NATIONAL ENGINEERING CENTER
University of the Philippines
Diliman, Quezon City

4.0 Regression Methodologies

Eugene Rex L. Jalao, Ph.D.

Associate Professor

Department Industrial Engineering and Operations Research

University of the Philippines Diliman

@thephdataminer

Module 3 of the Business Intelligence and Analytics Certification of UP NEC and the UP Center for Business Intelligence

Outline for This Training

1. Introduction to Data Mining
2. Data Preprocessing
   – Case Study on Big Data Preprocessing using R
3. Classification Methodologies
   – Case Study on Classification using R
4. **Regression Methodologies**
   – **Case Study: Regression Analysis using R**
5. Unsupervised Learning
   – Case Study: Social Media Sentiment Analysis using R

This Session's Outline

- Multiple Linear Regression
- Model Evaluation
- Variable Selection and Model Building
  - Best Subsets Regression
  - Stepwise Regression
  - Ridge Regression
  - Standardized Regression
- Indicator Variables
- Multicollinearity
- Logistic Regression
- Case Study

Regression

- Regression is a data mining task of predicting the value of target (numerical variable y) by building a model based on one or more predictors (numerical and categorical variables).

$$y = \beta_0 + \beta_1 x_1$$

- Not all observations will fall exactly on a straight line

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

where ε represents error

- it is a random variable that accounts for the failure of the model to fit the data *exactly*.
- $\varepsilon \sim N(0, \sigma^2)$

Required Dataset Structure

Attributes/Columns/Variables/Features $(p + 1)$

Rows/ Instances /Tuples /Objects (n)

Tid	**Refund**	**Marital Status**	**Taxable Income**
1	Yes	Single	125K
2	No	Married	100K
3	No	Single	70K
4	Yes	Married	120K
5	No	Divorced	95K
6	No	Married	60K
7	Yes	Divorced	220K
8	No	Single	85K
9	No	Married	75K
10	No	Single	90K

Predictor Variables/Independent Variables/Control Variables

Numeric Response Variable/ Dependent Variable/ Class Variable/ Label Variable/ Target Variable

Regression

- There are many uses of regression, including:
  - Data description
  - Parameter estimation
  - Prediction and estimation
  - Control
- Regression analysis is perhaps the most widely used statistical technique, and probably the most widely misused.

Multiple Linear Regression Models

- Multiple linear regression (MLR) is a method used to model the linear relationship between a target variable and more than one predictor variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- This is a multiple linear regression model in two variables.
- In general, the multiple linear regression model with k regressors is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Multiple Regression Models

- We define linear in terms of coefficients

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

- We can also model non-linear relationships
  - E.g.
  - Let $x_2' = x_2^2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$$

  - Then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2'$$

Estimation of the Model Parameters

- We use the Least Squares Estimation methodology to estimate Regression Coefficients
- Notation
  - n := number of observations available
  - k := number of regressor variables = $p = k + 1$
  - y := response or dependent variable
  - x_{ij} := i^{th} observation or level of regressor j.
- Some properties of Regression Models

$$E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$$

Least Squares Estimation of the Regression Coefficients

Observation, i	Response, y	Regressors			
		x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Least Squares Estimation of the Regression Coefficients

- Matrix notation is typically used:
- Let

$$y = \boldsymbol{X\beta} + \epsilon$$

- where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Least Squares Estimation of the Regression Coefficients

- To estimate β, we wish to minimize

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

- The solution is

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

- These are the least-squares normal equations.

The Delivery Time Data

Observation Number	Delivery Time (Minutes) y	Number of Cases x_1	Distance (Feet) x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

R Code to Run

```
> deliverytime =
  read.csv("deliverytime.csv")
> lrfit=lm(deltime ~ ncases + distance,
  data= deliverytime)
> summary(lrfit)
```

R Output

```
Call:
lm(formula = DelTime ~ Ncases + Distance, data = DeliveryTime)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7880 -0.6629  0.4364  1.1566  7.4197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.341231   1.096730   2.135 0.044170 *  
Ncases      1.615907   0.170735   9.464 3.25e-09 ***
Distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,	Adjusted R-squared:  0.9559 
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

This Session's Outline

- Multiple Linear Regression
- **Model Evaluation**
- Variable Selection and Model Building
  - Best Subsets Regression
  - Stepwise Regression
  - Ridge Regression
  - Standardized Regression
- Indicator Variables
- Multicollinearity
- Logistic Regression
- Case Study

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \bar{y}_i$
1	16.68	21.7081	−5.0281
2	11.50	10.3536	1.1464
3	12.03	12.0798	−0.0498
4	14.88	9.9556	4.9244
5	13.75	14.1944	−0.4444
6	18.11	18.3996	−0.2896
7	8.00	7.1554	0.8446
8	17.83	16.6734	1.1566
9	79.24	71.8203	7.4197
10	21.50	19.1236	2.3764
11	40.33	38.0925	2.2375
12	21.00	21.5930	−0.5930
13	13.50	12.4730	1.0270
14	19.75	18.6825	1.0675
15	24.00	23.3288	0.6712
16	29.00	29.6629	−0.6629
17	15.35	14.9136	0.4364
18	19.00	15.5514	3.4486
19	9.50	7.7068	1.7932
20	35.10	40.8880	−5.7880
21	17.90	20.5142	−2.6142
22	52.32	56.0065	−3.6865
23	18.75	23.3576	−4.6076
24	19.83	24.4028	−4.5728
25	10.75	10.9626	−0.2126

Model Evaluation: Questions

- Is at least one of the predictors, $x_1, x_2, \ldots, x_p$ useful in predicting the response?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
- Are there any outliers that might influence the coefficients?
- Do all the predictors help to explain y , or is only a subset of the predictors useful?

Testing the Global Significance of Regression

- To know if the x predictor variables influences y we consider the F Statistic from the ANOVA table output from R
- We usually test for:
  - H_0 : There is no relationship between all x and y.
  - H_a : There is some relationship between some x and y.
- p-Value Methodology
  - If $p < \alpha = 0.05$, Reject H_0
- F Test Methodology
  - Consider a Confidence Level, usually 95%
  - Lookup Critical Value $F_{\alpha,k,n-k-1}$ from Statistical F Tables
  - If F > $F_{\alpha,k,n-k-1}$, Reject H_0

Model Evaluation: Questions

- Is at least one of the predictors, $x_1, x_2, \ldots, x_p$ useful in predicting the response?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
- Are there any outliers that might influence the coefficients?
- Do all the predictors help to explain y , or is only a subset of the predictors useful?

Coefficient of Determination

- R^2 is called the coefficient of determination: proportion of variance (or information) explained by the predictor variables

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

- For the Delivery Time Data

$$R^2 = \frac{SS_R}{SS_T} = 95.96\%$$

Coefficient of Determination

- Some issues with R^2
  - R^2 can be inflated simply by adding more terms to the model (even insignificant terms)

```
Call:
lm(formula = DelTime ~ Ncases + Distance + Gibber, data = DeliveryTime)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6351 -0.7624  0.5539  1.2116  7.3706 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.579657   1.721687   1.498 0.148930    
Ncases       1.610432   0.177172   9.090    1e-08 ***
Distance     0.014470   0.003725   3.885 0.000855 ***
Gibber      -0.449819   2.464269  -0.183 0.856912    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.334 on 21 degrees of freedom
Multiple R-squared:  0.9597, Adjusted R-squared:  0.9539 
F-statistic: 166.5 on 3 and 21 DF,  p-value: 8.52e-15
```

Coefficient of Determination

- Adjusted R^2
  - Penalizes for added terms to the model that are not significant

$$R^2_{adj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p)$$

- For the Delivery Time Data

$$R^2_{adj} = 95.59\%$$

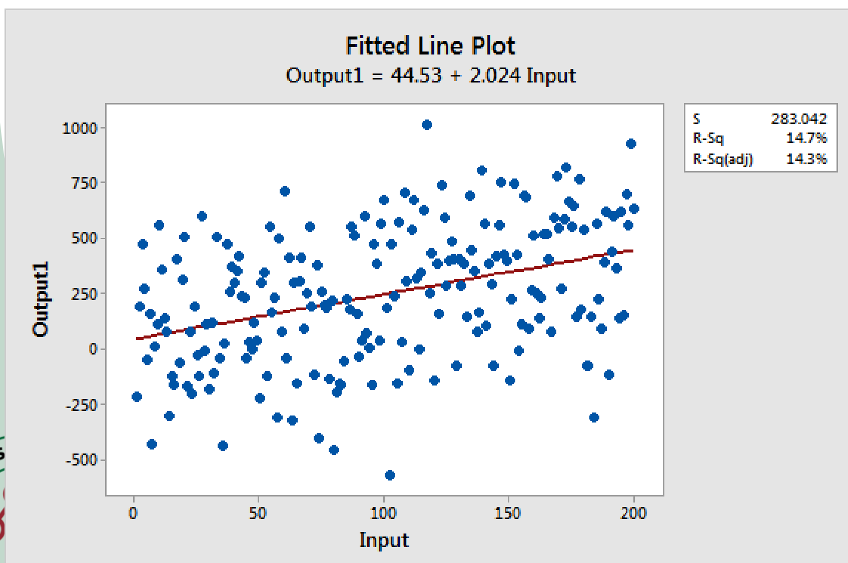
- With Gibberish

$$R^2_{adj} = 95.39\%$$

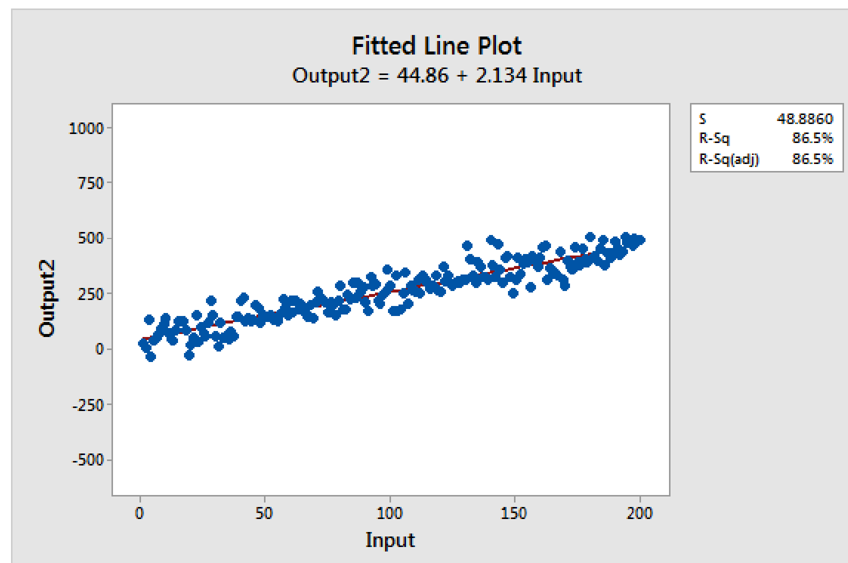
Limitations of R Squared

- **Similarities Between the Regression Models**
  - The two models are nearly identical in several ways:
  - Regression equations: Output = 44 + 2 * Input
  - Input is significant with P < 0.001 for both models

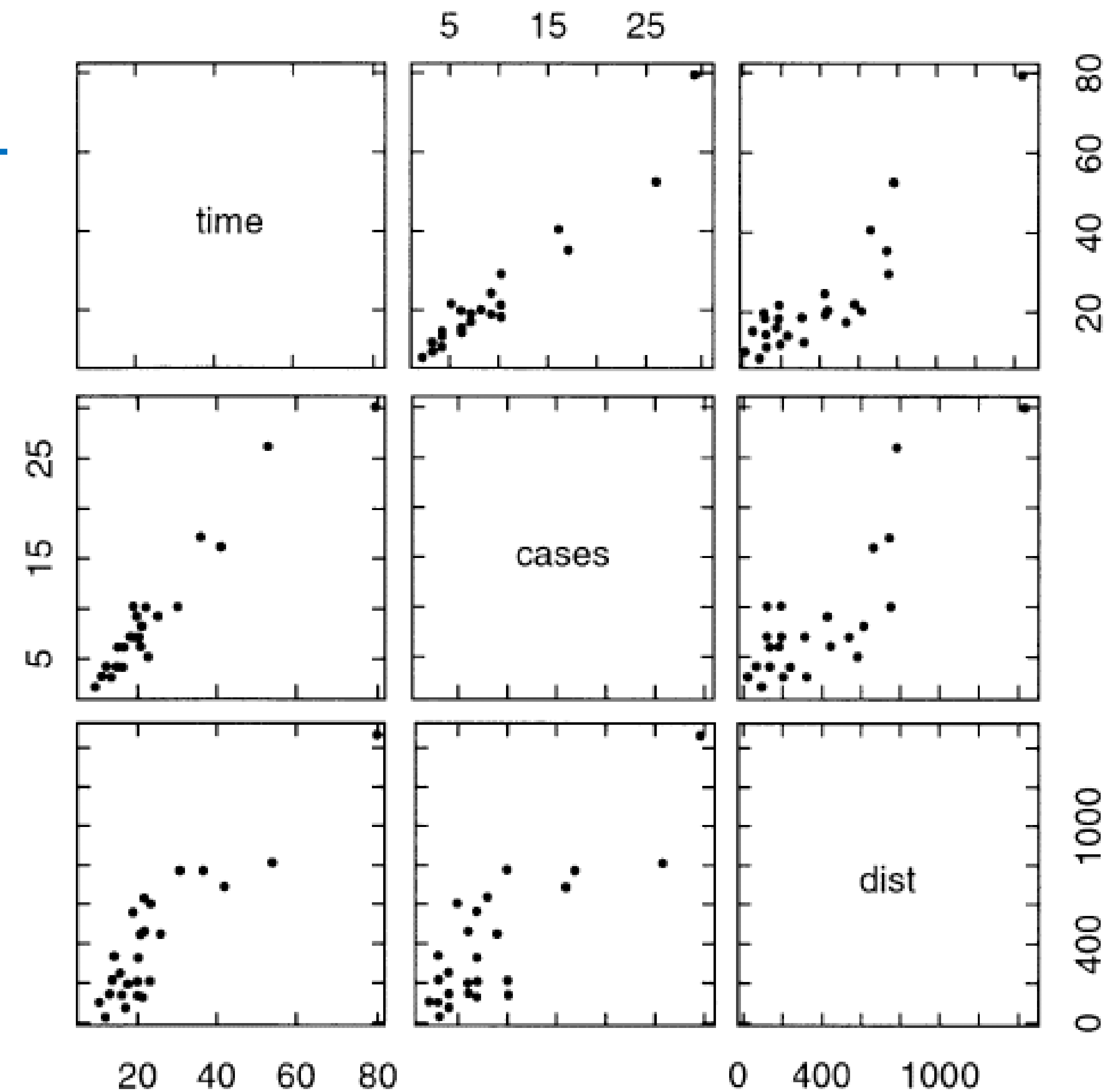
$R^2 = 14.3\%$



$R^2 = 86.5\ \%$



The Delivery Time Data

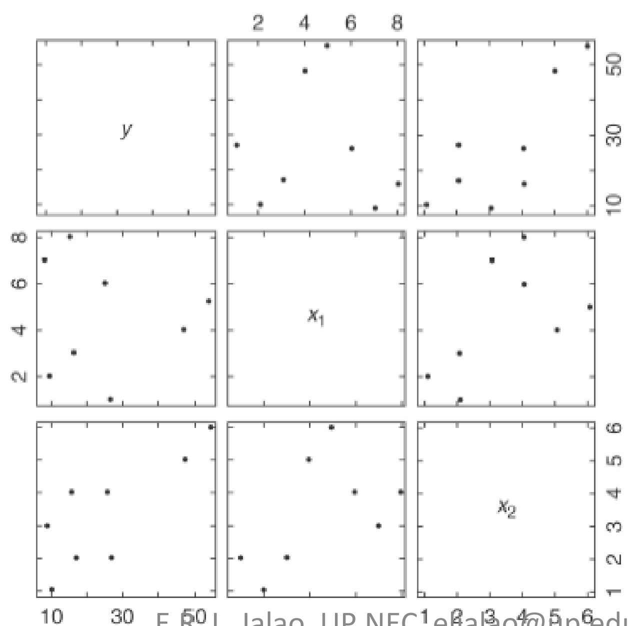


Scatterplot matrix for the delivery time data

Inadequacy of Scatter Diagrams in Multiple Regression

- Scatter diagrams of the regressor variable(s) against the response may be of little value in multiple regression.
  - These plots can actually be misleading
  - If there is an interdependency between two or more regressor variables, the true relationship between xi and y may be masked.

y	x_1	x_2
10	2	1
17	3	2
48	4	5
27	1	2
55	5	6
26	6	4
9	7	3
16	8	4



$$y = 8 - 5x_1 + 12x_2$$

Model Adequacy Checking

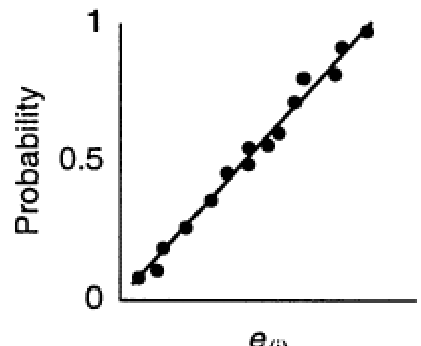
- Assumptions of Linear Regression that must be checked and passed before using the model
  - Relationship between response and regressors is linear (at least approximately).
  - Error term, ε has zero mean
  - Error term, ε has constant variance
  - Errors are uncorrelated
  - Errors are normally distributed (required for tests and intervals)
- Utilize Residual Plots to identify violations

Residual Plots

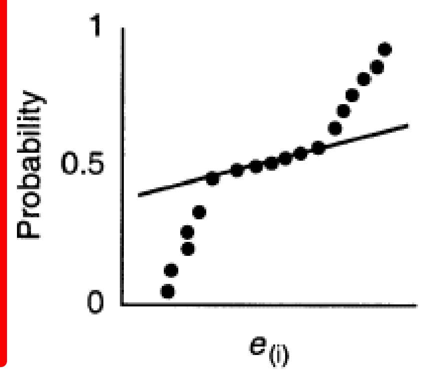
- Normal Probability Plot of Residuals/Q-Q Plot
  - Checks the normality assumption
- Residuals against Fitted values and Scale-Location Plot
  - Checks for nonconstant variance
  - Checks for nonlinearity
  - Looks for potential outliers
- Residuals Versus Leverage
  - Looks for potential outliers

Normal Probability Plot of Residuals

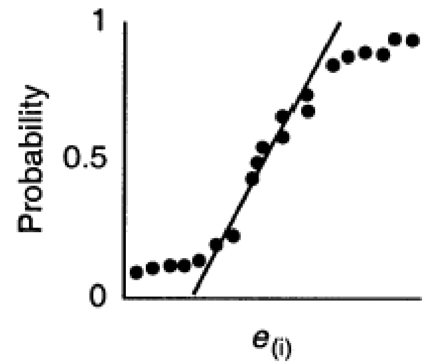
- Checks the normality assumption



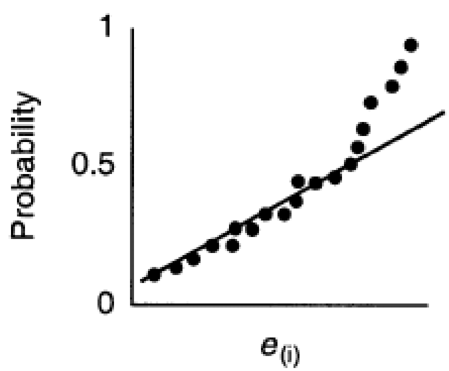
(a)



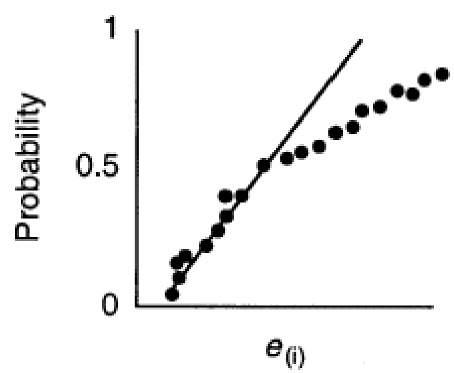
(b)



(c)



(d)

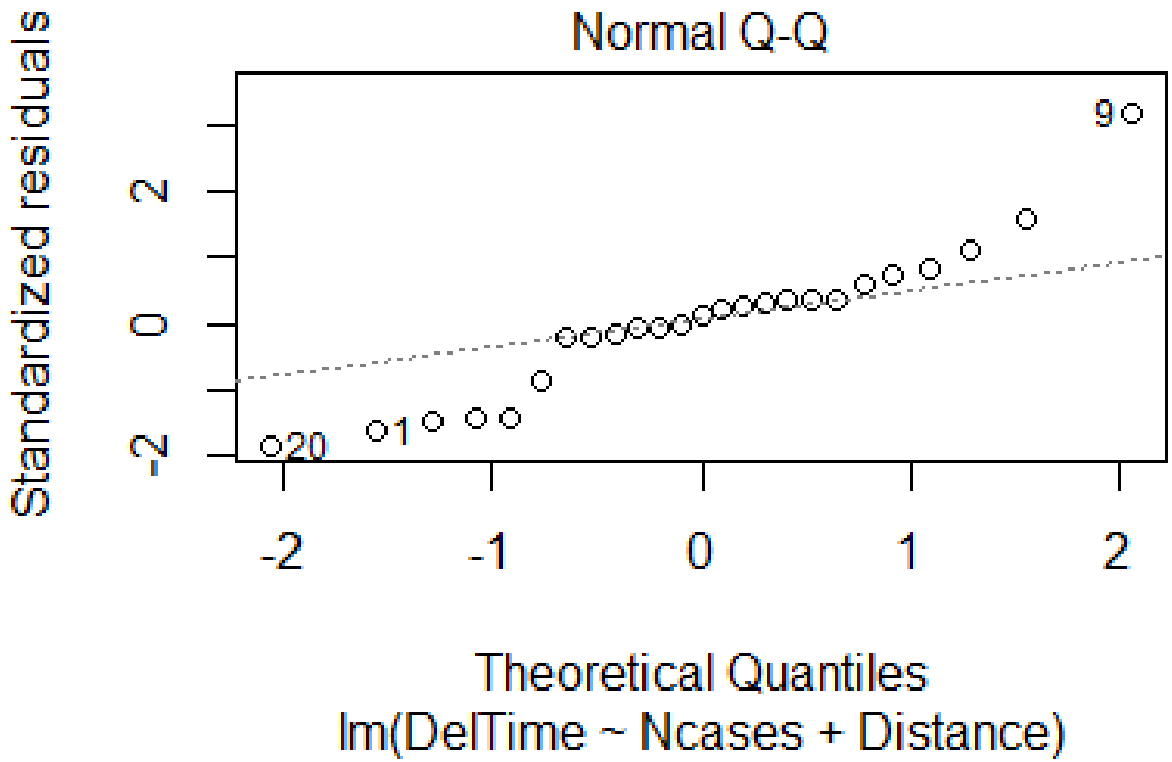


(e)

R Code to Run

```
> par(mfrow =c(2,2),mar=c(2,2,2,2))
> plot(lrfit)
```

Delivery Time Data: Normal Probability Plot

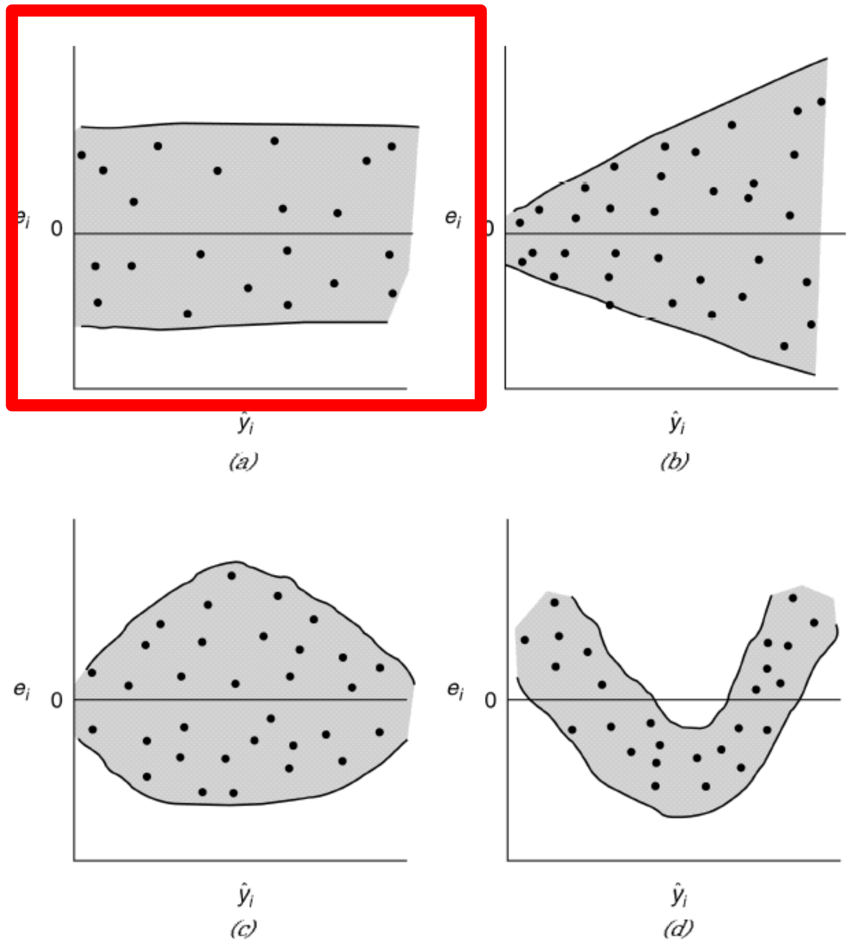


Variance Stabilizing Transformations

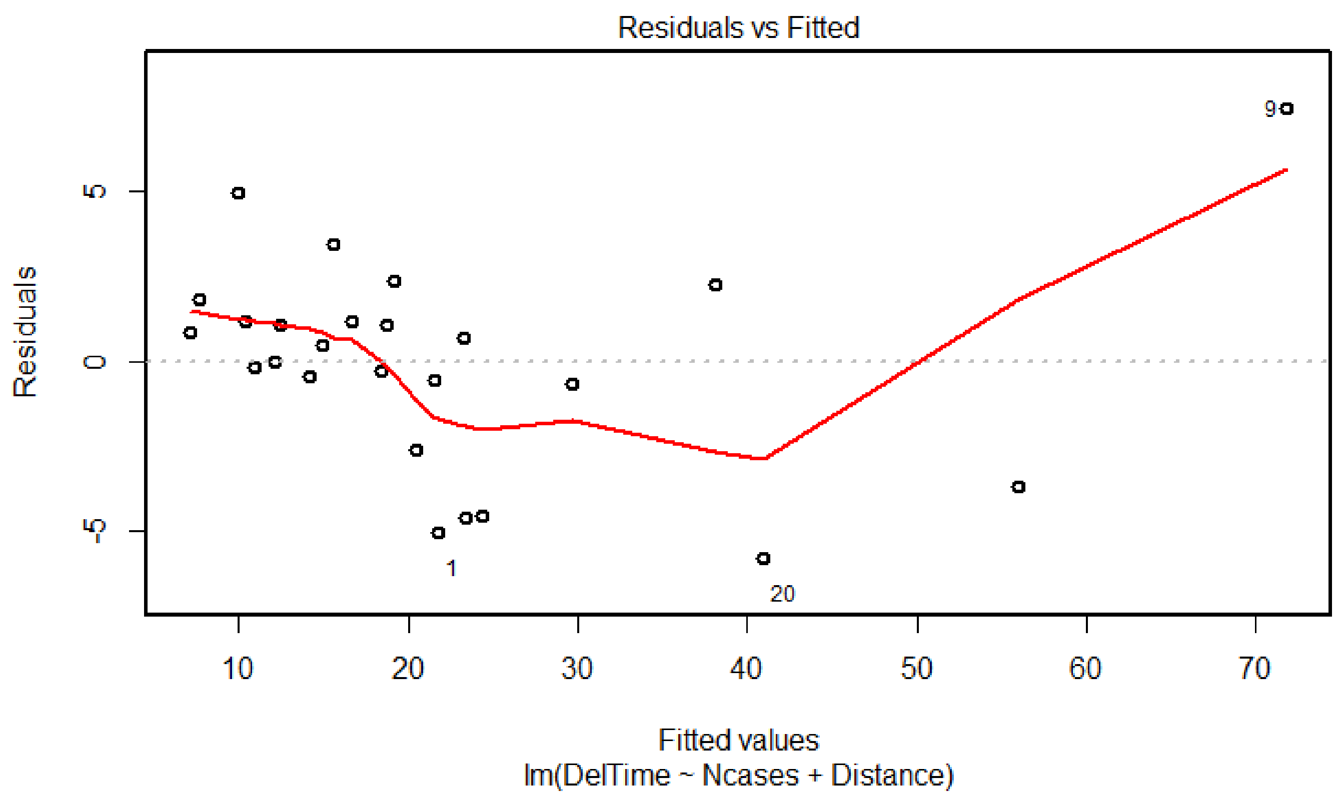
- Constant variance assumption
  - Often violated when the variance is functionally related to the mean.
  - Transformation on the response may eliminate the problem.
  - The strength of the transformation depends on the amount of curvature that is induced.
  - If not satisfied, the regression coefficients will have larger standard errors (less precision)

Residuals Versus Fitted Values Plot

- Checks for
  - Constant Variance Assumption
  - Outliers
  - Non Linearity



Delivery Time Data: Residuals Versus Fits



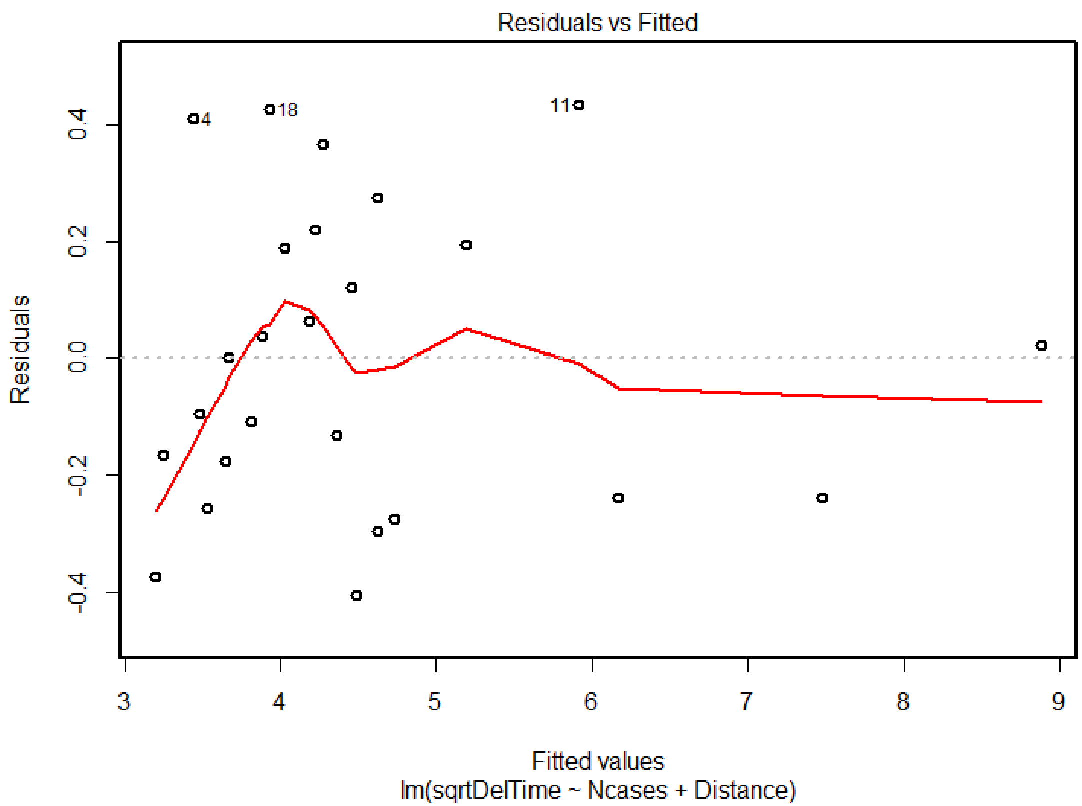
How to Solve?

- Do Transformations on Y

Relationship of σ^2 to $E(y)$	Transformation
$\sigma^2 \propto constant$	$y' = y$ (no transformation)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (square root; Poisson data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}(y)$ (arcsin; binomial proportions $0 \le y_i \le 1$)
$\sigma^2 \propto [E(y)]2$	$y' = \ln(y)$ (log)
$\sigma^2 \propto [E(y)]3$	$y' = y^{-\frac{1}{2}}$ (reciprocal square root)
$\sigma^2 \propto [E(y)]4$	$y' = y^{-1}$ (reciprocal)

Delivery Time Data: Residuals Versus Fits

```
› slrfit=lm(deltime^0.5~ncases+distance,data=deliverytime)
› plot(slrfit)
```



Model Evaluation: Questions

- Is at least one of the predictors, $x_1, x_2, \ldots, x_p$ useful in predicting the response?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
- Are there any outliers that might influence the coefficients?
- Do all the predictors help to explain y , or is only a subset of the predictors useful?

Predictions For New Orders

- Use the generated regression model to predict the mean response
- For delivery time data model is:
  $$\hat{y} = 2.34 + 1.616 \; * Ncases + 0.014 * Distance$$
- Using the Delivery Time Data For 2 Cases, 110 Feet Delivery Distance
  - Average Estimated Del Time: 7.15 Mins.
- For 10 Cases, 140 Feet Delivery Distance:
  - Average Estimated Del Time: 56.01 Mins.

R Code To Run

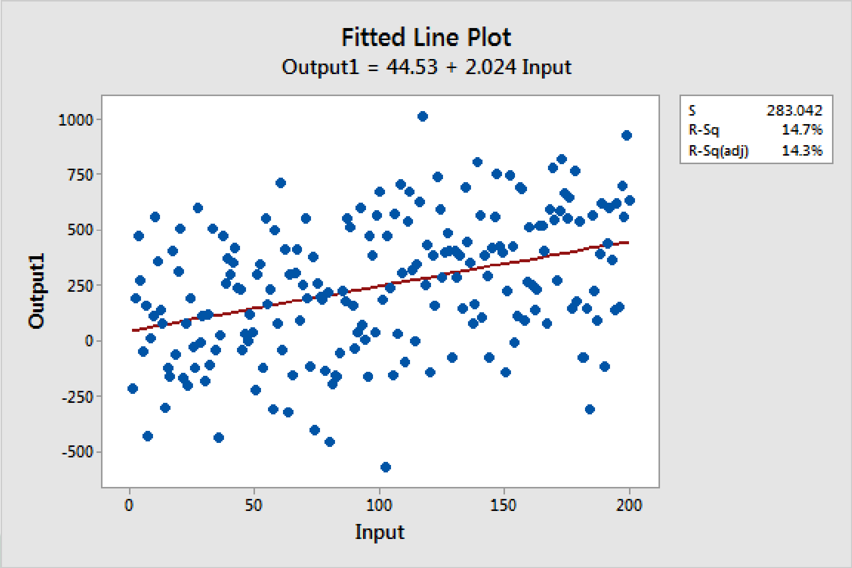
- ```
  deliverytimenewdata = read.csv(“deliverytimendata.csv")
  ```
- ```
  predict(lrfit, deliverytimenewdata , interval="confidence")
  ```

Confidence Intervals

- We use a confidence interval to quantify the uncertainty surrounding the average response
- Using the Delivery Time Data For 2 Cases, 110 Feet Delivery Distance
  - Average Estimated Del Time: 7.15 Mins.
  - Lower Limit: 5.22 Mins, Upper Limit: 9.08 Mins.
  - **Difference of ± 1.93**
- For 10 Cases, 140 Feet Delivery Distance:
  - Average Estimated Del Time: 20.51 Mins.
  - Lower Limit: 17.76 Mins. Upper Limit: 23.26 Mins.
  - **Difference of ± 2.75**

Recall

$R^2 = 14.3\%$



Prediction for Output1

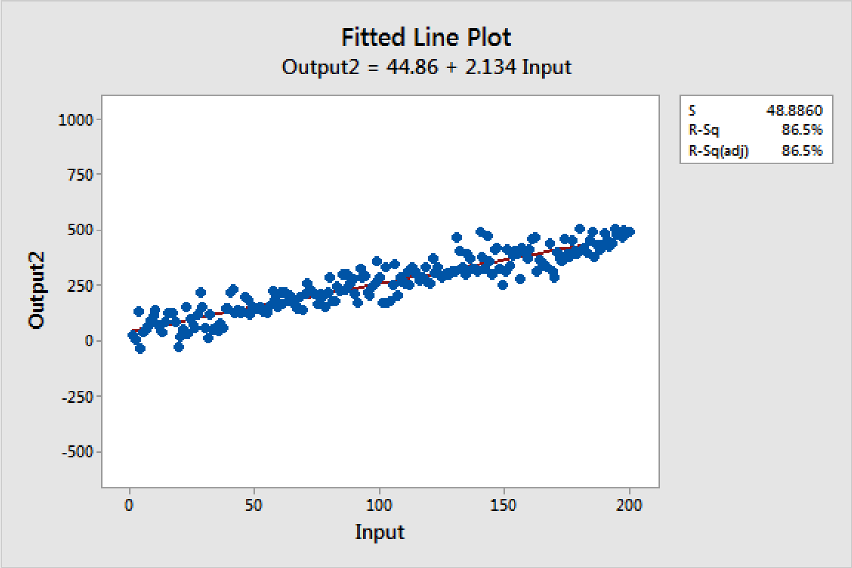
Regression Equation

Output1 = 44.5 + 2.024 Input

Variable	Setting
Input	10

Fit	SE Fit	95% CI	95% PI
64.7766	37.2129	(-8.60793, 138.161)	(-498.190, 627.743)

$R^2 = 86.5\ \%$



Prediction for Output2

Regression Equation

Output2 = 44.86 + 2.1343 Input

Variable	Setting
Input	10

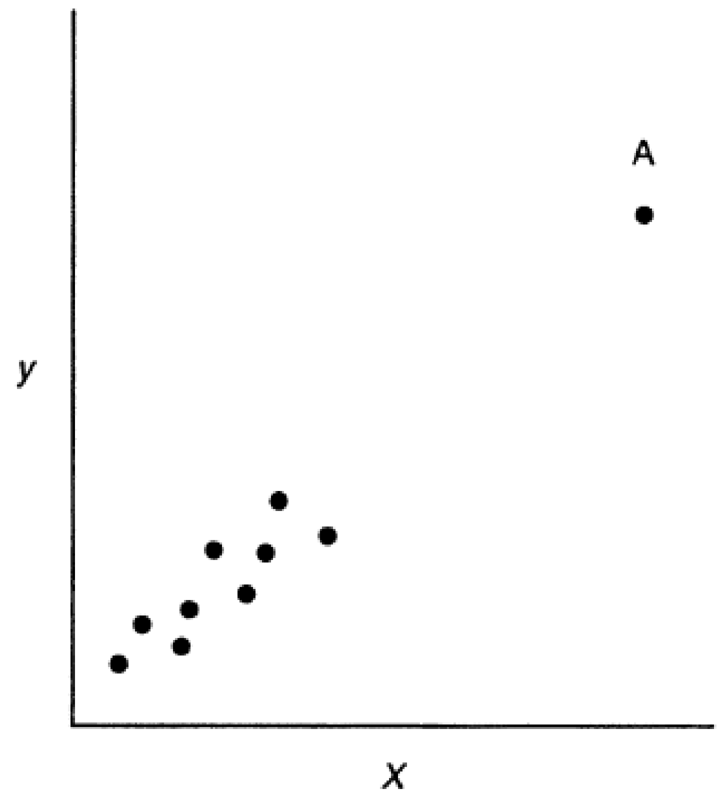
Fit	SE Fit	95% CI	95% PI
66.2076	6.42728	(53.5329, 78.8823)	(-31.0260, 163.441)

Model Evaluation: Questions

- Is at least one of the predictors, $x_1, x_2, \ldots, x_p$ useful in predicting the response?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
- Are there any outliers that might influence the coefficients?
- Do all the predictors help to explain y , or is only a subset of the predictors useful?

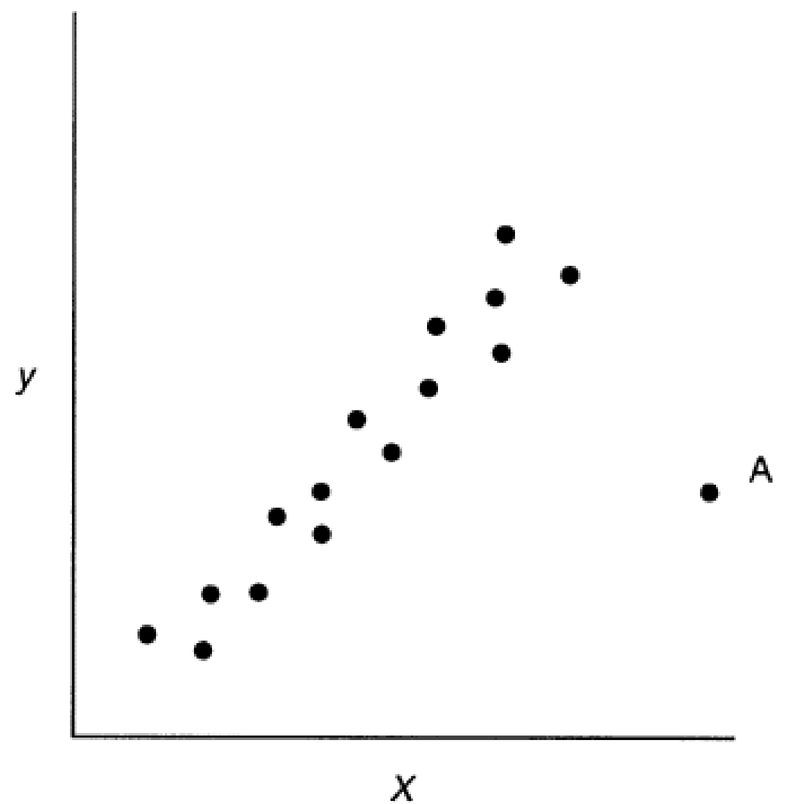
Importance of Detecting Influential Observations

- Leverage Point:
  - unusual x-value;
  - very little effect on regression coefficients.



Importance of Detecting Influential Observations

- Influence Point: unusual in y and x;



The Leverage Statistic

- h_i – **standardized measure** of the distance of the i^{th} observation from the center of the x-space.
- For simple regression

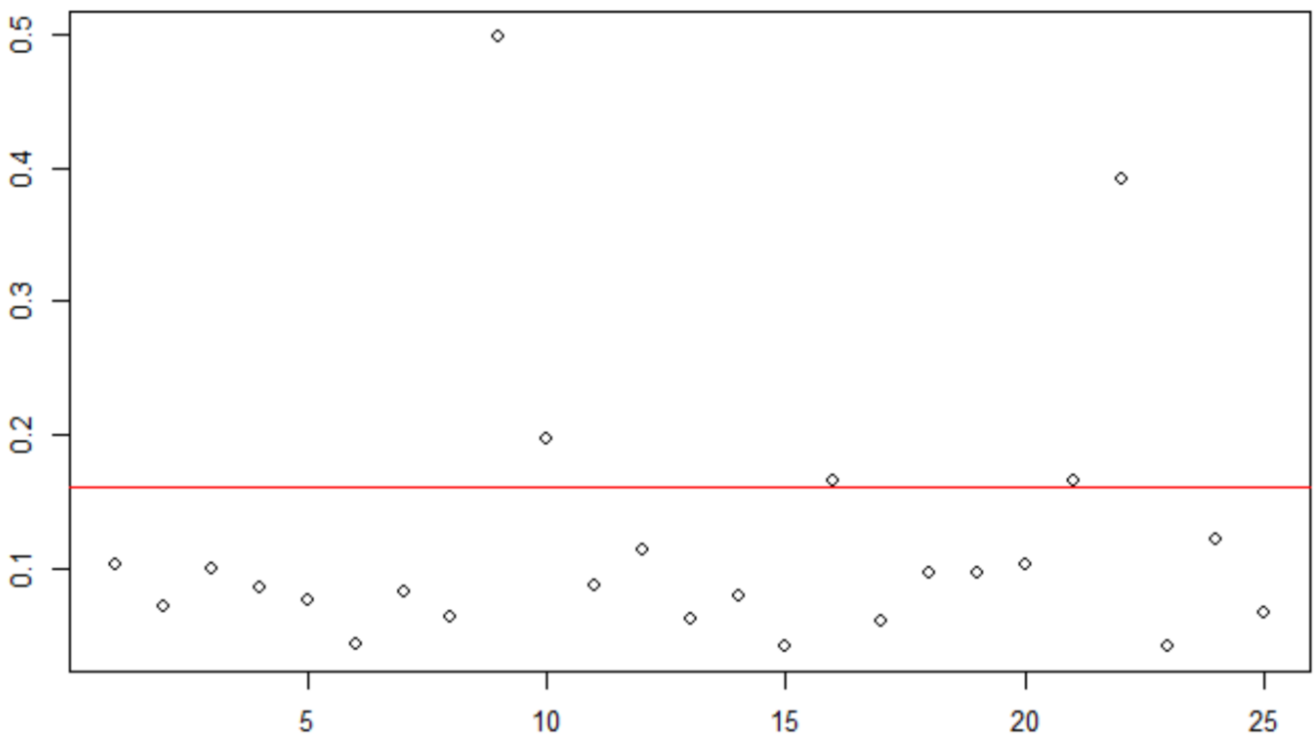
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- h_i increases with the distance of x_i from $\bar{x}$.
- If a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then that point is considered to be a leverage point.

Delivery Time Data

```
> plot(hatvalues(lrfit))
> abline(h=4/25, col="red")
```

$$Cutoff = \frac{(p+1)}{n} = \frac{4}{25} = 0.16$$



Outlier Detection: Studentized Residuals

- The plain residual ε_i and its plot is useful for checking how well the regression line fits the data, and in particular if there is any systematic lack of fit
- But, what value should be considered as a big residual?
  - ε_i retains the scale of the response variable.
  - standardize by an estimate of the variance of the residual.

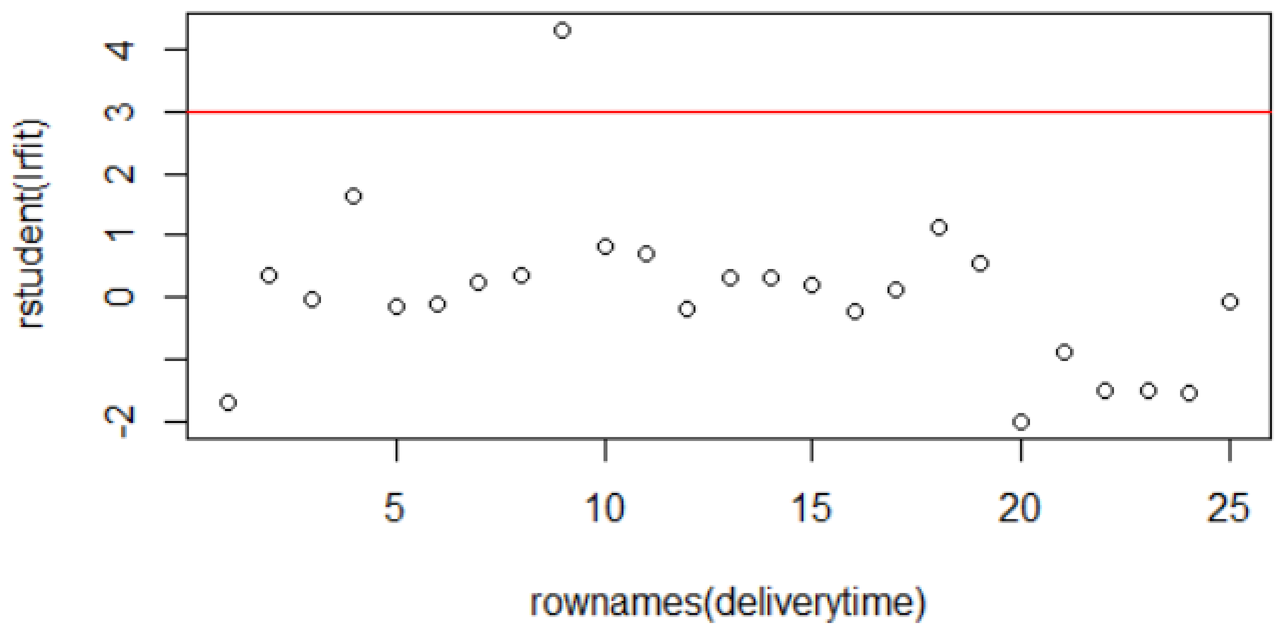
$$S_i = \frac{\varepsilon_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

- Observations whose studentized residuals are greater than 3 in absolute value are possible outliers

Delivery Time Data

```
› plot(rownames(deliverytime),
  rstudent(lrfit))
› abline(h=3, col="red")
› rstudent(lrfit)
```

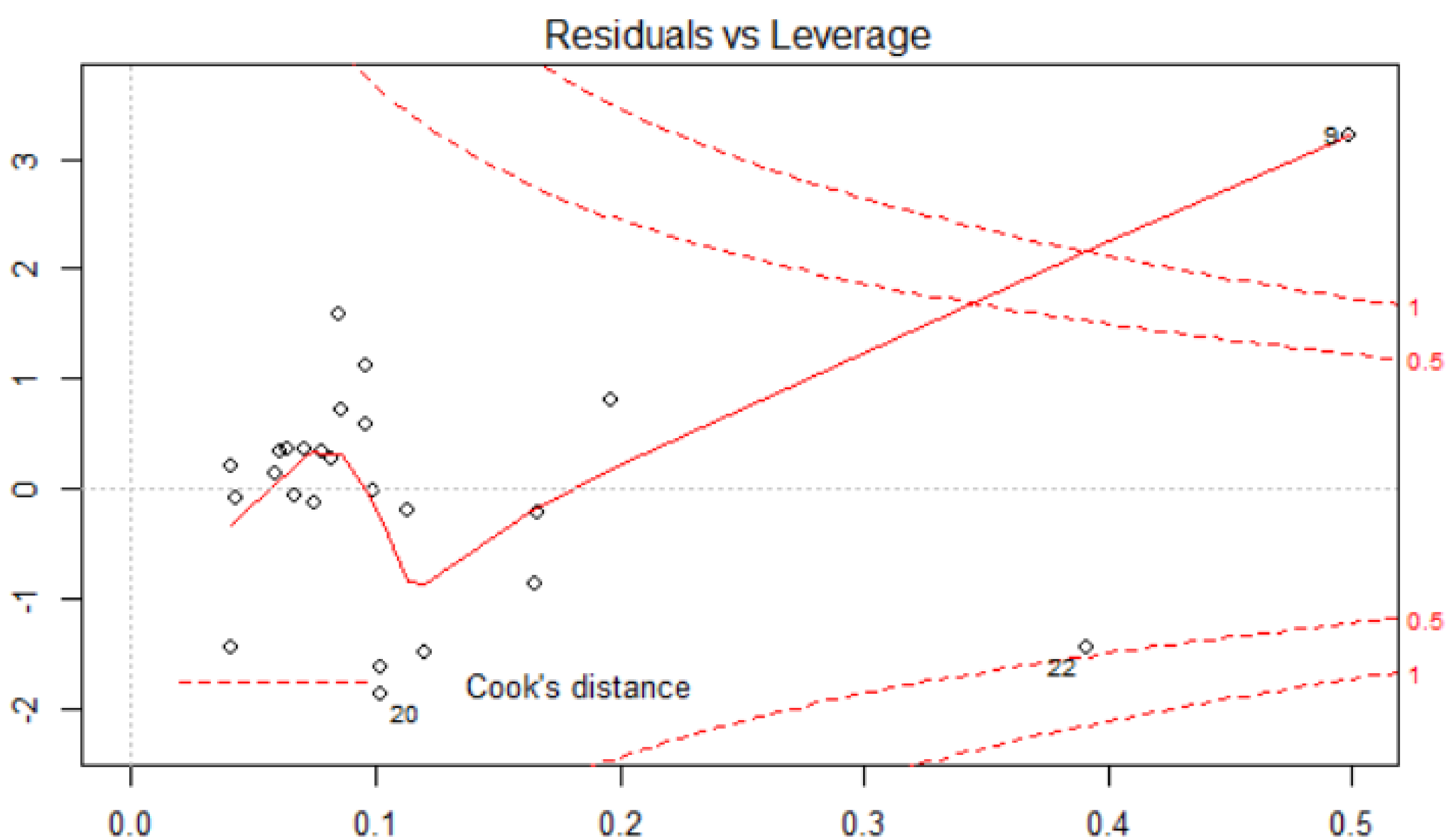
$Cutoff = \pm 3$



Row Values

```
> rstudent(lrfit)
          1           2           3           4           5           6           7 
-1.69562881  0.35753764 -0.01572177  1.63916491 -0.13856493 -0.08873728  0.26464769 
          8           9          10          11          12          13          14 
 0.35938983  4.31078012  0.80677584  0.70993906 -0.18897451  0.31846924  0.33417725 
         15          16          17          18          19          20          21 
 0.20566324 -0.21782566  0.13492400  1.11933065  0.56981420 -1.99667657 -0.87308697 
         22          23          24          25 
-1.48962473 -1.48246718 -1.54221512 -0.06596332 
```

Residuals Versus Leverage Plot



Model Evaluation: Questions

- Is at least one of the predictors, $x_1, x_2, \ldots, x_p$ useful in predicting the response?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
- Are there any outliers that might influence the coefficients?
- Do all the predictors help to explain y , or is only a subset of the predictors useful?

This Session's Outline

- Multiple Linear Regression
- Model Evaluation
- Variable Selection and Model Building
  - Best Subsets Regression
  - Stepwise Regression
  - Ridge Regression
  - Standardized Regression
- Indicator Variables
- Multicollinearity
- Logistic Regression
- Case Study

R Code To Run

```
› cardata= read.csv("cars.csv")
› rownames(cardata) =cardata[,1]
› cardata =cardata[,c(2:12)]
› mpglrfit= lm(mpg~.,data=cardata)
› summary(mpglrfit)
```

t-Test Using T Table

- If P value of variable x_i is (> 0.05) the variable in question is no longer needed since there are other variables already in the model that provides the same information as x_i

```
Call:
lm(formula = mpg ~ ., data = Car)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4506 -1.6044 -0.1196  1.2193  4.6271 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 12.30337   18.71788   0.657   0.5181  
cyl         -0.11144    1.04502  -0.107   0.9161  
disp         0.01334    0.01786   0.747   0.4635  
hp          -0.02148    0.02177  -0.987   0.3350  
drat         0.78711    1.63537   0.481   0.6353  
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739  
vs           0.31776    2.10451   0.151   0.8814  
am           2.52023    2.05665   1.225   0.2340  
gear         0.65541    1.49326   0.439   0.6652  
carb        -0.19942    0.82875  -0.241   0.8122  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869, Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

t-Test Using T Table

- However, it does not follow that if x_1 is not needed in a model that contains all other variables, it is not needed at all.

```
Call:
lm(formula = mpg ~ disp, data = Car)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8922 -2.2022 -0.9631  1.6272  7.2305 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
disp        -0.041215   0.004712  -8.747 9.38e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.251 on 30 degrees of freedom
Multiple R-squared:  0.7183,	Adjusted R-squared:  0.709 
F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

Variable Selection

- How to select the best model from multiple alternative Regression Models?
  - Concept of Overfitting and Underfitting
- All Possible Regressions
  - Assume the intercept term is in all equations considered. Then, if there are k regressors, we would investigate $2^k - 1$ possible regression equations.
  - Use the some criteria to determine some candidate models and complete regression analysis on them.

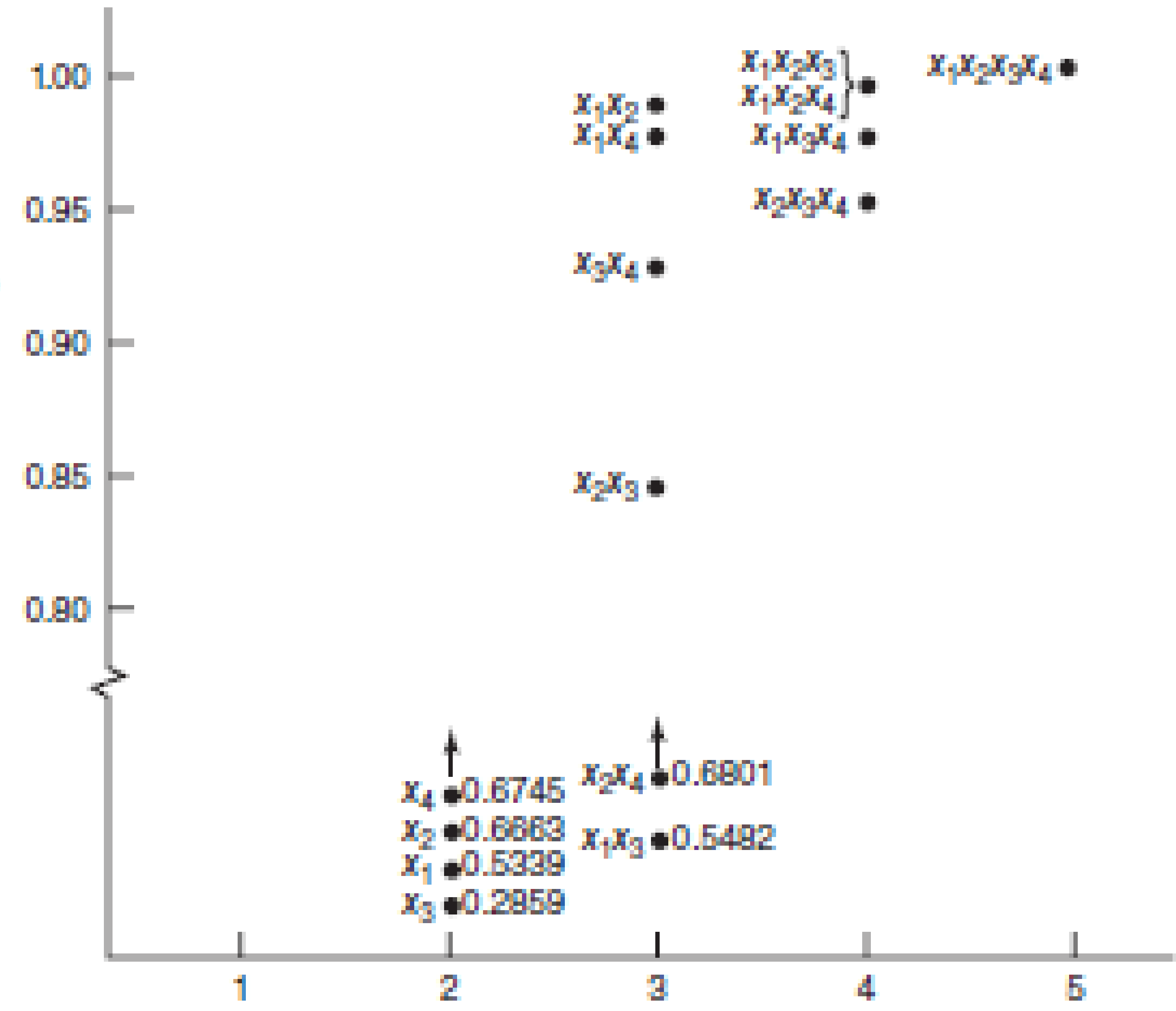
Hald Cement Data: Raw Data

Observation i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Hald Cement Data: All Possible Regressions

Variables in Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
x_1	81.479	1.869			
x_2	57.424		0.789		
x_3	110.203			−1.256	
x_4	117.568				−0.738
x_1x_2	52.577	1.468	0.662		
x_1x_3	72.349	2.312		0.494	
x_1x_4	103.097	1.440			−0.614
x_2x_3	72.075		0.731	−1.008	
x_2x_4	94.160		0.311		−0.457
x_3x_4	131.282			−1.200	−0.724
$x_1x_2x_3$	48.194	1.696	0.657	0.250	
$x_1x_2x_4$	71.648	1.452	0.416		−0.237
$x_2x_3x_4$	203.642		−0.923	−1.448	−1.557
$x_1x_3x_4$	111.684	1.052		−0.410	−0.643
$x_1x_2x_3x_4$	62.405	1.551	0.510	0.102	−0.144

Hald Cement Data: Size Versus R^2



Criteria for Evaluating Subset Regression Models

- Coefficient of Multiple Determination (R^2 and R^2_{adj})
- Mean Square Error
- AIC

R^2

- Say we are investigating a model with p terms,

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{\mathrm{Res}}(p)}{SS_T}$$

- Models with large values of R_p^2 are preferred, but adding terms will increase this value.

Adjusted R^2

- Say we are investigating a model with p terms,

$$R^2_{adj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p)$$

- This value will not necessarily increase as additional terms are introduced into the model.
- We want a model with the maximum adjusted R^2_{adj}

Residual Mean Square

- The MS_{res} for a subset regression model is

$$MS_{\mathrm{Re}\,s}(p) = \frac{SS_{\mathrm{Re}\,s}(p)}{n-p}$$

- $MS_{Res}(p)$ increases as p increases, in general.
- We want a model with a minimum $MS_{Res}(p)$.

Hald Cement Data

Number of Regressors in Model	p	Regressors in Model	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$
None	1	None	2715.7635	0	0	226.3136
1	2	x_1	1265.6867	0.53395	0.49158	115.0624
1	2	x_2	906.3363	0.66627	0.63593	82.3942
1	2	x_3	1939.4005	0.28587	0.22095	176.3092
1	2	x_4	883.8669	0.67459	0.64495	80.3515
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829

Akaike Information Criterion

- AIC is based on maximizing the expected entropy of the model. In case of OLS regression:

$$AIC = n \ln\left(\frac{SS_{Res}}{N}\right) + 2p$$

- The key insight to the AIC is similar to R^2_{adj}. As we add regressors to the model, SS_{Res} cannot increase.
- The issue whether the decrease in SS_{Res} justifies the inclusion of the extra terms
- We want a model with the lowest AIC

Computational Techniques for Variable Selection

- All Possible Regressions
- Step-Wise Regression

66

All Possible Regressions

- Once some candidate models have been identified, run regression analysis on each one individually and make comparisons
- Computationally expensive
- Recommended maximum ~ 15 variables = 32,768 Comparisons!

Hald Cement Data

Number of Regressors in Model	p	Regressors in Model	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

Stepwise Regression

- A heuristic methodology to select significant variables for a regression model
  - Starts with **no variables** in the model
  - Regressor variables are added one at a time starting with the variable with the **highest correlation** to y.
  - A regressor that makes it into the model, may also be removed it if is found to be **insignificant** with the addition of other variables to the model.

R Code to Run

```
› carbasefit =lm(mpg~1, data= cardata)
› Stepwise= step(carbasefit, scope =
  list(lower=~1,upper=~cyl+disp+hp+drat+w
  t+qsec+vs+am+gear+carb, direction =
  "both", trace=1))
```

Results

```
Start:  AIC=115.94
mpg ~ 1

       Df Sum of Sq     RSS     AIC
+ wt    1    847.73  278.32  73.217
+ cyl   1    817.71  308.33  76.494
+ disp  1    808.89  317.16  77.397
+ hp    1    678.37  447.67  88.427
+ drat  1    522.48  603.57  97.988
+ vs    1    496.53  629.52  99.335
+ am    1    405.15  720.90 103.672
+ carb  1    341.78  784.27 106.369
+ gear  1    259.75  866.30 109.552
+ qsec  1    197.39  928.66 111.776
<none>              1126.05 115.943

Step:  AIC=73.22
mpg ~ wt

       Df Sum of Sq     RSS     AIC
+ cyl   1     87.15  191.17  63.198
+ hp    1     83.27  195.05  63.840
+ qsec  1     82.86  195.46  63.908
+ vs    1     54.23  224.09  68.283
+ carb  1     44.60  233.72  69.628
+ disp  1     31.64  246.68  71.356
<none>               278.32  73.217
+ drat  1      9.08  269.24  74.156
+ gear  1      1.14  277.19  75.086
+ am    1      0.00  278.32  75.217
- wt    1    847.73 1126.05 115.943
```

```
Step:  AIC=63.2
mpg ~ wt + cyl

       Df Sum of Sq    RSS    AIC
+ hp    1    14.551 176.62 62.665
+ carb  1    13.772 177.40 62.805
<none>              191.17 63.198
+ qsec  1    10.567 180.60 63.378
+ gear  1     3.028 188.14 64.687
+ disp  1     2.680 188.49 64.746
+ vs    1     0.706 190.47 65.080
+ am    1     0.125 191.05 65.177
+ drat  1     0.001 191.17 65.198
- cyl   1    87.150 278.32 73.217
- wt    1   117.162 308.33 76.494

Step:  AIC=62.66
mpg ~ wt + cyl + hp

       Df Sum of Sq    RSS    AIC
<none>              176.62 62.665
- hp    1    14.551 191.17 63.198
+ am    1     6.623 170.00 63.442
+ disp  1     6.176 170.44 63.526
- cyl   1    18.427 195.05 63.840
+ carb  1     2.519 174.10 64.205
+ drat  1     2.245 174.38 64.255
+ qsec  1     1.401 175.22 64.410
+ gear  1     0.856 175.76 64.509
+ vs    1     0.060 176.56 64.654
- wt    1   115.354 291.98 76.750
```

Final Reduced Model

```
> carfinalfit = lm(mpg~wt + cyl + hp, data=cardata)
> summary(carfinalfit)
```

```
Call:
lm(formula = mpg ~ wt + cyl + hp, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9290 -1.5598 -0.5311  1.1850  5.8986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .  
hp          -0.01804    0.01188  -1.519 0.140015    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431, Adjusted R-squared:  0.8263 
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

As Compared to the Full Model

```
Call:
lm(formula = mpg ~ ., data = Car)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4506 -1.6044 -0.1196  1.2193  4.6271 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 12.30337   18.71788   0.657   0.5181  
cyl         -0.11144    1.04502  -0.107   0.9161  
disp         0.01334    0.01786   0.747   0.4635  
hp          -0.02148    0.02177  -0.987   0.3350  
drat         0.78711    1.63537   0.481   0.6353  
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739  
vs           0.31776    2.10451   0.151   0.8814  
am           2.52023    2.05665   1.225   0.2340  
gear         0.65541    1.49326   0.439   0.6652  
carb        -0.19942    0.82875  -0.241   0.8122  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869, Adjusted R-squared:  0.8066 
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Cautions

- No one model may be the "best"
- The techniques could result in different models
- Greedy Algorithm is used
- Inexperienced analysts may use the final model simply because the procedure spit it out.
- Needs lots of common sense.

Unit Normal Scaling

- Employs unit normal scaling for the regressors and the response variable. That is,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad for\ i = 1,2, \dots, n, \qquad j = 1,2 \dots, k$$

$$y_i^* = \frac{y_i - \bar{y}}{s_y}, \qquad for\ i = 1,2, \dots, n$$

- Where:

$$s_j^2 = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}\ \right)}{n-1}, s_y = \frac{\sum_{i=1}^{n} (y_i - \bar{y}\ )}{n-1}$$

Unit Normal Scaling

- All of the scaled regressors and the scaled response have sample mean equal to zero and sample variance equal to 1.
- The model becomes

$$y_i^* = \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik} + \epsilon$$

R Code to Run

```
› options(scipen=100)
› scardata = data.frame(scale(cardata,
  center = TRUE, scale = TRUE))
› scarfinalfit = lm(mpg~., data=scardata)
› summary(scarfinalfit)
```

Standardized R Coefficients

```
Call:
lm(formula = mpg ~ ., data = scardata)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5725 -0.2662 -0.0198  0.2023  0.7677 

Coefficients:
                          Estimate             Std. Error t value Pr(>|t|)
(Intercept) -0.0000000000000000296  0.0777330530182089585    0.00    1.000
cyl         -0.0330223456522466055  0.3096641664307349661   -0.11    0.916
disp         0.2742270553028448576  0.3672232189324069473    0.75    0.463
hp          -0.2443816814736877451  0.2476404680450327855   -0.99    0.335
drat         0.0698282938803363034  0.1450815898476484067    0.48    0.635
wt          -0.6031687597474482132  0.3075526378299181518   -1.96    0.063
qsec         0.2434321984378815806  0.2166897994811803340    1.12    0.274
vs           0.0265735795447262813  0.1759939311328732325    0.15    0.881
am           0.2086579003538392707  0.1702768859539114665    1.23    0.234
gear         0.0802340395579890508  0.1828011889671173895    0.44    0.665
carb        -0.0534436290472993166  0.2221026304107495130   -0.24    0.812
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4 on 21 degrees of freedom
Multiple R-squared:  0.869, Adjusted R-squared:  0.807 
F-statistic: 13.9 on 10 and 21 DF,  p-value: 0.000000379
```

This Session's Outline

- Multiple Linear Regression
- Model Evaluation
- Variable Selection and Model Building
  - Best Subsets Regression
  - Stepwise Regression
  - Ridge Regression
  - Standardized Regression
- **Indicator Variables**
- Multicollinearity
- Logistic Regression
- Case Study

Indicator Variables

- How to do we handle Qualitative Variables?
  - Red
  - Green
  - Blue
- Qualitative variables do not have a scale of measurement.
- We cannot assign numerical values as follows
  - Red = 1
  - Green =2
  - Blue =3
- Indicator variables – a variable that assigns levels to the qualitative variable (also known as dummy variables).

Example

- We like to relate the effective life of a cutting tool (y) used on a lathe to the lathe speed in revolutions per minute (x_1) and type of cutting tool used.

hours	rpm	tooltype
18.73	610	A
14.52	950	A
17.43	720	A
14.54	840	A
13.44	980	A
24.39	530	A
13.34	580	A
22.71	540	A
12.68	890	A
19.32	730	A
30.16	670	B
27.09	770	B
25.4	880	B
26.05	1000	B
33.49	760	B
35.62	590	B
26.07	910	B
36.78	650	B
34.95	810	B
43.67	500	B

Indicator Variables

- Tool type is qualitative and can be represented as:

$$x_2 = \begin{cases} 0 & ToolA \\ 1 & ToolB \end{cases}$$

- The regression model would be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Dataset With Indicator Variables

hours	rpm	tooltype	x2
18.73	610	A	0
14.52	950	A	0
17.43	720	A	0
14.54	840	A	0
13.44	980	A	0
24.39	530	A	0
13.34	580	A	0
22.71	540	A	0
12.68	890	A	0
19.32	730	A	0
30.16	670	B	1
27.09	770	B	1
25.4	880	B	1
26.05	1000	B	1
33.49	760	B	1
35.62	590	B	1
26.07	910	B	1
36.78	650	B	1
34.95	810	B	1
43.67	500	B	1

Example

- If Tool type A is used, model becomes:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- If Tool type B is used, model becomes:

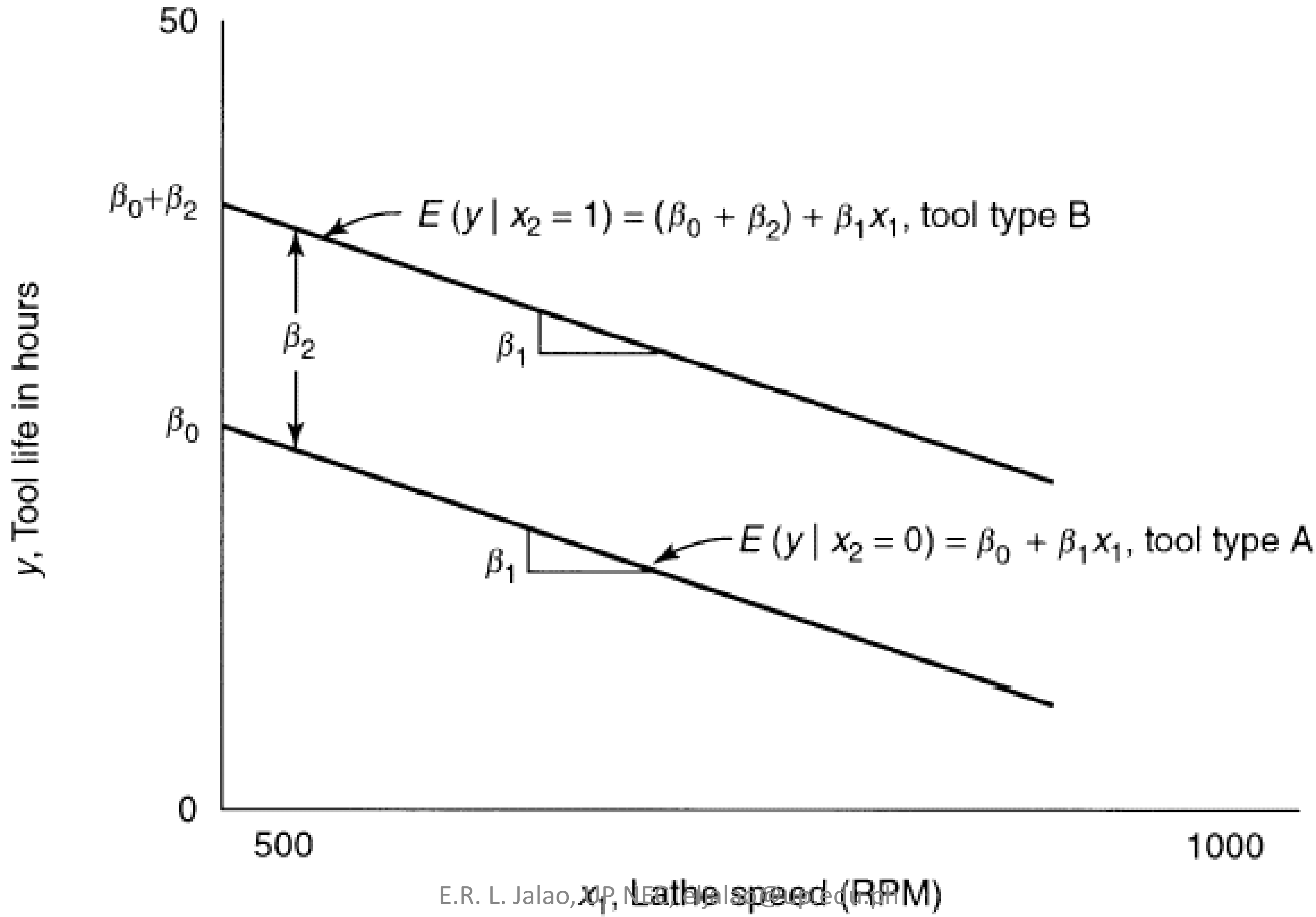
$$y = \beta_0 + \beta_1 x_1 + \beta_2 + \varepsilon$$

  – Then:

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

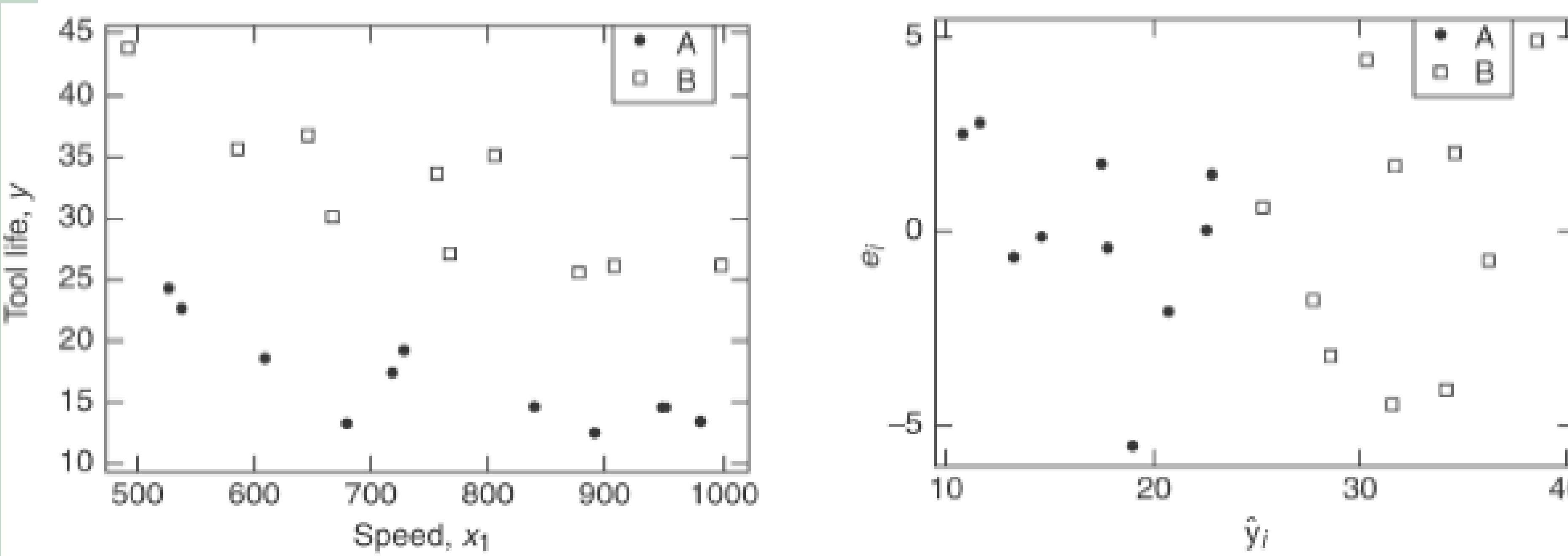
- Changing from A to B induces a change in the intercept (slope is unchanged and identical).
- We assume that the variance is equal for all levels of the qualitative variable.

Example



50
$\beta_0+\beta_2$
$E(y \mid x_2 = 1) = (\beta_0 + \beta_2) + \beta_1 x_1$, tool type B
β_2
β_1
β_0
$E(y \mid x_2 = 0) = \beta_0 + \beta_1 x_1$, tool type A
β_1
y, Tool life in hours
0
500
1000
x_1, Lathe speed (RPM)

Tool Life Data



Tool Life Data

```
› toollife = read.csv("toollife.csv")
› toollifefit=lm(hours~rpm+tooltype,data=toollife)
› summary(toollifefit)
Call:
lm(formula = Hours ~ RPM + ToolTypeB, data = ToolLife)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6255 -1.6308  0.0612  2.2218  5.5044 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.208726   3.738882   9.417 3.71e-08 ***
RPM         -0.024557   0.004865  -5.048 9.92e-05 ***
ToolTypeB   15.235474   1.501220  10.149 1.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.352 on 17 degrees of freedom
Multiple R-squared:  0.8787, Adjusted R-squared:  0.8645 
F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08
```

For Three More Levels

- For qualitative variables with a levels (specific categorical values), we would need $a-1$ indicator variables.
- For example, say there were three tool types, A, B, and C. Then two indicator variables (called x_2 and x_3) will be needed:

x_2	x_3	
0	0	if the observation is from tool type A
1	0	if the observation is from tool type B
0	1	if the observation is from tool type C

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Difference in Slope

- If we expect the slopes to differ, we can model this phenomenon by including an interaction term between the variables.
- Consider the tool life data again, and say we believe there may be different slopes for the two tools. The model we can fit to account for the change in slope is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

The Tool Life Data With Interactions

```
› toollifefit=lm(hours~rpm+tooltype+rpm*tooltype,data
  =toollife)
› summary(toollifefit)
```

```
Call:
lm(formula = Hours ~ RPM + ToolType + ToolType * RPM, data = ToolLife)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5534 -1.7088  0.3283  2.0913  4.8652 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   30.176013   4.724895   6.387 9.01e-06 ***
RPM           -0.017729   0.006262  -2.831  0.01204 *  
ToolTypeB     26.569340   7.115681   3.734  0.00181 ** 
RPM:ToolTypeB -0.015186   0.009338  -1.626  0.12345    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.201 on 16 degrees of freedom
Multiple R-squared:  0.8959, Adjusted R-squared:  0.8764 
F-statistic: 45.92 on 3 and 16 DF,  p-value: 4.37e-08
```

More than Two Indicator Variables

- Suppose that in the tool life data, a second qualitative factor, the type of cutting oil used, must be considered.
- Assuming that this factor has two levels, we may define a second indicator variable, x_3, as follows:

$$x_3 = \begin{cases} 0 & if\ low\ viscosity\ oil\ is\ used \\ 1 & if\ medium\ viscosity\ oil\ is\ used \end{cases}$$

More than Two Indicator Variables With Interactions

- Suppose that we consider interactions between cutting speed and the two qualitative factors.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$$

- Hence we can have the following models

Tool Type	Cutting Oil	Regression Model
A	Low viscosity	$y = \beta_0 + \beta_1 x_1 + \varepsilon$
B	Low viscosity	$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \varepsilon$
A	Medium viscosity	$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \varepsilon$
B	Medium viscosity	$y = (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \beta_4 + \beta_5) x_1 + \varepsilon$

This Session's Outline

- Multiple Linear Regression
- Model Evaluation
- Variable Selection and Model Building
  - Best Subsets Regression
  - Stepwise Regression
  - Ridge Regression
  - Standardized Regression
- Indicator Variables
- **Multicollinearity**
- Logistic Regression
- Case Study

Introduction

- Multicollinearity: the inflation of coefficient estimates due to interdependent regressors
- If all regressors are orthogonal (independent), with each other then multicollinearity is not a problem. However, this is a rare situation in regression analysis.
- More often than not, there are near-linear dependencies among the regressors such that

$$t_1x_1 + t_2x_2 + t_3x_3 + \cdots \approx 0$$

- is approximately true.

Effects of Multicollinearity

- Strong multicollinearity can result in large variances and covariances for the least squares estimates of the coefficients.
- This make the coefficient estimates very sensitive to minor changes in the model
- When severe multicollinearity is present, confidence intervals for coefficients tend to be very wide and t-statistics tend to be very small
- In other words, the variance of the least squares estimate of the coefficient will be very large.

Multicollinearity Diagnostics

- Ideal characteristics of a multicollinearity diagnostic:
  - We want the procedure to correctly indicate if multicollinearity is present; and,
  - We want the procedure to provide some insight as to which regressors are causing the problem.

Variance Inflation Factors

- Variance inflation factors are very useful in determining if multicollinearity is present.

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

- R_j^2 is the coefficient of determination of the regression model when regressor j is predicted from all other regressors
- VIFs > 5 to 10 are considered significant.

R Code

```
› library(car)
› wgmdata = read.csv("wgmdata.csv")
› wgmdatafit=lm(y~.,data=wgmdata)
› summary(wgmdatafit)
› vif(wgmdatafit)
```

Webster Gunst Mason Data

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = WGMdata)

Residuals:
         1          2          3          4          5          6 
-3.698e-15 -1.545e+00  1.545e+00  7.649e-01 -2.517e-01 -5.132e-01 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.6599    14.0465   1.186 0.288885    
x1           -0.5313     1.3418  -0.396 0.708482    
x2           -0.8385     1.4206  -0.590 0.580722    
x3           -0.7753     1.4094  -0.550 0.605914    
x4           -0.8440     1.4031  -0.601 0.573745    
x5            1.0232     0.3909   2.617 0.047247 *  
x6            5.0470     0.7277   6.936 0.000956 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.129 on 5 degrees of freedom
Multiple R-squared:  0.9457,	Adjusted R-squared:  0.8806 
F-statistic: 14.52 on 6 and 5 DF,  p-value: 0.004993

> VIF = vif(Mul)
> VIF
        x1         x2         x3         x4         x5         x6 
182.051943 161.361942 266.263648 297.714658   1.919992   1.455265 
```

R Code

```
› wgmdatafit=lm(y~x1+x2+x3+x5+x6,data=wgm
  data)
› summary(wgmdatafit)
› vif(wgmdataFit)
```

R Code

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x5 + x6, data = wgmdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.23934 -0.55281 -0.09346  0.26575  1.78622 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.22029    0.61750  13.312 1.11e-05 ***
x1           0.27280    0.10922   2.498 0.046671 *  
x2           0.01189    0.13216   0.090 0.931235    
x3           0.06943    0.11148   0.623 0.556321    
x5           1.05547    0.36608   2.883 0.027941 *  
x6           5.07257    0.68670   7.387 0.000316 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.068 on 6 degrees of freedom
Multiple R-squared:  0.9418,    Adjusted R-squared:  0.8933 
F-statistic: 19.42 on 5 and 6 DF,  p-value: 0.00121

> vif(reducedwgmfit)
      x1       x2       x3       x5       x6 
1.349819 1.562620 1.864258 1.883934 1.450319 
```

This Session's Outline

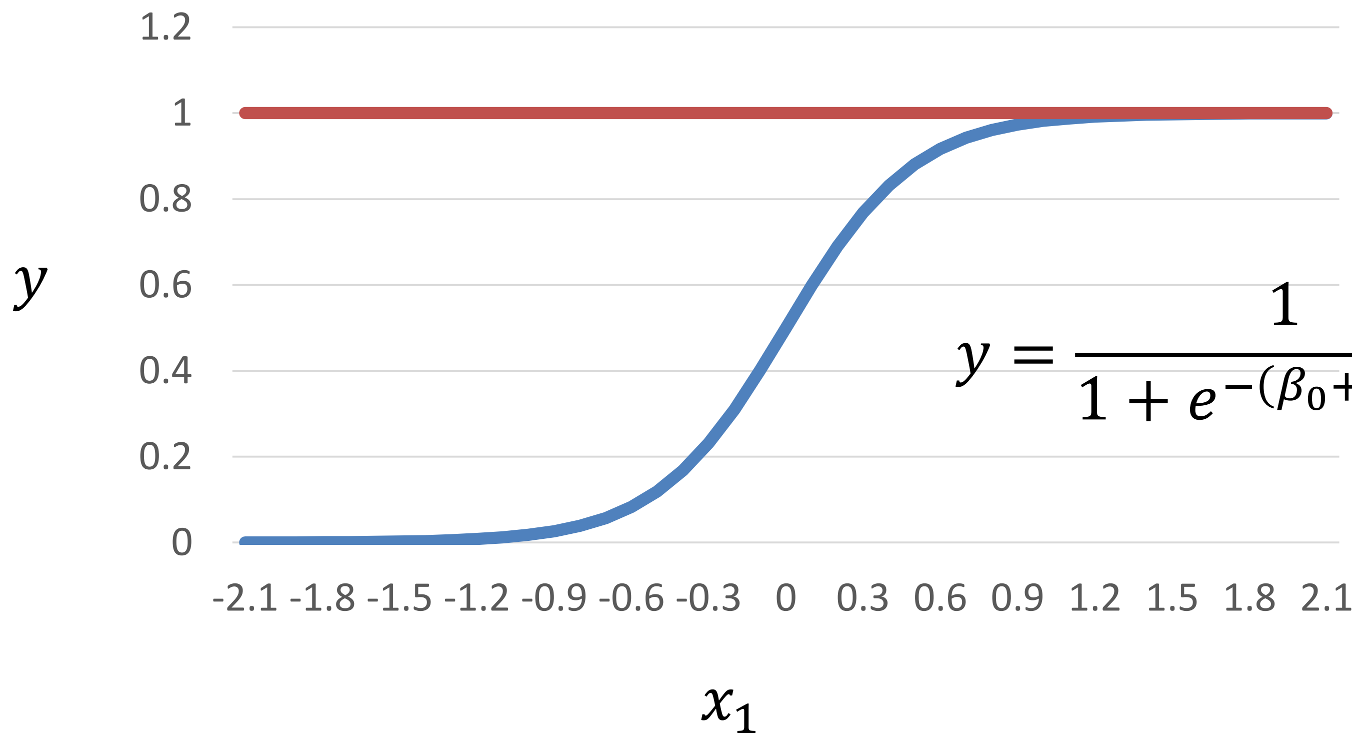
- Multiple Linear Regression
- Model Evaluation
- Variable Selection and Model Building
  – Best Subsets Regression
  – Stepwise Regression
  – Ridge Regression
  – Standardized Regression
- Indicator Variables
- Multicollinearity
- **Logistic Regression**
- Case Study

Logistic Regression

- Logistic regression predicts the probability of an outcome that can only have two values
- The prediction is based on the use of one or several predictors (numerical and categorical).
- Logistic regression produces a logistic curve, which is limited to values between 0 and 1.

Logistic Regression

- Logit Function



$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}}$$

Logistic Regression

- Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable.
- A linear regression is not appropriate for predicting the value of a binary variable for two reasons:
  - A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
  - Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.
- Predictors do not have to be normally distributed or have equal variance in each group.

Maximum Likelihood Estimation in Logistic Regression

- Logistic regression is a nonlinear model
  - Solving the ML score equations in logistic regression isn't quite as easy
- Solution is based on iteratively reweighted least squares or IRLS
  - An iterative procedure is necessary because parameter estimates must be updated from an initial "guess" through several steps
  - Weights are necessary because the variance of the observations is not constant
  - The weights are functions of the unknown parameters

E.R. L. Jalao, UP NEC, eljalao@up.edu.ph

Example: Menarche Data

- Data contains:
  - "Age" (average age of age homogeneous groups of girls),
  - "Total" (number of girls in each group),
  - "Menarche" (number of girls in the group who have reached menarche)
- Sources: (Milicer, H. and Szczotka, F., 1966, Age at Menarche in Warsaw girls in 1965, Human Biology, 38, 199-203)

R Code

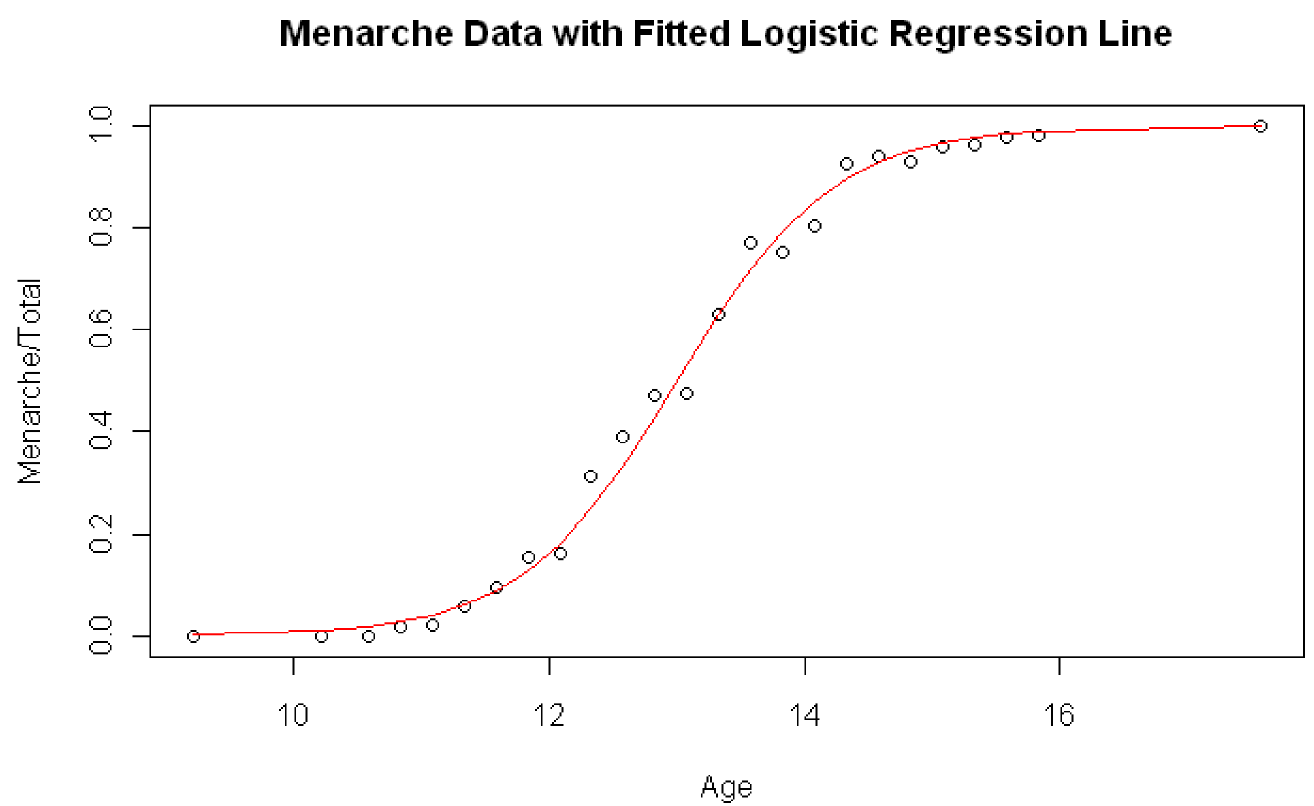
```
› library("MASS")
› menarchedata =
  read.csv("menarchedata.csv")
› menarchedata.fit = glm(cbind(menarche,
  total-menarche) ~ age,
  family=binomial(logit), data=menarchedata)
› summary(menarchedata.fit)
› plot(menarche/total ~ age,
  data=menarchedata)
› lines(menarchedata$age,
  menarchedata.fit$fitted, type="l",
  col="red")
```


E.R. L. Jalao, UP NEC, eljalao@up.edu.ph
108

R Output

```
Call:
glm(formula = cbind(Menarche, Total - Menarche) ~ Age, family = binomial(logit),
    data = menarche)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0363  -0.9953  -0.4900   0.7780   1.3675

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.22639    0.77068  -27.54   <2e-16 ***
Age           1.63197    0.05895   27.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   26.703  on 23  degrees of freedom
AIC: 114.76

Number of Fisher Scoring iterations: 4
```

Example: Menarche Data



$$Probability\ of\ Menarchy = \frac{1}{1 + e^{-(-21 + 1.63\ Age)}}$$

Example: Menarche Data

- Generated Model

$$Probability\ of\ Menarchy = \frac{1}{1 + e^{-(-21+1.63\ Age)}}$$

- The coefficient of "Age" can be interpreted as "for every one year increase in age the odds of having reached menarche increase by exp(1.632) = 5.11 times."
- Prediction for Age = 12

$$Probability\ of\ Menarchy = \frac{1}{1 + e^{-(-21+1.63\ *12)}}$$

$$Probability\ of\ Menarchy = 15.71\%$$

Global Model Validation

- To know if any of the x predictor variables influences y we consider the Deviance Statistic
- We usually test for:
  - H_0 : There is no significant difference between the actual and the predicted values
  - H_a : There is a significant difference between the actual and the predicted values
- p-Value Methodology
  - If $p < \alpha = 0.05$, Reject H_0

Global Model Validation

- 1-pchisq(3693.884,24)
- 1-pchisq(26.703,23)

```
> 1-pchisq(3693.884,24)
[1] 0
> 1-pchisq(26.703,23)
[1] 0.2688152
```

Recall the Credit Scoring Data

- Credit scoring is the practice of analyzing a persons background and credit application in order to assess the creditworthiness of the person
- The variables *income* (yearly), *age*, *loan* (size in euros) and *LTI*(the loan to yearly income ratio) are available.
- Our goal is to devise a model which *predicts*, whether or not a default will occur within 10 years..

http://www.r-bloggers.com/using-neural-networks-for-credit-scoring-a-simple-example/

R Code

```
› creditdata =
  read.csv("creditsetnumeric.csv")
› creditdata.fit = glm(default10yr ~
  income + age +loan+ LTI,
  family=binomial(logit),
  data=creditdata)
› summary(creditdata.fit)
```

R Output

```
Call:
glm(formula = default10yr ~ income + age + loan + LTI, family = binomial(logit),
    data = CreditData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1103  -0.0627  -0.0073  -0.0003   2.6102

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)  1.2068714  1.7236849    0.70                 0.48
income      -0.0000463  0.0000375   -1.24                 0.22
age         -0.3726547  0.0282724  -13.18 < 0.0000000000000002 ***
loan         0.0003079  0.0002595    1.19                 0.24
LTI         68.8527642 12.4780318    5.52          0.000000034 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1630.71  on 1999  degrees of freedom
Residual deviance:  400.58  on 1995  degrees of freedom
AIC: 410.6

Number of Fisher Scoring iterations: 9
```

Example: Interpretation

- Generated Model

$$Probability\ of\ Default = \frac{1}{1 + e^{-(1.2 - 4\times10^{-5}\ income - 0.37 age + 3\times10^{-4} loan\ + 68 LTI)}}$$

- The coefficient of "Age" can be interpreted as "for every one year increase in age the odds of defaulting increase by exp(-0.37) = 0.69 times.“
- Prediction for a new Client with Income = 66000, Age = 18, Loan = 8770, LTI = 0.000622

$$Probability\ of\ Default = \frac{1}{1 + e^{-(1.2 - 4\times10^{-5}\ (66k) - 0.37(18) + 3\times10^{-4}(8770)\ + 68(0.00062))}}$$

$$Probability\ of\ Default = 0.794$$

Model Validation

- To know if the x predictor variables influences y we consider the Deviance Statistic
- We usually test for:
  - H_0 : There is no significant effect when adding x_i in the model
  - H_a : There is a significant effect when adding x_i in the model
- p-Value Methodology
  - If $p < \alpha = 0.05$, Reject H_0

E.R. L. Jalao, UP NEC, eljalao@up.edu.ph

Testing Null and Residual Deviance

› anova(creditdata.fit,test="Chi")

```
> anova(creditdata.fit,test="Chi")
Analysis of Deviance Table

Model: binomial, link: logit

Response: default10yr

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    1999    1630.71
income  1     0.01      1998    1630.70    0.9186
age     1   478.57      1997    1152.13 < 2.2e-16 ***
loan    1   711.43      1996     440.70 < 2.2e-16 ***
LTI     1    40.12      1995     400.58 2.386e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This Session's Outline

- Multiple Linear Regression
- Model Evaluation
- Variable Selection and Model Building
  - Best Subsets Regression
  - Stepwise Regression
  - Ridge Regression
  - Standardized Regression
- Indicator Variables
- Multicollinearity
- Logistic Regression
- **Case Study**

Case 3: TV Advertising Revenue Dataset

- Jalao (2012) proposed a regression model to predict the revenue of advertising for a 30 second primetime TV show slot.
- Significant factors that affect the revenue of advertising where also determined.
- Data was obtained and compiled from multiple websites that provide information that could potentially affect the revenue of advertising.
- Moreover, the effect of several social media websites on the revenue of advertising was also studied.

References

- James ,Witten, Hastie, & Tibshirani, *An Introduction to Statistical Learning with Applications in R,* 1st Ed Springer, 2013
- Montgomery, Peck & Vining , *Linear Regression Analysis 5E*, Springer, 2012
- Data Mining Overview: http://www.saedsayad.com/
- Milicer, H. and Szczotka, F., 1966, Age at Menarche in Warsaw girls in 1965, Human Biology, 38, 199-203
- G. Runger, ASU IEE 578
- http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values