STEVENS INSTITUTE of TECHNOLOGY
School of Business

Amir H Gandomi; PhD
Assistant Professor
Stevens Institute of Technology
a.h.gandomi@stevens.edu

Multivariate Data Analysis – BIA 652

Class 4 – Multivariate Regression Analysis

Outline – Class 4

- SAS Help Desk Contact Zeyuan Wang, on CANVAS
- Visit http://www.ats.ucla.edu/stat/sas/output/reg.htm
- Assign second HW Due Next class
- Multiple Regression Lecture
  - Chapters 7.1-7.8, 7.9-7.11, 8, 9.2
  - Go over Multiple Regression Example – 7.1
- Office Hours:
  - Tuesdays 2-6 PM
  - Or By Appointment
- Grader:
  - Haochen Liu (hliu56@stevens.edu)

Multiple Linear Regression

Aims

- Extend simple linear regression to multiple dependent variables.
- Describe a linear relationship between:
  - A single continuous Y variable, and
  - Several X variables
- Draw inferences regarding the relationship
- Predict the value of Y from X_1, X_2, …, X_p.
- Research Questions: To what extent does some combination of the IVs predict the DV?
- E.g. To what extent does age, gender, type/amount of food consumption predict low density lipid level.

Assumptions

- Level of Measurement:
  - IVs – two or more, Continuous or dichotomous
  - DV - continuous
- Sample Size – Enough cases per IV
- Linearity: Are bivariate relationships linear
- Constant Variance (about line of best fit) – Homoscedasticity
- Multicollinearity: Between the IVs
- Multivariate outliers
- Normality of residuals about predicted value

Approaches

- Direct: All IVs entered simultaneously
- Forward: IVs entered one by one until there are no significant IVs to be entered.
- Backward: IVs removed one by one until there are no significant IVs to be removed.
- Stepwise: Combination of Forward and Backward
- Hierarchical: IVs entered in steps.

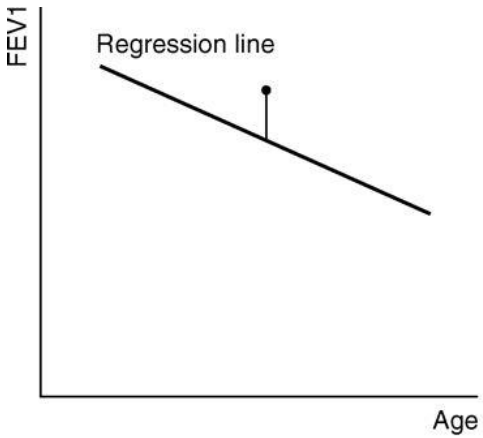
Write ups

- Assumptions: How tested, extent met
- Correlations: What are they, what conclusions
- Regression coefficients: Report and interpret
- Conclusions and Caveats

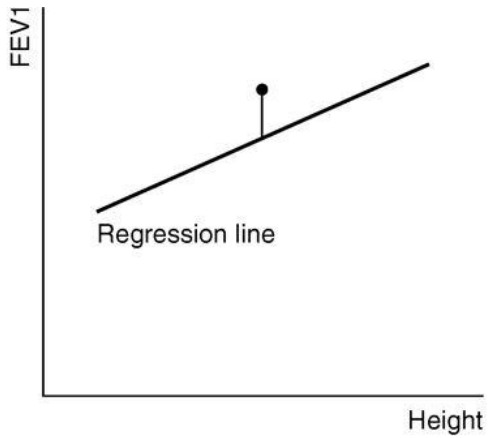
Steps in Multiple Regression

1. State the research hypothesis.
2. State the null hypothesis
3. Gather the data
4. Assess each variable separately first (obtain measures of central tendency and dispersion; frequency distributions; graphs); is the variable normally distributed?
5. Assess the relationship of each independent variable, one at a time, with the dependent variable (calculate the correlation coefficient; obtain a scatter plot); are the two variables linearly related?
6. Assess the relationships between all of the independent variables with each other (obtain a correlation coefficient matrix for all the independent variables); are the independent variables too highly correlated with one another?
7. Calculate the regression equation from the data
8. Calculate and examine appropriate measures of association and tests of statistical significance for each coefficient and for the equation as a whole
9. Accept or reject the null hypothesis
10. Reject or accept the research hypothesis
11. Explain the practical implications of the findings

Example (p 121)

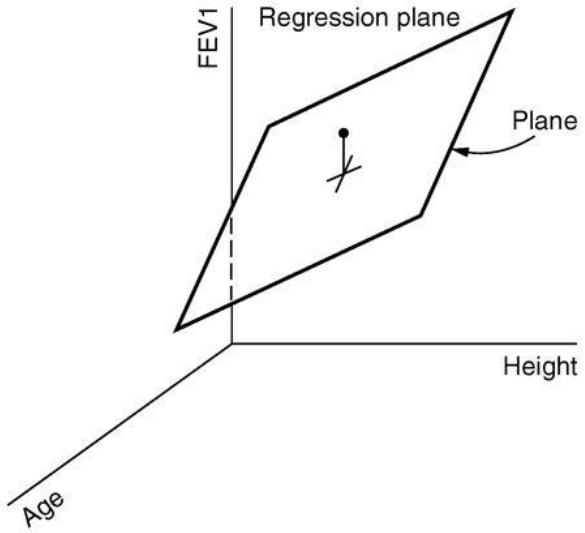


a. Simple Regression of FEV1 on Age



b. Simple Regression of FEV1 on Height

Figure 7.1 Hypothetical Representation of Simple and Multiple Regression Equations of FEV1 on Age and Height



c. Multiple Regression of FEV1 on Age and Height

Example (p 122)

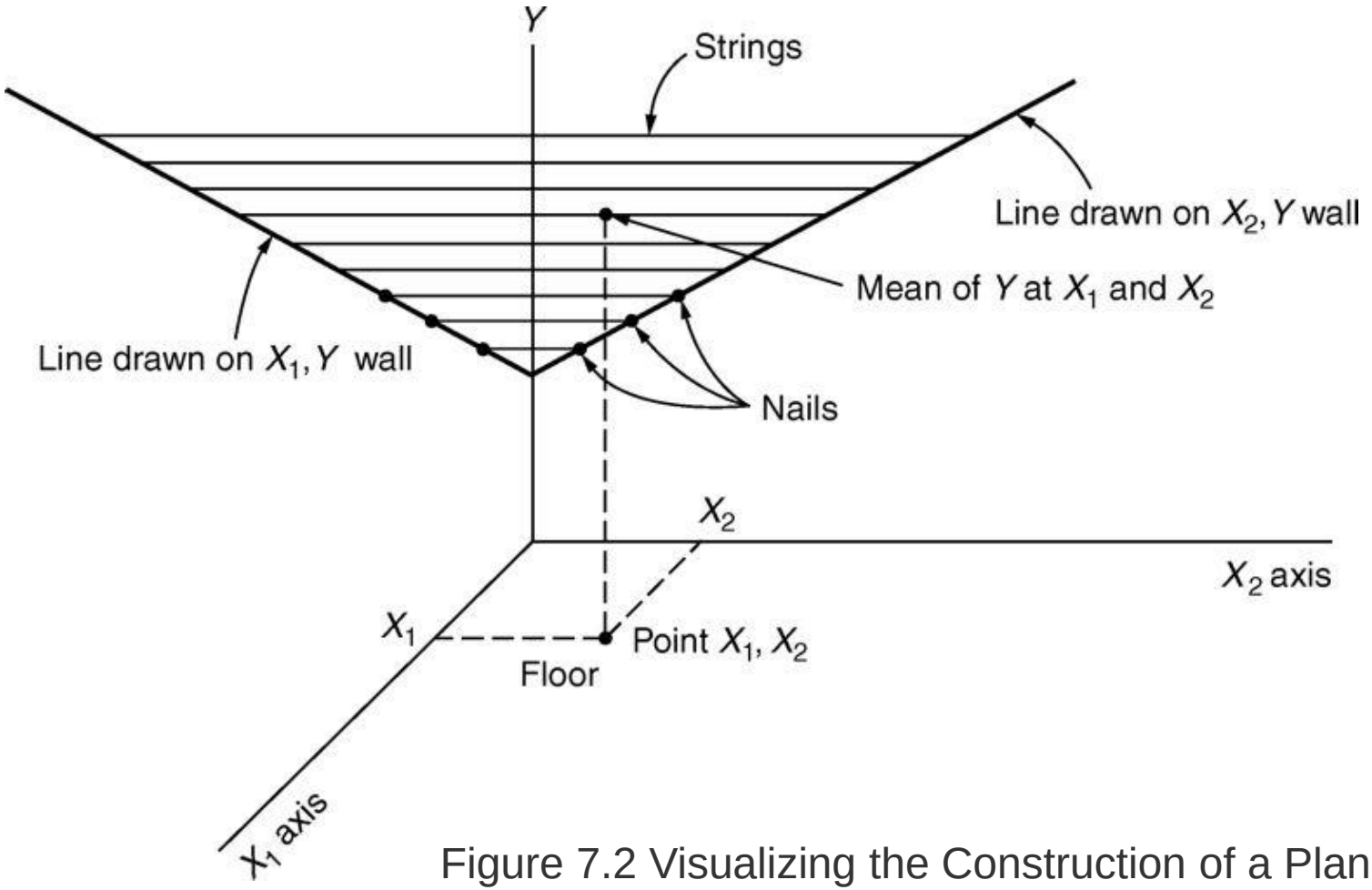


Figure 7.2 Visualizing the Construction of a Plane

Mathematical Model

- The mean of Y values at a given X is:

$$a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- Variance of Y values at any set of X's is σ^2

  (For all X)

- Y values are normally distributed at each X (needed for inference)

Types of X (independent) variables

- Fixed: selected in advance
- Variable: as in most studies

- X's can be continuous or discrete (categorical)
- X's can be transformations of other X's, e.g., polynomial regression.
  - Example: $X_1 \rightarrow \log(X_1)$

Computer Analysis

- Estimates of: $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ using least-squares.
- Residual mean square (S^2) is estimate of variance σ^2
- $S^2 = \sum(Y - \hat{Y})^2$ /(N-P-1) = Res. Mean Sq AKA: Standard Error of the Estimate Squared.
- Confidence intervals for mean of Y
- Prediction intervals for individual Y

Bonferroni correction

- In multiple hypotheses, the chance of incorrectly rejecting a null hypothesis increases
- Example of Bonferroni Correction: Test 3 hypotheses
- *p*-values are: 0.014, 0.036, 0.075
- Let nominal significance level = 0.15
- Bonferroni Adjusted *p*-values: *p*/m = 0.15/3 = 0.05
  - ∴ First two are significant
  - Probability of at rejecting at least 1 out of m hypotheses
- Familywise error rate (FWER): the probability of rejecting at least one true hypothesis

Analysis of variance (p 132)

- Does regression plane help in predicting values of Y?
- Test hypothesis that all β_i's = 0

Table 7.1: ANOVA Table for multiple regression

Source of variation	Sums of squares	df	Mean square	F
Regression	$\sum(\hat{Y}-\bar{Y})^2$	P	$\mathrm{SS}_{\mathrm{reg}}/P$	$\mathrm{MS}_{\mathrm{reg}}/\mathrm{MS}_{\mathrm{res}}$
Residual	$\sum(Y-\hat{Y})^2$	$N-P-1$	$\mathrm{SS}_{\mathrm{reg}}/(N-P-1)$	
Total	$\sum(Y-\bar{Y})^2$	$N-1$		

Example: Reg of FEV1 on height and weight (p 132)

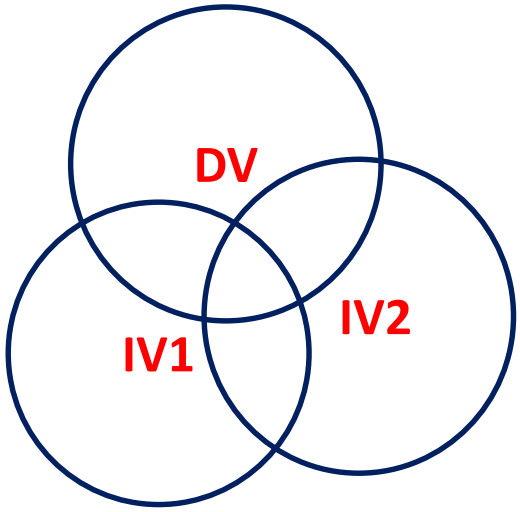
Table 7.2: ANOVA example from the lung function data (fathers)

Source of variation	Sums of squares	df	Mean square	F
Regression	21.0570	2	10.5285	36.81
Residual	42.0413	147	0.2860	
Total	63.0983			

- F = 36.81; df = 2, 147; p-value <0.0001
- Use percentile link from web site: http://faculty.vassar.edu/lowry/tabs.html#f

Venn Diagrams

- Multiple R^2 : between the DV and a linear combination of IVs
  - Bivariate Correlation between IV1 and DV
  - Bivariate Correlation between IV2 and DV
  - Correlation between IV1 and IV2
- Target: IV's that highly correlate with the DV, but don't highly correlate with each other



Correlation Coefficient

- The multiple correlation coefficient (R) measures the strength of association between Y, and the set of X's in the population.
- It is estimated as the simple correlation coefficient between the Y's and their predicted values ($\hat{Y}$'s)

Coefficient of Determination

- R^2 = Coefficient of determination
  = SS due to regression/SS total
- R^2 = (reduction in variance of Y due to X's) / (original variance of Y).
- Therefore $100R^2$ = % of variance of Y "explained by X's".
- And $100(1 - \rho^2)^{1/2}$ = % of Standard Deviation NOT "explained" by X's

Interpretation of R

R	% of variance "explained"	% of variance not "explained"	% of SD "explained"	% of SD not "explained"
±0.3	9%	91%	5%	95%
±0.5	25%	75%	13%	87%
±0.71	50%	50%	29%	71%
±0.95	90%	10%	69%	31%

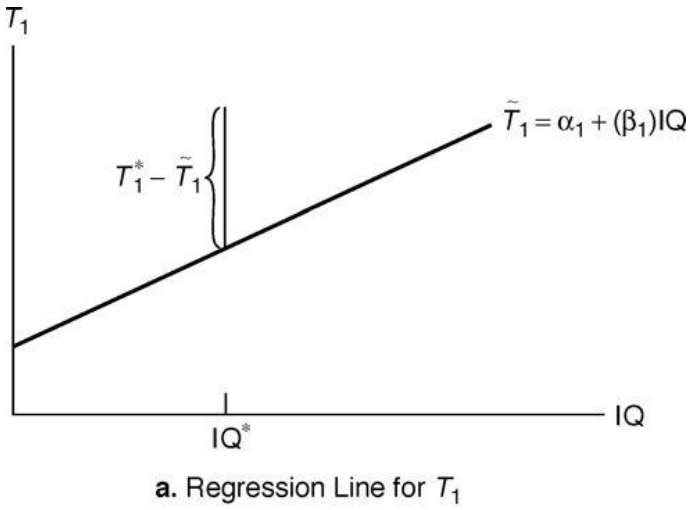
Partial Correlation Coefficient

- The correlation coefficient measuring the degree of dependence between two variables **after adjusting for the linear effect of one or more of the other X variables**

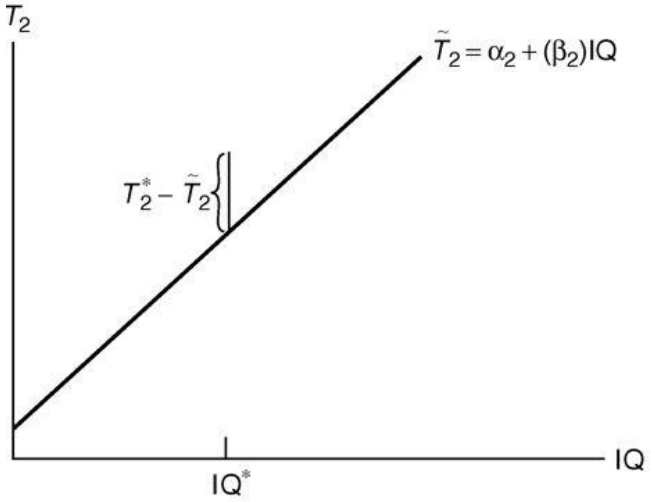
Example: T_1 and T_2 are test scores

- Find partial R between T_1 and T_2 after adjusting for IQ since the scores are related to students' IQ

Visually (p 130)



a. Regression Line for T_1



b. Regression Line for T_2

Figure 7.4 Hypothetical Population Regressions of T_1 and T_2 Scores, Illustrating the Computation of a Partial Correlation Coefficient

It is reasonable to first remove the linear effect of IQ from both test! Which results in Partial R

- Partial R = simple R between the two residuals

Interpretation of regression coefficients

- In the model: $a + b_1X_1 + b_2X_2 + \ldots + b_pX_p$ if ρ is the partial correlation between Y and X_1 , given $X_1, X_2, \ldots, X_p$, then

- Testing that $b_1 = 0$, is equivalent to testing that $\rho = 0$

Hence, b_i is called the partial regression coefficient of Y on X_1 , given $X_1, X_2, \ldots, X_p$

Values of regression coefficients

- Problem: Values of b_i 's are not directly comparable
- Hence: Standardized coefficients:
  - Standardized $b_i = b_i * (SD(X_i) / SD(Y))$
- Standardized b_i are directly comparable.

Multicollinearity

- The case where some of the X variables are highly correlated

$$[\mathrm{SE}(B_i)]^2 = \frac{S^2}{(N-1)(S_i)^2} \times \frac{1}{1-(R_i)^2}$$

- This will impact estimates, and their SE's (p 143)
  - S^2 Residual Mean Square.
  - S_i Standard Deviation of ith X variable
  - R_i Multiple Correlation between ith X and all other X's

- Variance Inflation Factor (VIF) = SE

Variance Inflation Factor

- Consider Tolerance, and its inverse, Variance Inflation Factor (VIF)
- VIF = 1 / Tolerance
- Example: Target Tolerance < 0.1, or VIF > 10
- Rule of thump:
  - VIF = 1 : not correlated.
  - 1 < VIF ≤ 5 : moderately correlated.
  - 5 < VIF : highly correlated.
- Remedy: use variable selection to delete some X variables, or a dimension reduction techniques such as Principal Components.

Misleading Correlations

- Spurious relationship due to:
  - Coincidence
  - Unseen factor (or Confounding factor)

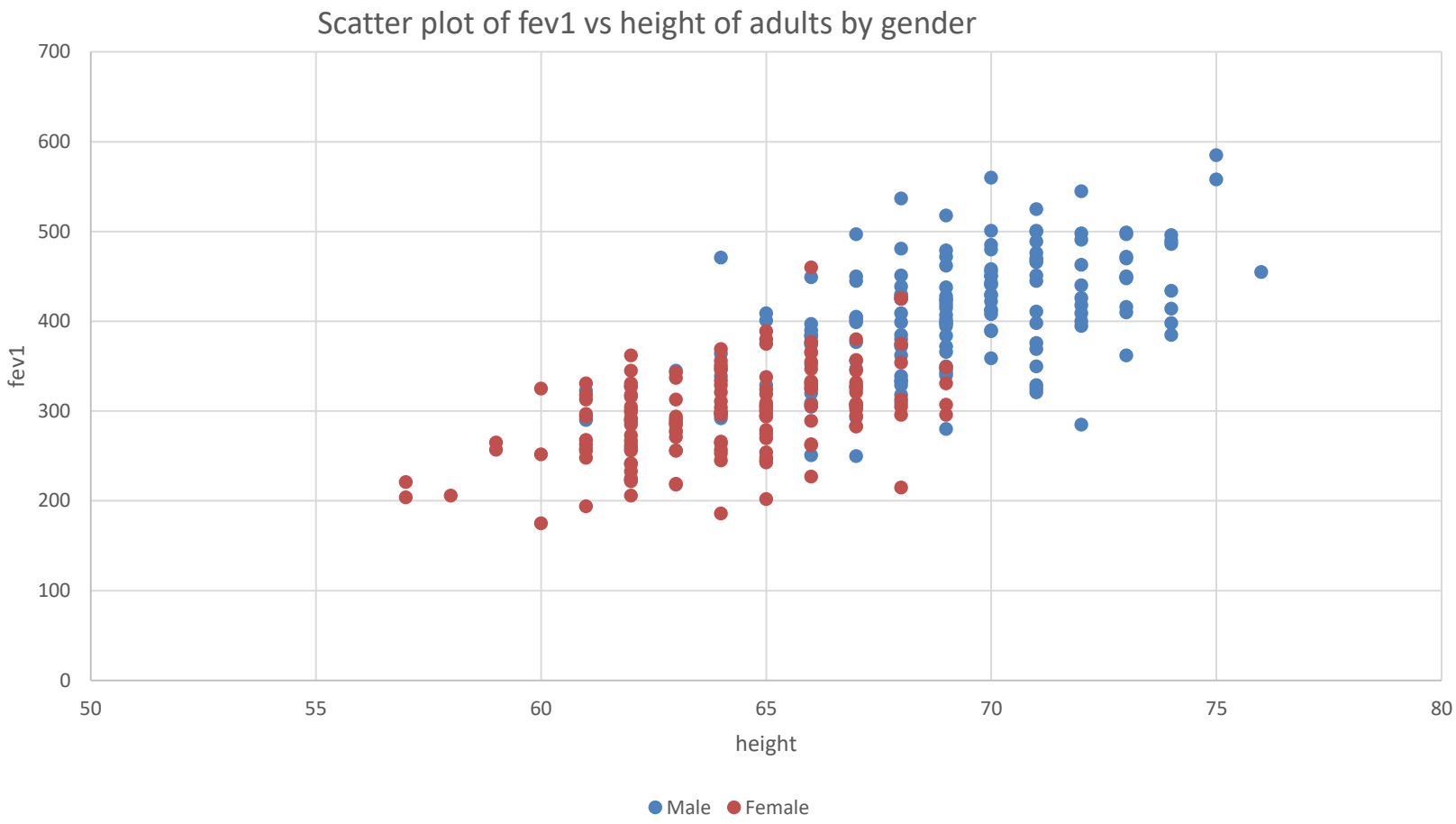
- Example (Lung Function data, Appendix A): FEV1 vs height and age

- Depends on gender

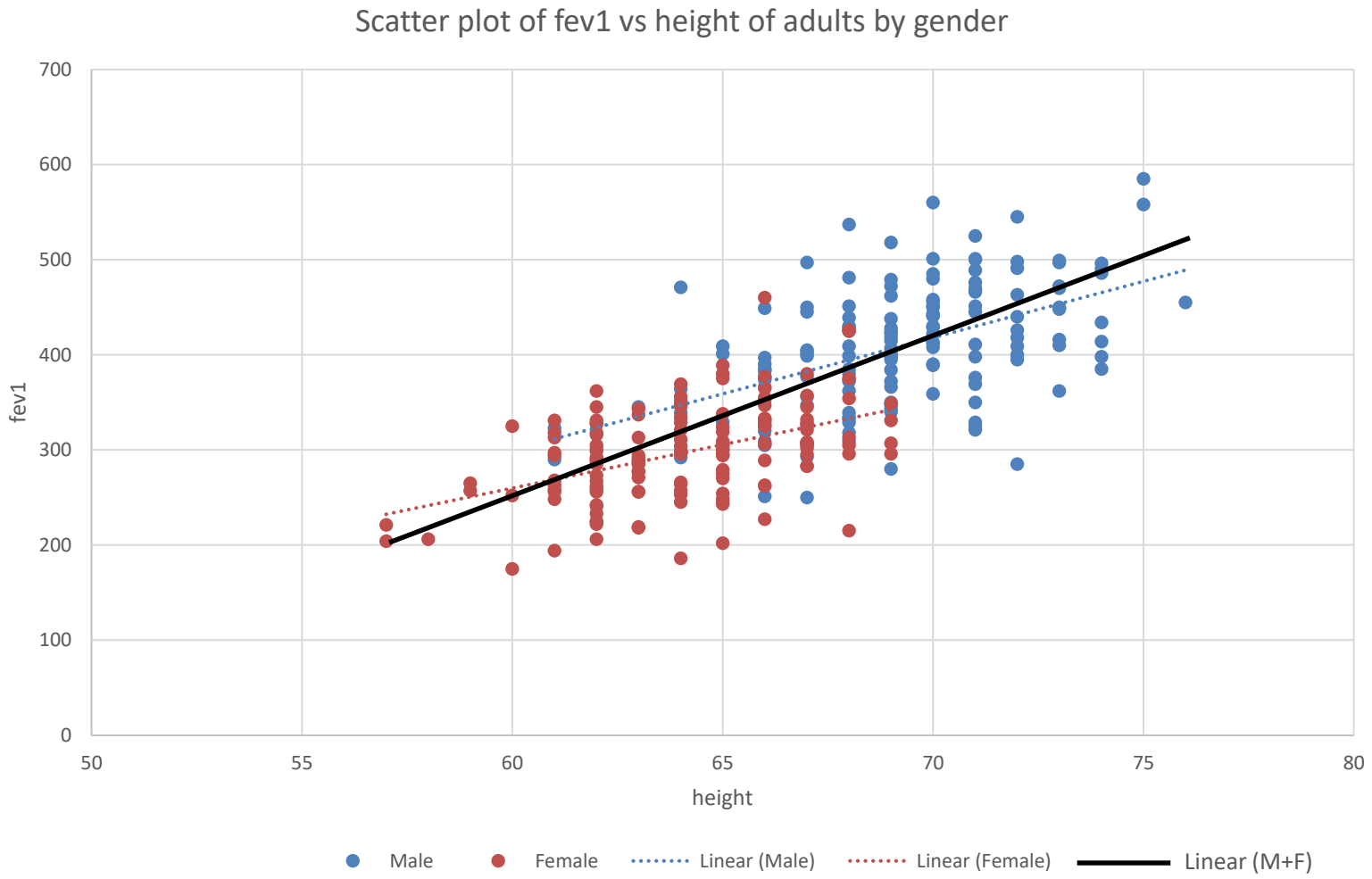
Total vs Stratified Correlation

Gender	Correlation between FEV1 and:	
	Height	Age
Total	0.739	-0.073
Male	0.504	-0.310
Female	0.465	-0.267

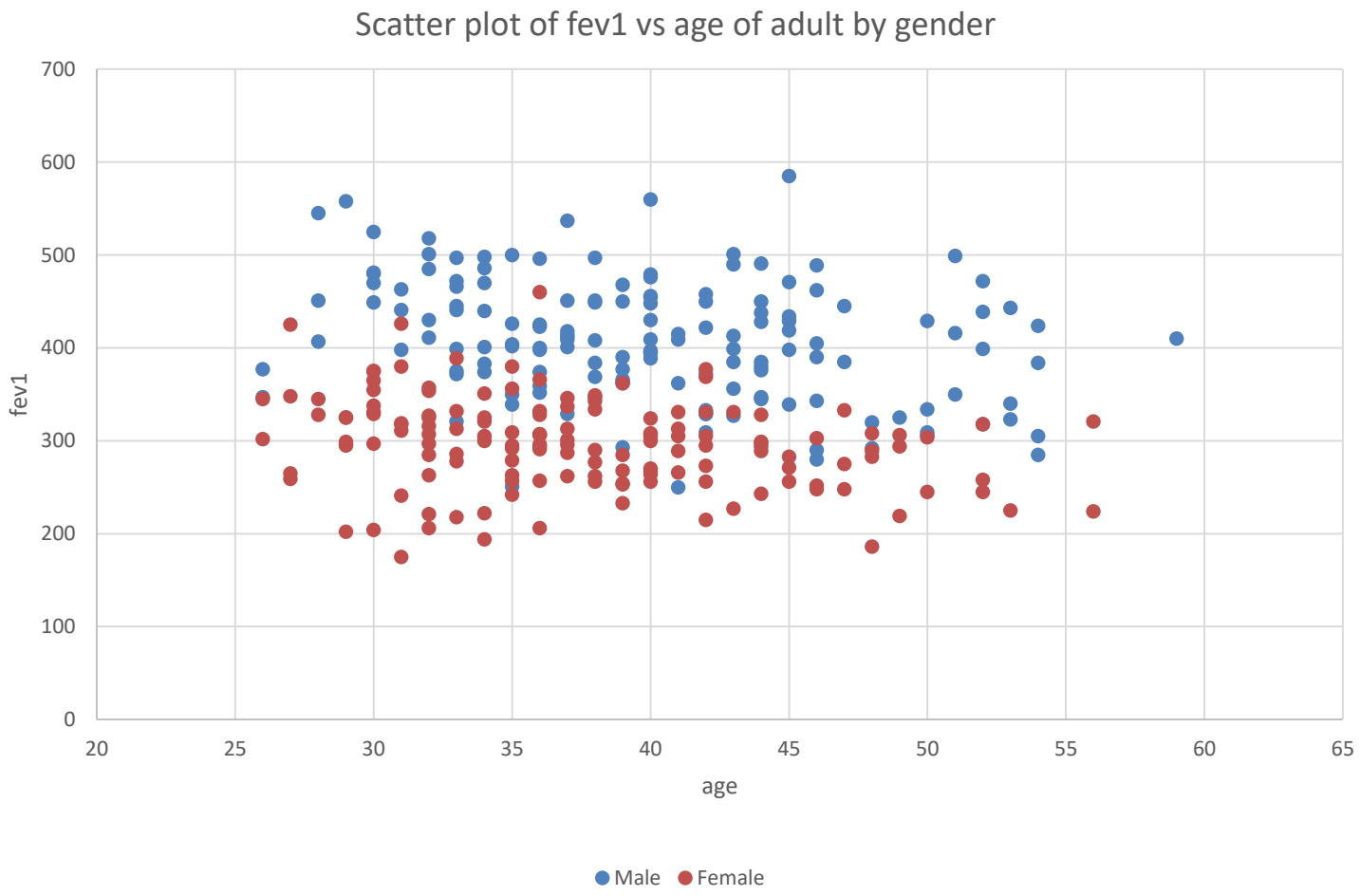
FEV1 vs height



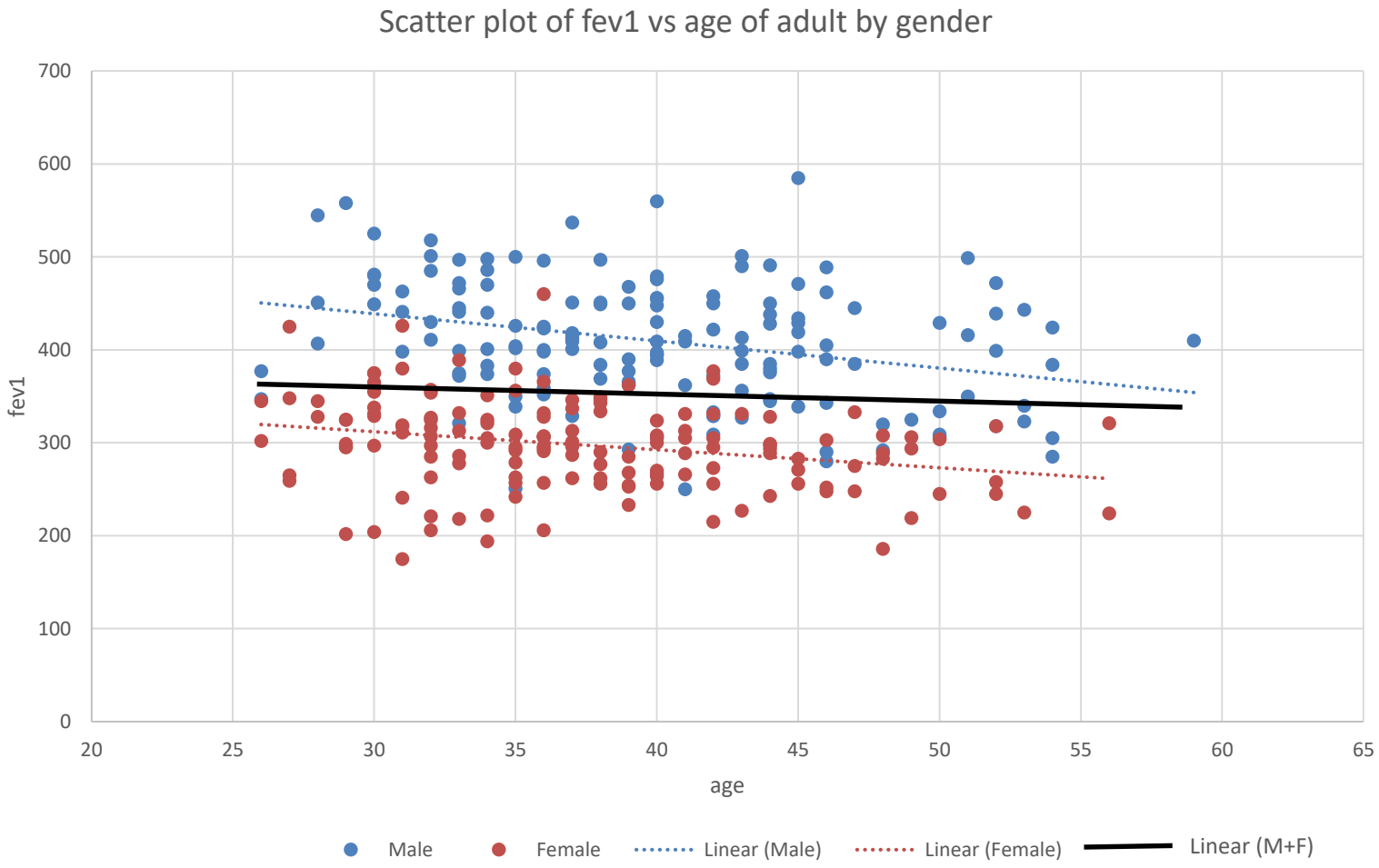
FEV1 vs height – Regression lines



FEV1 vs age


Scatter plot of fev1 vs age of adult by gender
fev1
700
600
500
400
300
200
100
0
20
25
30
35
40
45
50
55
60
65
age
Male
Female

FEV1 vs age– Regression lines



Variable interactions in Regression

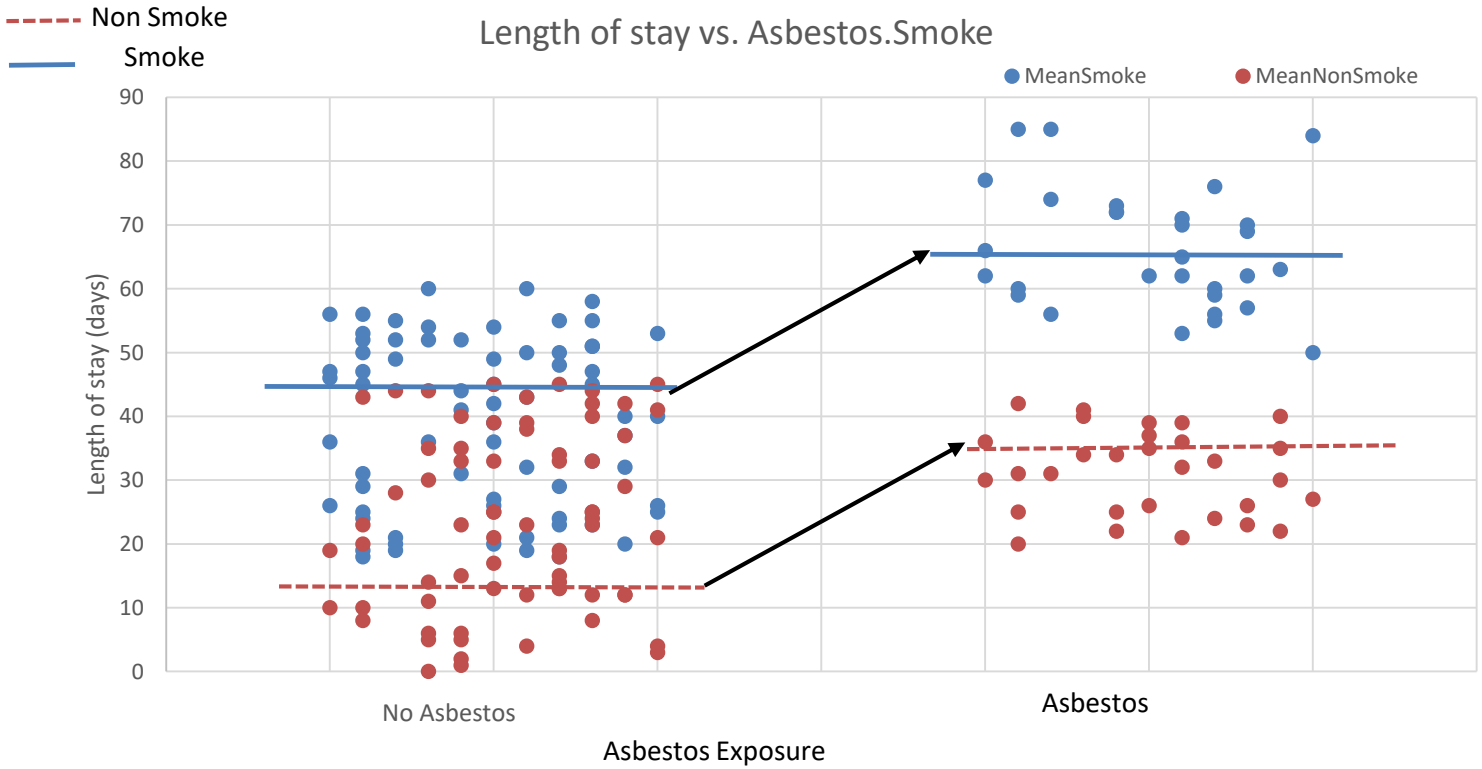
- Variables X1 and X2 have interactions if the effect of X1 on Y depend on X2 value (and vice versa)
- Joint effect of variables (the effect of one depends on the value of the other one)
- Correlation is about association between to variable but Interaction is about their effect on a third variable!

Dep	Ind1	Ind2
Output	Machines	No of workers
Sales	Products	Salesmen
Enrolment	Schools	Salesmen
House Price	No of rooms	Area

Ref: Analytics University

Interpreting Interaction in Linear Regression - 1

- (no interaction)

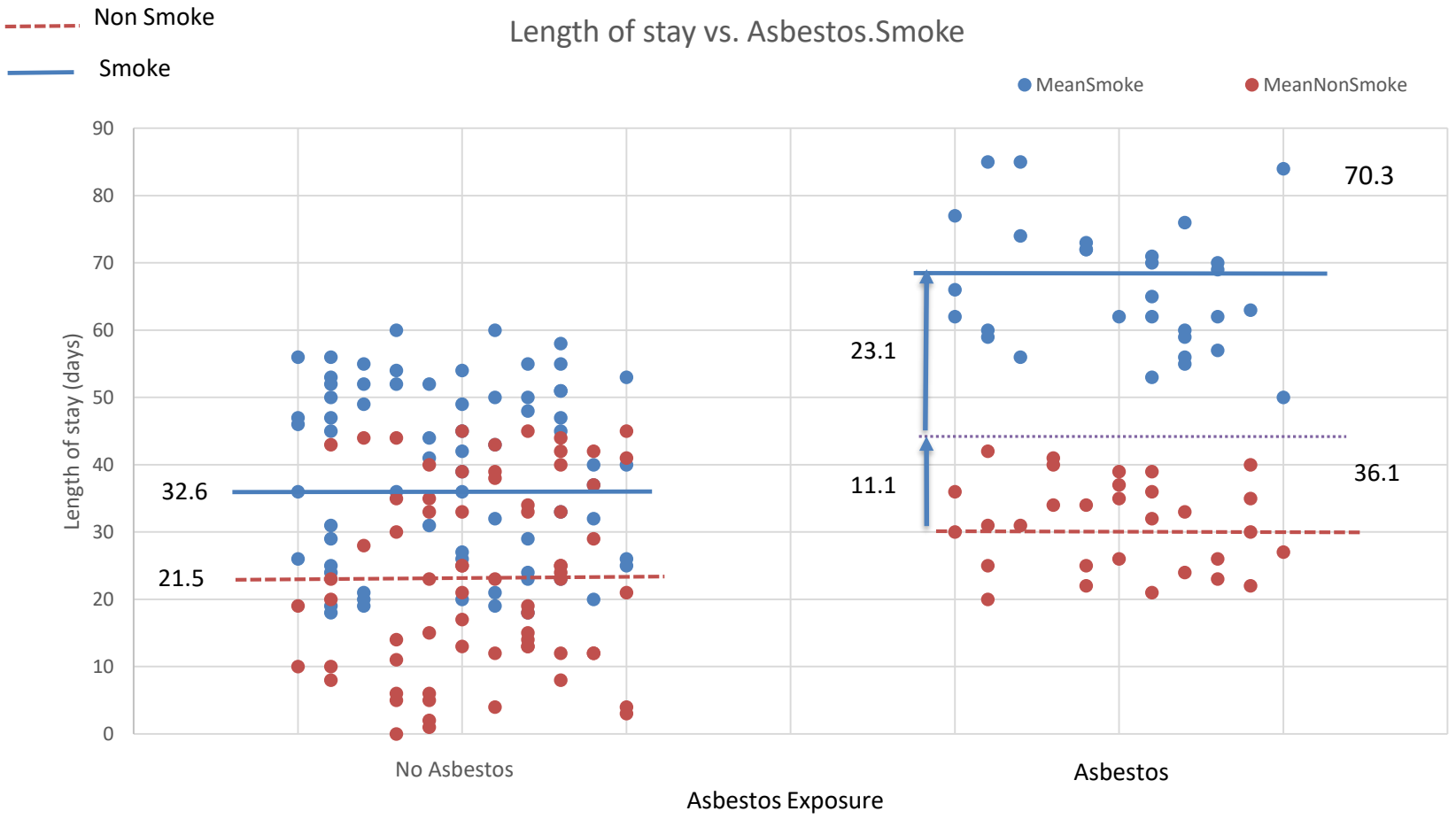


- Effect of Asbestos is same for s and n

Interpreting Interaction in Linear Regression - 2

$$\hat{\mu}_{y|x} = b_0 + b_1 \times smoke + b_2 \times Asbestos + b_{1,2} \times smoke \times Asbestos$$

$$\hat{\mu}_{y|x} = 21.5 + 11.1 \times smoke + 14.6 \times Asbestos + 23.1 \times smoke \times Asbestos$$



Residual Analysis

- Residual = $e = Y - \hat{Y}$
- Studentized residual = $\dfrac{\text{Residual}}{\text{estimate of residual standard deviation}} = \dfrac{e}{S\sqrt{1-h}}$
  - h called "leverage"

- Leverage: shows how far away is an observation from others

Outliers

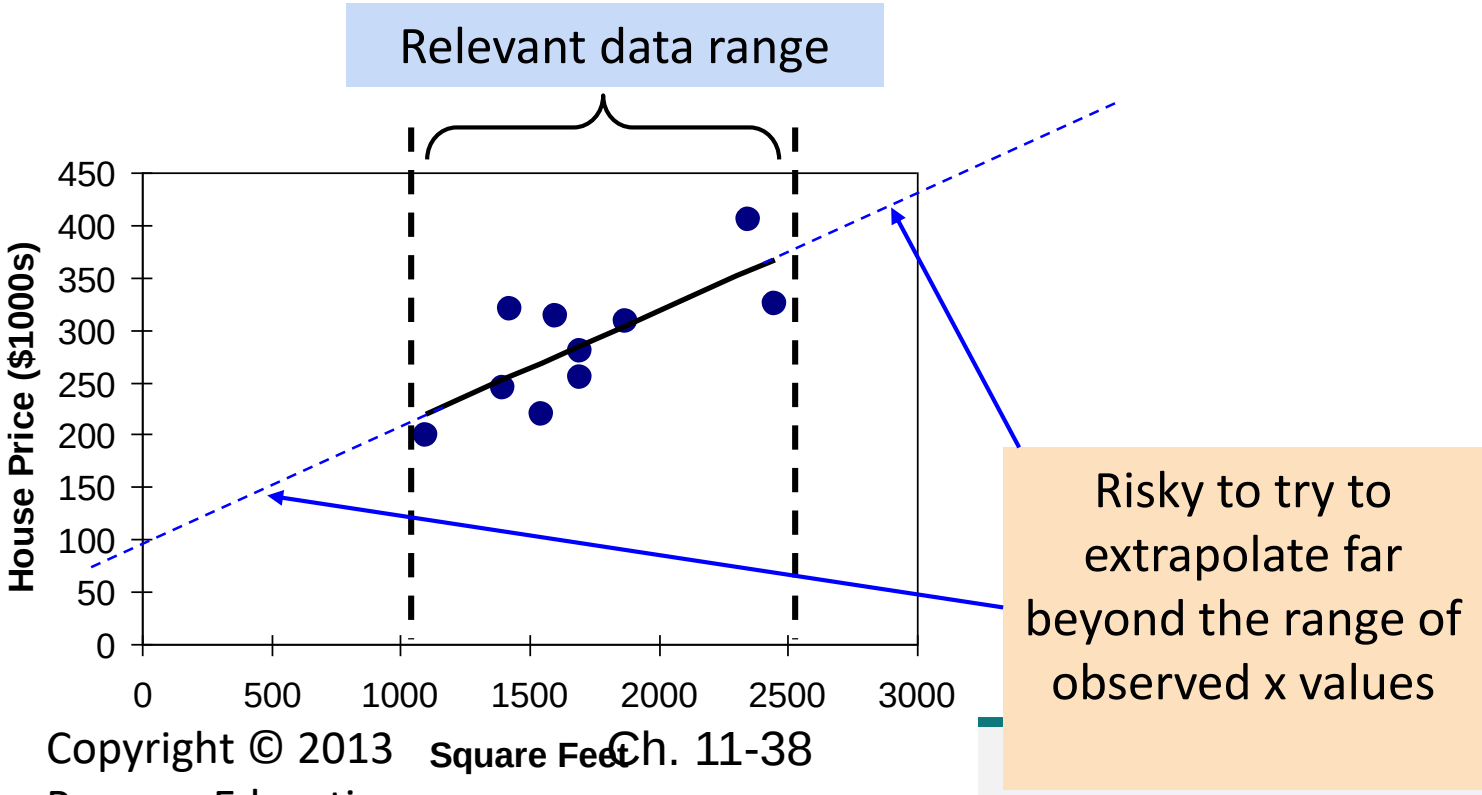
- Outlier in Y is studentized (or deleted studentized) residual >2

- Leverage = h = $\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum(X - \bar{X})^2}$
  - X's far from the mean of X have large leverage (h)
  - Observations with large leverage have large effect on the slope of the line.

- Outlier in X if h > 4/N

Relevant Data Range

- When using a regression model for prediction, only predict within the relevant range of data



Effect of Outliers (p 102)

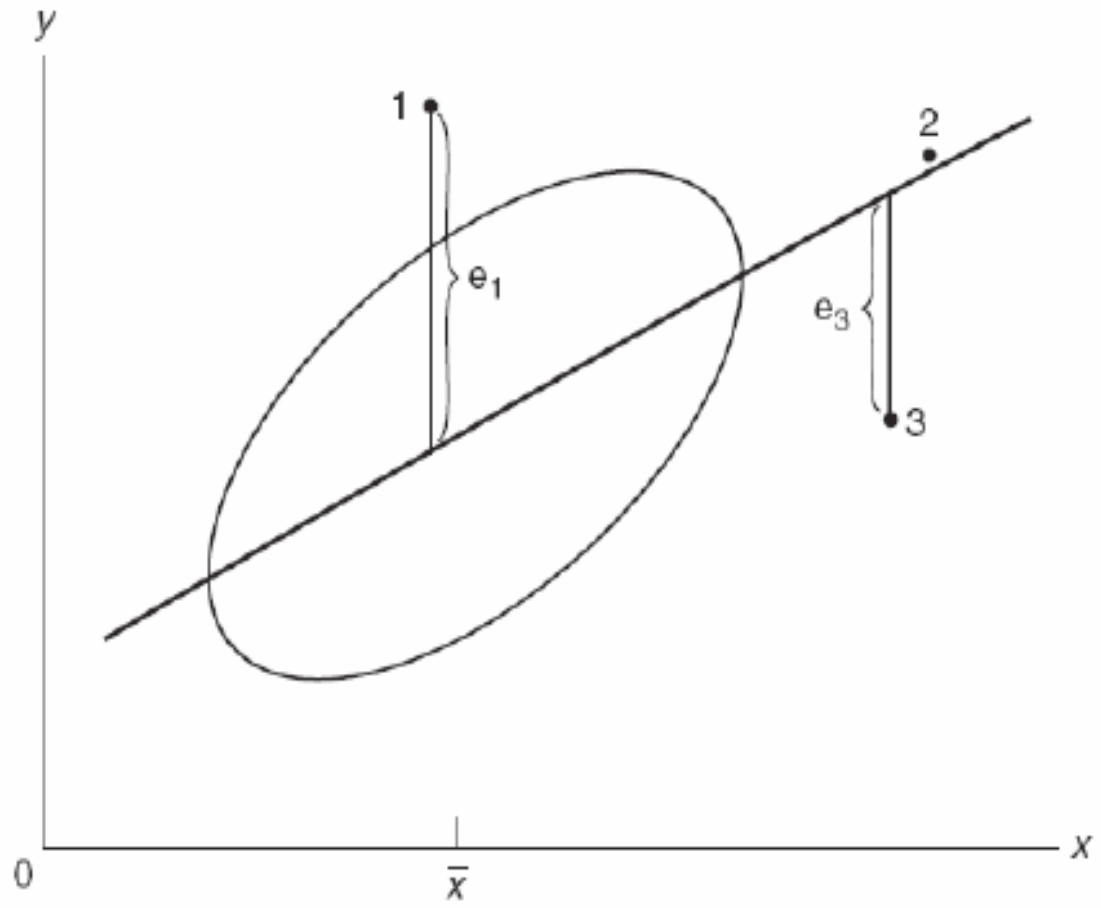


Figure 6.8: Illustration of the Effect of Outliers

Observations

- Point 1 is an outlier in Y with low leverage
  - impacts estimate of intercept but not slope
  - Tends to increase the estimates of S & SE of B
- Point 2 has high leverage; not an outlier in Y
  - doesn't impact estimate of B or A
- Point 3 has high leverage and is an outlier in Y
  - impacts the values of B, A, and S

Influential observations

An observation is influential if:

- It is an outlier in X and Y
- Cook's distance > $F_{0.5}(2,N-2)$
- DFFITS > $\frac{2\sqrt{p}}{\sqrt{N-2}}$

- Here, the number of parameters (p)=2 and N is the number of points

Try analysis with and without influential observations and compare results.

Some Caveats

- See list for simple regression
- Need representative sample
- Violations of assumptions, outliers
- Multicollinearity: coefficient of any one variable can vary widely, depending on what others are included in the model
- Missing values
- Number of observations in the sample should be large enough relative to number of variables in the model.

Overview of Variable Selection

Required for Variable Selection:

- A General Test
- Selection Criteria
- A Selection Process

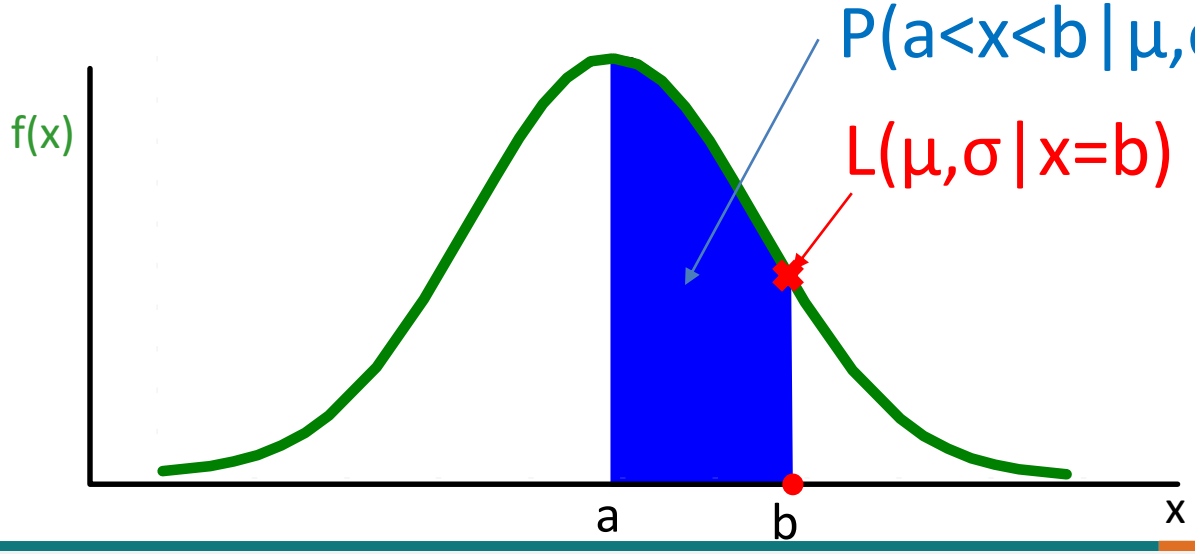
A General Test

If P variables are in the equation, Test if Q additional variables are useful:

- Test H_0 : Q additional variables are useless, i.e. their β's all = 0;
- H_a : Q additional variables are useful

likelihood

- ***Probability*** is used before data are available to describe possible future outcomes given a fixed value for the parameter.
- ***Likelihood*** is used after data are available to describe a function of a parameter for a given *outcome*.



$$\mathcal{L}(\theta \mid x) = P(x \mid \theta)$$

Likelihood Ratio (Deviance) Test

- Use the LR (deviance) Test:
  - Deviance = (-2)log (likelihood)
  - Under H_0 , deviance = D_0 , df_0 = N-P-1
  - Under H_a , deviance = D_a , df_a = N-P-1

- LR (deviance) test statistic is:
  - $D_0 - D_a$ is distributed approximately as χ^2 with Q degrees of freedom under H_0 for large N

Likelihood Ratio (deviance) Test Exact F Test

- If we assume normally distributed residuals, the LR Test becomes an Exact F Test

$$F = \frac{(ResSS_{reduced} - ResSS_{full})/Q}{ResSS_{full} \Big/ Resdf_{full}}$$

$$= \frac{(ResSS_{full} - ResSS_{reduced})/Q}{ResSS_{full} \Big/ Resdf_{full}}$$

$$= \frac{(R^2_{full} - R^2_{reduced})/Q}{(1 - R^2_{full})/Resdf_{full}}$$

$$df = Q, Resdf_{full}$$

Example: Chemical Companies

- Y = P/E = price to earnings ratio
- X_1 = D/E = debt to earnings ratio
- X_2 , X_3 , …, X_6 = other variables (p.160)
- Test:
  - H_0 : X_1 is as good as X_1 , … , X_6 in predicting Y,
  - Or equivalently,
  - H_0 : X_2 , X_3 , …, X_6 are useless in predicting Y when included with X_1

Computation (p 167)

- Full Model has all six X variables

  Reduced Model has X_1 only
  - $ResSS_{full} = 103.06$, $ResSS_{reduced} = 176.08$
  - Q = 6-1 =5, $Resdf_{full} = 30 - 6 - 1 = 23$
  - → F = 3.26 , df = 5 , 23
  - →p-value < 0.025

- Conclusion: reject H_0 - conclude that X_1 , … , X_6 are better than X_1 alone

Variable Selection

- Choose the subset of X variables to best "explain" or predict the dependent variable

- Criteria include balancing:
  - Conciseness – no unnecessary variables
  - Goodness of Fit, i.e. R^2

Multiple R^2

- $0 \leq R^2 \leq 1$
- The higher value, the better predictability of the DV from the IVs

$$R^2 = c^T R_{xx}^{-1} c$$
$$c = (r_{x_1 y}, r_{x_2 y}, \ldots\ldots, r_{x_N y})$$

- R_{xx} is correlation matrix
- R^2 : Including more variables gives higher (or the same) multiple correlation.
- R^2 is biased: if population R^2 really 0, then $E(R^2) = P/(N-1)$
  - Example: $N = 100$, $P = 1$, $E(R^2) = 0.01$
  - Example: $N = 21$, $P = 10$, $E(R^2) = 0.5$

Adjusted (multiple) R^2

- R^2 often not effective at identifying useless variables.
- R^2 is biased (Note R^2 = 1 - ResSS/TotSS)

- Address Bias by using MS instead of SS
  - Adjusted R^2 = 1 – ResMS/TotMS (see p 164)
  - Adjusted R^2 is approximately unbiased.

$$\bar{R}^2 = R^2 - \frac{P(1-R^2)}{N-P-1}$$

Mallows C_p

- *Mallows* C_p is a measure of the relative quality of statistical models for a given set of data
- Named after *Colin Lingwood Mallows*
- Compares MSE of reduced model to full model.

$$C_p = (N - P - 1)\left(\frac{ResMS_{reduced}}{ResMS_{full}} - 1\right) + (P + 1)$$

Akaike Information Criterion (AIC)

- Similar to C_p
- Tries to find a parsimonious model
- Uses an information function, likelihood function

AIC = -2 log (likelihood) + 2 * (number of parameters)

$$AIC = N\ ln\frac{RSS}{N} + 2(P+1)$$

Variable Selection Procedures: Stepwise Regression

- **Forward selection:** X variables added one at a time until optimal model is reached.
- **Backward elimination:** X variables removed one at a time until optimal model is reached.
- **Stepwise selection:** combines Forward and Backward.

Forward Selection:

- Start with no X variables
- Step 0: compute simple R's of Y with each X_i , and corresponding p-values.
- Step 1: enter variable with highest |R| (smallest *p*-value)
  - Compute partial R (or B) of Y with each other X
- Steps 2 – N: enter variable with smallest *p*
    - Compute partial R (or B) of Y with each other X
- Until smallest *p* > cutoff value.
  - Cutoff for *p* to enter = 0.15
  - Cutoff for *p* to remove = 0.30

Example: Forward Selection (p 171)

Table 8.3: Forward selection of variables for chemical companies

Variables added	Computed F-to-enter	Multiple R^2	Multiple $\bar{R}^2$
1. D/E	8.09	0.224	0.197
2. PAYOUTR1	2.49	0.290	0.237
3. NPM1	9.17	0.475	0.414
4. SALESGR5	3.39	0.538	0.464
5. EPS5	0.38	0.545	0.450
6. ROR5	0.06	0.546	0.427

Example: Forward Selection (cont)

- *p* - values
  - Step 1 = 0.008,
  - Step 2 = 0.13
  - Step 3 = 0.006
  - Step 4 = 0.08
  - Step 5, 6 >0.15
- Select the first 4 variables: D/E, PAYOUTR1, NPM1, SALESGR5

Backward Selection:

- Start with all X variables
- Step 0: compute partial R of Y with each X_i , given all other X's, and corresponding *p*-values (*p* to remove).
- Step 1: remove variable with lowest |partial R| (highest *p*-value)
  - Compute partial R of Y with each remaining X, given the other X's left in, and corresponding *p* to remove
- Steps 2, …,N: remove variable with highest *p*
    - Compute partial R (or B) of Y with other X's, given X's remaining, and corresponding p to remove
- Until highest *p* to remove < cutoff value.
  - Cutoff for *p* to enter = 0.15
  - Cutoff for *p* to remove = 0.30

Example: Backward Selection (p 172)

Table 8.3: Forward selection of variables for chemical companies

Variables added	Computed F-to-enter	Multiple R^2	Multiple $\bar{R}^2$
1. D/E	8.09	0.224	0.197
2. PAYOUTR1	2.49	0.290	0.237
3. NPM1	9.17	0.475	0.414
4. SALESGR5	3.39	0.538	0.464
5. EPS5	0.38	0.545	0.450
6. ROR5	0.06	0.546	0.427

- Step 1: Remove ROR5
- Go to Step 2.

Full Stepwise Selection:

- Because they are one at the time they may not find the "optimal" model
- Several Combination
- For example:
  - Start with Forward Selection
  - At each step: Look at variables "in" as candidates for removal.
  - Repeat until no X can be added or removed.

Best Subset Selection:

- Select one X with highest simple R with Y
- Select two X's with highest multiple R with Y
- Select three X's with highest multiple R with Y
- Repeat – Computing C_p , adjusted R^2 , or AIC with each iteration
- Compare and choose among the "best" subsets of various sizes.

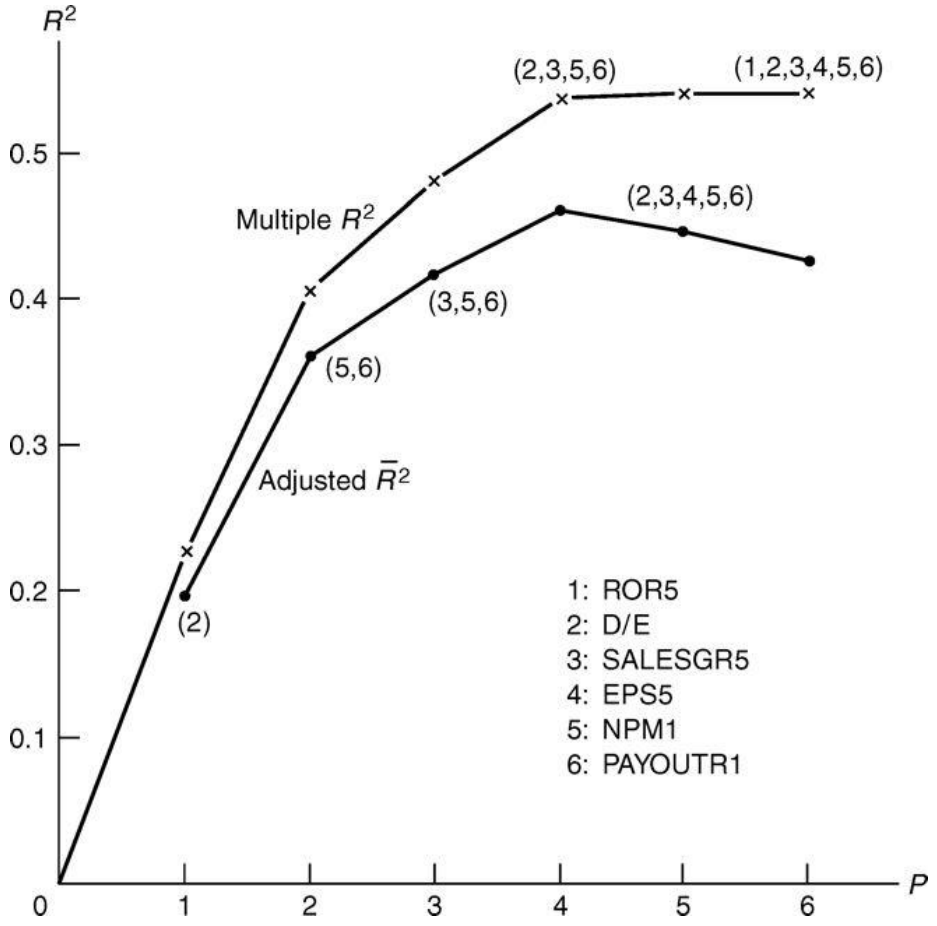
Example: Chemical Companies Data (p 175)

Table 8.6 Best three subsets for one, two, three, or four variables for chemical companies' data selected by stepwise procedure

Number of variables	Names of variables	R^2	$\bar{R}^2$	C_p	AIC
1	D/E	0.224	0.197	13.30	57.1
1	NPM1	0.123	0.092	18.40	60.8
1	PAYOUTR1	0.109	0.077	19.15	61.3
2	NPM1, PAYOUTR1	0.408	0.364	6.00	51.0
2	PAYOUTR1, ROR5	0.297	0.245	11.63	56.2
2	ROR5, EPS5	0.292	0.240	11.85	56.3
3	PAYOUTR1, NPM1, SALESGR5	0.482	0.422	4.26	49.0
3	PAYOUTR1, NPM1, D/E	0.475	0.414	4.60	49.4
3	PAYOUTR1, NPM1, ROR5	0.431	0.365	6.84	51.8
4	PAYOUTR1, NPM1, SALESGR5, D/E	0.538	0.464	3.42	47.6
4	PAYOUTR1, NPM1, SALESGR5, EPS5	0.498	0.418	5.40	50.0
4	PAYOUTR1, NPM1, SALESGR5, ROR5	0.488	0.406	5.94	50.6
Best 5	PAYOUTR1, NPM1, SALESGR5, D/E, EPS5	0.545	0.450	5.06	49.1
All 6		0.546	0.427	7.00	51.0

Example: Chemical Companies Data Multiple R^2 and Adjusted R^2 vs. P (p 176)

Figure 8.1
Multiple R^2 and Adjusted R^2 versus P for Best Subset with P Variables for Chemical Companies' Data



Example: Chemical Companies Data – C_p vs p (p 177)

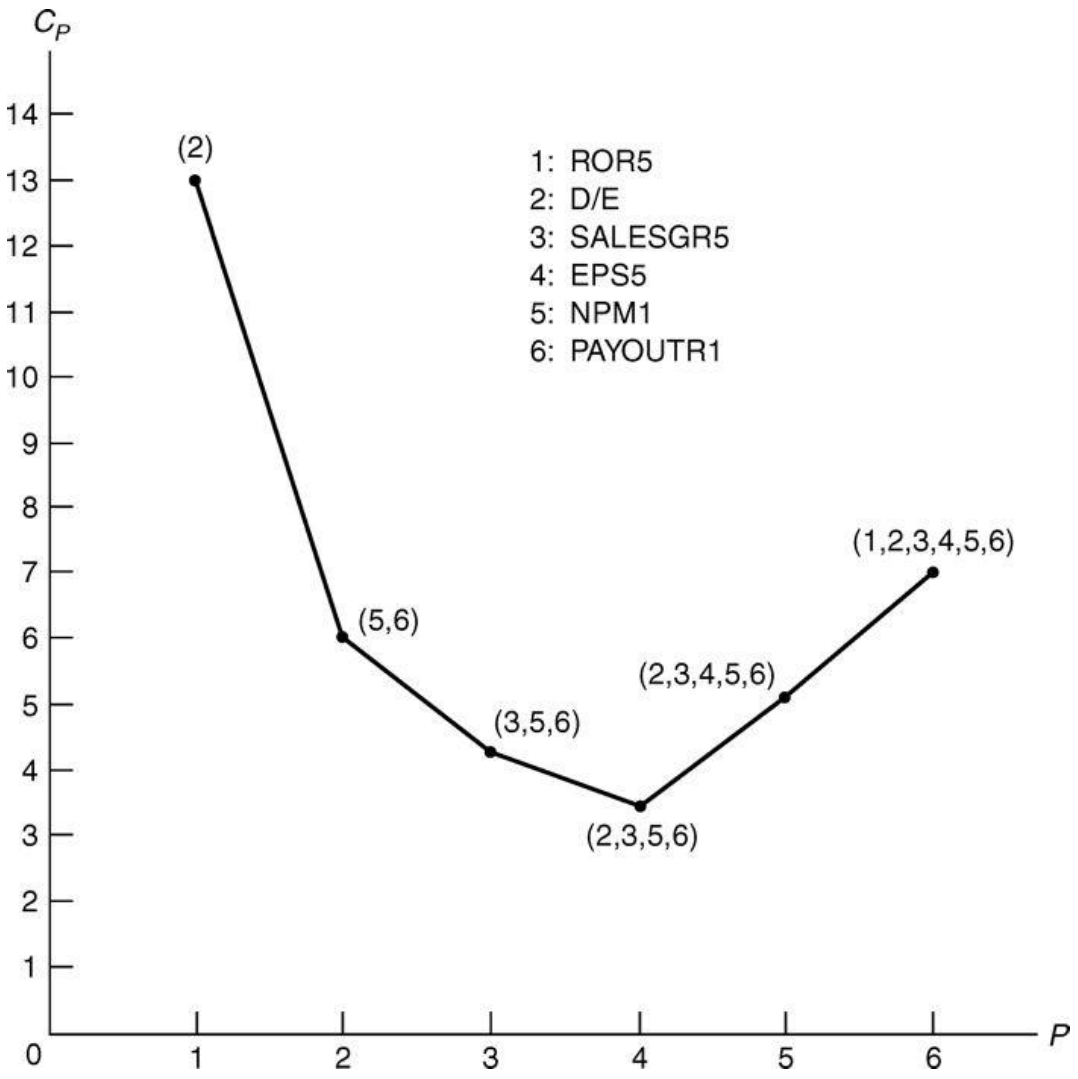


Figure 8.2
C_p versus P for Best Subset with P Variables for Chemical Companies' Data

Example: Chemical Companies Data – AIC vs. p (p 178)

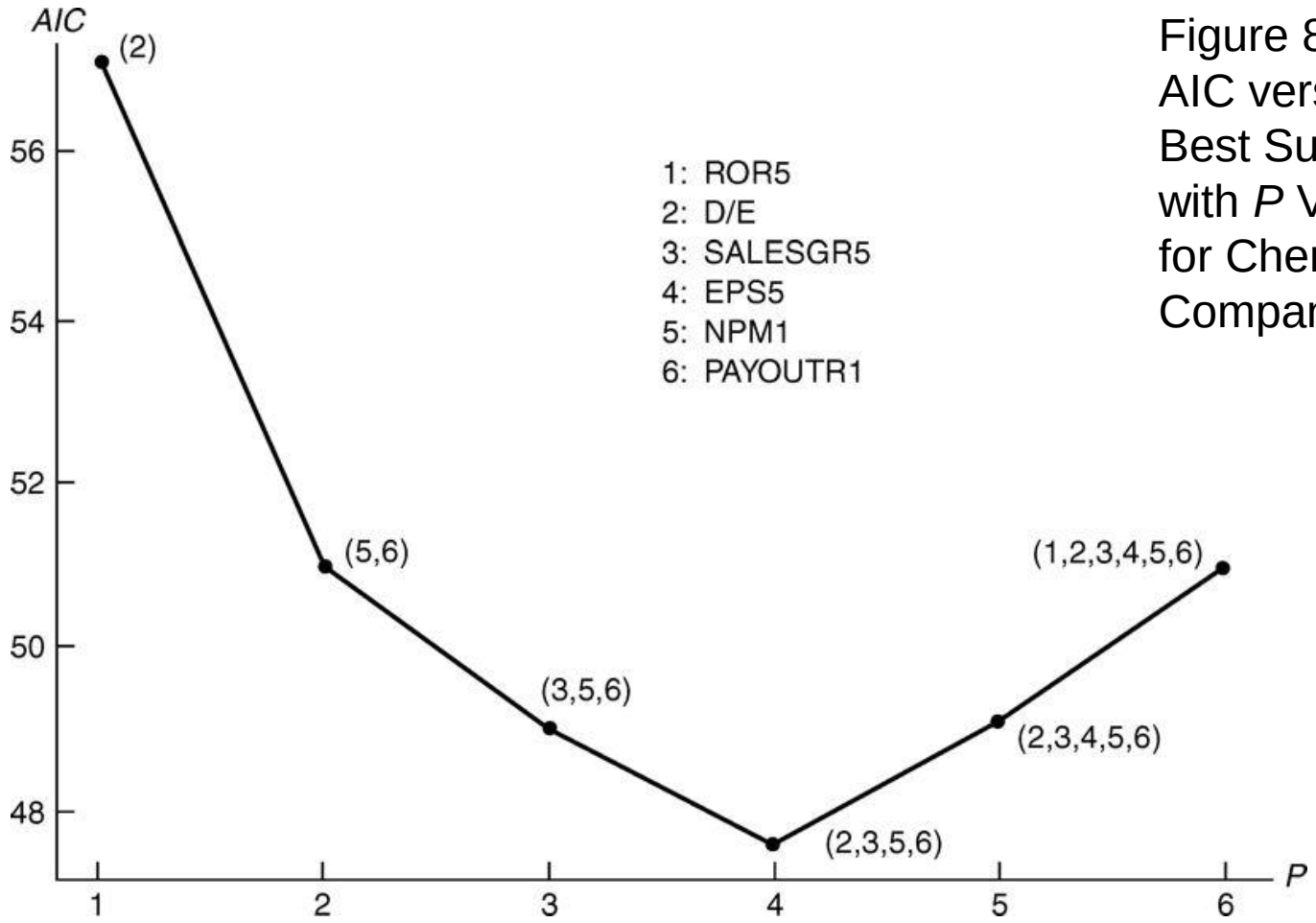


Figure 8.3
AIC versus *P* for Best Subset with *P* Variables for Chemical Companies' Data

Computer Output (p 179)

Table 8.7
Software commands and output for various variable selection methods

	S-PLUS/R	SAS	SPSS	Stata	STATISTICA
Variable selection statistic used					
F-to-enter or remove	stepwise (S-PLUS only)	REG	REGRESSION		General Regression Models
Alpha for F			REGRESSION	sw: regress	General Regression Models
R^2	leaps	REG			General Regression Models
Adjusted R^2	leaps	REG		vselect[a]	General Regression Models
C_P	leaps	REG		vselect[a]	General Regression Models
Printed variable selection criteria					
R^2	leaps	REG	REGRESSION	sw: regress	General Regression Models
Adjusted R^2	leaps	REG	REGRESSION	sw: regress	General Regression Models
C_P	leaps	REG	REGRESSION	vselect[a]	General Regression Models
AIC	step	REG	REGRESSION	vselect[a]	Generalized Linear/Nonlinear Models
Selection method					
Forward	step	REG	REGRESSION	sw: regress	General Regression Models
Backward	step	REG	REGRESSION	sw: regress	General Regression Models
Stepwise	step	REG	REGRESSION	sw: regress	General Regression Models
Forcing variable entry	leaps	REG	REGRESSION	sw: regress	General Regression Models
Best subsets	leaps	REG		vselect[a]	General Regression Models
All subsets	leaps	REG		vselect[a]	
Effects of variable plots					
Partial residual plots			REGRESSION	cprplot	Multiple Regression
Partial regression plots		REG	REGRESSION	avplot	
Augmented partial residual plot				acprplot	

[a] User written command

Model Validation

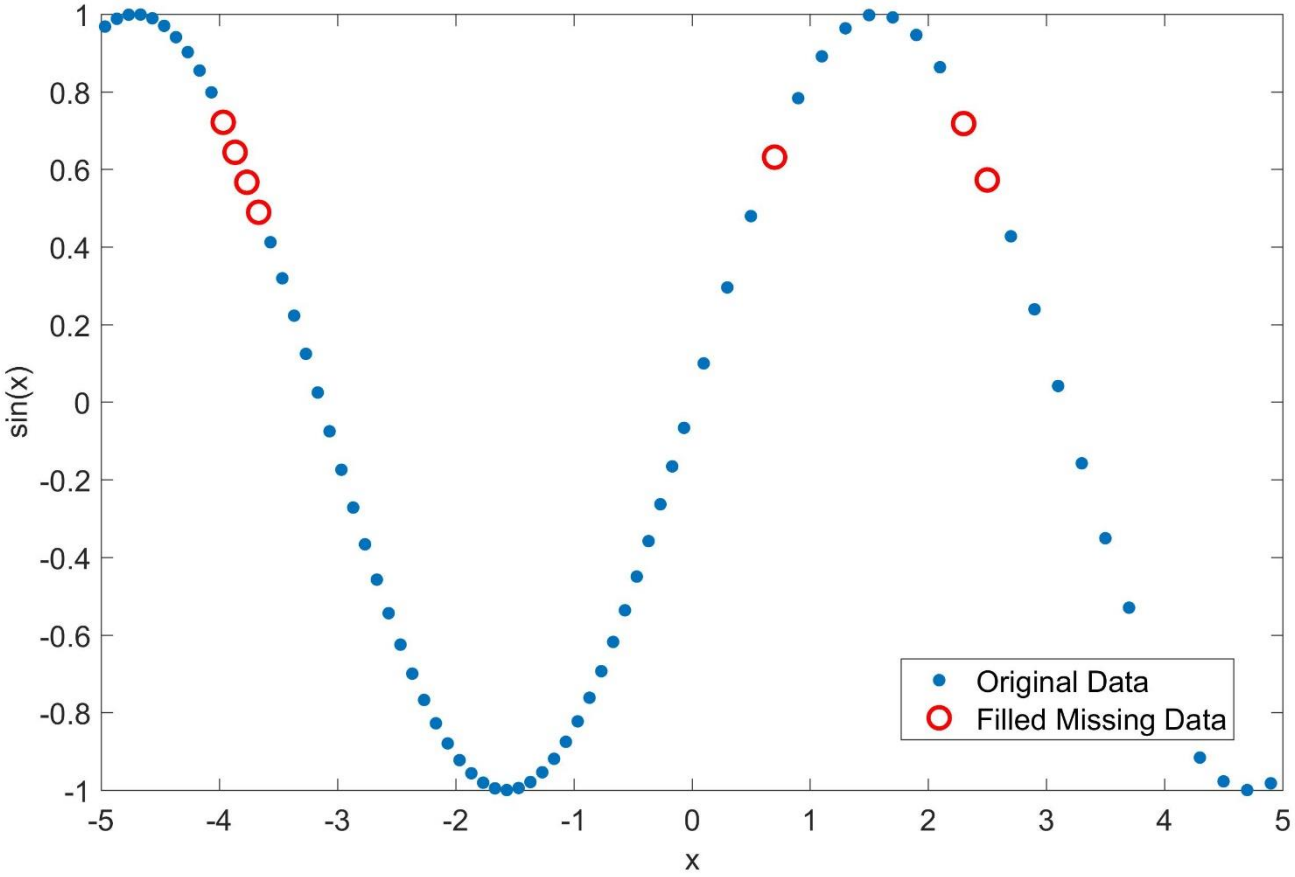
- Checking Selected Model:
  - Split into training sample (e.g. 3/4) and validation sample (e.g. 1/4). Use most convenient partition.
  - Compute regression equation from training sample
  - Use that equation to compute predicted values in the validation sample
  - Compute simple R of observed and predicted Y in validation sample, and compare it with multiple R in training sample.
  - Compare residuals in the two samples.

Improving Regression Analysis

- Missing Values
  - Imputation
  - Multiple Imputation

- Caveat: Missing values are very often present, and the theory of dealing with missing values is an active area of research. Most approaches introduce bias.

Imputation of Missing Values



Imputation Techniques

- Listwise/Pairwise Deletion
- Hot/Cold Deck
- Mean
- Regression
- Last Observation Carried Forward
- Stochastic
- Multiple
- Classified (e.g. Bus/Res)

Imputation Techniques (2)

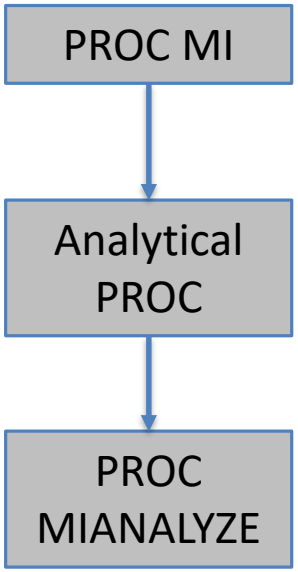
- Listwise Deletion:
  - Delete all cases (rows) with a significant missing value.
  - Effectiveness dependent on number of rows with missing values and randomness of missing data.
- Pairwise Deletion:
  - Delete only specific missing variables and cases if they are required in analysis.
  - Preserves more data that listwise for some analyses. N will vary across analyses.
- Hot Deck:
  - Missing value imputed from randomly selected similar record.
- Cold Deck:
  - Missing value selected from another data set.

Imputation Techniques (3)

- Mean:
  - Replace missing variables with the mean of the remaining variables.
  - Preserve mean for variable.
  - Creates problems with covariances – problem for multivariate analysis
  - Requires a second pass
- Regression / Stochastic:
  - Create a simple regression model to predict missing value.
  - Stochastic regression added average regression variance to regression value to introduce error, hence less bias.
- Last Observation Carried Forward:
  - Sort data set by other variables, then carry forward the value of the immediate predecessor.

Imputation Techniques (4)

- Multiple Imputation (e.g. MCMC):



Create multiply imputed data sets

Analyze complete data using standard procedures

Combine complete data results for inference

LOGCOST with and without MI

Table 2 Regression of LOGCOST on baseline and demographic variables before and after applying multiple imputation

	Without multiple imputation			Using multiple imputation		
LOGCOST	Estimate	Standard error	P-value	Estimate	Standard error	P-value
SEVERITY	0.66	0.07	<0.0001	0.65	0.07	<0.0001
Intercept	9.07	0.30	<0.0001	8.76	0.07	<0.0001
UNIT	−0.16	0.07	0.01	−0.17	0.07	0.01
MALE	0.08	0.07	0.22	0.08	0.07	0.21
LATINO	−0.07	0.09	0.46	−0.06	0.09	0.51
BLACK	0.13	0.09	0.14	0.11	0.09	0.23
ASIAN	−0.22	0.14	0.12	−0.21	0.14	0.15
OTHER	−0.32	0.12	0.007	−0.33	0.12	0.007
PRE-COST	0.80	0.08	<0.0001	0.80	0.08	<0.0001
AGE	0.12	0.07	0.09	0.14	0.07	0.08
DAYS	−0.003	0.002	0.29	−0.002	0.002	0.31

(*Afifi)

Python code: Multiple regression/Adj R-square/ AIC

```
import pandas as pd
import statsmodels.formula.api as smf
```

- Multiple regression without interaction

```
model = smf.ols(formula='mpg ~ wt + cyl', data=mtcars).fit()
summary = model.summary()
summary
```

OLS Regression Results

Dep. Variable:	mpg	**R-squared:**	0.830
Model:	OLS	**Adj. R-squared:**	0.819
Method:	Least Squares	**F-statistic:**	70.91
Date:	Tue, 19 Feb 2019	**Prob (F-statistic):**	6.81e-12
Time:	17:56:29	**Log-Likelihood:**	-74.005
No. Observations:	32	**AIC:**	154.0
Df Residuals:	29	**BIC:**	158.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	39.6863	1.715	23.141	0.000	36.179	43.194
wt	-3.1910	0.757	-4.216	0.000	-4.739	-1.643
cyl	-1.5078	0.415	-3.636	0.001	-2.356	-0.660

- Multiple regression with interactions

In [13]:
```
model_interaction = smf.ols(formula='mpg ~ wt + cyl + wt:cyl', data=mtcars).fit()
summary = model_interaction.summary()
summary.tables[1]
```

Out[13]:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	54.3068	6.128	8.863	0.000	41.755	66.858
wt	-8.6556	2.320	-3.731	0.001	-13.408	-3.903
cyl	-3.8032	1.005	-3.784	0.001	-5.862	-1.745
wt:cyl	0.8084	0.327	2.470	0.020	0.138	1.479

Python code: inputting missing value

- Dataset with missing value

	A	B	C
0	0.726668	-1.047753	0.086460
1	1.341999	1.245280	-0.921736
2	-0.965661	1.508397	-1.551891
3	NaN	-0.712767	-0.219648
4	NaN	NaN	-1.171752
5	0.542686	NaN	NaN
6	2.210098	-0.581692	NaN
7	0.382060	0.236265	NaN
8	0.671490	0.409287	-0.594266
9	1.419332	-0.815497	-1.702727

- Replace missing value with mean

```
dff.fillna(dff.mean())
```

	A	B	C
0	0.726668	-1.047753	0.086460
1	1.341999	1.245280	-0.921736
2	-0.965661	1.508397	-1.551891
3	0.791084	-0.712767	-0.219648
4	0.791084	0.030190	-1.171752
5	0.542686	0.030190	-0.867937
6	2.210098	-0.581692	-0.867937
7	0.382060	0.236265	-0.867937
8	0.671490	0.409287	-0.594266
9	1.419332	-0.815497	-1.702727

Ps: if you have a very large dataset you also can use df.dropna()

Quick Matrix Review

$(A - \lambda I) X = 0$

$A = \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}$ $\quad (A - \lambda I) = \begin{pmatrix} 3-\lambda & 1 \\ 2 & 2-\lambda \end{pmatrix}$

$\lambda = 1,4$

$\lambda = 1 \Rightarrow y = -2x$

$\lambda = 4 \Rightarrow y = x$

Quick Matrix Review

$(A - \lambda I) X = 0$

$A =$

3	1	3
2	2	5
1	3	2

$(A - \lambda I) =$

3-λ	1	3
2	2-λ	5
1	3	2-λ

$(3-\lambda)(2-\lambda)(2-\lambda)+1*5*1+3*3*2 - ((1*(2-\lambda)*3) + (2*1*(2-\lambda)) + ((3-\lambda)*5*3) = 0$

$(12 -16\lambda + 7\lambda^2 - \lambda^3 +5 +18) - ((6 - 3\lambda) + (4 - 2\lambda) + (45 - 15\lambda)) = 0$

$-20 + 4\lambda +7\lambda^2 - \lambda^3 = 0$

$\lambda = 7.17, -1.76, 1.59$

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

stevens.edu

Amir H Gandomi; PhD
Assistant Professor of Analytics & Information Systems
a.h.gandomi@stevens.edu