1. Linear Regression
   1. Other Considerations in the Regression Model
      1. Qualitative Predictors

Predictors with Two levels: Gender xi as dummy variable:

xi =



 1, the ith person is female

 0, the ith person is male

(2.1)

Model:

yi = β0 + β1 × xi + ǫi = β0 +



 β1 + ǫi, the ith person is female

 ǫi, the ith person is male

(2.2)

* + - * β0: average credit card balance among male
      * β0 + β1: average credit card balance among female
      * β1: average difference in credit card balance between male and female

Predictors with more than Two Levels: Ethnicity xi1, xi2 as dummy variable:

x1i =



 1, the ith person is Asian

 0, the ith person is not Asian

(2.3)

x2i =



 1, the ith person is Caucasian

 0, the ith person is not Caucasian

(2.4)

Model:

yi = β0 + β1 × x1i + β2 × x2i + ǫi (2.5)



 β1 + ǫi, the ith person is Asian





= β0 +



β2 + ǫi, the ith person is Caucasian

(2.6)

 ǫi, the ith person is African American



* + 1. Extensions of the Linear Model

Interaction Effect:

Consider the linear regression model with one quantitative predictor and one qualitative predictor. We are using the credit data as our example, suppose we wish to predict bal- ance using income(quantitative) and student(qualitative) variables. In the absence of an interaction term, the model takes the form:

balancei ≈ β0 + β1 × incomei +





 β2, the ith person is a student

 0, the ith person is not a student

(2.7)

= β1 × incomei +

 β0 + β2, the ith person is a student

 β0, the ith person is not a student

(2.8)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient | Std. Error | t-statistic | p-value |
| Intercept | 211.14 | 32.46 | 6.51 | 0 |
| Income | 5.98 | 0.56 | 10.75 | 0 |
| Student | 382.67 | 65.31 | 5.86 | 0 |
| R2 = 27.75% |  |  |  |  |

Table 1: Coefficients table

|  |  |  |  |
| --- | --- | --- | --- |
| Source | SS | df | MS F |
| Regression | 23400858 | 3-1 | 11700429 76.22 |
| Residual Error | 60939054 | 400-3 | 153499 |
| Total | 84339912 | 400-1 |  |

Table 2: ANOVA table

student non−student

Balance

1000

1400

Balance

1000

1400

0 50 100 150

200

600

200

600

0 50 100 150

Income

Income

Figure 1: Left: The model (2.8) with no interaction was fit, Right: The model (2.11) with interaction was fit

In the left panel of Figure 1, we notice that Model (2.8) fit two parallel lines to the data, one for students and one for non-students. The lines for students and non-students have different interceptions, β0 + β2 versus β0, but the same slope β0. The fact that the lines are parallel means that according to Model (2.8), the average effect on balance of a one-unit increase in income does not depend on whether or not the person is a student. This assumption on the model is simply wrong, since in fact a change in income may have a very different effect on the credit card balance of a student versus a non-student.

This issue can be solved by adding an interaction variable into our model:

yi ≈ β0 + β1xi1 + β2xi2 + β3xi1xi2 (2.9)

* + - * xi1: Income for the ith person (Quantitative variable)
      * xi2: whether the ith person is a student or not (Qualitative Dummy variable)
      * xi1xi2: Interaction term between Income(xi1) and Student(xi2) The model with interaction term can be interpreted as below:

balancei ≈ β0 + β1 × incomei +





 β2 + β3 × incomei, the ith person is a student

 0, the ith person is not a student (2.10)

=  (β0 + β2) + (β1 + β3) × incomei, the ith person is a student

 β0 + β1 × incomei, the ith person is not a student

(2.11)

In the right panel of Figure 1, we notice that when we added a interaction term in our model, we will get two different regression lines for the students and the non-students. Those lines have different intercepts, (β0 + β2) versus β0 and different slopes, (β1 + β3) versus β1. This allows for the possibility that changes in income may affect the credit card balances of students and non-students differently.

* + 1. Potential Problems

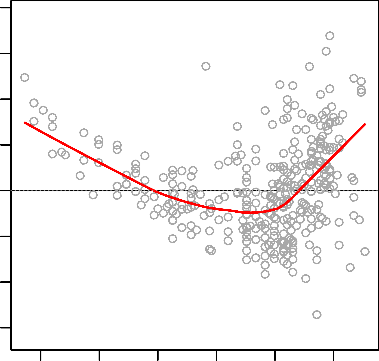
1. Non-linearity of the response-predictors relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High leverage points
6. Collinearity
7. Non-linearity of the Data:

Ideally, the residual plot will show no obvious pattern. The presence of a pattern may indicate a problem with non-linearity.

How to detect Non-linearity:

* + Plot the residuals(yi − yˆi) versus the fitted values yˆi as a residual plot (Figure 2) How to solve Non-linearity:
  + If the residual plot indicates non-linear in the data, then adding a non-linear transfor- mations of the predictors, such as logX, √X, X2, in the regression model.

Residual Plot for Linear Fit



10 15

Residual Plot for Quadratic Fit

334

Residuals

5

10 15 20

Residuals

5

323

330

334

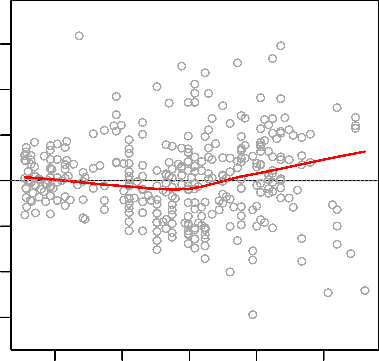
323

155

−15 −10 −5 0

−15 −10 −5 0

5 10 15 20 25 30



15 20 25 30 35

Fitted values

Fitted values

Figure 2: Residual Plot, Left: a strong pattern in the residual indicates non-linearity if we fit our data to a linear model. Right: There is little pattern in the residual when we fit a quadratic model to the data

1. Correlation of Error Terms:

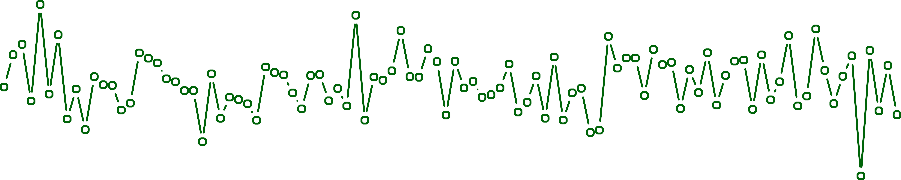
In the assumption of the linear regression model, the error terms ǫ1, ǫ1, . . . , ǫn are uncorre- lated. If in fact there is correlation among the error term, then the estimated standard

errors will tend to underestimate the true standard error. Such correlations fre- quently occur in time series data. In many cases, adjacent observations will have positive correlated errors.

How to detect Correlation of Error Terms:

* + Plot the residuals(yi − yˆi) versus the observation (Figure 3)
  + Identify obvious pattern in Figure 3

ρ=0.0



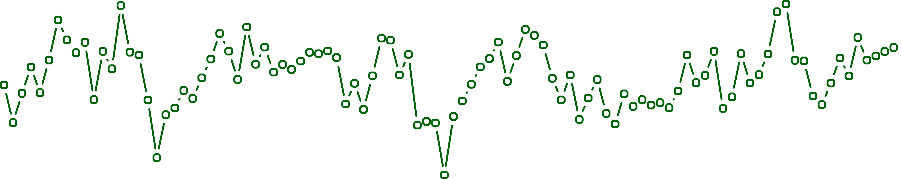
−1 0 1 2 3

0 20 40 60 80 100

Residual

−3

ρ=0.5



0 1 2

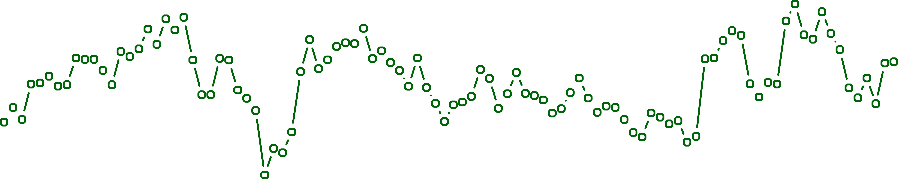
0 20 40 60 80 100

Residual

−4

−2

ρ=0.9



1.5

0 20 40 60 80 100

Residual

−1.5 −0.5 0.5

Observation

Figure 3: Plots of residuals versus simulated time series observations with different levels of ρ

How to solve Correlation of Error Terms:

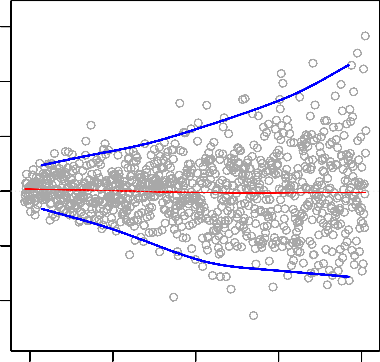
* + If the residual plot indicates a obvious patten, then there might be a problem in the design of experiments.

1. Non-Constant Variance of Error Terms(Heteroscedasticity):

In the assumption of the linear regression model, the variance of error terms are constant(Var(ǫi) = σ2). Unfortunately, it is often the case that the variances of the error terms are non-constant. How to detect Non-Constant Variance of Error Terms:

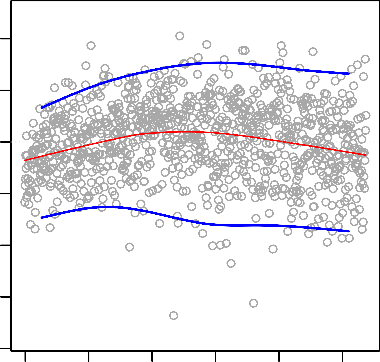
* + Plot residual plot to detect a funnel shape (Figure 4)

Response Y



Response log(Y)

998



10 15

0.2 0.4

975

845

5

605

Residuals

−10 −5 0

Residuals

−0.8 −0.6 −0.4 −0.2 0.0

437

671

10 15 20 25 30

2.4 2.6 2.8 3.0 3.2 3.4

Fitted values

Fitted values

Figure 4

How to solve Non-Constant Variance of Error Terms:

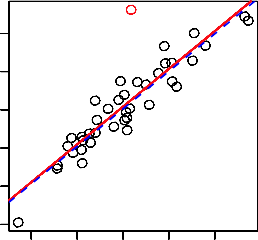
* + Transform the response Y using a concave function such as logY, √Y
  + Or, fit our model by weighted least squares or generalized least squares

1. Outliers:

How to detect Outliers:

* + Plot residual plot and Studentized residual plot to detect outliers (Figure 5)

20 20 20



Y

−4 −2 0 2 4 6

Residuals

−1 0 1 2 3 4

Studentized Residuals

2

4

6

−2 −1 0 1 2

0

X

−2 0 2 4 6

Fitted Values

−2 0 2 4 6

Fitted Values

Figure 5

How to solve Outliers:

* + Remove them

1. High Leverage Points:

Leverage points means the observations have an unusual value of xi. High leverage observa- tions tend to have a sizable impact on the estimated regression model.

How to detect High Leverage Points:

* + In order to identify an observation is a high leverage point, we compute the leverage statistic. A large value of this statistic indicates an observation with high leverage. (Figure 6)

The leverage statistic for a simple linear regression where yi = β0 + β1xi, (p = 2)

2

1 (xi − x¯) 1

hi = n +

n

. (xi x¯)2

−

i′=1

, where

< hi < 1 (2.12)

n

For the average leverage h¯ = p . So if a given observation has a leverage statistic that greatly exceeds p , then we suspect the corresponding point has high leverage (xi > p ). The proof

n

n n

of average leverage for a simple linear regression(p = 2) :

 

n n n 2

h¯ = . hi

= . 1 + .

(xi − x¯)

 /n = 2/n (2.13)

i=1 n



 i=1

n

i=1

n 

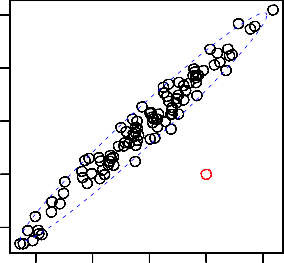
. (xi x¯)2 

−

′

i =1

41 20



2

−1 0 1 2 3 4

5

41

10

1

Studentized Residuals

20

Y

5

X 2

0

−1

−2 −1 0 1 2 3 4

0

−2

X

−2 −1 0 1 2

X 1

0.00 0.05 0.10 0.15 0.20 0.25

Leverage

Figure 6

The leverage for multiple predictors:

βˆ = (X′X)−1X′y

yˆ = Xβˆ = X(X′X)−1X′y

= Hy

H = X(X′X)−1X′

The leverage, hii, quantifies the influence that the observed response yi has on its predicted value yˆi. That is, if hii is small, then the observed response yi plays only a small role in the value of the predicted response yˆi. On the other hand, if hii is large, then the observed response yi plays a large role in the value of the predicted response yˆi. It’s for this reason that the hii are called the ”leverage”.

Here are some important properties of the leverage:

* The leverage hii is a measure of the distance between the xi value for the ith data point and the mean of the x values for all n data points.
* The sum of the hii equals p, the number of parameters (regression coefficients including the intercept).
* The leverage hii is a number between 0 and 1, inclusive.

Example

|  |  |  |
| --- | --- | --- |
| Obs. | X | Y |
| 1 | 0.1 | -0.0716 |
| 2 | 0.45401 | 4.1673 |
| 3 | 1.09765 | 6.5703 |
| 4 | 1.27936 | 13.815 |
| 5 | 2.20611 | 11.4501 |
| 6 | 2.50064 | 12.9554 |
| 7 | 3.0403 | 20.1575 |
| 8 | 3.23583 | 17.5633 |
| 9 | 4.45308 | 26.0317 |
| 10 | 4.1699 | 22.7573 |
| 11 | 5.28474 | 26.303 |
| 12 | 5.59238 | 30.6885 |
| 13 | 5.92091 | 33.9402 |
| 14 | 6.66066 | 30.9228 |
| 15 | 6.79953 | 34.11 |
| 16 | 7.97943 | 44.4536 |
| 17 | 8.41536 | 46.5022 |
| 18 | 8.71607 | 50.0568 |
| 19 | 8.70156 | 46.5475 |
| 20 | 9.16463 | 45.7762 |
| 21 | 4 | 40 |

R code:

x = matrix ( c ( 0 . 1 , 0 . 4 5 4 0 1 , 1 . 0 9 7 6 5 , 1 . 2 7 9 3 6 , 2 . 2 0 6 1 1 , 2 . 5 0 0 6 4 , 3 . 0 4 0 3 ,

3 . 2 3 5 8 3 , 4 . 4 5 3 0 8 , 4 . 1 6 9 9 , 5 . 2 8 4 7 4 , 5 . 5 9 2 3 8 , 5 . 9 2 0 9 1 , 6 . 6 6 0 6 6 , 6 . 7 9 9 5 3 ,

7 . 9 7 9 4 3 , 8 . 4 1 5 3 6 , 8 . 7 1 6 0 7 , 8 . 7 0 1 5 6 , 9 . 1 6 4 6 3 , 4 ) , ncol = 1 )

x = matrix ( cbind ( rep ( 1 , 2 1 ) , x ) , ncol =2) H = x %∗% s o l ve ( t ( x )%∗%x ) %∗% t ( x )

H = round (H, 4 )

1. Collinearity:

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. The importance of the predictor variables may be masked if we did not consider Collinearity.

How to detect Collinearity:

* + Compute correlation matrix for each predictors
  + Or, compute the variance inflation factor(VIF) How to solve Collinearity:
  + Drop one of the problematic variables from the regression model. Since the other variable provides enough information.
  + Or, combine the collinear variables together to get a single predictor.