_c (1 - \pi_c)\}.$$



Wald Test

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

For non-normal data, we can use the fact that asymptotically

$$\hat{\beta} \sim N(\beta, (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}).$$

Here $\mathbf{W} = \text{diag}\{n_1\pi_1(1 - \pi_1), n_2\pi_2(1 - \pi_2), \dots, n_c\pi_c(1 - \pi_c)\}$. If $n_i = 1$, the model is for binary data; if $n_i \geq 1$, the model is for Binomial data. In our textbook, the Hessian matrix H is denoted as \mathbf{G} , and

$$\text{cov}(\hat{\beta}) = (-\mathbf{G}(\beta))^{-1}.$$

Since \mathbf{W} depends on β , estimated $\text{cov}(\hat{\beta})$ is

$$\hat{\text{cov}}(\hat{\beta}) = (\mathbf{X}' \mathbf{W}(\hat{\beta}) \mathbf{X})^{-1}.$$



Wald Test

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Specifically, we test

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

using the test statistic

$$z_j^* = \frac{\hat{\beta}_j}{\sqrt{[(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]_{jj}}}$$

which is asymptotically $N(0, 1)$ under H_0 .

- If $|z_j^*| \leq z_{\alpha/2}$, conclude H_0 .
- If $|z_j^*| > z_{\alpha/2}$, conclude H_1 .

100(1 - α)% C.I. is $\hat{\beta}_j \pm z_{\alpha/2} \sqrt{[(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]_{jj}}$.



Likelihood Ratio Test

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The goal is test the hypothesis

$$H_0 : \beta_q = \beta_{q+1} = \dots = \beta_{p-1} = 0$$

H_a :not all of the β_k in H_0 equal zeros.

about two models

Full model: $\log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \mathbf{x}'\boldsymbol{\beta}_F = \beta_0 + X_1\beta_1 + \dots + X_{p-1}\beta_{p-1}$

Reduced model: $\log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \mathbf{x}'\boldsymbol{\beta}_R = \beta_0 + X_1\beta_1 + \dots + X_{q-1}\beta_{q-1}$.



Likelihood Ratio Test

DATA

564 / 494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Let $L(R)$ and $L(F)$ be the maximal likelihood function value for the reduced model and full model respectively. The likelihood ratio is

$$G^2 = -2 \log \frac{L(R)}{L(F)} = -2[\log L(R) - \log L(F)].$$

It can be shown that it follows a χ^2_{p-q} distribution.

- If $G^2 \leq \chi^2(1 - \alpha, p - q)$, conclude H_0 .
- If $G^2 > \chi^2(1 - \alpha, p - q)$, conclude H_a .



Outbreak

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Example

- Y : 1 if the disease was determined to have been present; 0 if not.
- X_1 : age of the individual;
- Socioeconomic status (upper, middle, lower): $X_2 = 1$ for middle class; and 0 other wise; $X_3 = 1$ for lower class; and 0 otherwise.
- City sector: $X_4 = 1$ for one sector; $X_4 = 0$ for another sector.



Example: Outbreak

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
outbreak<-read.table("Outbreak.txt",header=FALSE)
colnames(outbreak)<-c("Case","X1","X2","X3","X4","Y")
fit.full <- glm(Y~X1+X2+X3+X4,data=outbreak,family=binomial("logit"))
summary(fit.full)
fit.reduced <- glm(Y~X2+X3+X4,data=outbreak,family=binomial("logit"))
summary(fit.reduced)
anova(fit.reduced,fit.full)
```



Polynomial Logistic Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Just like the linear regression, we can use more sophisticated functions, such as polynomial functions of the predictors, in the regression model. For example, for logistic regression model

$$\log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \beta_0 + f_1(\mathbf{x})\beta_1 + f_2(\mathbf{x})\beta_2 + \dots + f_k(\mathbf{x})\beta_k.$$

Here $f_i(\mathbf{x})$ is some known function of the input values \mathbf{x} .

- Polynomial logistic regression:

$$\begin{aligned} \log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = & \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \beta_{12} X_1 X_2 + \dots + \beta_{p-1,p} X_{p-1} X_p \\ & + \beta_{11} X_1^2 + \dots + \beta_{pp} X_{pp}^2 + \dots \end{aligned}$$



Example: public IPOs

DATA

564-499

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Example

- Y : VC involvement in the IPO
- X : the log value of the face value of the company

```
IPO<-read.table('IPO.txt',header=FALSE)
colnames(IPO)<-c("ID","VC","Facevalue","Shares","Buyout")
X<-log(IPO$Facevalue)
fit<-glm(IPO$VC~X,family=binomial("logit"))
fit2<-glm(IPO$VC~X+I(X^2),family=binomial("logit"))
anova(fit1,fit2)
```



Example: public IPOs

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
pred<-predict(fit2,  
newdata=data.frame(X=seq(min(X),max(X),length=100)),type='response')  
plot(seq(min(X),max(X),length=100),pred,xlim=c(14,20),ylim=c(0,1),  
xlab="Log of Face Value", ylab="Probability",type="l")  
points(X,IPO$VC)
```



Criteria

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

AIC and BIC criteria can still be used here.

- $AIC_p = -2 \log L(\hat{\beta}) + 2p$
- $BIC_p = -2 \log L(\hat{\beta}) + p \log(n)$



Automatic Selection

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Best subsets procedures.
- Stepwise model selection: forward, backward, both

R function `step`, used in the same way as for `lm` objects.



Pearson Chi-Square Goodness of Fit Test

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Data Y_{ij} are binary data. They are independent and replicated. And $\sum_{i=1}^{n_j} Y_{ij} = Y_{\cdot j}$ follows binomial distribution.
- Hypothesis:

$$H_0 : E(Y) = [1 + \exp(-\mathbf{x}'\boldsymbol{\beta})]^{-1}$$

$$H_a : E(Y) \neq [1 + \exp(-\mathbf{x}'\boldsymbol{\beta})]^{-1}$$

- Compute observed outcomes:

$$O_{j1} = \sum_{i=1}^{n_j} Y_{ij} = Y_{\cdot j}$$

$$O_{j0} = \sum_{i=1}^{n_j} (1 - Y_{ij}) = n_j - Y_{\cdot j} = n_j - O_{j1}.$$



Pearson Chi-Square Goodness of Fit Test

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Compute the expected outcomes

$$E(Y_{ij}) = \pi_j = [1 + \exp(-\mathbf{x}'_j \boldsymbol{\beta})]^{-1}, \quad \hat{\pi} = [1 + \exp(-\mathbf{x}'_j \hat{\boldsymbol{\beta}})]^{-1}$$

$$E_{j1} = n_j \hat{\pi}_j, \quad E_{j0} = n_j (1 - \hat{\pi}_j) = n_j - E_{j1}.$$

- Compute the Chi-square statistic

$$\chi^2 = \sum_{j=1}^c \sum_{k=0}^1 \frac{(O_{jk} - E_{jk})^2}{E_{jk}}.$$

- The statistic $\chi^2 \sim \chi^2_{c-p}$ distribution.

- Hypothesis testing:

- If $\chi^2 \leq \chi^2(1 - \alpha, c - p)$, conclude H_0 .
- If $\chi^2 > \chi^2(1 - \alpha, c - p)$, conclude H_a .



Deviance Goodness of Fit

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic

Regression

Poisson
Regression

GLM

- Saturated model: $E(Y_{ij}) = \pi_j$ where π_j for $j = 1, \dots, c$ are the unknown parameters, not the linear coefficients.
- The maximum likelihood estimation for the saturated model is simply $p_j = \frac{Y_{.j}}{n_j}$ for $j = 1, \dots, c$. Its maximized likelihood is $L(F)$.
- The reduced model is the current model under the H_0 . Its maximized likelihood is $L(R)$.
- The likelihood ratio statistic is

$$\begin{aligned} G^2 &= -2 [\log L(R) - \log L(F)] \\ &= -2 \sum_{j=1}^c \left[Y_{.j} \log \left(\frac{\hat{\pi}_j}{p_j} \right) + (n_j - Y_{.j}) \log \left(\frac{(1 - \hat{\pi}_j)}{(1 - p_j)} \right) \right] \\ &= DEV(1, X_1, \dots, X_{p-1}) \end{aligned}$$



Deviance Goodness of Fit

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- $DEV(1, X_1, \dots, X_{p-1})$ is called *residual deviance*.
- Hypothesis testing:
 - If $DEV(1, X_1, \dots, X_{p-1}) \leq \chi^2(1 - \alpha, c - p)$, conclude H_0 .
 - If $DEV(1, X_1, \dots, X_{p-1}) > \chi^2(1 - \alpha, c - p)$, conclude H_a .



Example: Couple Study

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
coupon <- read.table("coupon.txt",header=FALSE)
colnames(coupon) <- c("X", "n", "Y", "p")
coupon$Y2<-coupon$n-coupon$Y
fit<-glm(cbind(Y, Y2)~X,data=coupon,family=binomial("logit"))
summary(fit)
O_1<-coupon$Y
O_0<-coupon$Y2
E_1<-coupon$n*fit$fitted.values
E_0<-coupon$n*(1-fit$fitted.values)
chi2<-sum((O_1-E_1)^2/E_1)+sum((O_0-E_0)^2/E_0)
qchisq(1-0.05,df=nrow(coupon)-2)
```



Logistic Residuals

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression
Probit and Logit
Transformation
Logistic Regression
Binomial Distribution
Inference on MLE
Polynomial Logistic
Regression
Variable Selection
Goodness-of-fit
**Logistic Regression
Diagnostics**
Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

If Y_i take on only the values of 0 and 1, the ordinary residual is defined as

$$e_i = Y_i - \hat{\pi}_i = \begin{cases} 1 - \hat{\pi}, & \text{if } Y_i = 1 \\ -\hat{\pi}_i, & \text{if } Y_i = 0. \end{cases}$$



Pearson Residuals

DATA
564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Divide the ordinary residual by the estimated standard error $\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}$,

$$r_{P,i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

The Pearson residuals are directly related to Pearson chi-square goodness of fit statistic.

$$\begin{aligned}\chi^2 &= \sum_{j=1}^c \sum_{k=0}^1 \frac{(O_{jk} - E_{jk})^2}{E_{jk}} = \sum_{j=1}^c \frac{(O_{j0} - E_{j0})^2}{E_{j0}} + \sum_{j=1}^c \frac{(O_{j1} - E_{j1})^2}{E_{j1}} \\ &= \sum_{i=1}^n \frac{[(1 - Y_i) - (1 - \hat{\pi}_i)]^2}{1 - \hat{\pi}_i} + \sum_{i=1}^n \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \\ &= \sum_{i=1}^n \frac{(Y_i - \hat{\pi}_i))^2}{1 - \hat{\pi}_i} + \sum_{i=1}^n \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i} = \sum_{i=1}^n \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)}\end{aligned}$$



Studentized Pearson Residual

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The Pearson residual is scale free, but its variance is not equal to 1. Approximately, the estimated standard error taking the inherent variation of the fitted value $\hat{\pi}_i$ into consideration is $\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)(1 - h_{ii})}$, where h_{ii} is the i th diagonal element of the $n \times n$ estimated hat matrix for logistic regression

$$\mathbf{H} = \hat{\mathbf{W}}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}' \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{W}}^{\frac{1}{2}}.$$

Here $\hat{\mathbf{W}}$ is the $n \times n$ diagonal matrix with elements $\hat{\pi}_i(1 - \hat{\pi}_i)$, \mathbf{X} is the usual $n \times p$ model matrix, and $\hat{\mathbf{W}}^{\frac{1}{2}}$ is a diagonal matrix with diagonal elements equal to the square roots of those in $\hat{\mathbf{W}}$. The resulting studentized Pearson residuals are defined as

$$r_{SP,i} = \frac{r_{P,i}}{\sqrt{1 - h_{ii}}}.$$



Deviance residuals

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Recall the definition of the deviance defined based on the likelihood ratio

$$G^2 = DEV(1, X_1, \dots, X_{p-1}) = -2 \sum_{j=1}^c \left[Y_{\cdot j} \log \left(\frac{\hat{\pi}_j}{p_j} \right) + (n_j - Y_{\cdot j}) \log \left(\frac{1 - \hat{\pi}_j}{1 - p_j} \right) \right].$$

For binary outcome data, disregarding the repeated measure, let $c = n$, $n_j = 1$ for all, $j = i$, $Y_{\cdot j} = Y_i$, $p_j = Y_{\cdot j}/n_j = Y_i$. Then G^2 becomes

$$\begin{aligned} G^2 &= -2 \sum_{i=1}^n \left[Y_i \log \left(\frac{\hat{\pi}_i}{Y_i} \right) + (1 - Y_i) \log \left(\frac{1 - \hat{\pi}_i}{1 - Y_i} \right) \right] \\ &= -2 \sum_{i=1}^n [Y_i \log \hat{\pi}_i + (1 - Y_i) \log(1 - \hat{\pi}_i) - Y_i \log Y_i - (1 - Y_i) \log(1 - Y_i)] \\ &= -2 \sum_{i=1}^n [Y_i \log \hat{\pi}_i + (1 - Y_i) \log(1 - \hat{\pi}_i)], \end{aligned}$$

since $Y_i \log Y_i = (1 - Y_i) \log(1 - Y_i) = 0$ for $Y_i = 0$ or $Y_i = 1$.



Deviance residuals

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- For binary data, the model deviance is

$$DEV(1, X_1, \dots, X_{p-1}) = -2 \sum_{i=1}^n [Y_i \log \hat{\pi}_i + (1 - Y_i) \log(1 - \hat{\pi}_i)].$$

- The deviance residual for the case i , denoted by dev_i , is defined by

$$dev_i = sign(Y_i - \hat{\pi}_i) \sqrt{-2[Y_i \log \hat{\pi}_i + (1 - Y_i) \log(1 - \hat{\pi}_i)]},$$

where the sign is positive if $Y_i \geq \hat{\pi}_i$ and negative when $Y_i < \hat{\pi}_i$.

- The sum of the squared dev_i is the model deviance.

$$\sum_{i=1}^n dev_i^2 = DEV(1, X_1, \dots, X_{p-1}).$$



Example: Outbreak

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
outbreak<-read.table("Outbreak.txt",header=FALSE)
colnames(outbreak)<-c("Case","X1","X2","X3","X4","Y")
fit.full <- glm(Y~X1+X2+X3+X4,data=outbreak,family=binomial("logit"))
summary(fit.full)

r_0 <- residuals(fit.full, type="response") # ordinary residual
r_P <- residuals(fit.full, type="pearson") # Pearson residual
X <- model.matrix(fit.full)
W <- diag(fit.full$weights)
W2 <- diag(sqrt(fit.full$weights))
H <- W2%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%W2
r_SP <- residuals(fit.full, type="pearson")/sqrt(1-diag(H))
r_D <- residuals(fit.full, type="deviance") # Deviance residual
```



Diagnostic Residual Plots

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Residual versus the predicted probabilities with lowess smooth: if the lowess smooth approximates a line having zero slope and intercept, we can conclude there is no significance model inadequacy is apparent.
- Half-normal probability plot with simulated envelope: see textbook for details, omitted here.



Example: Outbreak

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression
Probit and Logit
Transformation
Logistic Regression
Binomial Distribution
Inference on MLE
Polynomial Logistic
Regression
Variable Selection
Goodness-of-fit
Logistic Regression
Diagnostics
Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
plot(fit.full$fitted.values,r_0,main="ordinary residuals vs probabilities",  
xlab="Estimated Probability",ylab="Ordinary Residuals")  
plot(fit.full$fitted.values,r_P,main="Pearson residuals vs probabilities",  
xlab="Estimated Probability",ylab="Pearson Residuals")  
plot(fit.full$fitted.values,r_SP,main="Studentized Pearson residuals vs probabilities",  
xlab="Estimated Probability",ylab="Studentized Pearson Residuals")  
plot(fit.full$fitted.values,r_D,main="Deviance residuals vs probabilities",  
xlab="Estimated Probability",ylab="Deviance Pearson Residuals")
```



Example: Outbreak

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
plot(fit.full$fitted.values,r_SP,main="Studentized Pearson Residuals vs probabilities",  
xlab="Estimated Probability",ylab="Ordinary Residuals")  
lines(lowess(fit.full$fitted.values,r_0), col="blue")  
  
plot(fit.full$linear.predictors,r_SP,main="Studentized Pearson residuals vs Linear Predictor"  
xlab="Linear Predictor",ylab="Studentized Pearson Residuals")  
lines(lowess(fit.full$linear.predictor,r_SP), col="blue")  
  
plot(fit.full$fitted.values,r_D,main="Deviance residuals vs probabilities",  
xlab="Estimated Probability",ylab="Deviance Residuals")  
lines(lowess(fit.full$fitted.values,r_D), col="blue")  
  
plot(fit.full$linear.predictors,r_D,main="Deviance residuals vs probabilities",  
xlab="Linear Predictor",ylab="Deviance Residuals")  
lines(lowess(fit.full$linear.predictor,r_D), col="blue")
```



Detection of Influential Observations

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Influence on Pearson Chi-square and the Deviance Statistics.

- Let χ^2 and DEV denote the Pearson and deviance statistics based on the full data set.
- Let $\chi_{(i)}^2$ and $DEV_{(i)}$ denote the values of these tests statistics when case i is deleted.
- The i th *delta chi-square statistic* is defined as the change in the Pearson statistic when the i th case is deleted. Similarly, the i th *delta deviance statistic* is defined as the change in the deviance statistic when the i th case is deleted.

$$\Delta\chi_i^2 = \chi^2 - \chi_{(i)}^2 = r_{SP,i}^2$$

$$\Delta dev_i = DEV - DEV_{(i)} = h_{ii}r_{SP,i}^2 + dev_i^2$$



Detection of Influential Observations

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Influence on Pearson Chi-square and the Deviance Statistics.

- The rules of thumb for the linear regression is not applicable here, because the distribution of $\Delta\chi_i^2$ and Δdev_i is unknown.
- Visual assessment of an appropriate graphic: delta chi-square and delta deviance statistics are plotted against case number i , against $\hat{\pi}_i$, or against the linear predictor. Extreme values appear as spikes when plotted against case number, or as outliers in the upper corners of the plot when plotted against $\hat{\pi}_i$ or the linear predictor.



Detection of Influential Observations

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- Influence on the Fitted Linear Predictor: Cook's Distance, which measures the standardized change in the linear predictor $\hat{\pi}_i$ when the i th case is deleted.
Approximately,

$$D_i = \frac{r_{P,i}^2 h_{ii}}{p(1 - h_{ii})^2}.$$

- Visual assessment of an appropriate graphic.
- Leverage values h_{ii} are useful for identifying outliers in X space.



Example: Outbreak

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous

Logistic
Regression

Poisson
Regression

GLM

```
delta_chi2 <- r_SP^2
delta_dev <- diag(H)*delta_chi2+r_D^2
CookD<-r_P^2*diag(H)/(ncol(X)*(1-diag(H))^2)
plot(delta_chi2,type="b", xlab="Case", ylab="Delta Chi-Square",
main="Delta Chi-Square vs Index")
plot(delta_dev,type="b", xlab="Case",ylab="Delta Dev",
main="Delta Dev vs Index")
plot(fit.full$fitted.values,delta_chi2,xlab="Estimated Probability",
ylab="Delta Chi-Square", main="Delta Chi-Square vs Estimated Probability")
plot(fit.full$fitted.values,delta_dev,xlab="Estimated Probability",
ylab="Delta Dev", main="Delta Dev vs Estimated Probability")
plot(diag(H), type="b",ylab="Leverage",xlab="Case Index", main="leverage vs index")
plot(CookD, type="b",ylab="Cook's Distance",xlab="Case Index",
main="Cook's distance vs index")

library(ggplot2)
ggplot(data=data.frame(prob=fit.full$fitted.values, delta_dev2=delta_dev^2, CookD=CookD),
aes(x=prob, y=delta_dev2, size=CookD))+geom_point(alpha=0.2)
```



Inference about Mean Response

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- We want to estimate π for $\mathbf{x} = \mathbf{x}_h$, where $\mathbf{x}_h = [1, X_{h1}, \dots, X_{h,p-1}]'$.
- The mean response of interest $E(Y_h) = \pi_h = \frac{\exp(\mathbf{x}'_h \boldsymbol{\beta})}{1+\exp(\mathbf{x}'_h \boldsymbol{\beta})}$, and the estimated

$$\hat{\pi}_h = \frac{\exp(\mathbf{x}'_h \hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}'_h \hat{\boldsymbol{\beta}})}.$$

- The $100(1 - \alpha)\%$ confidence interval for $\hat{\pi}_h$ is $[L^*, U^*]$.

$$s^2(\hat{\boldsymbol{\beta}}) = \text{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}' \mathbf{W}(\hat{\boldsymbol{\beta}}) \mathbf{X})^{-1}, \quad s^2(\mathbf{x}'_h \hat{\boldsymbol{\beta}}) = \mathbf{x}'_h s^2(\hat{\boldsymbol{\beta}}) \mathbf{x}_h$$

$$L = \mathbf{x}'_h \hat{\boldsymbol{\beta}} - z_{\alpha/2} s(\mathbf{x}'_h \hat{\boldsymbol{\beta}}), \quad U = \mathbf{x}'_h \hat{\boldsymbol{\beta}} + z_{\alpha/2} s(\mathbf{x}'_h \hat{\boldsymbol{\beta}})$$

$$L^* = \frac{\exp(L)}{1 + \exp(L)}, \quad U^* = \frac{\exp(U)}{1 + \exp(U)}$$

- Use Bonferroni simultaneous confidence interval for g simultaneous confidence interval.



Prediction of a New Observation

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

We want to predict a binary outcome given $x = x_h$. Choice of Prediction Rule

- 1 Use 0.5 as cutoff, i.e., if $\hat{\pi}_h$ exceeds 0.5, predict 1; otherwise predict 0.
- 2 Find the best cutoff for the data set on which the multiple logistic regression model is based. Try different cutoff values and compare the proportion of the cases incorrectly predicted of the training data.
- 3 Use prior probabilities and costs of incorrect predictions in determine the cutoff.

The reliability of the prediction error rate observed in the model-building data set is examined by applying the chosen prediction rule to a validation data set.



Example: Outbreak

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
rule1<-as.integer(fit.full$fitted.values>=0.316)
table(rule1, outbreak$Y)
rule2<-as.integer(fit.full$fitted.values>=0.325)
table(rule2, outbreak$Y)
```



Data with Polytomous Responses

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Example (Pregnancy Duration)

Output

Y_i	Pregnancy Duration Category
1	Preterm (less than 36 weeks)
2	Intermediate term (36 to 37 weeks)
3	Full term (38 weeks or greater)

Input: nutritional status (X_1), age (categorized into three groups, X_2 and X_3), alcohol use history (X_4), and smoking history (X_5).



For Nominal (qualitative) Response

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

**Polytomous
Logistic
Regression**

Poisson
Regression

GLM

If Y_i is a qualitative variable (or nominal) of J categories, we can create J binary responses.

Example (Pregnancy Duration)

$$Y_{i1} = \begin{cases} 1, & \text{if case } i \text{ response is category 1} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{i2} = \begin{cases} 1, & \text{if case } i \text{ response is category 2} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{i3} = \begin{cases} 1, & \text{if case } i \text{ response is category 3} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{i1} + Y_{i2} + Y_{i3} = 1 \text{ or } Y_{i3} = 1 - Y_{i1} - Y_{i2}.$$



For Nominal (qualitative) Response

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

In general we will assume there are J categories. Then for the i th observation, there will be J binary responses Y_{i1}, \dots, Y_{iJ} , where

$$Y_{ij} = \begin{cases} 1, & \text{if case } i \text{ response is category } j. \\ 0, & \text{otherwise.} \end{cases}$$

Let π_{ij} denote the probability that category j is selected for the i th response, so $\pi_{ij} = P(Y_{ij} = 1)$.

For the J polytomous categories, there are $J(J - 1)/2$ pairs of odds ratio, $\frac{\pi_{il}}{\pi_{ik}}$ for each x_i , that compares every pair of the probability $P(Y_{il} = 1)$ and $P(Y_{ik} = 1)$. One category should be chosen as the baseline or referent category, and then all other categories will be compared to it.



For Nominal (qualitative) Response

We choose the J th category as the baseline, and assume the logistic model

$$\log \frac{\pi_{i1}}{\pi_{iJ}} = \log \frac{\pi_{i1}}{1 - \sum_{l=1}^{J-1} \pi_{il}} = \mathbf{x}'_i \boldsymbol{\beta}_1 \Leftrightarrow \pi_{i1} = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta}_1)}{1 + \sum_{l=1}^{J-1} \exp(\mathbf{x}'_i \boldsymbol{\beta}_l)}$$

⋮

$$\log \frac{\pi_{ij}}{\pi_{iJ}} = \log \frac{\pi_{ij}}{1 - \sum_{l=1}^{J-1} \pi_{il}} = \mathbf{x}'_i \boldsymbol{\beta}_j \Leftrightarrow \pi_{ij} = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta}_j)}{1 + \sum_{l=1}^{J-1} \exp(\mathbf{x}'_i \boldsymbol{\beta}_l)}$$

⋮

$$\log \frac{\pi_{i,J-1}}{\pi_{iJ}} = \log \frac{\pi_{i,J-1}}{1 - \sum_{l=1}^{J-1} \pi_{il}} = \mathbf{x}'_i \boldsymbol{\beta}_{J-1} \Leftrightarrow \pi_{i,J-1} = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta}_{J-1})}{1 + \sum_{l=1}^{J-1} \exp(\mathbf{x}'_i \boldsymbol{\beta}_l)}$$

If we want to compare any other pair of the probabilities,

$$\log \frac{\pi_{il}}{\pi_{ik}} = \mathbf{x}'_i (\boldsymbol{\beta}_l - \boldsymbol{\beta}_k), \quad \text{for } l, k = 1, \dots, J-1.$$



For Nominal (qualitative) Response

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

The likelihood of all the data is

$$l(Y_1, \dots, Y_n) = \prod_{i=1}^n \left[\prod_{j=1}^J (\pi_{ij})^{Y_{ij}} \right].$$

The log-likelihood is

$$\log l(\beta_1, \dots, \beta_{J-1}) = \sum_{i=1}^n \left(\sum_{j=1}^{J-1} (Y_{ij} \mathbf{x}'_i \boldsymbol{\beta}_j) - \log \left[1 + \sum_{l=1}^{J-1} \exp(\mathbf{x}'_i \boldsymbol{\beta}_l) \right] \right).$$

Maximizing the log-likelihood with respect to the parameters, we obtain the MLE $\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_{J-1}$.



Example: Pregnancy Duration

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE
Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
library(nnet)
pregnancy<-read.table('Pregnancy.txt',header=FALSE)
colnames(pregnancy)<-c("Case","Y", "Y1", "Y2", "Y3", "X1",
"X2", "X3", "X4", "X5")
fit <- multinom(cbind(Y3,Y2,Y1)~X1+X2+X3+X4+X5, data=pregnancy)
summary(fit)
train <-sample(size=80,x=1:nrow(pregnancy))
test <- setdiff(1:nrow(pregnancy), train)
fit2 <- multinom(Y~X1+X2+X3+X4+X5, data=pregnancy,subset=train)
pred <- predict(fit2,newdata=pregnancy[test,])
table(pred,pregnancy$Y[test])
```



Poisson Distribution

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- The Poisson distribution can be utilized for outcomes that are counts 0, 1, 2, ..., with a large count or frequency being a rare event.
- The probability mass function for a Poisson random variable is

$$f(y) = \frac{\mu^y \exp(-\mu)}{y!}, \quad y = 0, 1, 2, \dots$$

- $E(Y) = \mu$ and $\text{var}(Y) = \mu$. $\mu > 0$



Poisson Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Assume that the mean of the Poisson random variable depends on some explanatory variable X_1, \dots, X_{p-1} , then we can assume the relationship between $\mu(\mathbf{x})$ and the linear predictor $\mathbf{x}'\boldsymbol{\beta}$ to be,

$$\log \mu(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}.$$

The joint likelihood for data $\{\mathbf{x}_i, Y_i\}$ is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \frac{[\mu(\mathbf{x}_i, \boldsymbol{\beta})]^{Y_i} \exp[-\mu(\mathbf{x}_i, \boldsymbol{\beta})]}{Y_i!}.$$

The log-likelihood is

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^n Y_i \mathbf{x}'_i \boldsymbol{\beta} - \sum_{i=1}^n \exp(\mathbf{x}'_i \boldsymbol{\beta}) - \sum_{i=1}^n \log(Y_i!).$$



Poisson Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- $\hat{\beta}$ is the maximum likelihood estimate of β .
- Deviance is defined through likelihood ratio statistic between the likelihood of the current model and the saturated model. The saturated model would assume n parameters (μ_1, \dots, μ_n) for the n mean values of Y_1, \dots, Y_n . The only possible estimate for μ_i is Y_i , since there is only one observation Y_i for μ_i .

$$\begin{aligned}DEV(1, X_1, \dots, X_{p-1}) &= G^2 = -2 [\log L(R) - \log L(S)] \\&= -2 \left[\sum_{i=1}^n Y_i \log \mu(\mathbf{x}_i, \boldsymbol{\beta}) - \sum_{i=1}^n \mu(\mathbf{x}_i, \boldsymbol{\beta}) - \sum_{i=1}^n Y_i \log Y_i + \sum_{i=1}^n Y_i \right] \\&= -2 \left[\sum_{i=1}^n Y_i \log \frac{\mu(\mathbf{x}_i, \boldsymbol{\beta})}{Y_i} + \sum_{i=1}^n (Y_i - \mu(\mathbf{x}_i, \boldsymbol{\beta})) \right]\end{aligned}$$



Poisson Regression

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- The deviance residual for the i th case is

$$dev_i = \text{sign}(Y_i - \hat{\mu}_i) \left[-2Y_i \log \frac{\hat{\mu}_i}{Y_i} - 2(Y_i - \hat{\mu}_i) \right]^{1/2}.$$

- Inferences for a Poisson regression is done in the same way as for logistic regression.



Example: Store Visit

DATA
564-494
Simple Linear
Regression

Lulu Kang

Logistic
Regression
Probit and Logit
Transformation
Logistic Regression
Binomial Distribution
Inference on MLE
Polynomial Logistic
Regression
Variable Selection
Goodness-of-fit
Logistic Regression
Diagnostics
Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

```
visit<-read.table("visit.txt",header=FALSE)
colnames(visit) <- c("Y", "X1","X2","X3","X4","X5")
fit <-glm(Y~., data=visit, family=poisson("log"))
summary(fit)
dev<-residuals(fit,type="deviance")
plot(fit$fitted, dev)
plot(dev, type="b")
```



Exponential Family

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Most of the commonly used statistical distributions, e.g., Normal, Binomial and Poisson, are members of the *exponential family of distributions* whose densities can be written in the form

$$f(y, \theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

where ϕ is the dispersion parameter and θ is the canonical parameter.

It can be shown that

$$E(Y) = b'(\theta) = \mu$$

$$\text{var}(Y) = \phi b''(\theta) = \phi V(\mu)$$



Exponential Family

DATA

564/494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

- For normal distribution,

$$f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right) = \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right)$$

so that $\theta = \mu$, $\phi = \sigma^2$, and $a(\phi) = \phi$, $b(\theta) = \theta^2/2$,

$$c(y, \theta) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right).$$

- For binomial distribution, the pmf of the proportion $Z = Y/n$ is

$$f(z, n, \pi) = \binom{n}{nz} \pi^{nz} (1 - \pi)^{n-nz} = \exp\left(\frac{z \log \frac{\pi}{1-\pi} - \log \frac{1}{1-\pi}}{n^{-1}} + \log \binom{n}{nz}\right)$$

$\theta = \log \frac{\pi}{1-\pi}$, $\mu = \pi = \frac{\exp(\theta)}{1+\exp(\theta)}$, $b(\theta) = \log(1 + \exp(\theta))$, $a(\phi) = \phi = n^{-1}$ and

$$c(z, \theta) = \log \binom{n}{nz}.$$



Generalized Linear Model

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

A general linear model is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_{p-1} x_{i,p-1}$$

and two functions

- a link function that describes how the mean, $E(Y_i) = \mu_i$ depends on the linear predictor

$$g(\mu_i) = \eta_i$$

- a variance function that describes how the variance, $\text{var}(Y_i)$ depends on the mean

$$\text{var}(Y_i) = \phi V(\mu)$$

where the dispersion parameter ϕ is a constant.



Normal General Linear Model as a Special Case

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

For the general linear model with $\epsilon \sim N(0, \sigma^2)$, we have the linear predictor

$$\eta_i = \mathbf{x}'_i \boldsymbol{\beta}$$

the link function

$$g(\mu_i) = \mu_i$$

and the variance function

$$V(\mu_i) = 1.$$



Modeling Binomial Data

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression
Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Suppose

$$Y_i \sim \text{Binomial}(n_i, \pi_i)$$

and we wish to model the proportion Y_i/n . Then

$$E(Y_i/n_i) = \pi_i \quad \text{var}(Y_i/n_i) = \frac{1}{n_i} \pi_i (1 - \pi_i).$$

So our variance function is

$$V(\mu_i) = \mu_i(1 - \mu_i).$$

Our link function must map from $(0, 1) \rightarrow (-\infty, \infty)$. Common choices are

$$g(\mu_i) = \text{logit}(\mu_i) = \log \frac{\mu_i}{1 - \mu_i} \quad \text{or } g(\mu_i) = \Phi^{-1}(\mu_i).$$



Modeling Poisson Data

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

Suppose

$$Y_i \sim \text{Poisson}(\mu_i).$$

Then

$$E(Y_i) = \mu_i \quad \text{var}(Y_i) = \mu_i.$$

So our variance function is

$$V(\mu_i) = \mu_i.$$

Our link function must map from $(0, \infty) \rightarrow (-\infty, \infty)$. A natural choice is

$$g(\mu_i) = \log(\mu_i).$$



Transformation vs. GLM

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

In some situations a response variable can be transformed to improve linearity and homogeneity of variance so that a general linear model can be applied.

This approach has some drawbacks

- response variable has changed!
- transformation must simultaneously improve linearity and homogeneity of variance.
- transformation may not be defined on the boundaries of the sample space.



Transformation vs. GLM

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

For example, a common remedy for the variance increasing with the mean is to apply the log transform, e.g.

$$\begin{aligned}\log(y_i) &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \rightarrow E(\log(y_i)) &= \beta_0 + \beta_1 x_i.\end{aligned}$$

This is a linear model for the mean of $\log Y$ which may not always be appropriate. If Y is income perhaps we are really interested in the mean income of population subgroups, in which case it would be better to model $E(Y)$ using `glm`:

$$\log E(Y_i) = \beta_0 + \beta_1 x_i$$

with $V(\mu) = \mu$. This also avoids difficulty with $y = 0$.



Canonical Links

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection

Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

For a glm where the response follows a distribution from the exponential family, we have

$$g(\mu_i) = g(b'(\theta_i)) = \mathbf{x}'_i \boldsymbol{\beta}.$$

The canonical link is defined as

$$\begin{aligned} g(\cdot) &= (b'(\cdot))^{-1} \text{the inverse function of } b'(\theta) \\ \rightarrow g(\mu_i) &= \theta_i = \mathbf{x}'_i \boldsymbol{\beta}. \end{aligned}$$

Canonical links lead to desirable statistical properties of the glm hence tend to be used by default. However, there is no a priori reason why the systematic effects in the model should be additive on the scale given by this link.



Estimation of the Model Parameters

DATA

564-494

Simple Linear
Regression

Lulu Kang

Logistic
Regression

Probit and Logit
Transformation

Logistic Regression

Binomial Distribution

Inference on MLE

Polynomial Logistic
Regression

Variable Selection
Goodness-of-fit

Logistic Regression
Diagnostics

Mean and Prediction

Polytomous
Logistic
Regression

Poisson
Regression

GLM

A single algorithm can be used to estimate the parameters of an exponential family glm using maximum likelihood.

The log-likelihood for the sample y_1, \dots, y_n is

$$l = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i).$$

The maximum likelihood estimates are obtained by solving the score equations using Fisher's score methods.

$$s(\beta_j) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} = 0, \text{ for } \beta_j.$$