

MSAN 601

Linear Regression Analysis

Paul Intrevado

Multiple Linear Regression

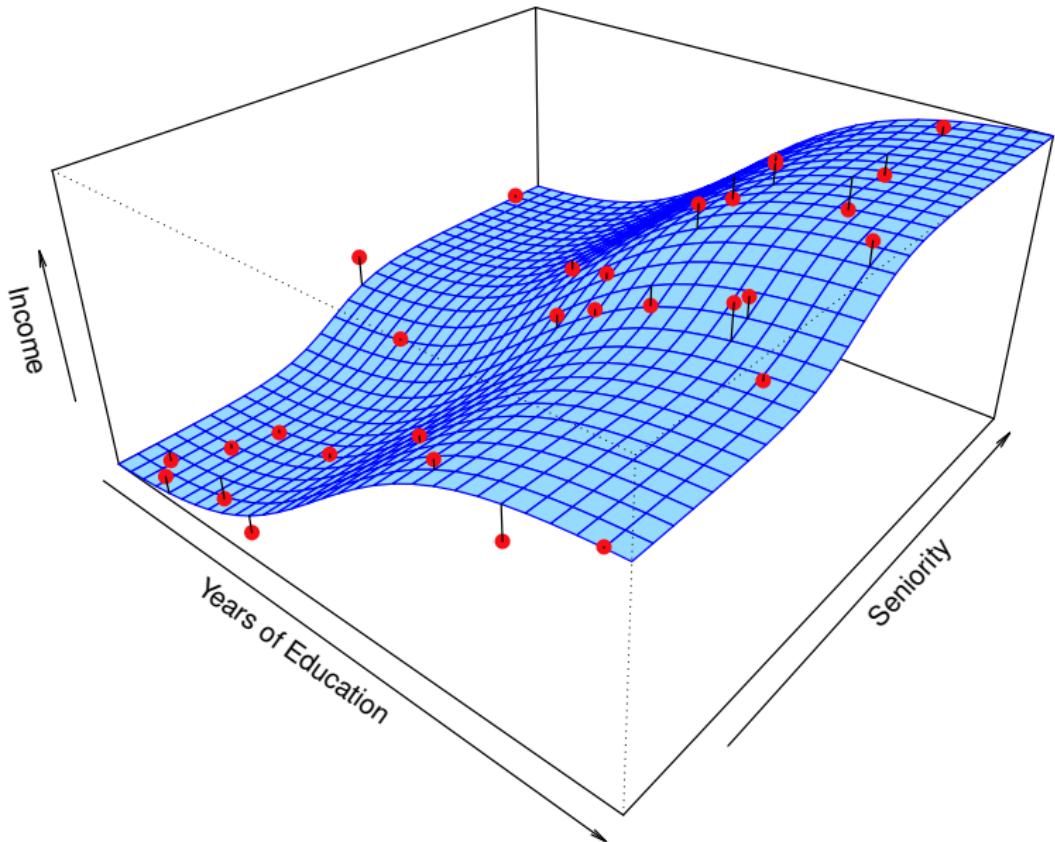
September 12, 2016



UNIVERSITY OF
SAN FRANCISCO

Master of Science
in Analytics

True Functional Relationship of Income Data



Why estimate the functional form?

- Prediction
 - ① Predict Y using $\hat{Y} = \hat{f}(x)$
 - ② $\hat{f}(x)$ often treated as a black box, i.e., we are not concerned with the exact form of \hat{f} , so long as it yields accurate predictions of Y

Why estimate the functional form?

- Prediction
 - ① Predict Y using $\hat{Y} = \hat{f}(x)$
 - ② $\hat{f}(x)$ often treated as a black box, i.e., we are not concerned with the exact form of \hat{f} , so long as it yields accurate predictions of Y
- Inference
 - ① We want to understand how Y changes as a function of the X 's
 - ② \hat{f} can no longer be treated as a *black-box*; we need to know its exact form to determine: what predictors are associated with the response; the relationship between the response and each predictor; and whether or not a linear approximation is appropriate (this is a non-exhaustive list)

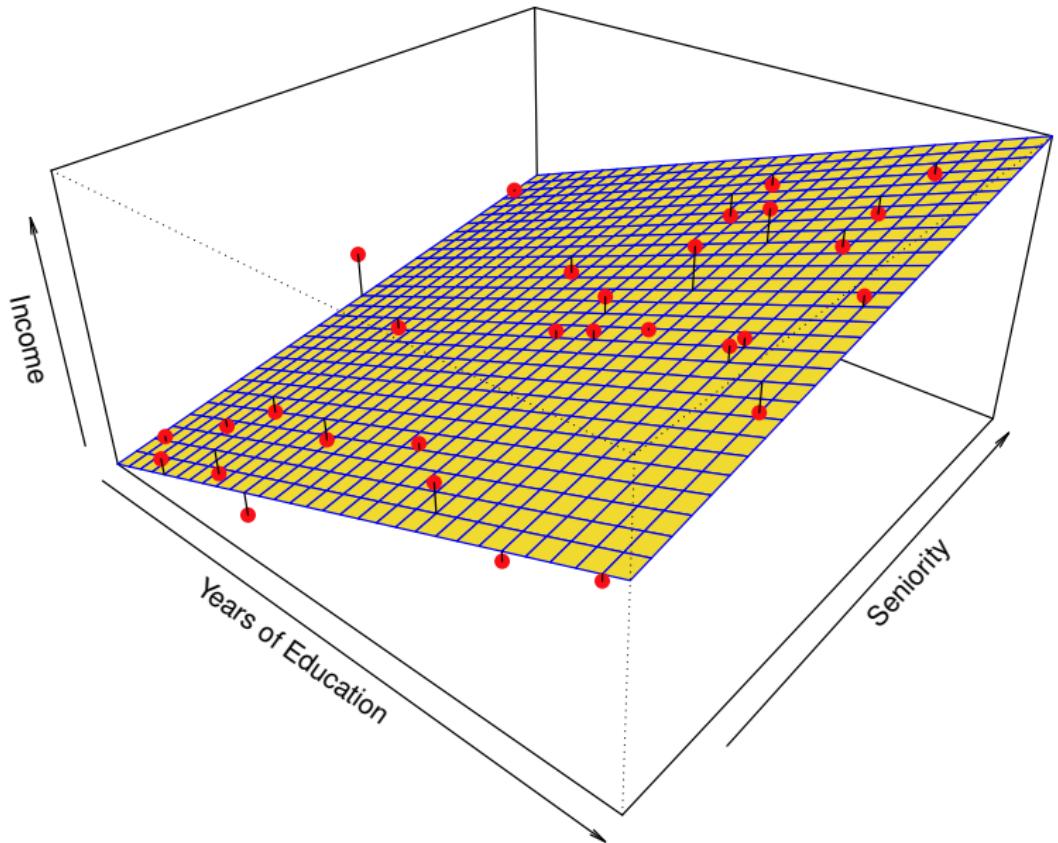
It is possible to choose a model that is both reasonably well-calibrated for both prediction and inference

Parametric vs Non-Parametric Methods

How to Estimate f ?

- Parametric
 - ① Two-step approach which involves (1) making an assumption (or assumptions) about the functional form of f , and (2) using training data to train or fit the model, e.g., in linear regression, estimating the β_k 's
 - ② This method is called parametric, as it reduces the problem to the estimation of a set of parameters
 - ③ This method is often easier as estimating a set of parameters is often less onerous than estimating an entire function f
 - ④ Conversely, parametric models can result in overly-simplistic models which are a poor representation of f

Linear Fit of Income Data

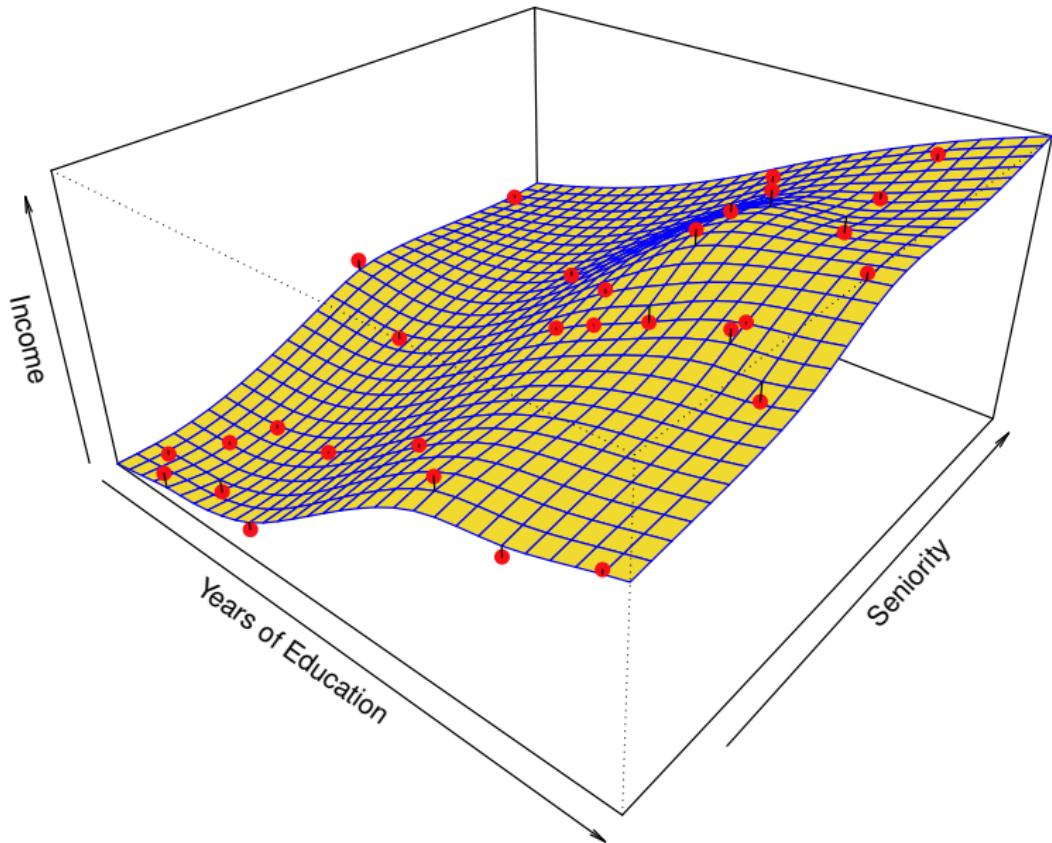


Parametric vs Non-Parametric Methods

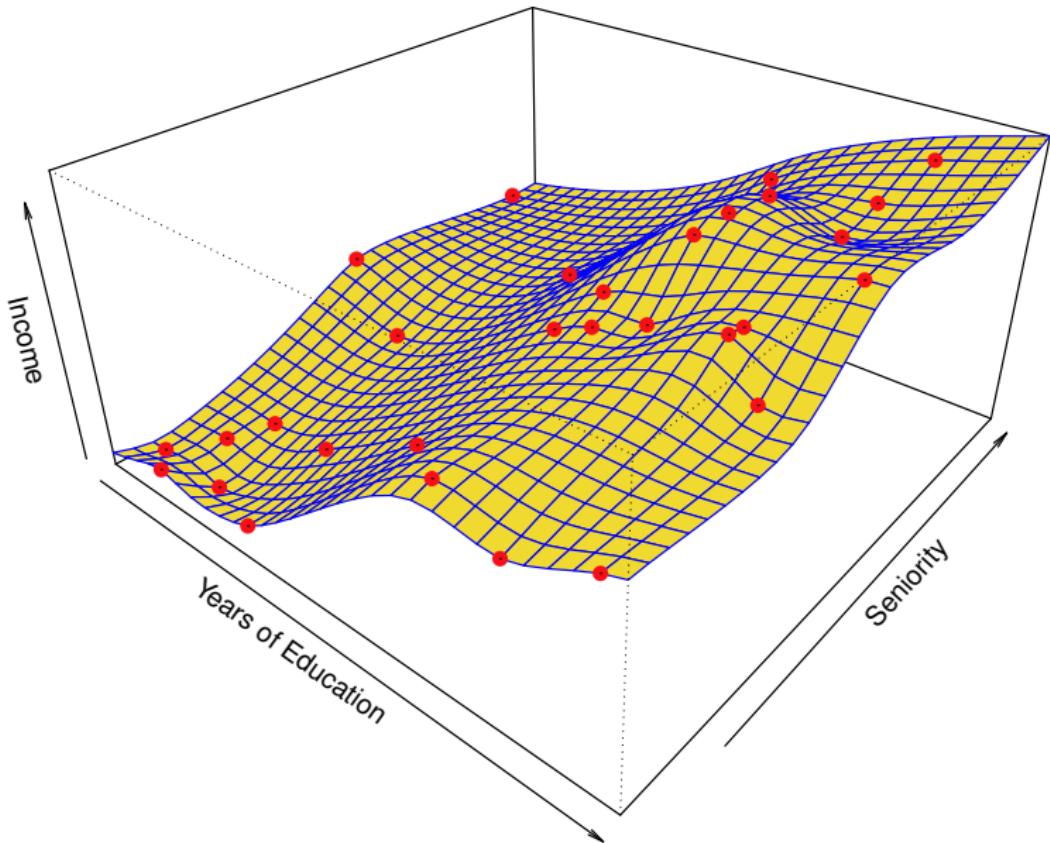
How to Estimate f ?

- Parametric
 - ① Two-step approach which involves (1) making an assumption (or assumptions) about the functional form of f , and (2) using training data to train or fit the model, e.g., in linear regression, estimating the β_k 's
 - ② This method is called parametric, as it reduces the problem to the estimation of a set of parameters
 - ③ This method is often easier as estimating a set of parameters is often less onerous than estimating an entire function f
 - ④ Conversely, parametric models can result in overly-simplistic models which are a poor representation of f
- Non-Parametric
 - ① These models are useful as they are not restricted to assumptions about the functional form of f , and they can accurately fit a wider range of possible shapes for f
 - ② Often requires a very large number of observations, as we are not reducing the model to a set of parameters

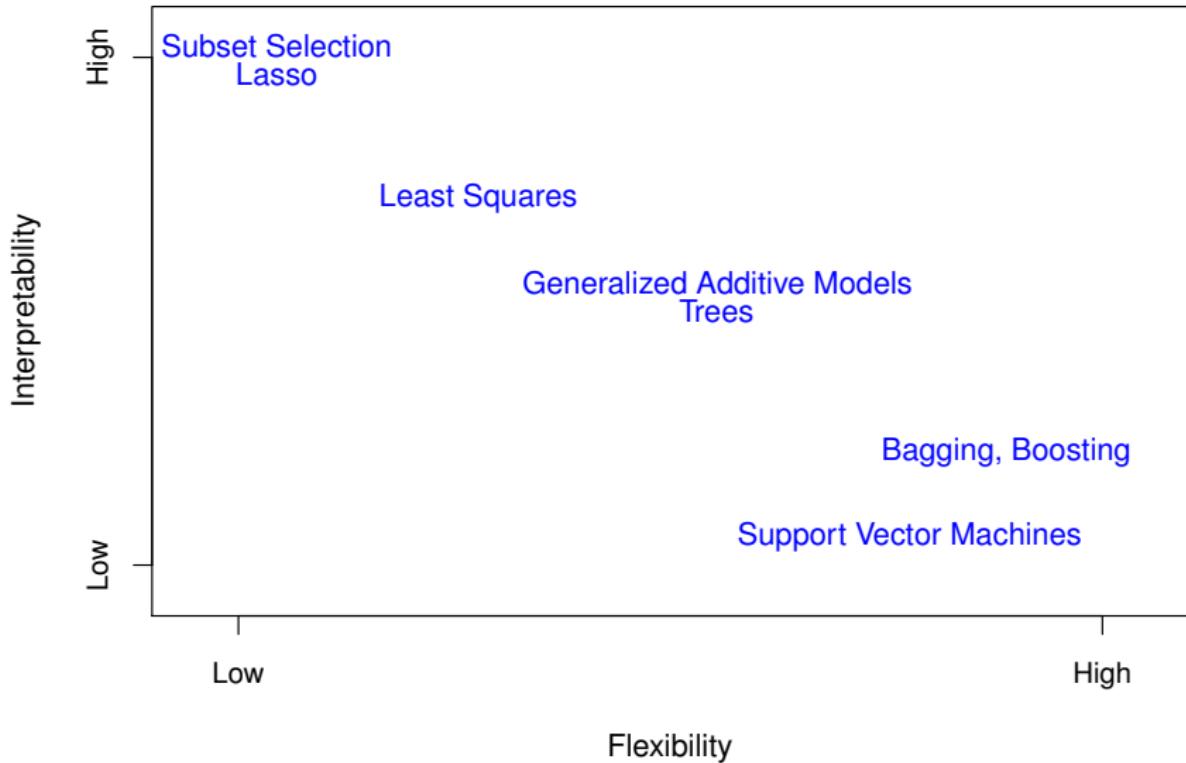
Smooth Thin-Plate Spline Fit of Income Data



Rough Thin-Plate Spline Fit of Income Data



Interpretability vs Flexibility



Why Restrict Ourselves to Linear Models?

- ① If we are interested in inference, linear models are much more interpretable, i.e., establishing the relationship between the response variable Y and the predictors X_1, \dots, X_{p-1} item
- ② Lasso regression employs a linear model, but uses an alternative fitting procedure for estimating the coefficients $\beta_0, \dots, \beta_{p-1}$; a number of coefficients are set exactly to zero (making it additionally restrictive), but the final model is easily interpreted as the response is only related to a small subset of predictors
- ③ Generalized Additive Models (GAMs) extend the linear model to include non-linear relationships but are less interpretable due to the non-linearity of the relationship between response and predictors

Supervised vs Unsupervised Learning

Most statistical learning problems fall into one of two categories:
supervised or unsupervised

Supervised

- Data where each observation of the predictor variables $X_{i1}, \dots, X_{i,p-1}$ is associated with a response variable Y_i
- Techniques includes linear regression, logistic regression, GAMs, boosting, support vector machines, etc.

Unsupervised

- Data where each observation of the predictor variables $X_{i1}, \dots, X_{i,p-1}$ is **not** associated with a response variable Y_i
- Techniques include cluster analysis

Selecting the Best Statistical Learning Model

- There is no one *best* method, there is no universal selection algorithm
- Each case is contextualized and the model you choose will be stylized
- This is where you, the Statistical Learning Professional, will employ your deep set of skills to select and refine the best possible statistical learning model for the application at hand

Assessing Model Accuracy

- A commonly-used measure for assessing the extent to which the predicted response value for a given observation is close to the true response value for that observation is the mean squared error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \quad (1)$$

- We fit our model using **training data**, $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ and compute the **training MSE** (1) to assist in evaluating model quality
- But we are not as interested in how well $\hat{f}(x_i) \approx y_i$ so much as how well $\hat{f}(x_0) \approx y_0$, where (x_0, y_0) is an observation **not used to train the model**, i.e., we seek to develop a model that will minimize the **test MSE**

$$\frac{1}{n} \sum (y_0 - \hat{f}(x_0))^2 \quad (2)$$

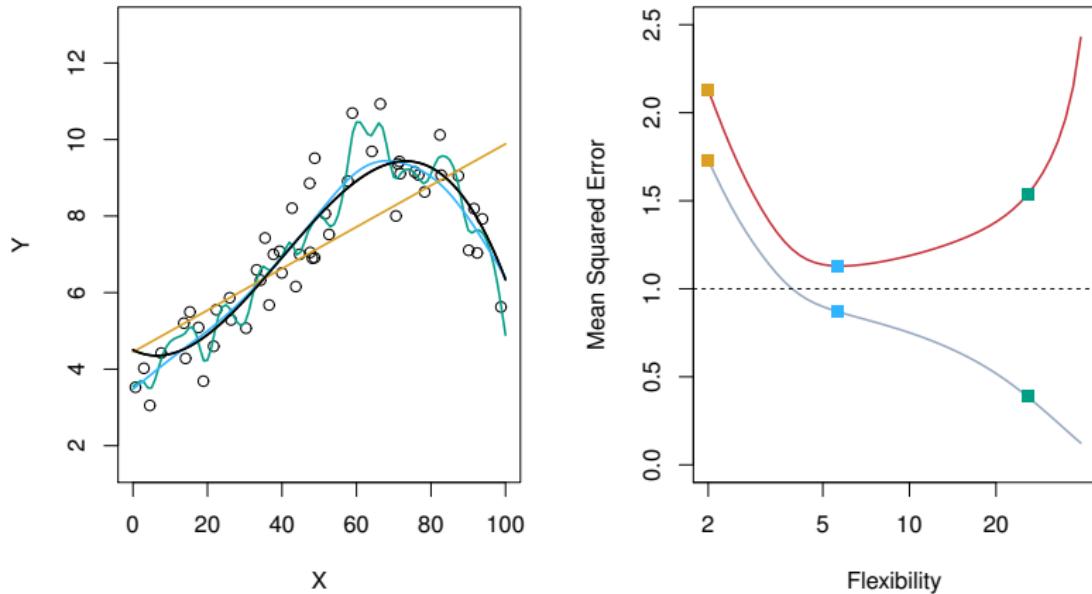
Assessing Model Accuracy cont'd

- What if we have no test data?
- Should we train a model that generates the lowest possible MSE we can find? Why or why not?

Assessing Model Accuracy cont'd

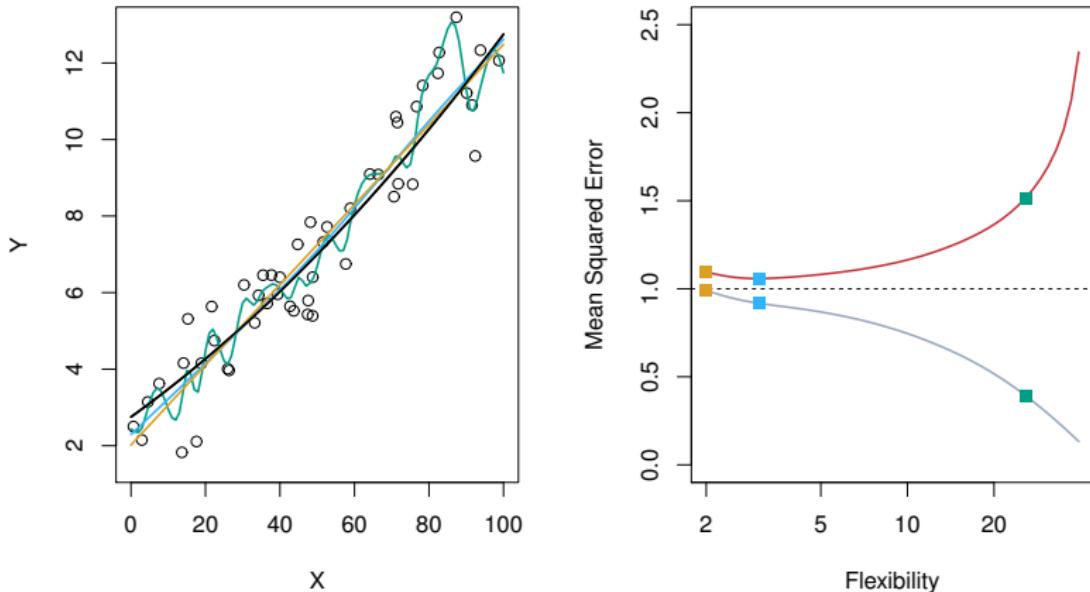
- What if we have no test data?
- Should we train a model that generates the lowest possible MSE we can find? Why or why not?
- There is no guarantee that a model that generates the smallest training MSE will also generate the smallest test MSE

Assessing Model Accuracy cont'd



- Simulated data from f in black
- Training MSE in gray (monotone decreasing in dof)
- Test MSE in red
- Dashed line represents the absolute minimum possible test MSE over all methods

Assessing Model Accuracy cont'd

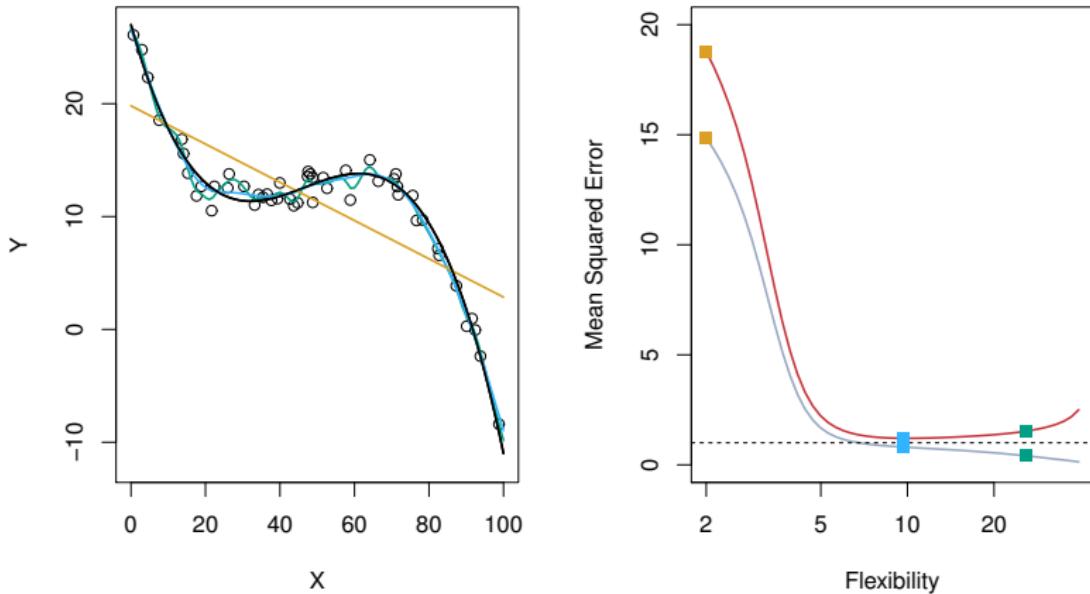


- Simulated data from f in black
- Training MSE in *gray* (monotone decreasing in *dof*)
- Test MSE in *red*
- Dashed line represent the absolute minimum possible test MSE over all methods

Assessing Model Accuracy cont'd

What phenomenon is occurring as we observe an increasing gap between the training and test MSE?

Assessing Model Accuracy cont'd



- Simulated data from f in black
- Training MSE in *gray* (monotone decreasing in *dof*)
- Test MSE in *red*
- Dashed line represents the absolute minimum possible test MSE over all methods

Section 1

MLR w/ Two Quantitative Predictor Variables

Multiple Regression Models

- Multiple regression, as opposed to simple regression, allows us to simultaneously incorporate the information of several predictor variables in an effort to explain the response variable
- When there are **two** predictor variables, X_1 and X_2 , the first-order regression model with **two predictor variables** is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

which is linear in the predictor variables

- X_{i1} and X_{i2} represent the response in the i^{th} trial for predictor variables 1 and 2
- The parameters of the model are β_0 , β_1 and β_2
- Under the assumption of $E\{\varepsilon_i\} = 0$, the regression function is

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Linear Response Surface with Two Predictor Variables

$$E\{Y\} = 10 + 2X_1 + 5X_2$$

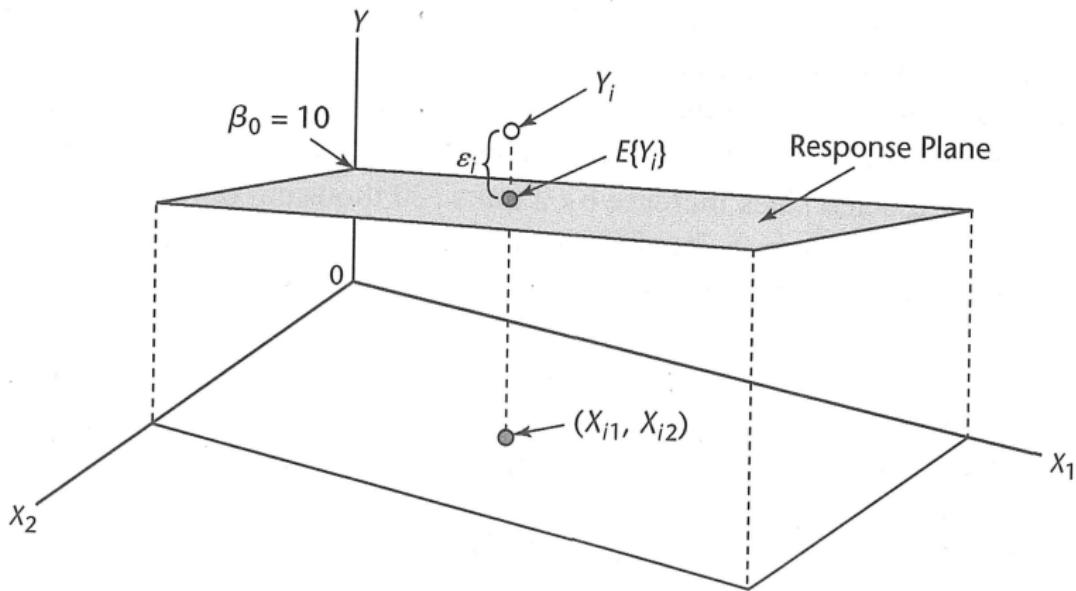


Figure 2: Planar Response Surface w/ Single Observations

Linear Response Surface with Two Predictor Variables

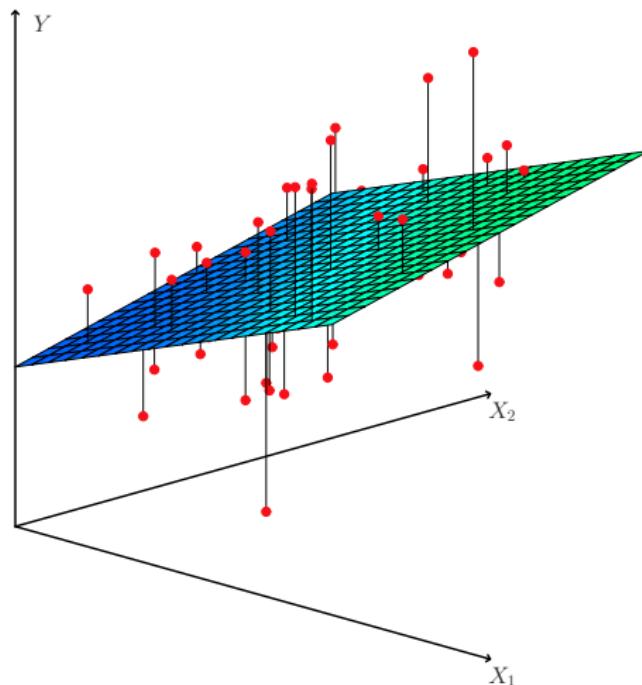


Figure 3: Planar Response Surface w/ Multiple Observations

Interpretation of Regression Coefficients

For a regression model with two predictor variables X_1 and X_2 , the interpretation of the regression coefficients **changes**

- β_0 Unless data is collected around $X_1 = 0$ and $X_2 = 0$, β_0 serves only the mathematical purposes of anchoring the plane on the y -axis; if data is collected around $X_1 = 0$ and $X_2 = 0$, then the value of β_0 is the $E\{Y\}$ when $X_1 = 0$ and $X_2 = 0$
- β_1 Indicates the change in the mean response $E\{Y\}$ per unit increase in X_1 **when X_2 is held constant**
- β_2 Indicates the change in the mean response $E\{Y\}$ per unit increase in X_2 **when X_1 is held constant**

Interpretation of Regression Coefficients cont'd

- Observe that the effect of X_1 on the mean response does not depend on the level of X_2 , and correspondingly, the effect of X_2 does not depend on the level of X_1 , thus the two variables are said to have an *additive effects*, i.e., they do not interact
- In sum, the first-order regression model is designed for predictor variables whose effects on the mean response are additive or do not interact
- In a two-variable multiple-regression setting, the parameters β_1 and β_2 may be referred to as *partial regression coefficients* as they reflect the partial effect of one predictor variable when the other predictor variable included in the model is held constant

Interpretation of Regression Coefficients [EXAMPLE]

A photo studio specializing in children's portraits operates studios in 21 cities. It collected data on sales (Y) in \$1,000's, the number of people aged 16 or younger in the community (X_1) in 1,000's of people, and per capita disposable income (X_2) in \$1000's.

- ① What is the equation of the fitted regression function?
- ② Interpret the coefficient estimates b_1 and b_2 .

Interpretation of Regression Coefficients [R CODE]

```
> dwaine <- read.table("~/Desktop/dwaine.txt",sep="",header=TRUE)

> lm_dwaine <- lm(Sales ~ targetPop + dispInc, data=dwaine)

> summary(lm_dwaine)
```

Call:

```
lm(formula = Sales ~ targetPop + dispInc, data = dwaine)
```

Residuals:

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

Coefficients:

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

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 ? 1

Residual standard error: 11.01 on 18 degrees of freedom

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

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

Interpretation of Regression Coefficients [EXAMPLE]

What does it mean to interpret a coefficient while holding all other variables constant?

- The equation of the fitted regression line is

$$\widehat{\text{sales}} = -68.86 + 1.4546 \text{targetPop} + 9.3655 \text{displnc}$$

- An interpretation of the coefficient of *targetPop* would be:
For each additional increase in the target population of 1,000 people, we expect sales to increase \$1,454 holding disposable income constant
- We can interpret by plugging any numeric value into *displnc* and noting that the intercept will change but the coefficient of *targetPop* will not

Section 2

MLR w/ $p - 1$ Predictor Variables

Regression with $p - 1$ Predictor Variables

- Given $p - 1$ predictor variables X_1, \dots, X_{p-1} , the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (3)$$

is called a **first-order regression model** with $p - 1$ predictor variables

- (3) can be written more compactly as

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i \quad \text{where } X_{i0} \equiv 1 \quad (4)$$

- Under the assumption of $E\{\varepsilon_i\} = 0$, the regression function is

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1} \quad (5)$$

Interpretation of Regression Coefficients

- The response function from (5) is a **hyperplane**
- In this context, when we refer to a hyperplane, we are speaking of a plane that exists in more than three dimensions, and therefore cannot be visualized
- The parameter β_k indicates a change in the mean response $E\{Y\}$ with a unit increase in the predictor variable X_k , given **all other predictor variables in the regression model are held constant**

General Linear Regression Model

- In general, the variables X_1, \dots, X_{p-1} in a regression model do not need to represent different predictor variables (as we shall see shortly)
- We therefore define the general linear regression model (GLRM) with normal error terms as follows

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (6)$$

where

- $\beta_0, \beta_1, \dots, \beta_{p-1}$ are parameters
- $X_{i1}, X_{i2}, \dots, X_{i,p-1}$ are known constants
- ε_i are independent $\sim \mathcal{N}(0, \sigma^2)$
- $i = 1, \dots, n$

The Meaning of 'Linear' in GLRM

- A general linear regression model is not restricted to linear response surfaces
- The term *linear model* refers to the fact that (6) is linear *in the parameters*: **it does not refer to the shape of the response surface**
- We say that a regression model is linear in the parameters when it can be written in the form

$$Y_i = c_{i0}\beta_0 + c_{i1}\beta_1 + c_{i2}\beta_2 + \dots + c_{i,p-1}\beta_{p-1} + \varepsilon_i \quad (7)$$

where the c_i terms are coefficients involving the predictor variables

The Meaning of ‘Linear’ in GLRM cont'd

$$Y_i = c_{i0}\beta_0 + c_{i1}\beta_1 + c_{i2}\beta_2 + \dots + c_{i,p-1}\beta_{p-1} + \varepsilon_i \quad (7)$$

- For example, the first-order model with two predictor variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

is linear in the parameters, with $c_{i0} = 1$, $c_{i1} = X_{i1}$ and $c_{i2} = X_{i2}$

- An example of a non-linear regression model is

$$Y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i$$

which cannot be expressed in the form of (7)

The GLRM in Matrix Form

To express

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (6)$$

in matrix terms, we require the following matrices

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X}_{n \times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\boldsymbol{\beta}_{p \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \boldsymbol{\varepsilon}_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The GLRM in Matrix Form cont'd

In the GLRM matrix representation

- Y represents an $n \times 1$ column vector of the observations of the dependent variable
- X represents an $n \times p$ matrix of the observations of the independent variable
 - Observe the first column of X contains 1's
 - The X matrix is often referred to as the *design matrix*
- β represents a $p \times 1$ column vector of the regression coefficients
- ε represents an $n \times 1$ column vector of error terms

The matrix representation of (6) is

$$\underset{n \times 1}{\mathbf{Y}} = \underset{n \times p}{\mathbf{X}} \underset{p \times 1}{\beta} + \underset{n \times 1}{\varepsilon} \quad (8)$$

Estimated Regression Coefficients

The least squares normal equations for the general linear regression model are

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}'$$

The least squares estimated regression coefficients b_0, b_1, \dots, b_{p-1} , fitted values \hat{Y}_i and residuals e_i , are represented by the column vectors \mathbf{b} , $\hat{\mathbf{Y}}$ and \mathbf{e} respectively

$$\mathbf{b}_{p \times 1} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \quad \hat{\mathbf{Y}}_{n \times 1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \quad \mathbf{e}_{n \times 1} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Fitted Values and Residuals

The fitted values are represented by

$$\hat{\mathbf{Y}}_{n \times 1} = \mathbf{X}\mathbf{b}$$

and the residual terms by

$$\mathbf{e}_{n \times 1} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

The vector of fitted values $\hat{\mathbf{Y}}$ can be expressed in terms of the hat matrix \mathbf{H}

$$\hat{\mathbf{Y}}_{n \times 1} = \mathbf{H}\mathbf{Y}$$

where

$$\mathbf{H}_{n \times n} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

MLR ANOVA Table

Source of Variation	SS	df	MS
Regression	$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right) \mathbf{Y}'\mathbf{J}\mathbf{Y}$	$p - 1$	$MSR = \frac{SSR}{p - 1}$
Error	$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	$n - p$	$MSE = \frac{SSE}{n - p}$
Total	$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right) \mathbf{Y}'\mathbf{J}\mathbf{Y}$	$n - 1$	

Figure 4: MLR ANOVA Table

Note here that \mathbf{J} is an $n \times n$ matrix of ones

F Test for a Regression Relation

To test whether there is a linear regression relation between the response variable Y and **the collection of predictor variables** X_1, X_2, \dots, X_{p-1}

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$$H_a : \text{not all } \beta_k = 0 : k = 1, \dots, p - 1$$

the F^* statistic is employed

$$F^* = \frac{MSR}{MSE}$$

and the decision rule to control for Type I error at α is

If $F^* \leq F(1 - \alpha; p - 1; n - p)$, do not reject H_0

If $F^* > F(1 - \alpha; p - 1; n - p)$, reject H_0

Coefficient of Multiple Determination

- The coefficient of multiple determination, R^2 is defined as follows

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

and measure the proportionate reduction in of total variation in Y associated with the use of the set of variables

X_1, X_2, \dots, X_{p-1}

Coefficient of Multiple Determination

- The coefficient of multiple determination, R^2 is defined as follows

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

and measure the proportionate reduction in of total variation in in Y associated with the use of the set of variables

X_1, X_2, \dots, X_{p-1}

- How many variables should our model have?
 - As many as possible?
 - A rule of thumb for the approximate number it takes to get a good model?
 - Other ideas?

Adjusted Coefficient of Multiple Determination

- **You can't unexplain error.** What does this statement mean?

Adjusted Coefficient of Multiple Determination

- **You can't unexplain error.** What does this statement mean?
- As you add more variables to the model, R^2 will never get smaller, as adding more variables to the model can only explain more error, not less
- Examining the formula for R^2 , we observe

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

as SSE decreases—which occurs when adding additional variables— R^2 invariably increases

- This behavior may incentivize a modeler to include as many explanatory variables as possible
- The coefficient of multiple determination is $r_{y\hat{y}}^2$, the squared correlation between observed response values (Y_i) and fitted values (\hat{Y}_i)

Adjusted Coefficient of Multiple Determination cont'd

- To offset the mathematical desire to include additional variables and increase the R^2 value, an alternative measure, the Adjusted Coefficient of Multiple Determination, R_a^2 , is computed as follows

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$$

- The adjusted coefficient of multiple determination adjusts R^2 by dividing each sum of squares by its associated degrees of freedom
- R_a^2 explicitly creates a conflict between two terms when an additional variable is added to the model: adding a variable will decrease SSE , but will it **sufficiently** decrease SSE to account for an increase in p ?

Adjusted Coefficient of Multiple Determination cont'd

$$R_a^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$$

- ① Adding an additional variable to the GLRM increases p
- ② Increasing p decreases $n - p$
- ③ Decreasing $n - p$ increases the fraction $\left(\frac{n-1}{n-p} \right)$
- ④ Increasing the fraction $\left(\frac{n-1}{n-p} \right)$ increases the entire term $\left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$, as the terms are mutiplicative
- ⑤ Increasing $\left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$ decreases $1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$, which is the definition of R_a^2

Adjusted Coefficient of Multiple Determination cont'd

$$R_a^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$$

- ① Adding an additional variable to the GLRM decreases SSE
- ② Decreasing SSE decreases the fraction $\left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$
- ③ Decreasing the fraction $\left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$ increases $1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$, which is the definition of R_a^2

The question we need answer: will an addition variable, when added to model, sufficiently decrease SSE to counter the increase in p ?

Adjusted Coefficient of Multiple Determination cont'd

$$0 \leq R_a^2 \leq 1$$

- When adding variables to a model, the R_a^2 value can either go up or down
- Iterating through multiple variations of MLR models, observe and record the R^2 and R_a^2 values not only across models, but also track the difference between both of these values for any given modeling variation : if the gap increases, it may be a sign that unnecessary/unhelpful variables have been added to the model

Interval Estimation for β_k

- For the normal error regression model

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim t_{(n-p)}$$

- The confidence limits for β_k with $1 - \alpha$ confidence are

$$b_k \pm t_{(1-\alpha/2;n-p)} s\{b_k\}$$

Tests for β_k

To test individual parameters β_k

$$H_0 : \beta_k = 0$$

$$H_a : \beta_k \neq 0$$

the t^* statistic is employed

$$t^* = \frac{b_k}{s\{b_k\}}$$

and the decision rule is

If $|t^*| \leq t(1 - \alpha/2; n - p)$, do not reject H_0

If $|t^*| > t(1 - \alpha/2; n - p)$, reject H_0

n.b. An F test can also be conducted to determine whether or not $\beta_k = 0$ (to be discussed shortly)

Confidence & Prediction Limits for $E\{Y_h\}$

The $1 - \alpha$ confidence limits for $E\{Y_h\}$ are

$$\hat{Y}_h \pm t_{(1-\alpha/2;n-p)} s\{\hat{Y}_h\}$$

where

$$s^2\{\hat{Y}_h\} = MSE(\mathbf{X}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h) = \mathbf{X}'_h s^2\{\mathbf{b}\} \mathbf{X}_h$$

The $1 - \alpha$ prediction limits for a new observation $Y_{h(new)}$ corresponding to the vector \mathbf{X}_h are

$$\hat{Y}_h \pm t_{(1-\alpha/2;n-p)} s\{pred\}$$

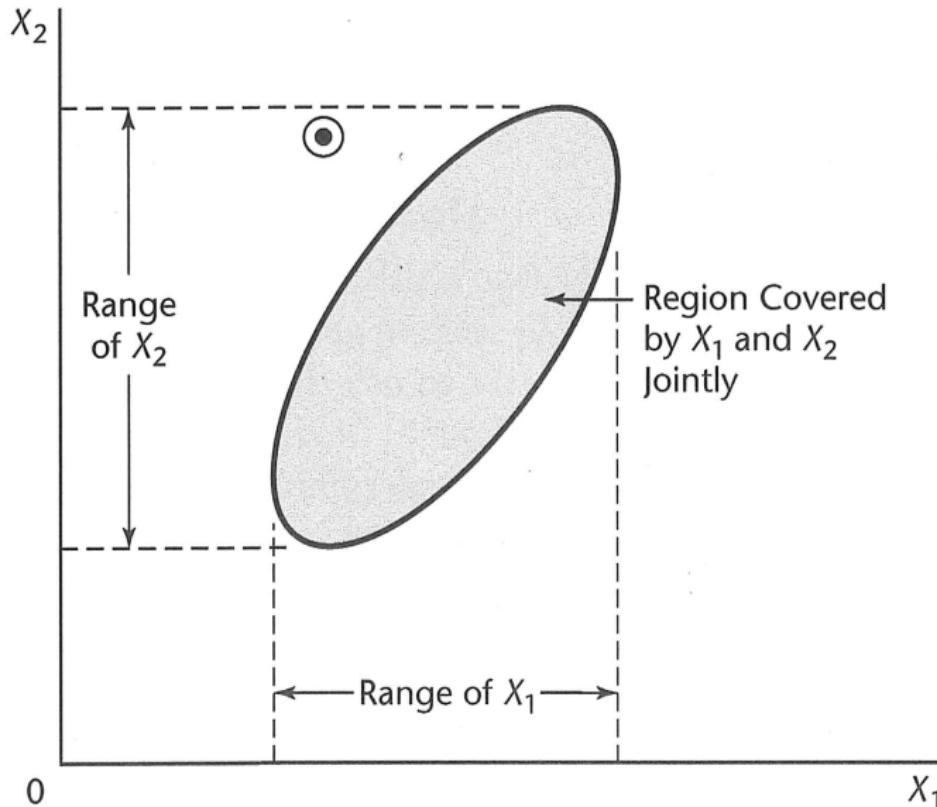
where

$$s^2\{pred\} = MSE + s^2\{\hat{Y}_h\} = MSE(1 + \mathbf{X}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

Be Cautious of Hidden Extrapolations

- In multiple regression, taking care not to extrapolate outside the scope of the model can be particularly difficult as there are multiple predictor variables X_1, X_2, \dots, X_{p-1} **jointly** define the region
- One cannot solely look at the ranges of each predictor variable
- E.g., in a multiple linear regression with two predictor variables, one can choose a value at which they would like predictions to be made, X_{h1} and X_{h2} , which lie within the range of predictor variables X_1 and X_2 yet lies outside of the **joint** region of observations

Be Cautious of Hidden Extrapolations cont'd



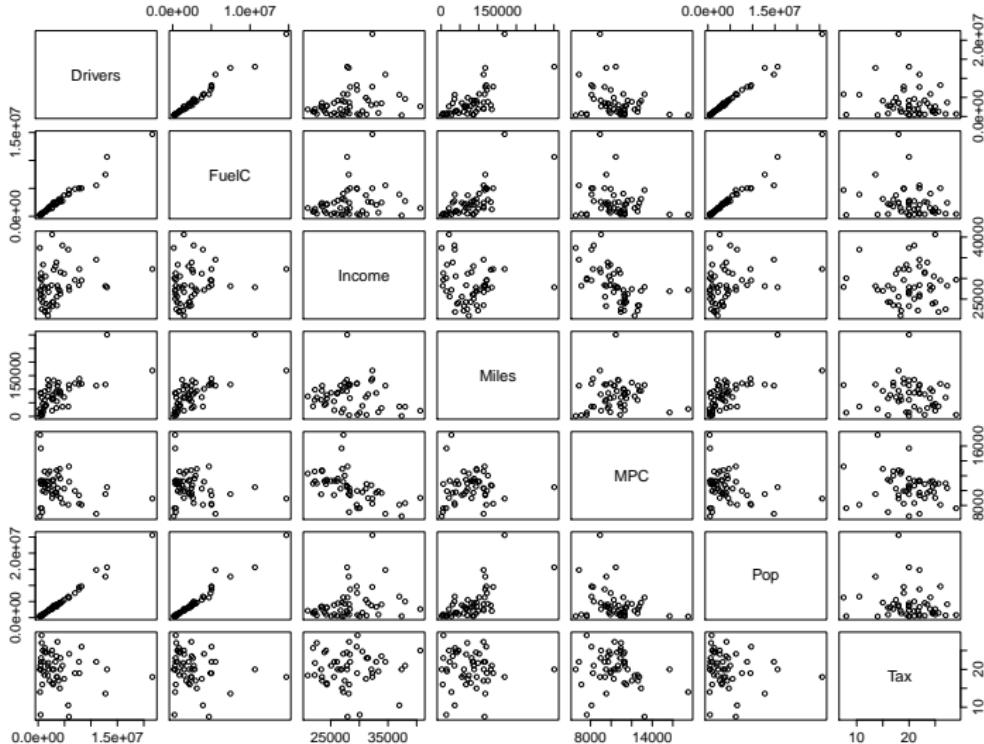
- Scatter Plot Matrices are useful in identifying relationships not only between the dependent and various independent variables, but also within the different independent; strong relationships between predictor variables; a quantitative version of the scatter plot matrix is the correlation matrix which is equally as useful

$$\begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} & \cdots & r_{yx_{p-1}} \\ r_{x_1y} & 1 & r_{x_1x_2} & \cdots & r_{x_1x_{p-1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{x_{p-1}y} & r_{x_{p-1}x_1} & r_{x_{p-1}x_2} & \cdots & 1 \end{bmatrix}$$

```
> cor(fuel2001) # fuel2001 data from alr4 package
```

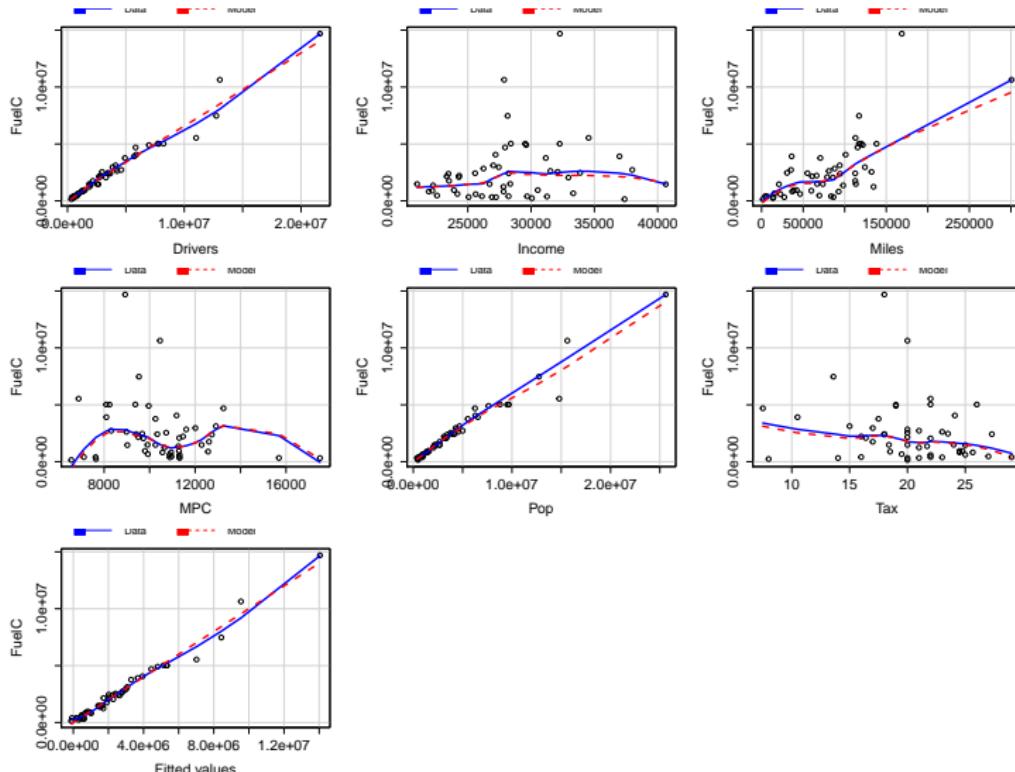
	Drivers	FuelC	Income	Miles	MPC	Pop	Tax
Drivers	1.0000000	0.9850793	0.25617832	0.66864482	-0.26937062	0.9949681	-0.16540553
FuelC	0.9850793	1.0000000	0.21069894	0.72427675	-0.20501689	0.9792750	-0.19988992
Income	0.2561783	0.2106989	1.00000000	-0.13522291	-0.55696630	0.2650850	-0.01068494
Miles	0.6686448	0.7242767	-0.13522291	1.00000000	0.04544132	0.6711723	-0.06450685
MPC	-0.2693706	-0.2050169	-0.55696630	0.04544132	1.00000000	-0.2870154	-0.18217185
Pop	0.9949681	0.9792750	0.26508498	0.67117226	-0.28701537	1.0000000	-0.14586581
Tax	-0.1654055	-0.1998899	-0.01068494	-0.06450685	-0.18217185	-0.1458658	1.00000000

Scatterplot Matrix of fuel2001 Data from alr4



```
> pairs(~ . , data=fuel2001)
```

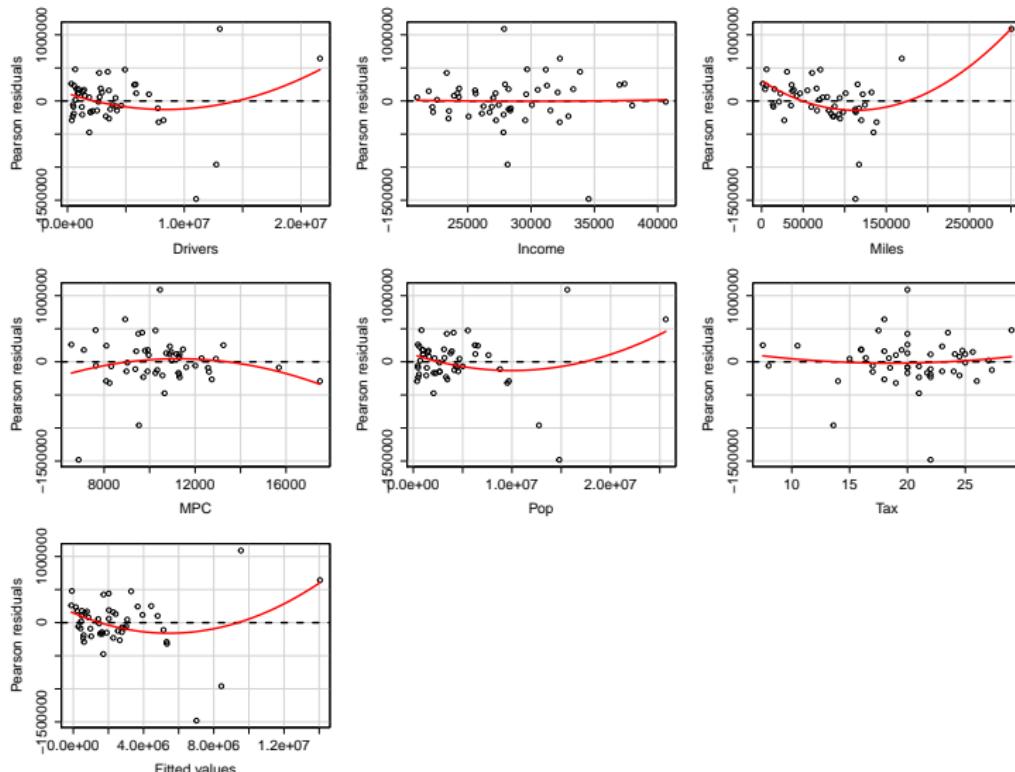
Marginal Model Plots of fuel2001 Data from alr4



```
> lm_fuel <- lm(FuelC ~ . , data = fuel2001)
> marginalModelPlots(lm_fuel)
```

- ② A plot of residuals (or studentized) against **fitted values** is useful for assessing the appropriateness of the multiple regression function and the constancy of error variance of the error terms, as well as for providing information on outliers
- ③ Residuals should also be plotted against **each** of the predictor variables to identify potential issues such as curvature effects or heteroskedasticity with respect to an individual variable

Residual Plots of fuel2001 Data from alr4



```
> residualPlots(lm_fuel) # fuel2001 data from alr4 package  
# residualPlots function from car package
```

Notes on residualPlots from car Package

- Along with generating a matrix of residuals against all predictors as well as against fitted values, the `residualPlots` function also generates a *lack-of-fit* test for each numeric predictor, where

$$H_0 : \text{no lack-of-fit}$$
$$H_a : \text{lack-of-fit}$$

```
> residualPlots(lm_fuel)
```

	Test stat	Pr(> t)
Drivers	2.013	0.050
Income	0.086	0.932
Miles	4.600	0.000
MPC	-1.768	0.084
Pop	1.986	0.053
Tax	0.508	0.614
Tukey test	2.733	0.006

Notes on residualPlots from car Package cont'd

- When testing residuals against the **predictors**, this type of test consists of adding the predictor of interest to the original regression model, albeit to the power 2, rerunning the regression model, and subsequently testing whether or not the coefficient of the squared variable is statistically significantly different from zero

E.g. In a model where Y is regressed on X_1 and X_2 , the term X_1^2 is added to the regression equation, the model is refit and the coefficient of X_1^2 , $\beta_{X_1^2}$ is tested to be statistically significantly different from 0; if yes, conclude that sufficient curvature exists in the residuals to be in violation of the assumptions of linear regression

- The abovementioned process is repeated for each variable, resulting in the t^* and associated p -values from the previous slide
- When testing **residuals against fitted values**, we must use different approach called *Tukey's test for non-additivity*

Notes on residualPlots from car Package cont'd

- The red line added to each of the residual plots against predictors is quadratic
 - To obtain a LOESS curve instead of a quadratic, set `smooth=T` and `quadratic=F` in the `residualPlots` function
- n.b.** Even if you select a LOESS curve to be displayed on your residual plots, the *t*-test for *lack-of-fit* is always run to test the statistical significance of the respective quadratic predictor variable
- Set `type="rstudent"` to generate residual plots with studentized residuals instead of the default Pearson residuals
 - Other conditions for the `residualPlots` function exist that allow you to plot a singular residual plot or your choice of a subset of plots

- ④ Sequence plots of residuals in sequence (or against time) may provide additional diagnostic information
- ⑤ QQ-Plots are also useful for examining the normality of residuals
 - I do not know of a function that generates QQ-Plots for residuals against all predictor variables in a matrix or sequence
- ⑥ Residuals can also be plotted against important omitted variables to identify endogeneity, and also against interaction terms—whether such interaction terms were included in the regression model or not—in an effort to identify potential usefulness of adding an interaction term

- ⑦ The Brown-Forsythe test for heteroskedasticity extends easily to the MLR case
- ⑧ The Breusch-Pagan test for heteroskedasticity can be applied in a similar fashion as the SLR case when testing for heteroskedasticity of residuals
 - ① Regress Y on the X (SLR) or X 's (MLR) that are responsible for heteroskedasticity
 - ② Regress the squared error terms from the aforementioned model on a the same set of predictor variables, and obtain the SSR , which we SSR_{BP}
 - ③ Regress Y on **all** predictor variables to obtain SSE
 - ④ Employ SSR_{BP} and SSE to construct the test statistic, which is now distributed according to a chi-square distribution with q degrees of freedom, where q is the number of predictor variables on which the squared residuals were regressed

- The remainder of this course will focus on various modeling techniques and remedial measures for MLR
- Box-Cox transformations: the method developed by Box & Cox to transform the Y still apply from the SLR chapter; Box & Tidwell also developed an iterative approach to ascertain the appropriate power transformations for each predictor variable in a MLR model

Section 3

Extra Sums of Squares

Extra Sums of Squares

- An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model
- Equivalently, one can view the extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model

Extra Sums of Squares

- When regressing Y on X in an SLR, the SSR for the model is the SSR for the only predictor variable in the model
- When regressing Y on X_1, X_2, \dots, X_{p-1} in an MLR, the SSR for the model is the SSR for **all** the predictor variables in the model, i.e., $SSR(X_1, X_2, \dots, X_{p-1})$, where the comma is to be read statistically as a joint/intersection (\cap)
- The following is an **incorrect** interpretation of the model SSR

$$SSR(X_1, X_2, \dots, X_{p-1}) \neq SSR(X_1) + SSR(X_2) + \dots + SSR(X_{p-1})$$

Extra Sums of Squares

- When regressing Y on X in an SLR, the SSR for the model is the SSR for the only predictor variable in the model
- When regressing Y on X_1, X_2, \dots, X_{p-1} in an MLR, the SSR for the model is the SSR for **all** the predictor variables in the model, i.e., $SSR(X_1, X_2, \dots, X_{p-1})$, where the comma is to be read statistically as a joint/intersection (\cap)
- The following is an **incorrect** interpretation of the model SSR

$$SSR(X_1, X_2, \dots, X_{p-1}) \neq SSR(X_1) + SSR(X_2) + \dots + SSR(X_{p-1})$$

- The correct interpretation of the model SSR is

$$\begin{aligned} SSR(X_1, X_2, \dots, X_{p-1}) &= SSR(X_1) + \\ &\quad SSR(X_2 | X_1) + \\ &\quad SSR(X_3 | X_1, X_2) + \\ &\quad \dots \\ &\quad SSR(X_{p-1} | X_p, \dots, X_1) \end{aligned}$$

SLR ANOVA Table

Source of Variation	SS	df	MS
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n - 2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n - 1$	

Figure 5: SLR ANOVA Table

MLR Type I ANOVA Table

Source of Variation	SS	df	MS
Regression	$\text{SSR}(X_1, X_2, X_3)$	3	$\text{MSR}(X_1, X_2, X_3)$
X_1	$\text{SSR}(X_1)$	1	$\text{MSR}(X_1)$
$X_2 X_1$	$\text{SSR}(X_2 X_1)$	1	$\text{MSR}(X_2 X_1)$
$X_3 X_1, X_2$	$\text{SSR}(X_3 X_1, X_2)$	1	$\text{MSR}(X_3 X_1, X_2)$
Error	$\text{SSE}(X_1, X_2, X_3)$	$n - 4$	$\text{MSE}(X_1, X_2, X_3)$
Total	SSTO	$n - 1$	

- Also known as a sequential ANOVA table, as variables are sequentially added to the regression model
- Call `anova` function from basic R functionality

Extra Sums of Squares: Body Fat EXAMPLE

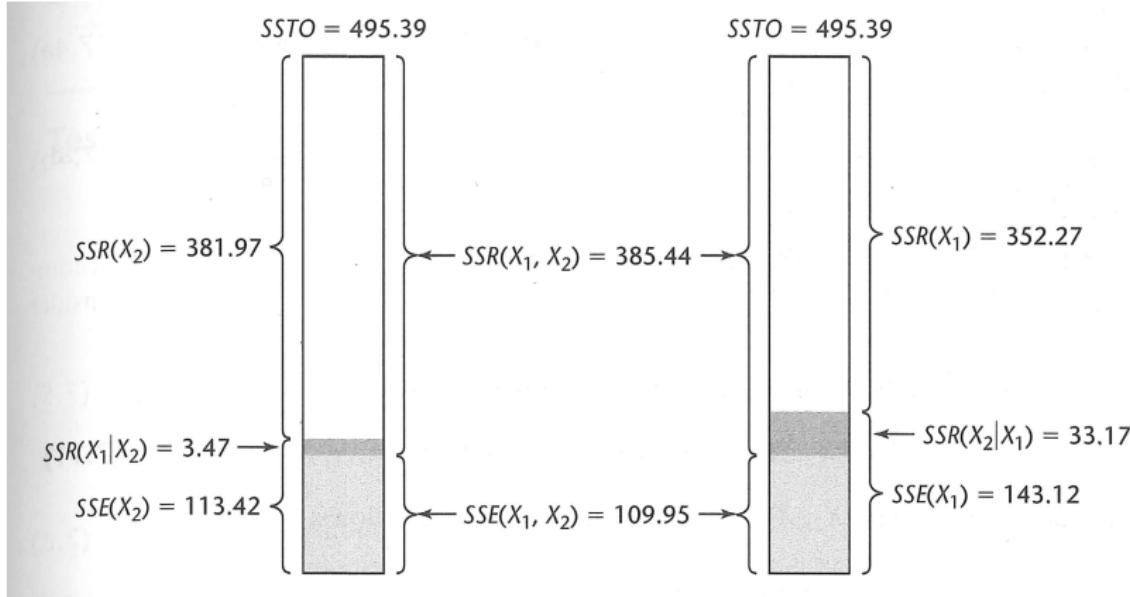


Figure 6: Extra Sum of Squares Schematic

MLR Type II ANOVA Table

Source of Variation	SS	df	MS
Regression	$\text{SSR}(X_1, X_2, X_3)$	3	$\text{MSR}(X_1, X_2, X_3)$
$X_1 X_2, X_3$	$\text{SSR}(X_1 X_2, X_3)$	1	$\text{MSR}(X_1 X_2, X_3)$
$X_2 X_1, X_3$	$\text{SSR}(X_2 X_1, X_3)$	1	$\text{MSR}(X_2 X_1, X_3)$
$X_3 X_1, X_2$	$\text{SSR}(X_3 X_1, X_2)$	1	$\text{MSR}(X_3 X_1, X_2)$
Error	$\text{SSE}(X_1, X_2, X_3)$	$n - 4$	$\text{MSE}(X_1, X_2, X_3)$
Total	SSTO	$n - 1$	

- For each variable, we get a test for adding one of the predictors to a model *that includes all other variables*
- Call Anova, **noting the capital A**, from the car package
- We will soon show that F^* in a Type II ANOVA table is equal to the square of the t^* value for individual coefficients in the model summary

Section 4

Testing Regression Coefficients using Extra
Sums of Squares

Using Extra SS to Test Regression Coefficients

- We already know how to test for the statistical significance of a single $\beta_k = 0$ in a MLR model, namely, by calculating the t statistic
- We already know how to test for the statistical significance of the entire model, i.e., for all $\beta_k = 0$ in a MLR model, namely, by calculating the F statistic
- Extra sums of squares allow us to approach these tests differently
- Let us revisit these tests using Extra SS

Testing Whether a Single $\beta_k = 0$

$$H_0 : \beta_k = 0 \quad 1 \leq k \leq p - 1$$

$$H_a : \beta_k \neq 0$$

- ① Run the full MLR regression model and obtain the $SSE = SSE(X_1, X_2, \dots, X_{p-1})$, which we will call SSE_f , noting the degrees of freedom associated with are $n - p$
- ② Run a **reduced** MLR model, assuming H_0 holds, i.e., regress Y on $X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1}$, obtaining an $SSE = SSE(X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1})$, which we will call SSE_r , noting that there are now $n - p - 1$ degrees of freedom associated with the error of the reduced model
- ③ The F^* test statistic we use is computed as follows

$$F^* = \frac{SSE_r - SSE_f}{df_r - df_f} \div \frac{SSE_f}{df_f}$$

Testing Whether a Single $\beta_k = 0$ cont'd

For ease of exposition, let's assume a MLR model with 3 predictor variables, X_1 , X_2 and X_3 , where we are testing whether $\beta_3 = 0$

$$\begin{aligned} F^* &= \frac{SSE_r - SSE_f}{df_r - df_f} \div \frac{SSE_f}{df_f} \\ &= \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n - 3) - (n - 4)} \div \frac{SSE(X_1, X_2, X_3)}{n - 4} \end{aligned}$$

noting that

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

we rewrite

$$F^* = \frac{SSR(X_3 | X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n - 4}$$

Testing Whether a Single $\beta_k = 0$ cont'd

For ease of exposition, let's assume a MLR model with 3 predictor variables, X_1 , X_2 and X_3 , where we are testing whether $\beta_3 = 0$

$$\begin{aligned} F^* &= \frac{SSE_r - SSE_f}{df_r - df_f} \div \frac{SSE_f}{df_f} \\ &= \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n - 3) - (n - 4)} \div \frac{SSE(X_1, X_2, X_3)}{n - 4} \end{aligned}$$

noting that

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

we rewrite

$$F^* = \frac{SSR(X_3 | X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n - 4}$$

$$F^* = \frac{MSR(X_3 | X_1, X_2)}{MSE(X_1, X_2, X_3)} = (t^*)^2$$

MLR Type II ANOVA Table

```
> anova(lm_bodyFat)
Analysis of Variance Table #### Type I (basic R command)
```

Response: bodyFat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tricep	1	352.27	352.27	57.2768	1.131e-06 ***
thigh	1	33.17	33.17	5.3931	0.03373 *
midarm	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

```
# call the Anova NOT anova (small v capital A)
> Anova(lm-bodyFat) ##(from the 'car' package)
```

Anova Table (Type II tests)

Response: bodyFat

	Sum Sq	Df	F value	Pr(>F)
tricep	12.705	1	2.0657	0.1699
thigh	7.529	1	1.2242	0.2849
midarm	11.546	1	1.8773	0.1896
Residuals	98.405	16		

Making the Distinction Between F Tests

- When we test whether $\beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$, we are running an **overall** F test
- When we test whether $\beta_k = 0$, we are running an **partial** F test
- When we test whether **some** $\beta_k = 0$, we are running an **partial** F test, and the hypothesis test is

$$H_0 : \text{some } \beta_k = 0$$

$$H_a : \text{not all } \beta_k \text{ equal 0}$$

Section 5

Coefficients of Partial Determination

Coefficient of Partial Determination

- Recall that a coefficient of (single or multiple) determination, R^2 , measures the proportionate reduction in the variation of Y achieved by the introduction of the entire set of X variables considered in the model
- A coefficient of **partial determination**, in contrast, measures the **marginal** contribution of one X variable when all others are already included in the model

Coefficient of Partial Determination cont'd

- A simple example best suited to suss the underlying mechanics: assume a MLR with 2 predictor variables
 - $SSE(X_2)$ measures the variation in Y when only X_2 is included in the model
 - $SSE(X_1, X_2)$ measures the variation in Y when both X_1 and X_2 are included in the model
- Hence, the relative marginal reduction in the variation of Y associated with X_1 **when X_2 is already in the model** is

Coefficient of Partial Determination cont'd

- A simple example best suited to suss the underlying mechanics: assume a MLR with 2 predictor variables
 - $SSE(X_2)$ measures the variation in Y when only X_2 is included in the model
 - $SSE(X_1, X_2)$ measures the variation in Y when both X_1 and X_2 are included in the model
- Hence, the relative marginal reduction in the variation of Y associated with X_1 **when X_2 is already in the model** is

$$R_{Y1|2}^2 = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$

- Therefore, $R_{Y1|2}^2$ measures the proportionate reduction in the variation in Y remaining after X_2 is included in the model that is gained by also including X_1

Extra Sums of Squares: Body Fat EXAMPLE REVISITED

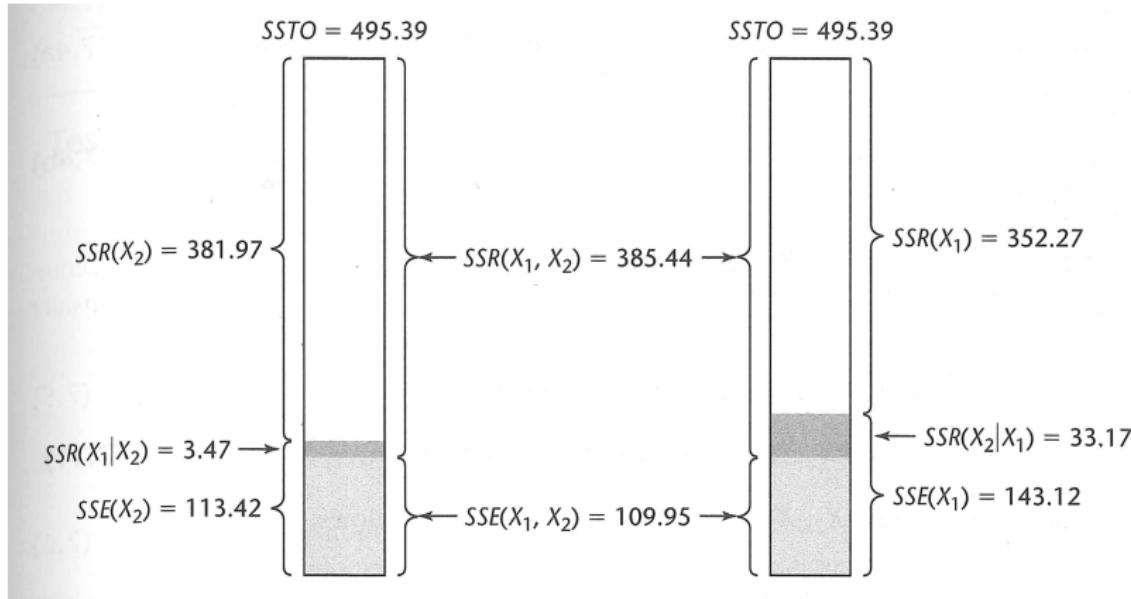


Figure 7: Extra Sum of Squares Schematic

Generate a Type I and Type II ANOVA table and compute all possible coefficients of partial determination.

- What do you conclude about which variables you would now include in your regression model?
- Is either the Type I or Type II ANOVA table useful for this purpose?

Section 6

Multicollinearity and Its Effects

Multicollinearity and Its Effects

In MLR, the nature and significance of the relations between the predictor variables and the response variable are often of particular interest. Common questions include:

- What is the relative importance of the effects of the different predictor variables?
- What is the magnitude of the effect of a given predictor variable on the response variable?
- Can any predictor variable be dropped from the model because it has little or no effect on the response variable?
- Should any predictor variables not yet included in the model be considered for possible inclusion?

Multicollinearity and Its Effects cont'd

If...

- ① ... the predictor variables included in the model are uncorrelated among each other
- ② ... the predictor variables included in the model are uncorrelated with any other predictor variables that are related to the response variable but have been omitted from the model

...then relatively simple answers to these questions can be had

In many non-experimental situations, predictor variables are correlated with other predictor variables in the model as well as predictor variables that are related to the response variable yet have been omitted from the model

Uncorrelated Predictor Variables

- When predictor variables are uncorrelated, the MLR ANOVA table and the table of regression coefficients demonstrate interesting properties, namely, that the marginal contribution of one predictor variable in reducing the error sum of squares when the other predictor variables are in the model is **exactly** the same as when his predictor variable is in the model, e.g.,
 - $SSR(X_2|X_1) = SSR(X_2)$
 - $SSE(X_3|X_1, X_2) = SSE(X_3)$

Uncorrelated Predictor Variables cont'd

The R^2 values are completely additive when the predictor variables are perfectly uncorrelated

```
> uncorrData <- read.table('~/DesktopuncorrelatedData.txt',
  sep="",header=TRUE)

> cor(uncorrData)

      x1          x2          y
x1  1.0000000  0.0000000 -0.1965703
x2  0.0000000  1.0000000 -0.4324546
y   -0.1965703 -0.4324546  1.0000000

> (cor(uncorrData$y,uncorrData$x1))^2 + (cor(uncorrData$y,uncorrData$x2))^2
[1] 0.2256569

> lm_uncorrData <- lm(y ~ . , data = uncorrData)

> summary(lm_uncorrData)$r.squared
[1] 0.2256569
```

Uncorrelated Predictor Variables cont'd

- Given that $\sigma\{X_1, X_2\} = 0$, to show that the regression coefficient of X_1 is unchanged when X_2 is added to the first-order regression model with only two predictors:

$$b_1 = \frac{\frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} - \left[\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2} \right]^{1/2} r_{Y2} r_{12}}{1 - r_{12}^2} \quad (9)$$

- If X_1 and X_2 are uncorrelated, the $r_{12} = 0$, reducing (9) to

$$b_1 = \frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} \quad (10)$$

which is the definition of b_1 for the SLR case

- We conclude, without loss of generality, that when $\sigma\{X_1, X_2\} = 0$, adding X_2 to the regression model does not change the regression coefficient for X_1

Perfectly Correlated Predictor Variables

Y	X_1	X_2
1.91	1	2
14.23	2	4
9.01	3	6
1.88	4	8

E.g., calling the `lm` function on a MLR model with two predictor variables that are perfectly correlated with each other yields the following output

```
> lm_corr <- lm(y ~ x_1 + x_2)

> lm_corr

Call:
lm(formula = y ~ x_1 + x_2)

Coefficients:
(Intercept)          x_1          x_2
               8.085       -0.531        NA
```

Perfectly Correlated Predictor Variables

- If you have perfectly correlated variables, \mathbf{X} is no longer full rank, which implies that one column of data can be expressed as a linear combination of the other variables
- This is **BAD** because $\mathbf{X}'\mathbf{X}$ is no longer invertible

Multicollinearity and Its Effects

E.g.,

- Regressing family food expenditures (Y) on family income (X_1), family savings (X_2) and age of head of household (X_3), clearly the explanatory variables will be correlated with themselves
- Moreover, X_1 , X_2 and X_3 will be correlated with other predictor variables not included in the model yet that have a relation with food expenditures such as family size
- When predictor variables are correlated amongst themselves, **multicollinearity** among them is said to exist

Effects of Multicollinearity cont'd

- Of course, both previous examples of perfectly *uncorrelated* and perfectly *correlated* predictor variables are extremes and seldom occur under non-experimental conditions
- Note the following about multicollinearity
 - ① Correlation between some or all predictor variables **does not generally inhibit our ability to obtain a good fit**, nor does it tend to affect inferences about mean responses or the prediction of new observations

Effects of Multicollinearity cont'd

- Of course, both previous examples of perfectly *uncorrelated* and perfectly *correlated* predictor variables are extremes and seldom occur under non-experimental conditions
- Note the following about multicollinearity
 - ① Correlation between some or all predictor variables **does not generally inhibit our ability to obtain a good fit**, nor does it tend to affect inferences about mean responses or the prediction of new observations
 - ② When predictor variables are highly correlated, *SE*'s for the regression coefficients may grow substantially, potentially causing many of the estimated regression coefficients not to be *statistically significant*, even though a definite statistical relation exists between there response variable and the set of predictor variables

Multicollinearity and Its Effects cont'd

- Observe that

$$s^2\{b_k\} = \frac{MSE}{\sum_i(X_{ij} - \bar{X}_k)^2(1 - R_k^2)} \quad (11)$$

where R_k^2 here is the R^2 value obtained by regressing X_k on all other independent variables

Multicollinearity and Its Effects cont'd

- Observe that

$$s^2\{b_k\} = \frac{MSE}{\sum_i(X_{ij} - \bar{X}_k)^2(1 - R_k^2)} \quad (11)$$

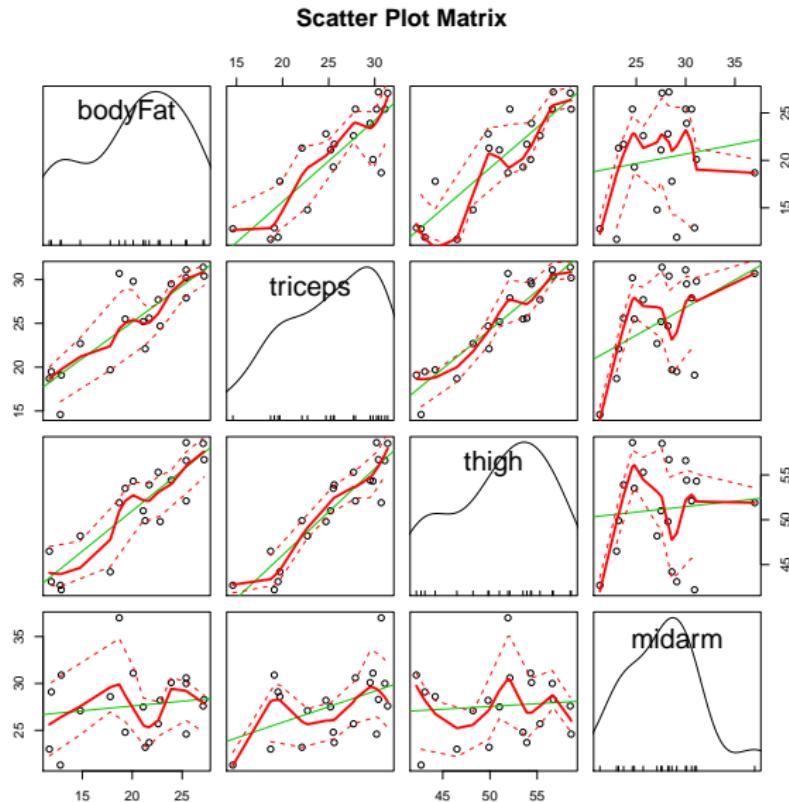
where R_k^2 here is the R^2 value obtained by regressing X_k on all other independent variables

- We conclude that increasing multicollinearity increases R_k^2 , thereby reducing the $1 - R_k^2$ term in the denominator and driving $s^2\{b_k\}$ asymptotically to infinity

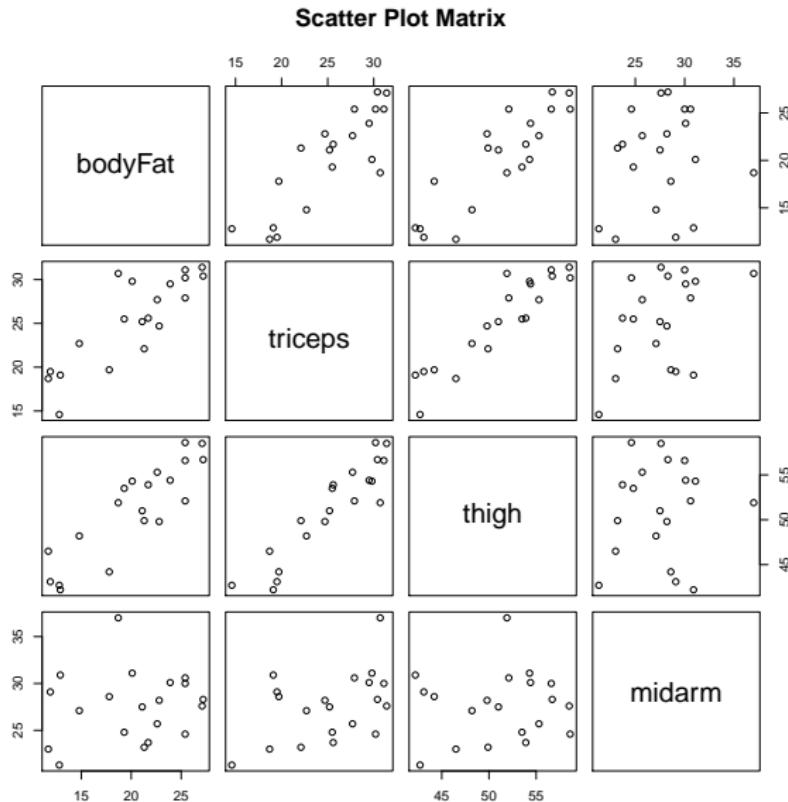
Effects of Multicollinearity cont'd

- ③ The common interpretation of a regression coefficient as measuring the change in the expected value of the response variable when the given predictor variable is increased by one unit while all other predictor variables are held constant is not fully applicable when multicollinearity exists
 - It may not be possible in practice to hold other predictor variables constant while varying one predictor variable in the presence of multicollinearity
 - Conclusion: the simple interpretation of the regression coefficients as measuring the marginal effects are often unwarranted with **highly** correlated predictor variables
- ④ In certain contexts, the term **intercorrelation** is used in reference to mildly correlated predictor variables, reserving the term **multicollinearity** for severe correlation between predictor variables; we will not adhere to this convention

Scatter Plot Matrix [BODYFAT EXAMPLE, car Package]



Scatter Plot Matrix [BODYFAT EXAMPLE, Basic R Functionality]



Scatter Plot Matrixes [R Code]

```
# w/ car package

library(car)
scatterplotMatrix(~ bodyFat + triceps + thigh + midarm, data= bodyFatData,
main="Scatter Plot Matrix")

###

# using basic R functionality

pairs(~ bodyFat + triceps + thigh + midarm, data= bodyFatData,
main="Scatter Plot Matrix")
```

Effects of Multicollinearity cont'd

Predictor Variables in the Model	b_1	b_2
X_1	0.8572	—
X_2	—	0.8565
X_1, X_2	0.2224	0.6594
X_1, X_2, X_3	4.334	-2.857

Table 1: Estimated Regression Coefficients: BODYFAT EXAMPLE

Effects of Multicollinearity cont'd

Predictor Variables in the Model	b_1	b_2
X_1	0.8572	—
X_2	—	0.8565
X_1, X_2	0.2224	0.6594
X_1, X_2, X_3	4.334	-2.857

Table 1: Estimated Regression Coefficients: BODYFAT EXAMPLE

- From the scatter plot matrix, we observe that X_1 and X_2 are highly correlated ($r_{12} = 0.924$)
- X_3 is not particularly strongly correlated with X_1 and X_2 individually ($r_{13} = 0.458, r_{23} = 0.085$)
- X_3 is highly correlated with X_1 and X_2 together; running a regression of X_3 on X_1 and X_2 yield an $R^2 = 0.998$
- b_2 changes sign when X_3 is added to the model

Effects of Multicollinearity cont'd

It can be shown that

$$\sigma^2\{b_1^*\} = \sigma^2\{b_2^*\} = \frac{(\sigma^*)^2}{1 - r_{12}^2}$$

where b_1^* and b_2^* are the standardized regression coefficients and σ^* is the standardized error term variance (standardized regression models will be discussed shortly; the interested reader is directed to Kutner et al. 4e/5e, Chapter 7, §7.5)

As the correlation between X_1 and X_2 approaches 1, the variance of b_1^* and b_2^* grows without bound.

Effects of Multicollinearity cont'd

- When predictor variables are correlated, the marginal contribution of any one predictor variable in reducing the error sum of squares can vary significantly
- In a regression model, when adding one explanatory variable that is highly correlated with an explanatory variable already in the model, the marginal contribution of the additional variable in reducing SSE is often comparatively small
- This can be observed with diminished or small extra sums of squares terms, as well as in diminished or small partial coefficients of determination

Don't be fooled

- There are cases when adding an explanatory variable that is strongly correlated to an explanatory variable in the model actually increases extra sum of squares term, i.e.,
 $SSR(X_2|X_1) > SSR(X_2)$: in this case, X_1 is considered a *suppressor variable*
- This result (above) would be the result of poor modeling, but makes the point that an algorithmic approach to model selection is difficult

Effects of Multicollinearity on $s\{b_k\}$

Predictor Variables in the Model	$s\{b_1\}$	$s\{b_2\}$
X_1	0.1288	—
X_2	—	0.1100
X_1, X_2	0.3034	0.2912
X_1, X_2, X_3	3.016	2.582

Table 2: SE's of Regression Coefficients: BODYFAT EXAMPLE

The high degree of multicollinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients.

Effects of Multicollinearity on Fitted Values & Predictions

Predictor Variables in the Model	MSE
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

Table 3: MSE's with Varying Predictor Variables: BODYFAT EXAMPLE

- We know that error cannot be unexplained, therefore adding additional variables—even highly correlated ones—to the model will usually not increase MSE so long as n is sufficiently large

$$MSE = \frac{SSE}{n - p}$$

- Fitted values and predictions may be materially affected by adding variables with high multicollinearity into the model; evaluate on a case-by-case basis

Section 7

Polynomial Regression Models

Polynomial Regression Models

- Polynomial regression models have two common uses:
 - ① When the true curvilinear response function is indeed a polynomial function
 - ② When the true curvilinear response function is unknown or complex, yet is well approximated by a polynomial function
- Polynomial regression models may contain one, two or more **predictor** variables, and each predictor variable may be present in various powers

E.g., A polynomial regression model with one predictor variable raised to both

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i \quad (12)$$

where $x_i = X_i - \bar{X}$

- This polynomial model is called a *second-order model with one predictor variable* because the single predictor variable is expressed in the model to the first and second powers

Centering the Predictor Variable

- Note that the predictor variable is centered, i.e., expressed as a deviation about its mean \bar{X} , and that the i^{th} centered observation is denoted x_i
- The reason for centering the predictor variable is that X and X^2 are often highly correlated, causing serious computational issues when the $\mathbf{X}'\mathbf{X}$ matrix is inverted for estimating regression coefficients in the normal equation calculations
- Centering the predictor variable often reduces the multicollinearity substantially, helping avoid computational and, therefore, estimation difficulties

Second-Order Single-Predictor Polynomial Regression

- The response function for the second-order regression model with one predictor variable is

$$E\{Y\} = \beta_0 + \beta_1 x + \beta_{11} x^2$$

- The response function is a parabola and is frequently referred to as a quadratic response function

- n.b.
- The regression coefficient β_0 represents the mean response of Y when $x = 0$, i.e., when $X = \bar{X}$
 - The regression coefficient β_1 is often referred to as the *linear effect coefficient* and β_{11} is called the *quadratic effect coefficient*

n^{th} -Order Single-Predictor Polynomial Regression

- The regression model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \varepsilon_i$$

where $x_i = X_i - \bar{X}$ is a third-order model with one predictor variable, and the response function is

$$E\{Y\} = \beta_0 + \beta_1 x + \beta_{11} x^2 + \beta_{111} x^3$$

- Employ polynomial models with predictor variables to powers higher than three with caution, as the interpretation of such coefficients can become **very** difficult for such models, e.g., models can become highly erratic for interpolations
- It can be shown that a polynomial model of sufficiently high order can always be found to fit data containing no repeat observations perfectly

Second-Order Two-Predictor Polynomial Regression

- The regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

where $x_{i1} = X_{i1} - \bar{X}_1$ and $x_{i2} = X_{i2} - \bar{X}_2$ is a second-order model with two predictor variables, and the response function is

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

- n.b.
- The above model contains separate linear and quadratic components for each of the two predictor variables and a cross-product term, $x_1 x_2$, which represents the interaction effect (more soon on that)
 - The coefficient β_{12} is often called the interaction effect coefficient

Quadratic Response Surface EXAMPLE

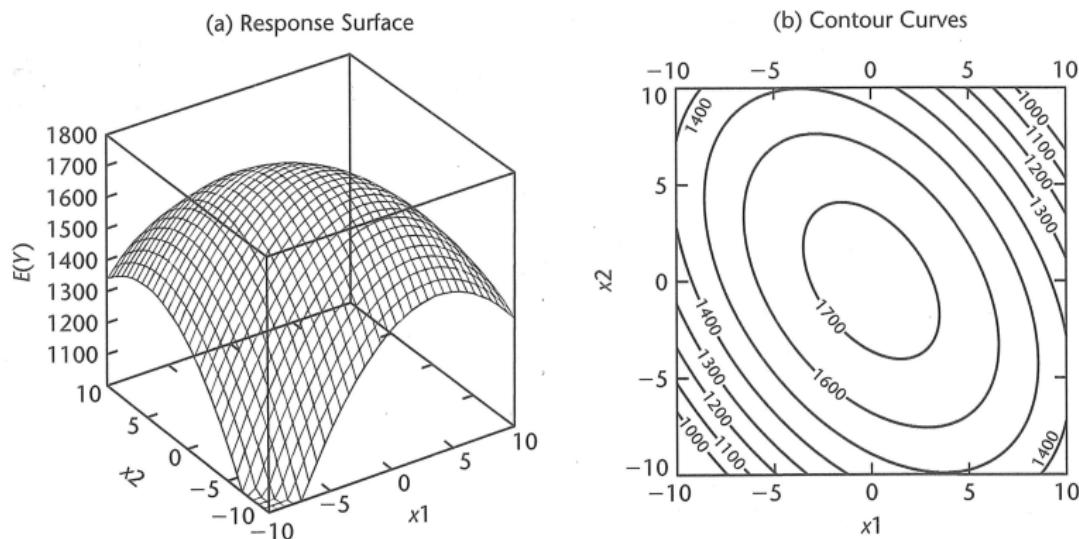


Figure 8: Quadratic Response Surfaces & Contour Plot

Hierarchical Approach to Fitting Polynomials

- When fitting polynomial regression models, statisticians will often fit the highest order model which they feel comfortable employing, either a second- or third-order polynomial regression model, and subsequently scale back, lowering powers to evaluate whether lower-order models are adequate
- With the hierachal approach, if a polynomial term of a given order is retained, then **all related terms of lower order are also retained in the model**

Hierarchical Approach to Fitting Polynomials

- When fitting polynomial regression models, statisticians will often fit the highest order model which they feel comfortable employing, either a second- or third-order polynomial regression model, and subsequently scale back, lowering powers to evaluate whether lower-order models are adequate
- With the hierarchical approach, if a polynomial term of a given order is retained, then **all related terms of lower order are also retained in the model**
- E.g., in a cubic polynomial regression function where the cubic term is statistically significant, the quadratic term is also kept in the model
 - The lower-order quadratic term is viewed as providing more basic information about the shape of the response function
 - The higher-order cubic term is viewed as providing refinements in the specification of the shape of the response function

Hierarchical Approach to Fitting Polynomials

- When fitting polynomial regression models, statisticians will often fit the highest order model which they feel comfortable employing, either a second- or third-order polynomial regression model, and subsequently scale back, lowering powers to evaluate whether lower-order models are adequate
- With the hierachal approach, if a polynomial term of a given order is retained, then **all related terms of lower order are also retained in the model**
- E.g., in a cubic polynomial regression function where the cubic term is statistically significant, the quadratic term is also kept in the model
 - The lower-order quadratic term is viewed as providing more basic information about the shape of the response function
 - The higher-order cubic term is viewed as providing refinements in the specification of the shape of the response function
- The same is true for interaction terms: keeping a significant interaction term in the model implies keeping all related first-order predictor variables

Ramsey RESET Test for Functional Misspecification

Regression Equation Specification Error Test (RESET)

- If you have a SLR, linear in the independent variables, e.g.,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and you wish to test whether or not a quadratic term is a suitable addition to the model, one can easily add a quadratic term to the model, as follows

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

and run a t -test or a partial F -test to see whether or not β_{11} , the coefficient of the quadratic independent variable is statistically significantly different from zero

- n.b.** This approach easily extends to cubic and higher powers of the independent predictor variables

Ramsey RESET Test for Functional Misspecification cont'd

- What if you have a more complex MLR, linear in the independent variables, e.g.,

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

and you wish to test whether or not quadratic, cubic (or other) terms might be suitable additions to the model

- Adding, testing, removing individual terms, let alone examining the relationship(s) among various additional quadratic, cubic, higher-order, and interaction terms would prove an onerous task

Ramsey RESET Test for Functional Misspecification cont'd

Ramsey proposes an interesting top-down approach

- ① Regress Y on all relevant X 's (linearly), as follows

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

which results in the following regression function

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_{p-1} X_{i,p-1}$$

Ramsey RESET Test for Functional Misspecification cont'd

Ramsey proposes an interesting top-down approach

- ① Regress Y on all relevant X 's (linearly), as follows

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

which results in the following regression function

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_{p-1} X_{i,p-1}$$

- ② Regress Y on all relevant X 's (linearly) **and** on the \hat{Y}_i from the aforementioned regression function, raised to a power, let's say squared, for the sake of exposition, resulting in the following regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \alpha_1 \hat{Y}_i^2 + \varepsilon_i$$

Ramsey RESET Test for Functional Misspecification cont'd

cont'd By including the \hat{Y}_i^2 as a predictor variable, Ramsey effectively incorporates all the of predictor variables $X_{i1}, X_{i2}, \dots, X_{i,p-1}$ raised to the second power, as well as all of the associated interaction effects in single term

Running a *t*-test on the associated coefficient α_1 evaluates whether or not quadratic or interaction terms should be included in the model

This is admittedly a broad stroke, but in the event that $H_0 : \alpha_1 = 0$ holds, it can save a modeler a lot of pain and time

Ramsey RESET Test for Functional Misspecification cont'd

Additionally complex models can also be run, e.g.,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \alpha_1 \hat{Y}_i^2 + \alpha_2 \hat{Y}_i^3 + \varepsilon_i$$

where one can run

- t -tests on individual coefficients, $t_{crit} = t_{(n-p;1-\alpha/2)}$
- restricted F -tests ($F_{crit} = F_{(\delta,n-p;1-\alpha)}$) to test the significance of all coefficients of the \hat{Y}_i^λ terms, where δ is the number of coefficients to be tested)

Ramsey RESET Test in R

$$H_0 : \alpha_1 = \dots = \alpha_k = 0$$

$$H_a : \text{not all } \alpha_k = 0$$

```
> install.packages("lmtest")

> library(lmtest)

## power is the power lambda to which you want to raise
# the fitted values, can be a sequence if you wish to
# to test multiple powers, e.g., power=2:4 to test
# quadratic, cubic and quartic terms

## type="fitted" to test fitted values

> resettest(lm_faithful , power=2 , type="fitted")

RESET test

data: lm_faithful
RESET = 13.9844, df1 = 1, df2 = 269, p-value = 0.0002252
```

Comments on Polynomial Regression

- Using polynomial models can consume multiple degrees of freedom; under certain circumstances, it may be more efficient to examine non-linear regression models or to transform variables
- Centering predictor variables for polynomial models is not a requirement, only a recommendation to reduce multicollinearity; you can also examine orthogonal polynomials, effectively eliminating multicollinearity (not discussed in this class)

Polynomial Regression in R

- Oddly, the `lm` function in R will not interpret the following model algebraically
- The following code generates the following unexpected output

```
> lm_faithful <- lm(waiting ~ eruptions + eruptions^2, data=faithful)
```

```
> summary(lm_faithful)$coefficients
            Estimate Std. Error t value    Pr(>|t|)
(Intercept) 33.47440  1.1548735 28.98534 7.136015e-85
eruptions   10.72964  0.3147534 34.08904 8.129959e-100
```

- We are required to employ the `I()` operator, where all elements within the parentheses are evaluated algebraically

```
> lm_faithful <- lm(waiting ~ eruptions + I(eruptions^2), data=faithful)
```

```
> summary(lm_faithful)$coefficients
            Estimate Std. Error t value    Pr(>|t|)
(Intercept) 17.200021  4.4957614 3.825831 1.620900e-04
eruptions   22.213151  3.0861604 7.197666 6.118255e-12
I(eruptions^2) -1.766202  0.4723004 -3.739573 2.252453e-04
```

Interpretation of β_k 's in Polynomial Regression

- Assume the following fitted regression function

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_{11} X_1^2 + b_{22} X_2^2$$

- It is important to notice that b_1 no longer measures the estimated change in \hat{Y} with respect to X_1 , as it makes no sense to hold X_1^2 constant while changing X_1
- If we want to use the compute the predicted change in Y , you can use the formula above, but if we are interested in the rate of change, we must compute the partial derivative

$$\frac{\partial \hat{Y}}{\partial X_1} \approx b_1 + 2b_{11}X_1$$

Polynomial Regression in R [EXAMPLE 1]

- Load `wage1.RData` and regress `wage` on `exper` and `exper^2`, generating the following fitted regression equation

$$\widehat{wage} = 3.72 + 0.298exper - 0.00613exper^2 \quad (13)$$

and the partial derivative is

$$\frac{\partial \widehat{wage}}{\partial exper} \approx 0.298 - (2)(0.00613)exper \quad (14)$$

- According to (14)
 - The first year of experience is worth approximately 0.298 \$/hr
 - The second year is worth less: $0.298 - (2)(0.00613)(1)$ \$/hr
 - In going from 10 to 11 years of experience, wage is predicted to increase by $0.298 - (2)(0.00613)(10) = 0.176$ \$/hr

Polynomial Regression in R [EXAMPLE 2]

- Load `women` dataset from the `datasets` package and run a Ramsey RESET test to determine if a quadratic term is appropriate. If not, what about a cubic?

Section 8

Interaction Regression Models

Interaction Regression Models

- When speaking of interaction effects, we leave the universe of regression models where the effects of X_1, X_2, \dots, X_{p-1} are **additive**, e.g.,

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_{11} X_1^2 + \beta_2 X_2$$

and begin to examine regression models where predictor variables no longer additive but interactive with each other, e.g.,

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- In the above equation, the cross-product term $X_1 X_2$ is called an interaction term; it may also be referred to as a linear-by-linear or bilinear interaction term

Interpretation of Interaction Regression Coefficients

- The regression model for two quantitative predictor variables with linear effects on Y and interacting effects effects of X_1 and X_2 on Y represented by a cross-sectional term is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \varepsilon_i$$

- The meaning of the regression coefficients here **is not the same** as that given earlier because of the interaction term $\beta_3 X_{i1}X_{i2}$
- The regression terms β_1 and β_2 no longer indicate the change in the mean response with a unit increase of the predictor variable, with the other predictor variable held constant at any given level

Interpretation of Interaction Regression Coefficients cont'd

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \varepsilon_i$$

- The **change** in the mean response with a unit increase in X_1 when X_2 is held constant is

$$\beta_1 + \beta_3 X_2$$

- Similarly, the **change** in the mean response with a unit increase in X_2 when X_1 is held constant is

$$\beta_2 + \beta_3 X_1$$

- Therefore, in this regression model, both the effect of X_1 for a given level of X_2 and the effect of X_2 for a given level of X_1 depend on the level of the other predictor variable

Conditional Effects Plots

Let's examine 3 different sample regression response functions

(a) $E\{Y\} = 10 + 2X_1 + 5X_2$

(b) $E\{Y\} = 10 + 2X_1 + 5X_2 + 0.5X_1X_2$

(c) $E\{Y\} = 10 + 2X_1 + 5X_2 - 0.5X_1X_2$

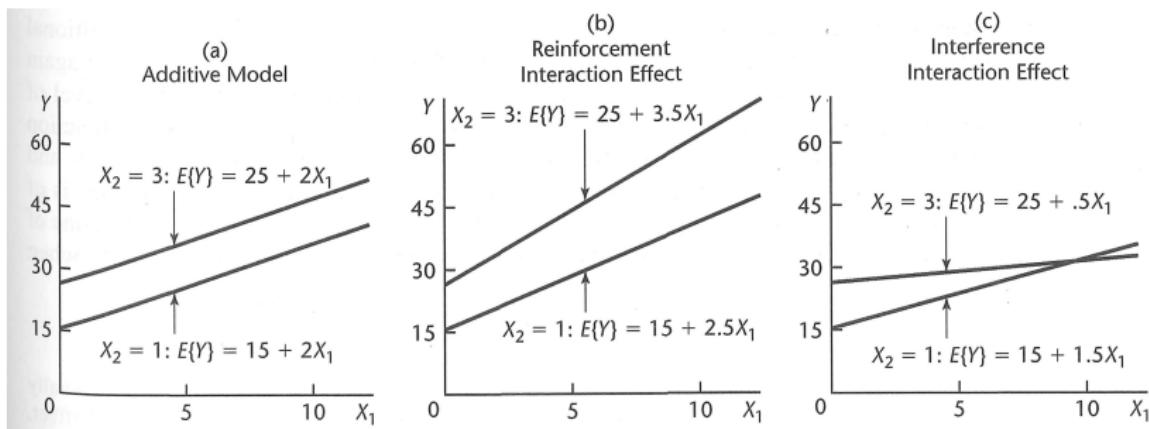


Figure 9: Conditional Effects Plots

Contour Plots [R Code]

```
universityAdmissions <- read.table('~/Dropbox/Public/MSAN601USF/...
  universityAdmissions.txt',sep=' ',header=FALSE)

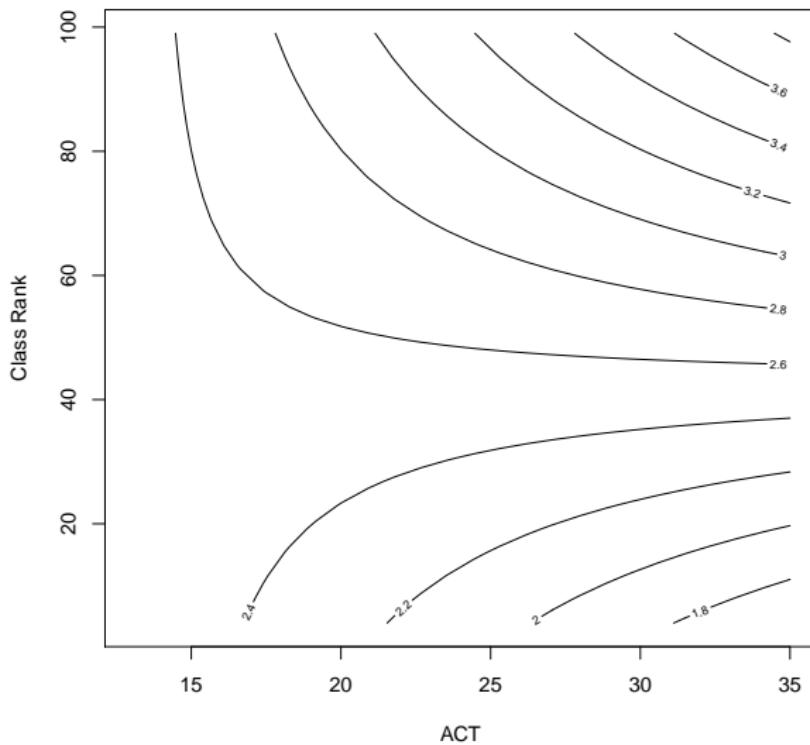
#here V2 is GPA (Y), V3 is class rank (X1), V4 is ACT score (X2)

univAd.lm <- lm(V2~V3*V4,data=universityAdmissions) #regresses V2 on ...
  V3,V4 and V3*V4

install.packages('rsm')
library('rsm')
```

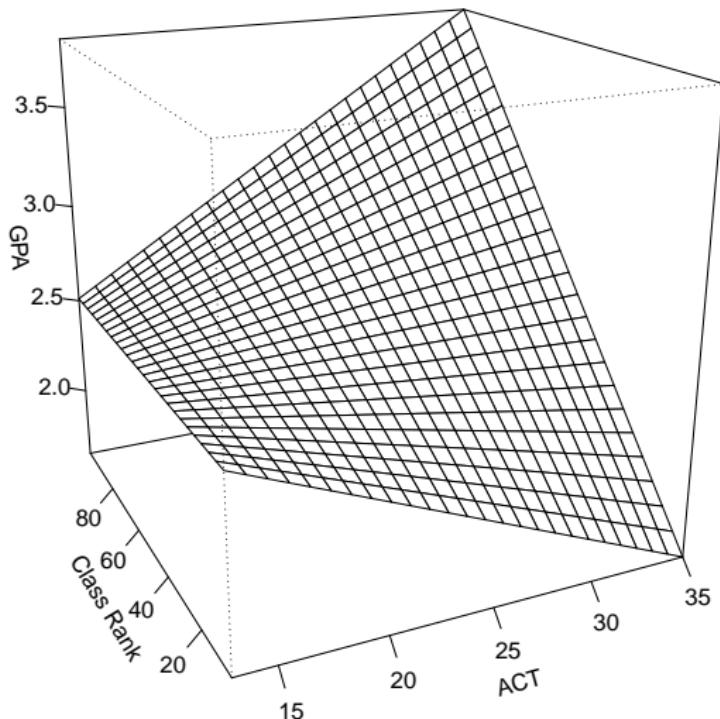
Contour Plot

```
contour(univAd.lm,V3~V4,xlabs=c('Class Rank', 'ACT'))
```



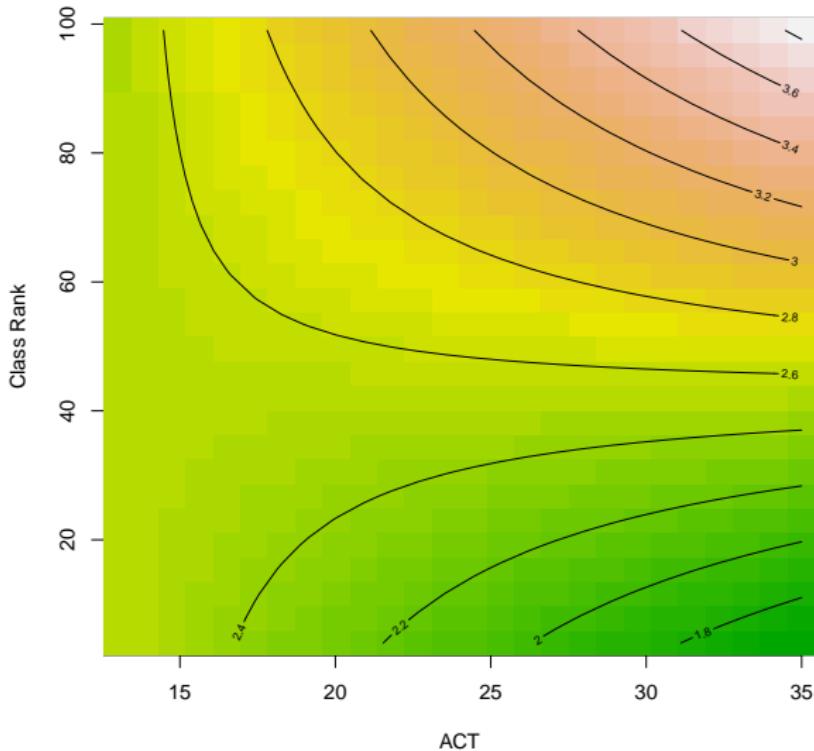
Perpective/Surface Plot

```
persp(univAd.lm, V3~V4,xlabs=c('Class Rank','ACT'),zlab='GPA')
```



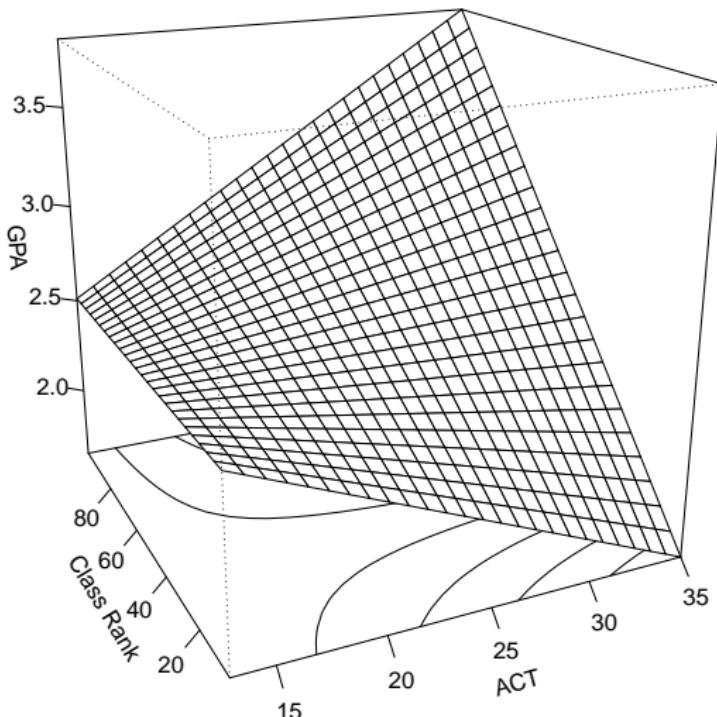
More Contour Plots

```
contour(univAd.lm,V3~V4,image=TRUE,xlabs=c('Class Rank','ACT'),zlab='GPA')
```



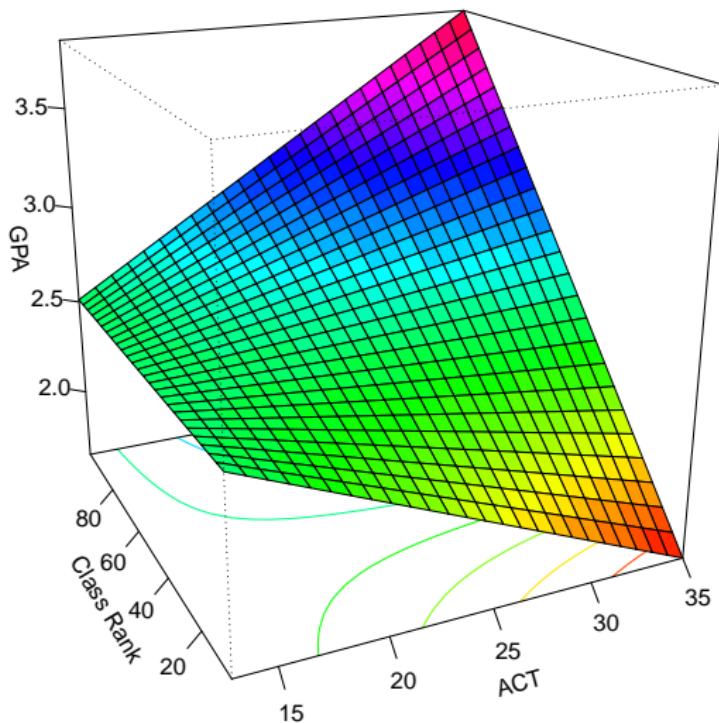
More Perspective/Surface Plots

```
persp(univAd.lm, V3~V4, contours=TRUE,xlabs=c('Class Rank','ACT'),zlab='GPA')
```



More Perspective/Surface Plots

```
persp(univAd.lm, V3~V4,image=TRUE, contours='colors',...  
xlabs=c('Class Rank','ACT'),zlab='GPA')
```



Final Notes on Interaction Regression Models

- ① When interaction terms are added to a regression model, high levels of collinearity may exist between some of the predictor variables as well as among some of the interaction terms. A partial remedy is to center the predictor variables, e.g.,
 $x_{ik} = X_{ik} - \bar{X}_{ik}$.
- ② Attempt to identify *a priori*, based on contextual issues, which interaction terms, if any, may be of interest to investigate; even with a small number of predictor variables, the number of pairwise interaction terms can be large
- ③ Use caution when including interactions among more than two variables
- ④ The hierarchical principle applies with interaction regressors, just as it does with polynomial regressors

Interaction Terms in R

- To denote an interaction term in an `lm()` call, type
`> lm(y ~ x + z + x:z)`
where the `{x:z}` represents the interaction between `x` and `z`

Interaction Terms in R

- To denote an interaction term in an `lm()` call, type
`> lm(y ~ x + z + x:z)`
where the {x:z} represents the interaction between x and z
- To denote all possible interaction terms in an `lm()` call, type
`> lm(y ~ x * z * w)`
*where the {x * z * w} expands to*
`{x + z + w + x:z + x:w + z:w + x:z:w}`

- To denote an interaction term in an `lm()` call, type
`> lm(y ~ x + z + x:z)`
where the {x:z} represents the interaction between x and z
- To denote all possible interaction terms in an `lm()` call, type
`> lm(y ~ x * z * w)`
*where the {x * z * w} expands to*
 $\{x + z + w + x:z + x:w + z:w + x:z:w\}$
- To denote interaction terms up to a specified degree in an `lm()` call, type `> lm(y ~ (x + z + w)^2)`
where the {(x + z + w)^2} expands to
 $\{x + z + w + x:z + x:w + z:w\}$

Advertising EXAMPLE in R

- Load Advertising dataset from Canvas
- Fit the best possible model using interaction and/or polynomials regressors.
- Do you have an intuition as to why the significant regressors are contextually relevant?

Section 9

Qualitative Predictors

Qualitative Predictor Variables

- We can model categorical information using qualitative predictor variables, e.g., gender, season, etc.
- These are typically called indicator, dummy or binary variables (the latter is reserved exclusively for a 0/1 coding scheme)
- There are **many** ways to quantitatively identify the classes of various categorical variables, but perhaps the most common is the binary coding scheme 0/1, e.g., if regressing department store sales (Y) on online marketing budget (X_1) and customer gender, one might code gender as follows

Qualitative Predictor Variables

- We can model categorical information using qualitative predictor variables, e.g., gender, season, etc.
- These are typically called indicator, dummy or binary variables (the latter is reserved exclusively for a 0/1 coding scheme)
- There are **many** ways to quantitatively identify the classes of various categorical variables, but perhaps the most common is the binary coding scheme 0/1, e.g., if regressing department store sales (Y) on online marketing budget (X_1) and customer gender, one might code gender as follows

$$X_2 = \begin{cases} 0 & \text{if male} \\ 1 & \text{otherwise} \end{cases} \quad X_3 = \begin{cases} 0 & \text{if female} \\ 1 & \text{otherwise} \end{cases}$$

resulting in the following regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

Binary Predictor Variables

Coding each variable with a binary predictor variable results in the following \mathbf{X} matrix

- Assuming, for the sake of exposition, we have 4 observations, i.e., $n=4$, where the first two observations are male, and the second two are female

$$\mathbf{X}_{4 \times 4} = \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

- What problem do you observe?

Binary Predictor Variables

Coding each variable with a binary predictor variable results in the following \mathbf{X} matrix

- Assuming, for the sake of exposition, we have 4 observations, i.e., $n=4$, where the first two observations are male, and the second two are female

$$\mathbf{X}_{4 \times 4} = \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

- What problem do you observe?
- A simple way (but not the only way) to address this issue is to code c categories with $c - 1$ indicator variables

Interpretation w/ Binary Predictor Variables

For the following equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

where Y is sales, X_1 is online marketing budget and

$$X_2 = \begin{cases} 0 & \text{if male} \\ 1 & \text{otherwise} \end{cases}$$

the response function for this regression model is

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Interpretation w/ Binary Predictor Variables cont'd

- If the customer is female ($X_2 = 0$), the response is

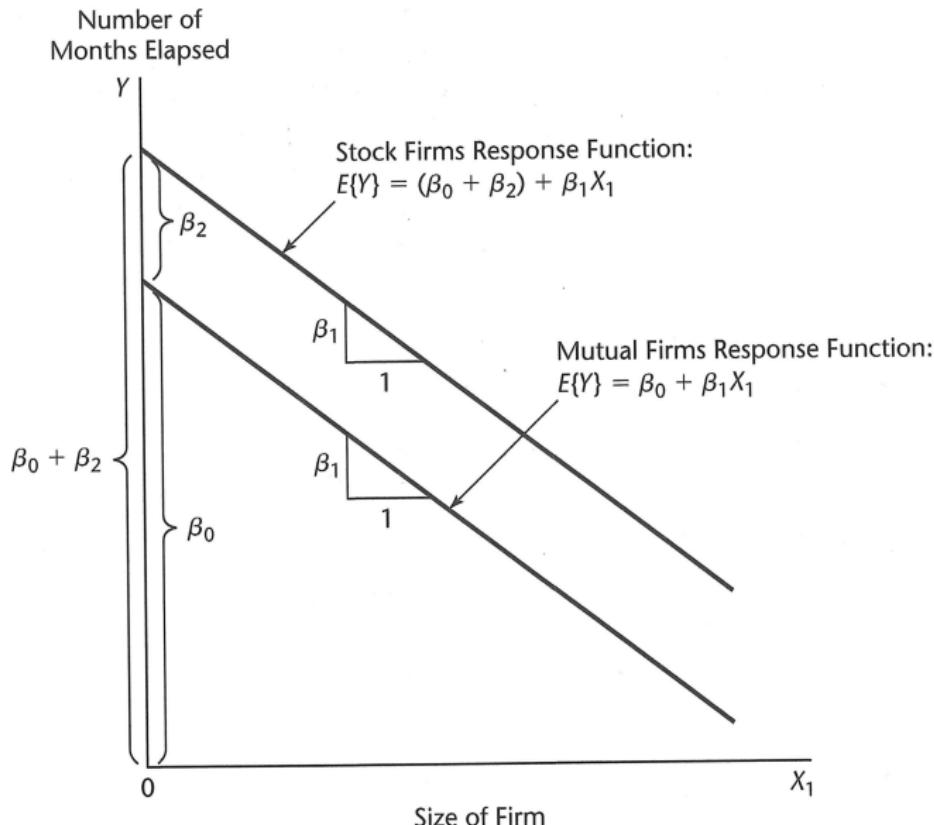
$$Y_i = \beta_0 + \beta_1 X_{i1} \quad \text{Female}$$

- If the customer is male ($X_2 = 1$), the response is

$$Y_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} \quad \text{Male}$$

- Observe what is happening: The slope of the response remains the same, but we shift the intercept by β_2 units

Interpretation w/ Binary Predictor Variables cont'd



Why Run A Single Model?

Why not fit separate regressions for each category of qualitative predictor?

Why Run A Single Model?

Why not fit separate regressions for each category of qualitative predictor?

- ① Since the model assumes equal slopes and the same constant error term variance for each category, the common slope β_1 can be best estimated by pooling categories
- ② Inferences for β_0 and other β 's associated with the $c - 1$ categories can be made more precisely by working with a single, common regression model since more degrees of freedom will be associated with MSE

Binary Predictor Variables EXAMPLE MLB Data

We have MLB data with player salaries (Y), years of experience (X_1) and whether or not the player is Japanese (X_2), coded as follows

$$X_2 = \begin{cases} 1 & \text{if Japanese} \\ 0 & \text{otherwise} \end{cases}$$

```
> load("~/Desktop/mlb.RData")  
  
> lm_mlb <- lm(Salary ~ Experience + Japanese , data=mlb)  
  
> lm_mlb
```

Coefficients:

(Intercept)	Experience	Japanese1
490007	733199	9693796

The estimated regression function is therefore

$$\hat{Y} = 490,007 + 733,199X_1 + 9,693,796X_2$$

Binary Predictor Variables EXAMPLE MLB Data

The interpretation of this estimated regression function is as follows:

- ① If the player **is not** Japanese, then $X_2 = 0$, and our estimated regression function is

$$\hat{Y} = 490,007 + 733,199X_1 + 9,693,796(0)$$

$$\hat{Y} = 490,007 + 733,199X_1$$

- ② If the player **is** Japanese, then $X_2 = 1$, and our estimated regression function is

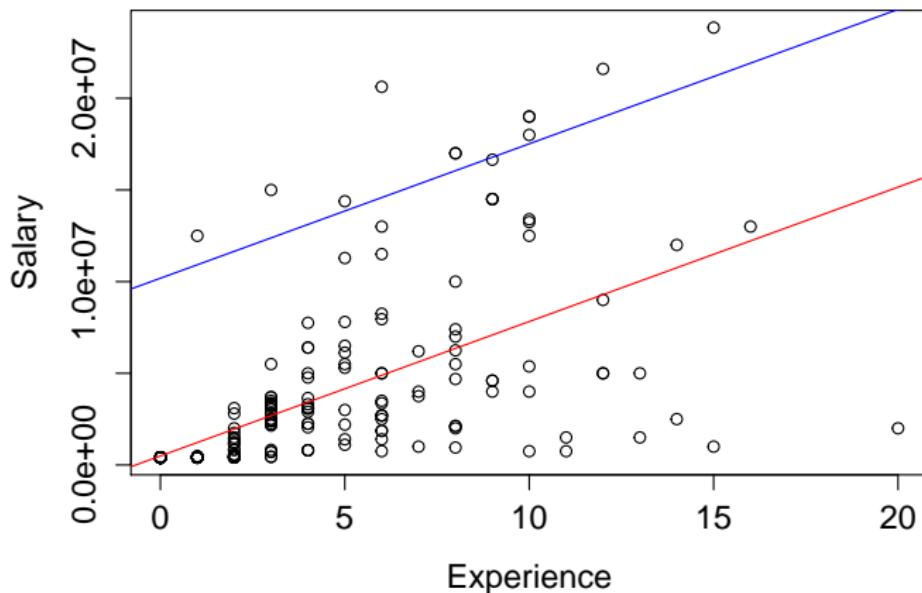
$$\hat{Y} = 490,007 + 733,199X_1 + 9,693,796(1)$$

$$\hat{Y} = (490,007 + 9,693,796) + 733,199X_1$$

$$\hat{Y} = 10,183,803 + 733,199X_1$$

Binary Predictor Variables EXAMPLE MLB Data

Salary vs. Experience



- Blue line is for Japanese players
- Red line is for non-Japanese players

Binary Predictor Variables EXAMPLE MLB Data

What if you decide to recode X_2 differently?

$$X_2 = \begin{cases} 0 & \text{if Japanese} \\ 1 & \text{otherwise} \end{cases}$$

Binary Predictor Variables EXAMPLE MLB Data

What if you decide to recode X_2 differently?

$$X_2 = \begin{cases} 0 & \text{if Japanese} \\ 1 & \text{otherwise} \end{cases}$$

The regression model will accommodate by adjusting the y -intercept

```
> lm_mlb_recoded
```

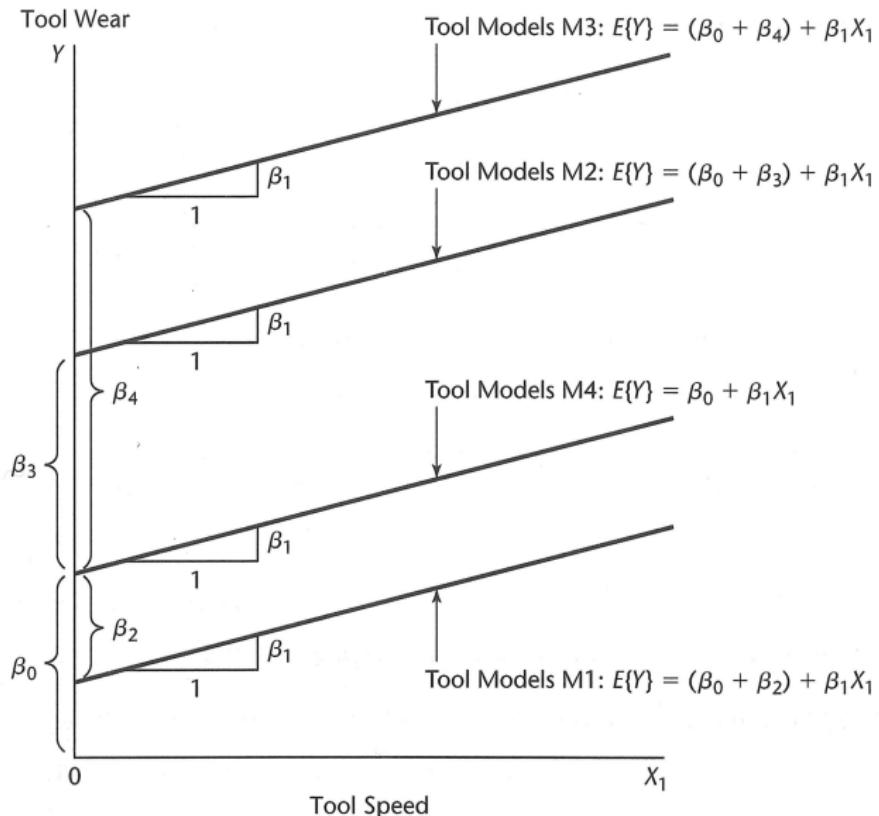
Coefficients:

(Intercept)	Experience	Japanese1
10183803	733199	-9693796

The estimated regression function is now

$$\hat{Y} = 10,183,803 + 733,199X_1 - 9,693,796X_2$$

More Than Two Categories of Binary Predictor Variables



Non-Binary Coding of Qualitative Predictor Variables

There is no rule that compels us to code each of the $c - 1$ categories with 0/1's

E.g. When coding frequency of use of a particular product one might use

$$X_k = \begin{cases} 3 & \text{frequent user} \\ 2 & \text{moderate user} \\ 1 & \text{infrequent user} \end{cases}$$

n.b. This coding scheme implies a change in the mean response through each of the categories, whereas binary coding schemes make no assumptions about the spacing of the classes and rely on the data to show the differential effects that occur

Non-Binary Coding of Qualitative Predictor Variables cont'd

There is no rule that compels us to code each of the $c - 1$ categories with 0/1's

- Alternatively, in a two-category coding, one might choose to employ a -1/1 coding scheme

$$X_k = \begin{cases} -1 & \text{Local Supermarket} \\ 1 & \text{National Supermarket Chain} \end{cases}$$

- n.b. This coding scheme implies that β_0 is the average response, from which the response will differ by the β_k in each direction based on the category

Quantitative/Qualitative Predictor Variable Interactions

A model containing both qualitative and quantitative predictor variables as well as their interactions could be modeled as follows

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \varepsilon;$$

where X_{i1} is some quantitative variable and

$$X_{i2} = \begin{cases} 1 & \text{category 1} \\ 0 & \text{otherwise} \end{cases}$$

The regression coefficients in this case take on a different meaning

- If $X_{i2} = 0$, then you end up with a SLR regression function
 - If $X_{i2} = 1$, the the intercept changes to $(\beta_0 + \beta_2)$ and the **slope also changes** to $(\beta_1 + \beta_3)$
- n.b. In this scenario, you can have **reinforcement** or **interference** interaction effects

Quantitative/Qualitative Predictor Variable Interactions

A model containing both qualitative and quantitative predictor variables as well as their interactions could be modeled as follows

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon;$$

where X_{i1} is some quantitative variable and

$$X_{i2} = \begin{cases} 1 & \text{category 1} \\ 0 & \text{otherwise} \end{cases}$$

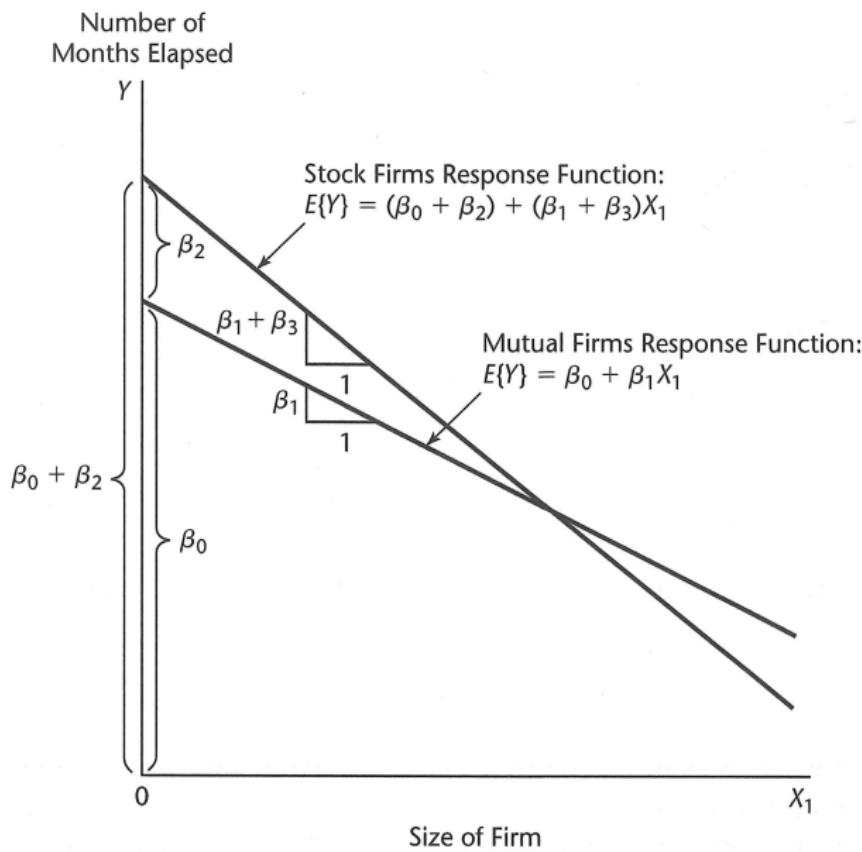
The regression coefficients in this case take on a different meaning

- If $X_{i2} = 0$, then you end up with a SLR regression function
- If $X_{i2} = 1$, the the intercept changes to $(\beta_0 + \beta_2)$ and the **slope also changes** to $(\beta_1 + \beta_3)$

n.b. In this scenario, you can have **reinforcement** or **interference** interaction effects

- When the response functions do not interact within the scope of the model, the interaction is said to be *ordinal*, whereas if the response functions intersect, it is called a *disordinal interaction*

Quantitative/Qualitative Predictor Variable Interactions



Credit EXAMPLE in R

- Load Credit.csv from Canvas
- Fit the best possible model you can using the variables income and student to predict balance
- Do you have an intuition as to why the significant regressors are contextually relevant?

Section 10

Model Selection

Exploratory Observational Studies

- In many fields including social, behavioral and health sciences, management and other fields, it is often not possible to conduct controlled experiments
- As a result, exploratory observational studies are run in an effort to search for explanatory variables that may be related to the response variable
- After what is often a lengthy list of variables has been compiled, some variables may be quickly screened out
- An explanatory variable may:
 - ① not be fundamental to the problem
 - ② be subject to large measurement errors
 - ③ may effectively duplicate other explanatory variables

Exploratory Observational Studies

- In many fields including social, behavioral and health sciences, management and other fields, it is often not possible to conduct controlled experiments
 - As a result, exploratory observational studies are run in an effort to search for explanatory variables that may be related to the response variable
 - After what is often a lengthy list of variables has been compiled, some variables may be quickly screened out
 - An explanatory variable may:
 - ① not be fundamental to the problem
 - ② be subject to large measurement errors
 - ③ may effectively duplicate other explanatory variables
- n.b. What is the ratio of observations to the quantity of predictor variables? Kutner et al. 4e/5e suggest 6 to 10 observations for every variable in the model

Preliminary Model Investigation

- ① What functional form should each explanatory variable take to enter the regression model?
- ② What important interactions should be included in the model?

Once a first pass has eliminated superfluous variables, the subsequent step is often a further ‘tightening’ of the model, i.e., what subsets of variables can perform almost as well (or as well) as the full model?

Too Few Predictor Variables

Elimination of key explanatory variables can seriously damage the explanatory power of the model and lead to

- ① biased estimates of regression coefficients
- ② biased mean responses
- ③ biased prediction intervals
- ④ biased MSE

This bias is a function of underfitting the data, and the error therefore reflecting non-random effects of excluded (latent) variables

Too Many Predictor Variables

Too many predictor variables overfits the data, often inflating variances, as well as

- ① generating possible multicollinearity issues
- ② difficulty in model management

Surgical Unit EXAMPLE

Load surgicalUnitData.txt from Canvas (use $n = \{1, \dots, 54\}$)

- Y : survival time in days
- X_1 : blood clotting score
- X_2 : prognostic index
- X_3 : enzyme function test score
- X_4 : liver function test score
- X_5 : age in years

$$X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

Criteria for Model Selection

- From any set of $p - 1$ predictor variables, 2^{p-1} alternative models can be constructed (that are linear in the predictor variables)
- One approach to model selection is an exhaustive search of the linear (in the predictors) models and the subsequent evaluation of those models according to various criteria, of which we will examine six

Criteria for Model Selection

- From any set of $p - 1$ predictor variables, 2^{p-1} alternative models can be constructed (that are linear in the predictor variables)
- One approach to model selection is an exhaustive search of the linear (in the predictors) models and the subsequent evaluation of those models according to various criteria, of which we will examine six
 - ① R_p^2 or SSE_p Criterion
 - ② $R_{a,p}^2$ or MSE_p Criterion
 - ③ Mallow's C_p Criterion
 - ④ Akaike's Information Criterion (AIC_p)
 - ⑤ Schwartz's Bayesian Criterion (SBC_p)
 - ⑥ $PRESS_p$ Criterion

Foundational Nomenclature

- We shall denote the number of potential X variables in the pool by $P - 1$
- The number of X variables in a subset will be denoted as $p - 1$
- Therefore

$$1 \leq p \leq P$$

- We will assume that the number of observations exceeds the maximum number of potential parameters, i.e.,

$$n > P$$

with the hope that $n \gg P$

R_p^2 or SSE_p Criterion

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

- Evaluating when diminishing returns to scale are achieved is subjective
- Recall R_p^2 is monotonically non-decreasing in p

$R_{a,p}^2$ or MSE_p Criterion

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE_p}{SSTO}$$

- Evaluating when diminishing returns to scale are achieved is also subjective
- Recall $R_{a,p}^2$ can increase or decrease in p

- Mallow's C_p criterion is concerned with the total mean squared error of the n fitted values for each subset regression model
- MSE involves the total error in each fitted value, $\hat{Y}_i - \mu_i$, where μ_i is the true mean response when the levels of the predictor variables X_k are those for the i^{th} case
- The total error is made up of two components
 - ① The bias component for the i^{th} fitted value \hat{Y}_i

$$E\{\hat{Y}_i\} - \mu_i$$

where $E\{\hat{Y}_i\}$ is the expectation of the i^{th} fitted value for a given regression model; if the fitted model is not correct, i.e., if it is biased, $E\{\hat{Y}_i\}$ will differ from the true mean response μ_i , representing the bias

- ② The random error component for \hat{Y}_i

$$\hat{Y}_i - E\{\hat{Y}_i\}$$

which represents the deviation of the fitted value \hat{Y}_i for the given sample from the expected value when the i^{th} fitted value is obtained by fitting the same regression model to all possible samples

Mallow's C_p Criterion cont'd

The MSE for \hat{Y}_i is defined as the expected value of the squared total error, i.e.,

$$(\hat{Y}_i - \mu_i)^2 = [(E\{\hat{Y}_i\} - \mu_i) + (\hat{Y}_i - E\{\hat{Y}_i\})]^2$$

which can be shown to be

$$E\{\hat{Y}_i - \mu_i\}^2 = (E\{\hat{Y}_i\} - \mu_i)^2 + \sigma^2\{\hat{Y}_i\}$$

which is the sum of the squared bias and the variance of \hat{Y}_i
Total mean square error for all n fitted values is

$$\sum E\{\hat{Y}_i - \mu_i\}^2 = \sum(E\{\hat{Y}_i\} - \mu_i)^2 + \sum \sigma^2\{\hat{Y}_i\}$$

Mallow's C_p Criterion cont'd

The criterion measure, denoted Γ_p , is simply the total mean squared error divided by σ^2

$$\Gamma_p = \frac{1}{\sigma^2} \left[\sum (E\{\hat{Y}_i\} - \mu_i)^2 + \sum \sigma^2\{\hat{Y}_i\} \right]$$

Proof can be found in Kutner 4e/5e, Chapter 9, §9.3

- The model includes all $P - 1$ potential predictor variables is assumed to have been carefully chosen such that $MSE(X_1, X_2, \dots, X_{P-1})$ is an unbiased estimator of σ^2
- It can then be shown that an estimator of Γ_p is C_p

$$C_p = \frac{SSE_p}{MSE(X_1, X_2, \dots, X_{P-1})} - (n - 2p)$$

- When there is no bias in the regression regression model with $p - 1$ predictor variables, $E\{\hat{Y}_i\} \equiv \mu_i$, then the expected value of C_p is approximately p , $E\{C_p\} \approx p$

With a model containing all $P - 1$ predictor variables, $C_p = P$

Proof.

$$\begin{aligned} C_p &= \frac{\text{SSE}(X_1, X_2, \dots, X_{P-1})}{\frac{\text{SSE}(X_1, X_2, \dots, X_{P-1})}{n-P}} - (n - 2P) \\ &= (n - P) - (n - 2P) \\ &= P \end{aligned}$$



Mallow's C_p Criterion cont'd

Therefore, when using $p - 1$ predictor variables, the calculation of C_p breaks down as follows

$$C_p = \frac{\frac{SSE(X_1, X_2, \dots, X_{p-1})}{SSE(X_1, X_2, \dots, X_{P-1})}}{n-P} - (n - 2p)$$

$$C_p = \left[\frac{SSE(X_1, X_2, \dots, X_{p-1})}{SSE(X_1, X_2, \dots, X_{P-1})} \right] (n - P) - (n - 2p)$$

As p gets smaller, $SSE(X_1, X_2, \dots, X_{p-1})$ increases, which implies an increase in

$$\left[\frac{SSE(X_1, X_2, \dots, X_{p-1})}{SSE(X_1, X_2, \dots, X_{P-1})} \right] (n - P)$$

which, if not offset by subtracting the increasing $n - 2p$ term, will grow larger

- Therefore, a search using Mallow's C_p identifies a C_p value closest to p across the various models
 - Models with $C_p \gg p$ are assumed to have substantial bias
 - Models with $C_p < p$ are assumed to have no bias, being below p due to sampling error
- n.b. Effective use of the Mallow's C_p criterion requires careful development of the pool of $P - 1$ potential X variables, with the predictor variables expressed in appropriate form (linear, quadratic, transformed) and important interactions included, such that $MSE(X_1, X_2, \dots, X_{P-1})$ is unbiased

Akaike's Information Criterion

$$AIC_p = n \ln SSE_p - n \ln n + 2p$$

Schwartz's Bayesian Criterion

$$SBC_p = n \ln SSE_p - n \ln n + [\ln n]p$$

- n.b. We are searching for models with small AIC_p & SBC_p
- Both measures share a first term, $n \ln SSE_p$ that decrease in p
 - Both measures share a second, constant term, $n \ln n$
 - Both models penalize large p values with varying penalties
 - If $n \geq 8$, the penalty for $SBC_p > AIC_p$, indicative of SBC_p favoring highly parsimonious models ($\ln(8) \approx 2.079442$)

$PRESS_p$ Criterion

- The prediction sum of squares ($PRESS$) criterion is a measure of how well the use of the fitted values for a subset model can predict the observed responses Y_i ; the error sum of squares SSE , is also such a measure
- $PRESS$ differs from $SSE = \sum(Y_i - \hat{Y}_i)^2$ in that each fitted value \hat{Y}_i is obtained by estimating the regression function with $n - 1$ observations, i.e., the i^{th} observation is removed when estimating the regression function, and that regression is then used to obtain a predicted value $\hat{Y}_{i(i)}$ for the i^{th} case
- The $PRESS_p$ criterion is the sum of the squared prediction errors over all n cases

$$PRESS_p = \sum(Y_i - \hat{Y}_{i(i)})^2$$

- Models with small $PRESS$ values are considered good candidate models, the implication being that small prediction errors are indicative of a good model
- n.b. We will shortly learn a technique that avoids having to compute n different regression models for each value of p

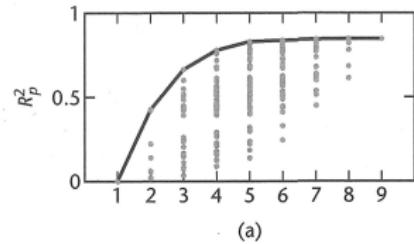
Criteria for Model Selection EXAMPLE

X Variables in Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	p	SSE_p	R_p^2	$R_{a,p}^2$	C_p	AIC_p	SBC_p	$PRESS_p$
None	1	12.808	0.000	0.000	151.498	-75.703	-73.714	13.296
X_1	2	12.031	0.061	0.043	141.164	-77.079	-73.101	13.512
X_2	2	9.979	0.221	0.206	108.556	-87.178	-83.200	10.744
X_3	2	7.332	0.428	0.417	66.489	-103.827	-99.849	8.327
X_4	2	7.409	0.422	0.410	67.715	-103.262	-99.284	8.025
X_1, X_2	3	9.443	0.263	0.234	102.031	-88.162	-82.195	11.062
X_1, X_3	3	5.781	0.549	0.531	43.852	-114.658	-108.691	6.988
X_1, X_4	3	7.299	0.430	0.408	67.972	-102.067	-96.100	8.472
X_2, X_3	3	4.312	0.663	0.650	20.520	-130.483	-124.516	5.065
X_2, X_4	3	6.622	0.483	0.463	57.215	-107.324	-101.357	7.476
X_3, X_4	3	5.130	0.599	0.584	33.504	-121.113	-115.146	6.121
X_1, X_2, X_3	4	3.109	0.757	0.743	3.391	-146.161	-138.205	3.914
X_1, X_2, X_4	4	6.570	0.487	0.456	58.392	-105.748	-97.792	7.903
X_1, X_3, X_4	4	4.968	0.612	0.589	32.932	-120.844	-112.888	6.207
X_2, X_3, X_4	4	3.614	0.718	0.701	11.424	-138.023	-130.067	4.597
X_1, X_2, X_3, X_4	5	3.084	0.759	0.740	5.000	-144.590	-134.645	4.069

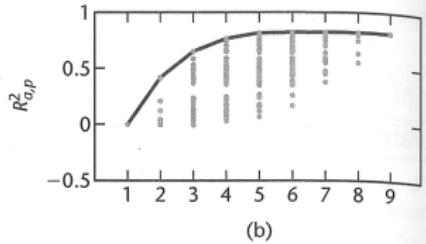
Section 11

Automatic Search Procedures for Model Selection

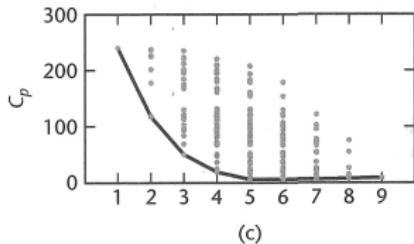
Best Subsets



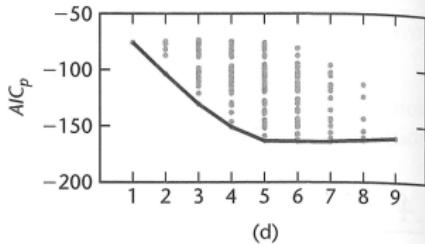
(a)



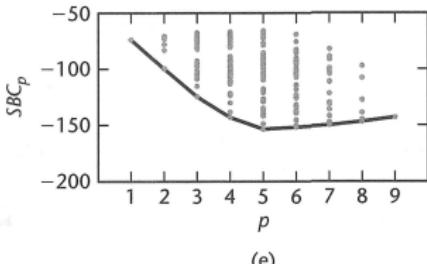
(b)



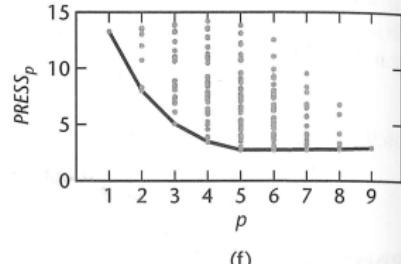
(c)



(d)



(e)



(f)

Best Subsets cont'd

p	(1) SSE_p	(2) R_p^2	(3) $R_{a,p}^2$	(4) C_p	(5) AIC_p	(6) SBC_p	(7) $PRESS_p$
1	12.808	0.000	0.000	240.452	-75.703	-73.714	13.296
2	7.332	0.428	0.417	117.409	-103.827	-99.849	8.025
3	4.312	0.663	0.650	50.472	-130.483	-124.516	5.065
4	2.843	0.778	0.765	18.914	-150.985	-143.029	3.469
5	2.179	0.830	0.816	5.751	-163.351	<u>-153.406</u>	<u>2.738</u>
6	2.082	0.837	0.821	<u>5.541</u>	-163.805	-151.871	2.739
7	2.005	0.843	<u>0.823</u>	<u>5.787</u>	<u>-163.834</u>	-149.911	2.772
8	1.972	0.846	<u>0.823</u>	7.029	-162.736	-146.824	2.809
9	1.971	<u>0.846</u>	0.819	9.000	-160.771	-142.870	2.931

The use of the 'Best Subsets' approach should be used to identify a set of 'good' models which one can use as a basic filtering approach, after which, further analysis is required to examine all of diagnostic measures and, if applicable, remedial measures

Stepwise Regression Methods

- When the number of predictor variables grows large, solving all possible instances for a given number of $p - 1$ predictor variables can be difficult, time consuming or too computationally intensive, especially with large n , which usually (and should) grow with $p - 1$
 - In this case, a Stepwise Regression Method can be employed, significantly reducing computational time
- n.b.** Although 'Best Subsets' is computationally more intensive, it can identify a maximizing or minimizing model (or models) for a given criterion, generating multiple 'good' models for further evaluation, whereas 'Stepwise Methods' will result in the selection of a single model, detracting from that fact that other models with more or fewer variables may also be 'good'
- If using 'Stepwise Methods,' it is suggested to use the model algorithmically selected and explore variations on that model to confirm whether or not it truly is the best

Forward Stepwise Regression

- ① Fit an SLR for **each** predictor of the $P - 1$ potential predictor variables, finding the predictor variable that has the maximal t^* , i.e.,

$$\min p\text{-value} \quad \forall k \in \{1, \dots, P - 1\}$$

- n.b this algorithm should have a predetermined threshold α such that the p -value must be below said threshold to be included in the model
- ② Once the first variable is selected, let's call this variable $X_{<1>}$, $P - 2$ 2-variable MLR regressions are run, where one variable is fixed to be $X_{<1>}$ and the other is selected from the remaining $P - 2$ remaining variables; an additional variable with the lowest p -value is added so long as its p -value is below α , otherwise the algorithm terminates

Forward Stepwise Regression cont'd

- ③ Assuming a second variable, let's call it $X_{<2>}$, is added to the model, the algorithm will then re-evaluate the updated p -values of the variables (plural) that were included in the model **prior** to the addition of $X_{<2>}$; if those p -values are greater than the threshold α value, they are dropped from the model
- ④ Repeat steps 2 and 3 until the algorithm terminates due to either
 - No remaining variables exhibit p -values that are below the threshold α value
 - All variables have been included in the model and no additional variables remain

Notes on Forward Stepwise Regression

- Simulation studies have shown that using too large of a threshold α for the inclusion of variables in a model often results in regression functions with too many predictor variables, suffering from overfitting (amongst other things)
- Alternatively, simulation studies have also shown that using too small of a threshold α for the inclusion of variables in a model results in underfitting and an overestimation of σ^2
- Ensure $\alpha_{add} < \alpha_{drop}$ to avoid potential (infinite) cycling of adding and removing of the same predictor variable
- The order in which variables enter the regression model does not reflect their importance

Other Stepwise Procedures

- ① Forward Selection: this is a simplified version of Forward Stepwise Regression that only sequentially adds variables based on a threshold α value, omitting the ability to re-test and drop existing variables
- ② Backward Elimination: the opposite of forward selection, beginning with all $P - 1$ predictor variables and sequentially dropping variables that do not meet a threshold α value, without any chance of variable reintroduction
- ③ Stepwise Backward Regression: the opposite of forward stepwise regression selection, or alternatively, similar to Backward Elimination with additional ability to reintroduce previously dropped variables

Final Notes on Algorithmic Model Selection

- Some statisticians argue for backward stepwise searches or forward stepwise searches. Why?
- Don't forget in all of the aforementioned procedures, only drop or add **one variable at a time**, then rerun the regression and reevaluate
- What issues have we not broached with respect to algorithmic model selection?
- (Human) Judgement needs to play an important role in model building for exploratory studies

R & Algorithmic Model Selection

- The step function in the car package uses AIC_p criterion for “backward” (default) or “forward” stepwise regression
- The step function also allows for a “both” option, that at any given step considers both an addition or a removal

```
> install.package("car")
> library(car)
> lmObject <- lm(Y ~ X1 + X2 + X3, data=myData)
> backwardRegression <- step(lmObject, scope=list(lower=~X2))

# where the scope=list(lower=~X2) forces X2 to be in every model run

# this package is robustly designed, dropping all related factors if a
# single factor level is dropped

# an AIC value for all P-1 predictors is generated, and then the
# algorithm drops the variable that will maximally decrease the AIC

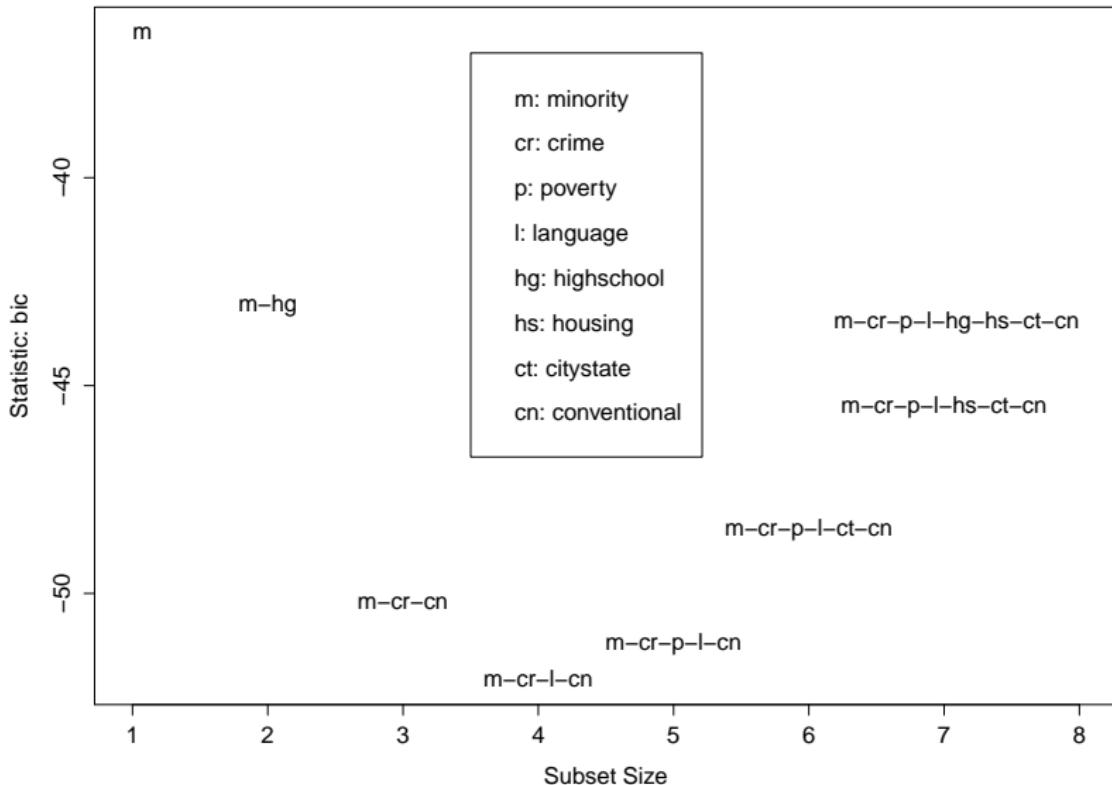
# termination condition: removal of a variable increases AIC

# an interesting way to evaluate the algorithmically selected model
# is to plot fitted values with P-1 predictors verses fitted
# values for the algorithmically selected model
```

R & Algorithmic Model Selection cont'd

- The `leaps` package contains the `regsubsets` function for linear models which finds the `nbest` subset to optimize Mallow's C_p
- The `regsubsets` function works directly on regressors and not on predictors...what's the difference, and why can that cause serious problems?
- What about optimizing Mallow's C_p , is that a good choice?

subsets, regsubsets & leaps EXAMPLE



Section 12

Model Validation

Model Validation Techniques

The final step in the model-building process is the validation of the selected regression models, which usually involved checking a candidate model against independent data

- ① Collect new data and verify the model's predictive ability
- ② Comparison of results with theoretical expectations, earlier empirical results, and/or simulation results
- ③ Use of a holdout sample to check the model and its predictive ability

This is most easily done in a controlled experiment and far more difficult to do in an observational study

- ① Run a regression on the new data and reestimate the model form
- ② A second method is to calculate the mean squared prediction error (*MSPR*)

$$MSPR = \frac{\sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2}{n^*}$$

where

- Y_i and \hat{Y}_i are the response and predicted values for the i^{th} validation case based on the model-building data set
- n^* is the number of cases in the validation data set

If *MSPR* and *MSE* are close, then *MSE* for the original model is not terribly biased and therefore adequate

- If your data set is permissively large, split the data you have collected into two sets, a model-building set and a validation (or prediction) set
 - This technique is also known as cross-validation
- n.b. Make sure the model-building set is sufficiently large, e.g., roughly 6 to 10 observations per predictor variable (Kutner et al. recommendation)
- If you can't make an equal data split, reserve enough data for your model-building set and use what is left as a validation set

- Splitting data can be random, or alternatively, one can match pairs of collected data and send one of each to the model-building and validation set respectively
- Note that by splitting data, variance of the estimated regression coefficients can increase if the data set is not large, however, once the model has been validated, it is customary to use the entire data set for estimating the final regression model

Best Subsets cont'd

Statistic	(1) Model 1 Training Data Set	(2) Model 1 Validation Data Set	(3) Model 2 Training Data Set	(4) Model 2 Validation Data Set	(5) Model 3 Training Data Set	(6) Model 3 Validation Data Set
p	5	5	6	6	7	7
b_0	3.8524	3.6350	3.8671	3.6143	4.0540	3.4699
$s\{b_0\}$	0.1927	0.2894	0.1906	0.2907	0.2348	0.3468
b_1	0.0733	0.0958	0.0712	0.0999	0.0715	0.0987
$s\{b_1\}$	0.0190	0.0319	0.0188	0.0323	0.0186	0.0325
b_2	0.0142	0.0164	0.0139	0.0159	0.0138	0.0162
$s\{b_2\}$	0.0017	0.0023	0.0017	0.0024	0.0017	0.0024
b_3	0.0155	0.0156	0.0151	0.0154	0.0151	0.0156
$s\{b_3\}$	0.0014	0.0020	0.0014	0.0020	0.0014	0.0021
b_5	—	—	—	—	-0.0035	0.0025
$s\{b_5\}$	—	—	—	—	0.0026	0.0033
b_6	—	—	0.0869	0.0731	0.0873	0.0727
$s\{b_6\}$	—	—	0.0582	0.0792	0.0577	0.0795
b_8	0.3530	0.1860	0.3627	0.1886	0.3509	0.1931
$s\{b_8\}$	0.0772	0.0964	0.0765	0.0966	0.0764	0.0972
SSE_p	2.1788	3.7951	2.0820	3.7288	2.0052	3.6822
$PRESS_p$	2.7378	4.5219	2.7827	4.6536	2.7723	4.8981
C_p	5.7508	6.2094	5.5406	7.3331	5.7874	8.7166
MSE_p	0.0445	0.0775	0.0434	0.0777	0.0427	0.0783
$MSPR$	0.0773	—	0.0764	—	0.0794	—
$R^2_{a,p}$	0.8160	0.6824	0.8205	0.6815	0.8234	0.6787

Section 13

Added Variable Plots

- We have examined a number of diagnostics to date, many of which can be traced back to SLR models, and include
 - ① residual plots against predictors in the model
 - ② residual plots against predictors **not** in the model
- A limitation of these residual plots is that they may not properly show the nature of the marginal effect of a predictor variable, given the other predictor variables are in the model
- Added-Variable Plots or Partial Regression Plots or Adjusted Variable Plots are a refined form of residual plot that provide information about the marginal importance of a predictor variable X_k , given the other predictor variables are already in the model

- In an added variable plot, both the response variable and the predictor variable X_k of interest are both regressed against the remaining predictor variables and the residuals obtained for each
- These residuals reflect the part of each variable that is not linearly associated with the other predictor variables already in the regression model
- Plotting these residuals against each other
 - ① shows the marginal importance of this variable in reducing the variability
 - ② may provide information about the nature of the marginal regression relation for the predictor variable X_k under consideration for possible inclusion in the regression model

- Say we have Y , X_1 and X_2 , and we are curious to know if adding X_1 to an SLR model regressing Y on X_2 will be useful
 - ① Regress Y on X_2 to obtain the fitted values and residuals

$$\hat{Y}_i(X_2) = b_0 + b_2 X_{i2}$$

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

- ② Regress X_1 on X_2 to obtain the fitted and residual values

$$\hat{X}_{i1}(X_2) = b_0^* + b_2^* X_{i2}$$

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

- ③ Generate a plot of $e_i(Y|X_2)$ (y -axis) on $e_i(X_1|X_2)$ (x -axis)

- Say we have Y , X_1 and X_2 , and we are curious to know if adding X_1 to an SLR model regressing Y on X_2 will be useful
 - ① Regress Y on X_2 to obtain the fitted values and residuals

$$\hat{Y}_i(X_2) = b_0 + b_2 X_{i2}$$

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

- ② Regress X_1 on X_2 to obtain the fitted and residual values

$$\hat{X}_{i1}(X_2) = b_0^* + b_2^* X_{i2}$$

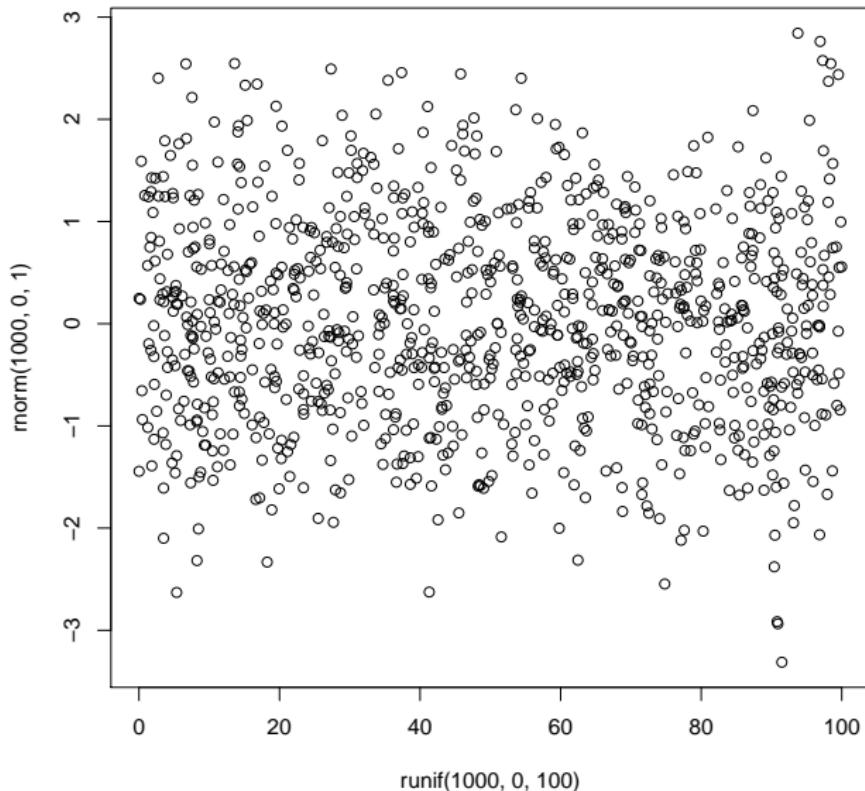
$$e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

- ③ Generate a plot of $e_i(Y|X_2)$ (y -axis) on $e_i(X_1|X_2)$ (x -axis)

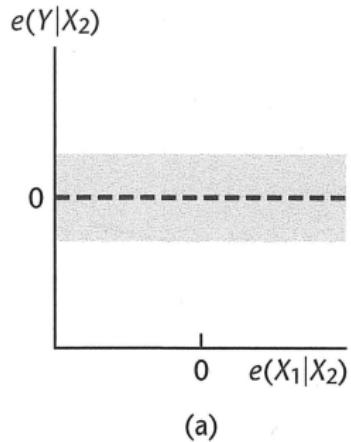
What do you think the scatter will look like if the marginal value of the added variable is negligible?

Added-Variable Plots cont'd

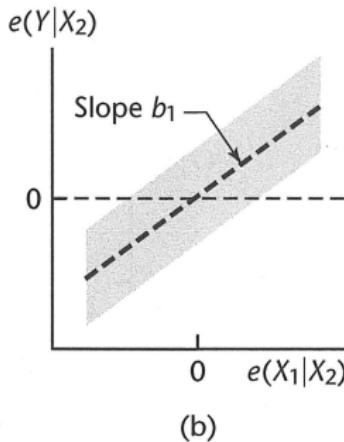
```
> plot(runif(1000,0,100),rnorm(1000,0,1))
```



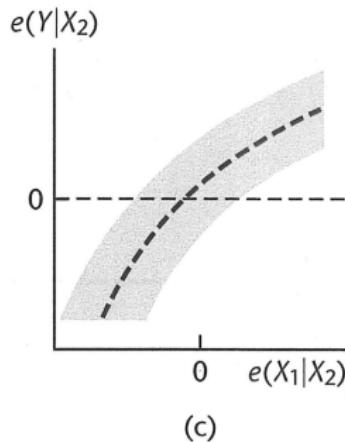
Added-Variable Plots cont'd



(a)



(b)



(c)

Added-Variable Plots [R Code]

```
> universityAdmissions <- read.table('~/Dropbox/Public/MSAN601USF/
  universityAdmissions.txt',sep=' ',header=FALSE)

#here V2 is GPA (Y), V3 is class rank (X1), V4 is ACT score (X2)

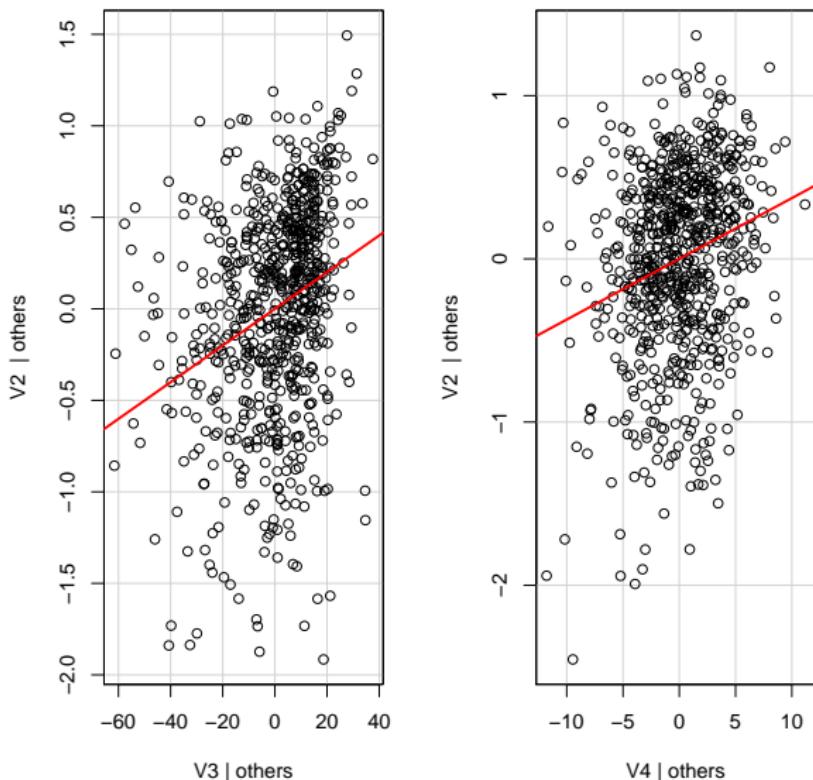
> univAd.lm <- lm(V2~V3+V4,data=universityAdmissions)
  #regresses V2 on V3 and V4

> install.packages('car')
> library('car')

> avPlots(univAd.lm)
```

Added-Variable Plots [R Output]

Added-Variable Plots



Final Notes on Added-Variable Plots

- ① An Added-Variable Plot only suggests the nature of the functional relation in which a predictor variable should be added to the regression model, but it does not provide an analytical expression of the relation
- ② The relationship shown in an Added-Variable Plot is conditional on other variables being included in model: it says nothing of variable in question directly
- ③ If you decide the marginal value is sufficiently significant to add the variable to the model, don't forget to run the basic diagnostics and remedial measures once added, i.e., additional residual plots, attempting transforms, polynomial effects, etc.

Final Notes on Added-Variable Plots cont'd

- ④ Be cautious when interpreting Added-Variable Plots: if the functional form of the variables already in the model are misspecified, e.g., they are linear in the model but should be polynomial, the perceived marginal effect of the added variable may be obfuscated (this is also true for other issues such as high levels of multicollinearity between predictor variables)
- ⑤ Slope of the fitted regression line, when regressing $e_i(Y|X_2)$ on $e_i(X_1|X_2)$ is exactly the coefficient of X_2 if you regressed Y on both X_1 and X_2
- ⑥ The R^2 value obtained when regressing $e_i(Y|X_2)$ on $e_i(X_1|X_2)$ is coefficient of partial determination $R^2_{YX_2|X_1}$
- ⑦ AV plots always go through which point?

Section 14

Component + Residual Plots

Component + Residual Plots

An effective way to detect whether or not a variable X_k needs to be transformed is called the Component + Residual Plot or the Partial Residual Plot

- ① Regress Y on all predictor variables including X_k , and obtain the residuals e_i ;
- ② Compute the partial residuals as follows

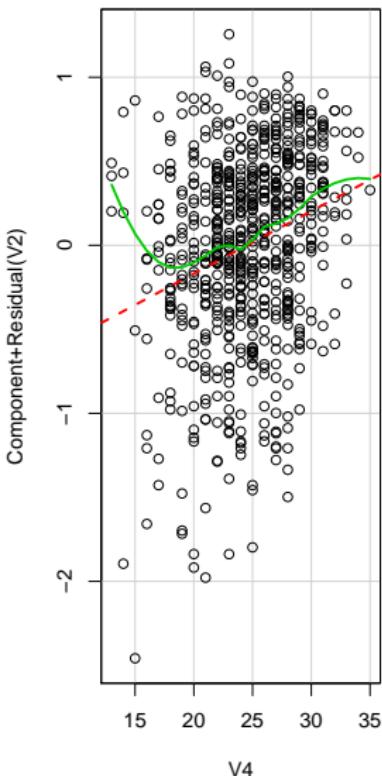
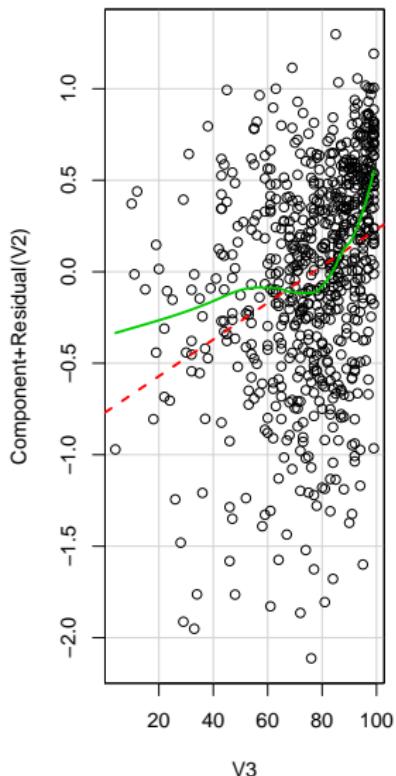
$$p_i(X_k) = e_i + b_k X_{ik}$$

- ③ Generate a plot of $p_i(X_k)$ (y-axis) on X_k (x-axis)
 - If $p_i(X_k)$ regressed on X_k is linear, we are in good shape
 - If $p_i(X_k)$ regressed on X_k is non-linear, we may want to attempt transformations based on the shape of the scatter

Component + Residual Plots [R Output]

```
> crPlots(univAd.lm)
```

Component + Residual Plots



- Component + Residual Plots may not be effective in recovering nonlinear partial relationships between the response and the predictors
- Combining Conditional Expectations and Residuals (CERES)
- Uses nonparametric regression smoothers rather than polynomial regression to adjust for nonlinear relationships among predictors
- Use `ceresPlots` function in the `car` package
- See Cook (1993) for more information

Section 15

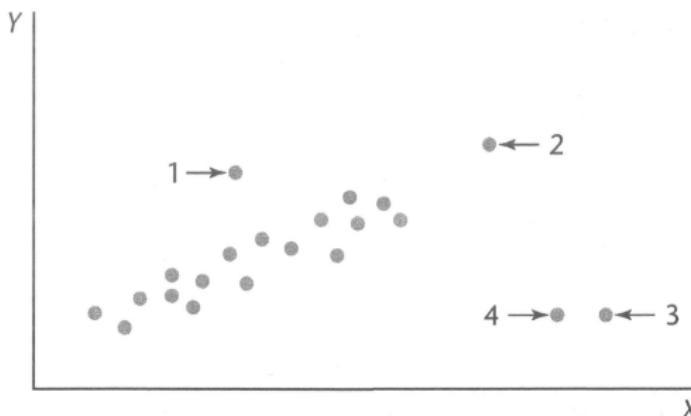
Identifying Outlying Y Observations

Outlying Cases

- Frequently, data sets contain cases that are outlying or extreme
- These outlying cases often involve a large residual and often have dramatic effects of the fitted least squares regression function
- An observation may be outlying or extreme with respect to its Y value, its X value(s) or both

Outlying Cases

- Frequently, data sets contain cases that are outlying or extreme
- These outlying cases often involve a large residual and often have dramatic effects of the fitted least squares regression function
- An observation may be outlying or extreme with respect to its Y value, its X value(s) or both



Outlying Cases cont'd

Classify these cases as being outliers in Y , X or both. Which have the least influence on the regression line and which have the most? Why?

Classify these cases as being outliers in Y , X or both. Which have the least influence on the regression line and which have the most? Why?

- ① Outlier in Y , minimal influence on regression line

Classify these cases as being outliers in Y , X or both. Which have the least influence on the regression line and which have the most? Why?

- ① Outlier in Y , minimal influence on regression line
- ② Outlier in X and Y , minimal influence on regression line

Classify these cases as being outliers in Y , X or both. Which have the least influence on the regression line and which have the most? Why?

- ① Outlier in Y , minimal influence on regression line
- ② Outlier in X and Y , minimal influence on regression line
- ③ Outlier in X , highly influential on regression line

Classify these cases as being outliers in Y , X or both. Which have the least influence on the regression line and which have the most? Why?

- ① Outlier in Y , minimal influence on regression line
- ② Outlier in X and Y , minimal influence on regression line
- ③ Outlier in X , highly influential on regression line
- ④ Outlier in X , highly influential on regression line

Studentized Residuals

- An estimator of the standard deviation of residuals e_i is

$$s\{e_i\} = \sqrt{MSE(1 - h_{ii})}$$

where h_{ii} is the i^{th} diagonal element of the hat matrix \mathbf{H} , and can be calculated as

$$h_{ii} = \mathbf{X}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$$

where

$$\mathbf{X}_i = \begin{bmatrix} 1 \\ X_{i,1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}_{p \times 1}$$

- The ratio of e_i to $s\{e_i\}$ is called the **studentized residual** and is denoted by r_i

$$r_i = \frac{e_i}{s\{e_i\}}$$

Studentized Residuals cont'd

- Whereas residuals will have substantially different sampling variations if their standard deviations differ markedly, the studentized residuals r_i have constant variance (when the model is appropriate)
- Studentized residuals are often called *internally studentized residuals*

- To make residuals additionally effective in detecting outlying Y observations, measure the i^{th} residual $e_i = Y_i - \hat{Y}_i$ when the fitted regression is based on all of the cases except the i^{th} observation
- If Y_i is far outlying, the fitted regression line with the i^{th} observation removed should differ sufficiently from the fitted regression line with the i^{th} observation included, making the deleted residual large and therefore easier to detect
- To compute a deleted residual
 - ① Delete the i^{th} case
 - ② Fit a regression function to the remaining $n - 1$ cases
 - ③ Obtain the point estimate of the expected value when the X levels are those of the i^{th} case, denoted $\hat{Y}_{i(i)}$
 - ④ Compute the deleted residual d_i

$$d_i = Y_i - \hat{Y}_{i(i)}$$

n.b. this is the same as the *PRESS* prediction error seen earlier

Deleted Studentized Residuals

- Combining both aforementioned refinements to residuals, we obtain the studentized deleted residual, a strong method by which to identify residuals
- We obtain studentized deleted residuals by studentizing the deleted residual as follows

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

where $MSE_{(i)}$ is the MSE calculated when the i^{th} case omitted

- t_i can be rewritten as

$$t_i = e_i \sqrt{\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2}}$$

to avoid having to recalculate $MSE_{(i)}$ for each omitted case

Testing for Large Deleted Studentized Residuals

- A Bonferroni simultaneous test procedure with a family significance level of α requires critical t value of

$$t_{(1-\alpha/2n; n-p-1)}$$

where all $|t_i| > t_{(1-\alpha/2n; n-p-1)}$ are considered to be influential outliers

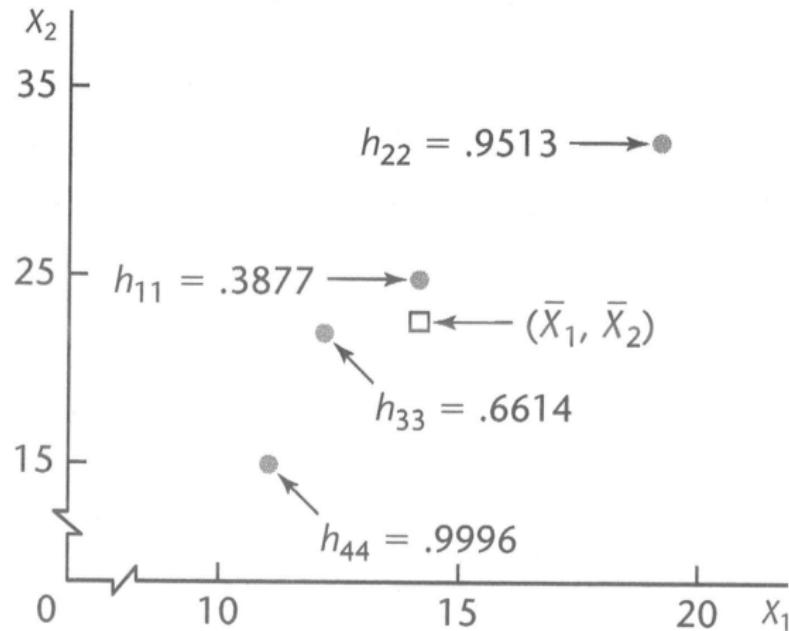
n.b. We have not covered the Bonferroni test in this course

Section 16

Identifying Outlying X Observations

- The hat matrix \mathbf{H} is helpful in directly identifying outlying X observations, in particular, the diagonal elements h_{ii}
- The diagonal elements h_{ii} of the hat matrix have some useful properties
 - ① $0 \leq h_{ii} \leq 1$
 - ② $\sum h_{ii} = p$
- h_{ii} is also a measure of the distance between the X values for the i^{th} case and the mean X values for all n cases
- A large h_{ii} value implies that the i^{th} case is distant from the center of all X observations
- In this context, the diagonal element h_{ii} is called the **leverage** of the i^{th} case

Illustration of Leverage Values



If the i^{th} case is outlying in terms of its X observations and therefore has a large leverage value h_{ii} , it exercises substantial leverage in determining the fitted value \hat{Y}_i for the following reasons

- ① The fitted value \hat{Y}_i is a linear combination of the observed Y values

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

with h_{ii} being the weight of observation Y_i in determining this fitted value

Thus the larger h_{ii} the more important Y_i is in determining \hat{Y}_i

- n.b h_{ii} is solely a function of the X values, so h_{ii} measures the role of the X values in determining how important Y_i is in affecting the fitted value \hat{Y}_i

If the i^{th} case is outlying in terms of its X observations and therefore has a large leverage value h_{ii} , it exercises substantial leverage in determining the fitted value \hat{Y}_i for the following reasons

- ② The larger h_{ii} , the smaller the variance of the residual e_i

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$$

Hence the larger h_{ii} , the closer the value Y_i will be to the fitted value \hat{Y}_i

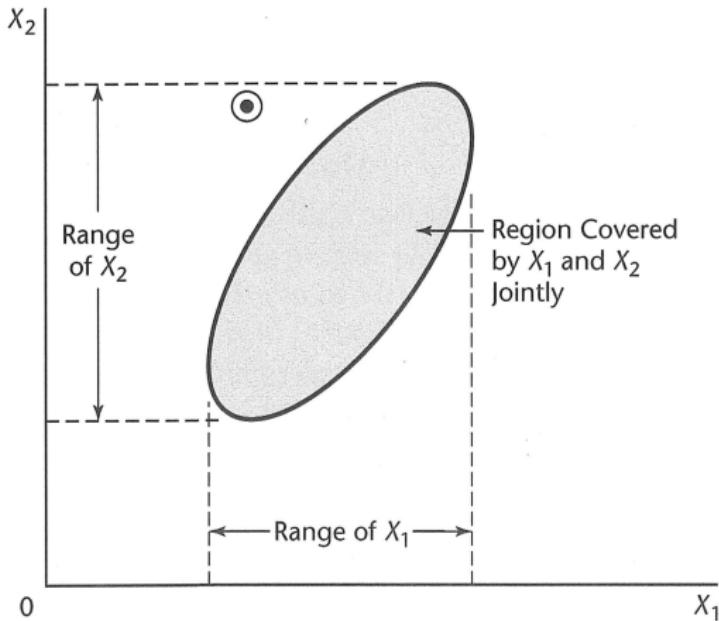
In the extreme case where $h_{ii} = 1$, $Y_i = \hat{Y}_i$

- A leverage value is considered large if it is more than twice as large as the mean leverage value, denoted \bar{h}

$$\bar{h} = \frac{\sum h_{ii}}{n} = \frac{p}{n}$$

- Hence leverage values greater than $2p/n$ are considered by this rule to indicate outlying cases with respect to their X values
- Another rule of thumb is $h_{ii} > 0.5$ are considered to have high leverage, and $0.2 < h_{ii} < 0.5$ are considered to have moderate leverage

Leverage Values & Hidden Extrapolations



Leverage Values & Hidden Extrapolations cont'd

- With more than two predictor variables, there is no graphical method available to detect whether or not extrapolations are present
- Formulaically, one can compute

$$h_{new,new} = \mathbf{X}'_{new} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_{new}$$

where \mathbf{X}_{new} is the vector containing the X values for which an inference about a mean response or new observation is to be made

- If $h_{new,new}$ is well within the range of leverage values h_{ii} for the cases of the data set, no extrapolation is involved
- If $h_{new,new}$ is much larger than the leverage values for the cases in the data set, an extrapolation is indicated

Section 17

Identifying Influential Observations

Identifying Influential Observations

- After identifying cases that are outlying with respect to their Y or X values, we subsequently want to ascertain whether or not these outlying cases are **influential**, where influential is defined as a resulting significant change in the fitted regression function if an influential observation is omitted
- Not all outlying cases need be influential
- We will examine three measures of influence
 - ① $DFFITS$
 - ② Cook's Distance
 - ③ $DFBETAS$

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

- *DFFITS* is an acronym for the difference between the fitted values
- Specifically, $(DFFITS)_i$ measures the difference between the fitted Y values for the i^{th} case when all n observations are used in fitting the regression function \hat{Y}_i , and the fitted values for the i^{th} case when omitting the i^{th} case $\hat{Y}_{i(i)}$
- The denominator provides a standardization so that the value $(DFFITS)_i$ for the i^{th} case represents the number of estimated standard deviations of \hat{Y}_i that the fitted value \hat{Y}_i increases or decreases with the inclusion of the i^{th} case in fitting the regression model

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

- n.b. The final equality above expresses *DFFITS* as the studentized deleted residual, increased or decreased by a factor that is a function of the leverage value
- Large *DFFITS* values are indicative of influential outliers
 - Rule of thumb
 - For small- to medium-size data sets, *DFFITS* values exceeding 1 are indicative of influential outliers
 - For large-size data sets, *DFFITS* values exceeding $2\sqrt{p/n}$ are indicative of influential outliers

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE} = \frac{e_i^2}{pMSE} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

- In contrast to *DFFITS* which considers the influence of the i^{th} case on the fitted value \hat{Y}_i , Cook's Distance measure considers the influence of the i^{th} case on all n fitted values
- Note that the differences in the numerator are squared and summed, so that the aggregate influence of the i^{th} case is measured without regard to the sign of the effects
- The denominator acts to standardize the measure
- Interpret Cook's D values, the rule of thumb is that the percentile value of D_i should be less than the 10 to 20th percentile of the $F_{(p, n-p)}$ distribution
- Small D_i values indicate no/little influence, large values indicate influential observations

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

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

- $DFBETAS$ is a measure of the influence of the i^{th} case on each regression coefficient b_k is calculated as the difference between the estimated regression coefficient b_k based on all n cases and the regression coefficient obtained when the i^{th} case is omitted
- Cook's Distance measure is algebraically equivalent to $DFBETAS$

Diagnostic Plots [R Code]

```
> universityAdmissions <- read.table('~/Dropbox/Public/MSAN601USF/
  universityAdmissions.txt',sep=' ',header=FALSE)

#here V2 is GPA (Y), V3 is class rank (X1), V4 is ACT score (X2)

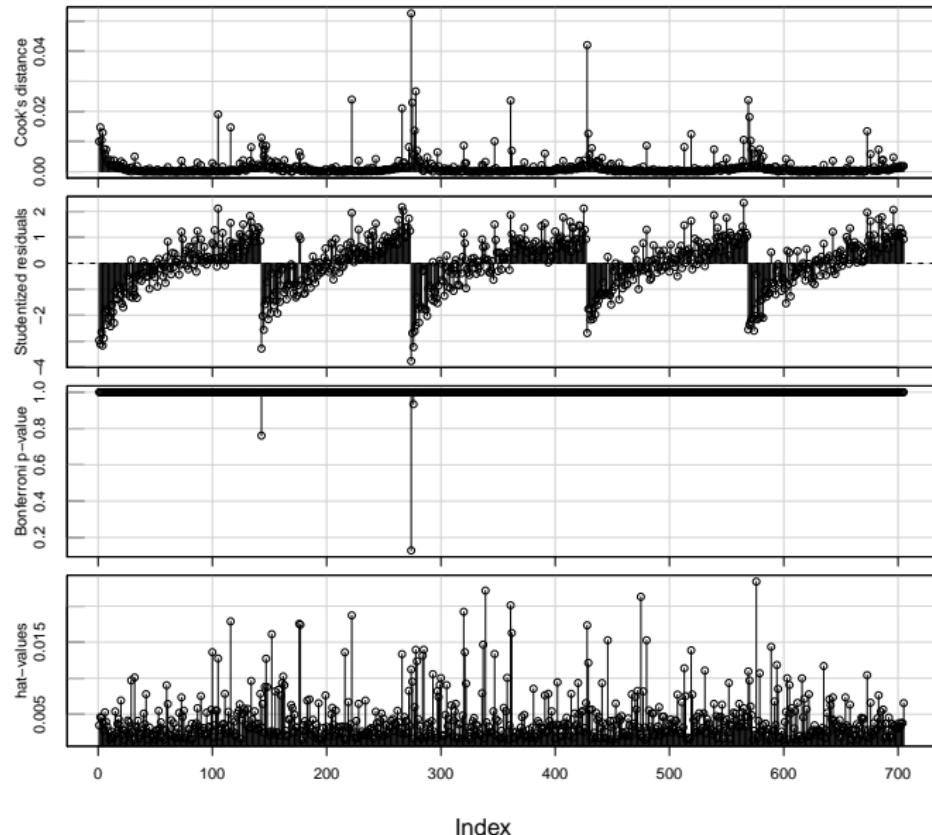
> univAd.lm <- lm(V2~V3+V4,data=universityAdmissions)
  #regresses V2 on V3 and V4

> install.packages('car')
> library('car')

> influenceIndexPlot(univAd.lm)
```

Diagnostic Plots [R Output]

Diagnostic Plots



Section 18

Multicollinearity Diagnostics

Several issues arise when there is high multicollinearity between variables being considered for a regression model

- ① Adding or deleting variables changes regression coefficients
 - ② The extra sum of squares changes as a function of the variables already in the model
 - ③ The $s\{b_k\}$ become large when the predictor variables in the predictor variables in the regression model are highly correlated with each other
 - ④ The estimated regression coefficients themselves may not be statistically significant even though a definite statistical relation exists between the response variable and the set of predictor variables
- n.b.** The above can occur without multicollinearity being present, but only under unusual circumstances

Informal Diagnostics for Multicollinearity

The presence of multicollinearity can be detected using the following observational techniques

- ① Large changes in $s\{b_k\}$'s when a predictor variable is added or removed from the model, or when an observation is altered or deleted
- ② Nonsignificant results in individual tests on the regression coefficients for important predictor variables
- ③ Estimated regression coefficients with algebraic signs that are opposite from expectations based on theoretical considerations or prior experience
- ④ Large coefficients of simple correlation between pairs of predictor variables (found in the correlation matrix r_{xx})
- ⑤ Wide confidence intervals for the regression coefficients representing important predictor variables

- The Variance Inflation Factor (*VIF*) is a formalized, widely accepted quantitative technique to assess multicollinearity
 - VIFs measure how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related
- n.b. The theory underlying *VIFs* is based on the matrix representation of MLR models as well as the standardized MLR model, both of which we have skirted around up to this point, therefore the final results are presented herein, with a reference to Kutner 4e/5e, Chapter 10, §10.5

- The VIF for a given predictor variable X_k , VIF_k is

$$VIF_k = \frac{1}{1 - R_k^2} \quad k = 1, \dots, p - 1$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p - 2$ other X variables in the model

- The VIF_k is equal to 1 when $R_k^2 = 0$, i.e., when X_k has no linear relationship with the other X variables in the model
- When $R_k^2 \neq 0$, $VIF_k > 1$, indicating an inflated $s\{b_k\}$
- When $R_k^2 = 1$, VIF_k is unbounded
- Rule of Thumb: a maximally acceptable VIF among all X variables exceeding 10 is often used as an indicator of severe multicollinearity and therefore exerting undue influence of the least squares estimates

Variance Inflation Factor cont'd

- In the aggregate, one might also calculate the mean VIF , \overline{VIF} , for all predictor variables X_1, \dots, X_{p-1}

$$\overline{VIF} = \frac{\sum VIF_k}{p - 1}$$

where $\overline{VIF} \gg 1$ is typically interpreted as cause for concern

Final Notes of the Variance Inflation Factor

- ① Some computer packages will use the reciprocal of the Variance Inflation Factor, $1/VIF_k$, as criteria for allowing a variable into a model when employing a stepwise regression procedure, with typical tolerance limits being 0.01, 0.001, or 0.0001, below which the variable is not permitted entry into the model
- ② A limitation of the VIF is the inability to distinguish between several simultaneous multicollinearities

Variance Inflation Factors [R Code]

```
> universityAdmissions <- read.table('~/Dropbox/Public/MSAN601USF/
  universityAdmissions.txt',sep=' ',header=FALSE)

#here V2 is GPA (Y), V3 is class rank (X1), V4 is ACT score (X2)

> univAd.lm <- lm(V2~V3+V4,data=universityAdmissions)
  #regresses V2 on V3 and V4

> install.packages('car')
> library('car')

> vif(univAd.lm)
```

V3	V4
1.243492	1.243492

Section 19

Standardized MLR Model

Lack of Comparability

- We often have a difficult time comparing regression coefficients in an MLR. Which coefficient is the most important?

E.g.

$$\hat{Y} = 200 + 20,000X_1 + 0.2X_2$$

One may be inclined to assume X_1 is the only important predictor variable in the model and that X_2 is largely irrelevant...but remember, we don't know the units

- The units can be different orders of magnitude, e.g., X_1 could be in pennies and X_2 in \$10,000 of dollars
- X_1 and X_2 could also be in non-comparable units, e.g., m³ of CO₂ and dollars

Correlation Transformation

- The correlation transformation is a modification of the usual standardization variable
- A typical standardization of the variable includes centering and subsequently expressing the centered observations in units of the standard deviation of the variable
- The correlation transformation is a simple function of the standardized variables

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

and

$$X_{ik}^* = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{s_k} \right)$$

Standardized Regression Model

- The standardized regression model with the transformed variables Y^* and X_k^* as defined by the correlates on transformation is called a standardized regression model

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

- There is no intercept parameter in the standardized regression model as the least squares calculations would always lead to an estimated intercept term of zero
- The parameters in the standardized regression $\beta_1^*, \dots, \beta_{p-1}^*$ and the original parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$ in the ordinary multiple regression model are related as follows

$$\beta_k = \left(\frac{s_Y}{s_k} \right) \beta_k^* \quad \beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \dots - \beta_{p-1} \bar{X}_{p-1}$$

Matrix Components of the Standardized Regression Model

The matrix components of the standardized regression model have some interesting properties; it can be shown that

$$\mathbf{X}'\mathbf{X}_{(p-1) \times (p-1)} = \mathbf{r}_{XX}$$

where \mathbf{r}_{XX} is the simple correlation matrix between all pairs of X variables

$$\mathbf{b}_{(p-1) \times 1} = \mathbf{r}_{XX}^{-1} \mathbf{r}_{XY}$$

where \mathbf{r}_{XY} is a vector containing the coefficients of simple correlation between the response variable Y and each of the X variables and where

$$\mathbf{b}_{(p-1) \times 1} = \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{p-1}^* \end{bmatrix}$$

where b_1^*, \dots, b_{p-1}^* are the standardized regression coefficients

Photo Studio EXAMPLE

A photo studio specializing in children's portraits operates studios in 21 cities. It collected data on sales (Y) in \$1,000's, the number of people aged 16 or younger in the community (X_1) in 1,000's of people and the per capita disposable income (X_2) in \$1000's.

Photo Studio EXAMPLE

A photo studio specializing in children's portraits operates studios in 21 cities. It collected data on sales (Y) in \$1,000's, the number of people aged 16 or younger in the community (X_1) in 1,000's of people and the per capita disposable income (X_2) in \$1000's.

```
> dwaine <- read.table('~/Desktop/dwaine.txt',sep=' ',header=FALSE)
> colnames(dwaine) <- c("targetPop","perCapDispInc","Sales")

> dwaine.lm <- lm(Sales ~ targetPop + perCapDispInc, data= dwaine)
> summary(dwaine.lm)

<< output partially omitted for concision >>
```

Coefficients:

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

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared: 0.9167, Adjusted R-squared: 0.9075
F-statistic: 99.1 on 2 and 18 DF, p-value: 1.921e-10

Photo Studio EXAMPLE cont'd

To obtain the standardized coefficients, we run the following code

```
> install.packages('QuantPsyc')
> library('QuantPsyc')

> lm.beta(dwaine.lm)

targetPop perCapDispInc
0.7483670      0.2511039
```

Therefore standardized regression function becomes

$$\hat{Y}^* = 0.748X_1^* + 0.251X_2^*$$

The interpretation of the coefficients changes: an increase of one standard deviation of X_1 (target population) when X_2 (per capita disposable income) is fixed leads to a much larger increase in expected sales (in units of standard deviations of Y) than does an increase of one standard deviation of X_2 when X_1 is fixed.

Final Notes on Standardized Regression Coefficients

- Be cautious in jumping to conclusions regarding the magnitude of the standardized regression coefficients and their relative importance
- In our previous example $b_1^* > b_2^*$, which can be interpreted as X_1 having a much greater impact on sales than X_2 , but ...

Final Notes on Standardized Regression Coefficients

- Be cautious in jumping to conclusions regarding the magnitude of the standardized regression coefficients and their relative importance
- In our previous example $b_1^* > b_2^*$, which can be interpreted as X_1 having a much greater impact on sales than X_2 , but ...it turns out that X_1 and X_2 are highly correlated

```
> cor(dwaine)
            targetPop perCapDispInc      Sales
targetPop      1.0000000    0.7812993 0.9445543
perCapDispInc  0.7812993    1.0000000 0.8358025
Sales          0.9445543    0.8358025 1.0000000
```

- The reason we apply a correlation transformation as well as standardization of the variables is to ensure that the $\mathbf{X}'\mathbf{X}$ matrix contains values between -1 and 1
- Without applying the correlation transformation, we would obtain the same

Section 20

Remedial Measures for Multicollinearity

Method of Principal Components

- One remedial measure for multicollinearity is to form one or several composite indices based on highly correlated variables, an index being a linear combination of the correlated predictor variables
- The methodology of **principal components** provides for composite indices that are uncorrelated
- Often, just a few of these composite indices capture a substantial amount of information contained in the predictor variables, which are then used in regression analysis as predictor variables in lieu of the original highly correlated predictor variables
- The down side?

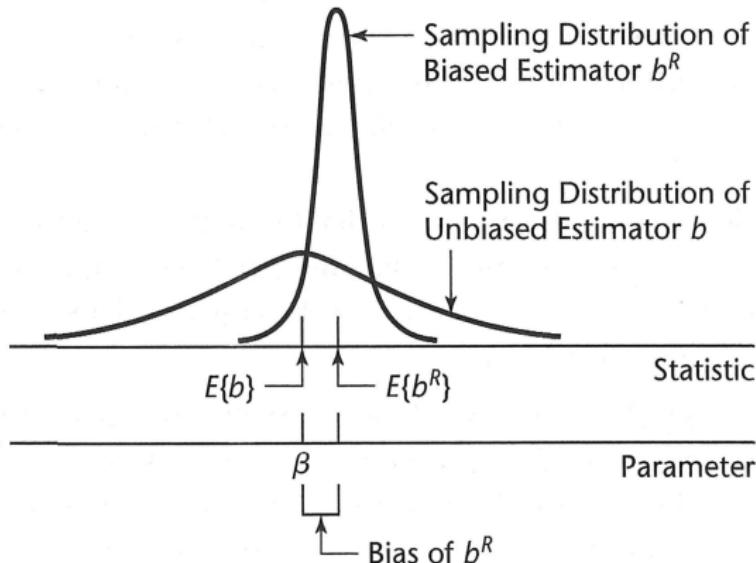
Method of Principal Components

- One remedial measure for multicollinearity is to form one or several composite indices based on highly correlated variables, an index being a linear combination of the correlated predictor variables
- The methodology of **principal components** provides for composite indices that are uncorrelated
- Often, just a few of these composite indices capture a substantial amount of information contained in the predictor variables, which are then used in regression analysis as predictor variables in lieu of the original highly correlated predictor variables
- The down side? It may be difficult to interpret the regression coefficients

- Ridge regression is one of several methods that have been proposed to remedy multicollinearity problems by modifying the method of least squares to allow for biased estimators of the regression coefficients
- When an estimator only has small bias and is substantially more precise than an unbiased estimator, it may well be the preferred side it will have a higher probability of being close to the true parameter value

Ridge Regression cont'd

- Estimator b is unbiased but imprecise, whereas estimator b^R is much more precise with small very little bias
- The probability that b^R falls near the true value of β is much greater than that for the unbiased estimator b



- A measure of the combined effect of bias and sampling variation is the mean squared error (recall from Mallow's C_p), which is the expected value of the squared deviation of the biased estimator b^R from the true parameter β
- Recall

$$E\{b^R - \beta\}^2 = \sigma^2\{b^R\} + (E\{b^R\} - \beta)^2$$

- In reference to the standardized regression model (transformed by the correlation transformation), the least squares normal equations are

$$\mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{XY}$$

- The ridge standardized regression estimators are obtained by introducing into the least squares normal equations a biasing constant $c \geq 0$ in the following form

$$(\mathbf{r}_{XX} + c\mathbf{I})\mathbf{b}^R = \mathbf{r}_{XY}$$

where \mathbf{b}^R is the vector of standardized ridge regression coefficients b_k^R

$$\mathbf{b}^R_{(p-1) \times 1} = \begin{bmatrix} b_1^R \\ b_2^R \\ \vdots \\ b_{p-1}^R \end{bmatrix}$$

where \mathbf{I} is the $(p - 1) \times (p - 1)$ identity matrix

- The solution of the normal equations yields the ridge standardized regression coefficients

$$\mathbf{b}^R = (\mathbf{r}_{XX} + c\mathbf{I})^{-1} \mathbf{r}_{XY}$$

- The constant c reflects the amount of bias in the estimators
- When $c = 0$, the ridge standardized regression coefficients reduce to the least squares regression coefficients in standardized form
- When $c > 0$, the ridge regression coefficients are biased but tend to be more stable, i.e., less variable than ordinary least squares estimators

The Choice of a Biasing Constant c

- It can be shown that the bias component of

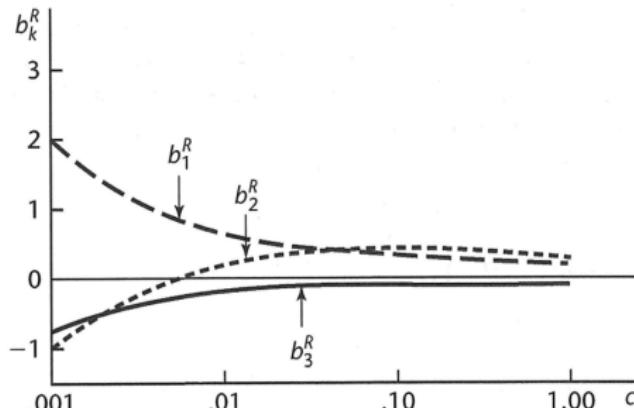
$$E\{b^R - \beta\}^2 = \sigma^2\{b^R\} + (E\{b^R\} - \beta)^2$$

increases as c gets larger, with all b_k^R tending toward zero while the variance component gets smaller

- It can also be shown that there always exists some value c for which the ridge regression estimator \mathbf{b}^R has a total mean squared error less than the ordinary least squares estimator \mathbf{b}
- The problem is that the optimum value of c changes from application to application
- A commonly used method of determining the biasing constant c is based on the *ridge trace* and the VIF_k 's

The Choice of a Biasing Constant c Using a Ridge Trace

- The ridge trace is a simultaneous plot of the $p - 1$ estimated ridge standardized regression coefficients for different values of c , usually between 0 and 1
- Empirical evidence suggests that the estimated ridge standardized regression coefficients can fluctuate wildly at small values of c , even changing signs, and eventually settle down and begin converging to zero as c increases, with VIF's dropping concomitantly



Body Fat EXAMPLE with Ridge Regression

- Recall for the body fat example, when including all three variables in the model, the result was all three predictors being statistically insignificant due to, we concluded, to very high levels of multicollinearity
- If we are insistent on keeping in all three variables, we can use ridge regression to mitigate the excessive multicollinearity effects
- At roughly $c = 0.02$, the VIF_k 's are all roughly equal to 1, and the resulting fitted model is

$$\hat{Y}^* = 0.5463X_1^* + 0.3774X_2^* - 0.1369X_3^*$$

which, when converted back to the unstandardized variables is

$$\hat{Y} = -7.39 + 0.55X_1 + 0.368X_2 - 0.1917X_3$$

Body Fat EXAMPLE with Ridge Regression

- If you recall the original (un-ridged) version of this model with all the variables included, b_2 was negative, which didn't make much sense
- Now that we have applied ridge regression, the sign of b_2 has been corrected to what we would expect
- Although it is not shown here, the SSE of the model, which increases with c , has only increased from 0.1986 at $c = 0$ to 0.2182 at $c = 0.02$, with R^2 decreasing from 0.8014 to 0.7818, changes that are (subjectively) modest

Final Notes on Ridge Regression

- ① The ridge regression estimates can be obtained by the method of *penalized least squares*, combining the usual sum of squares term with a penalty for large regression coefficients

$$Q = \sum_{i=1}^n [Y_i^* - (\beta_1^* X_{i1}^* + \dots + \beta_{p-1}^* X_{i,p-1}^*)]^2 + c \left[\sum_{j=1}^n (\beta_j^*)^2 \right]$$

where the penalty is a biasing constant c times the sum of squares of the regression coefficient

Large absolute regression coefficients lead to large penalties, thus for $c > 0$, the 'best' coefficients will be smaller in magnitude than their OLS counterpart estimates

For this reason, ridge estimators are sometimes referred to as *shrinkage* estimators

Final Notes on Ridge Regression cont'd

- ② A major limitation of ridge regression is that ordinary inference procedures are not applicable and exact distributional properties are not known
- ③ Choice of the biasing constant c is subjective
- ④ Ridge regression can be used to select variables to eliminate from the model by analyzing the ridge trace; typically variables with unstable traces, traces that are very close to zero or traces that do not tend toward zero are dropped

Section 21

Remedial Measures for Influential Cases

- We have tools such as the hat matrix and deleted studentized residuals to help us identify outlying X and Y values respectively
- We also have tools—*DFFITS*, *DFBETAS* and Cook's Distance—that facilitate our identification of influential outliers
- We are particularly concerned with influential cases and their effect on the model, this is why we examine Robust Regression

LAR or LAD Robust Regression

- Least absolute residuals (LAR) or least absolute deviations (LAD) regression, also called *minimum L_1 -norm regression*, is one of the most widely used robust regression procedures
- It is insensitive to both outliers and model inadequacies
- The regression coefficients are estimated by minimizing the sum of absolute deviations of the Y observations from their means
- The criterion to be minimized is

$$L_1 = \sum |Y_i - (\beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1})|$$

- Given absolute rather than squared deviations are sought to be minimized, less emphasis is naturally given to outlying cases

LAR or LAD Robust Regression cont'd

- This can be solved by linear programming techniques
- The LAR regression model loses some of its properties, e.g., the sum of residuals will not necessarily sum to zero
- The solution for the estimated regression coefficients may not be unique (a property of the solution method of linear programming)

Additional Robust Regression Procedures

- Iteratively reweighted least squares (IRLS) robust regression uses a weighted least squares procedure (see Kutner 4e/5e, Chapter 11, §11.3) to dampen the influence of outlying observations, by using a weighting procedure that is dependent on how far outlying cases are
- Least median of squares (LMS) regression replaces the sum of squared deviation in ordinary least squares by the median of the squared deviations, which is a robust estimator of location
- The criterion for this procedure is to minimize the median squared deviation

$$\text{median}\{[Y_i - (\beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1})]^2\}$$

with respect to the regression coefficients

- Thus, this procedure leads to estimated regression coefficients b_0, b_1, \dots, b_{p-1} that minimize the median of the squared residuals

Section 22

Autocorrelation in Time Series Data

Autocorrelation in Time Series Data

- The basic regression models considered so far have assumed that the random error terms ε_i are either uncorrelated random variables or independent normal random variables
- If we encounter data sets that are time-series based instead of cross-sectional data, the assumption of uncorrelated or independent error terms is often violated
- Error terms correlated over time are said to be autocorrelated or serially correlated
- In many economics and time-series applications, positively autocorrelated error terms can often be attributed to the omission of one or multiple critical explanatory variables

E.g. Regressing annual sales of a product against yearly price of the product over 30 years; if population size has an important effect on sales, its omission from the model may lead to the error terms being positively autocorrelated

The Problems with Autocorrelation

The consequences of autocorrelated data include

- ① Estimated regression coefficients that, although unbiased, lose their property of having minimum variance
- ② MSE may seriously underestimate the variance of the error terms
- ③ $s\{b_k\}$ calculated according to ordinary least squares procedures may seriously underestimate the true standard deviation of the estimated regression coefficient
- ④ CI's and t and F distributions for test are no longer strictly applicable

An Intuitive Explanation

- Consider a SLR with time series data (indices)

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

- Let's assume that the error terms are positively autocorrelated as follows

$$\varepsilon_t = \varepsilon_{t-1} + u_t$$

where the u_t , called disturbances, are independent normal random variables

- We will assume the u_t have mean 0 and variance of 1

An Intuitive Explanation

t	(1)		(2)	
	u_t		$\epsilon_{t-1} + u_t = \epsilon_t$	
0	—			3.0
1	.5		3.0 + .5 =	3.5
2	-.7		3.5 - .7 =	2.8
3	.3		2.8 + .3 =	3.1
4	0		3.1 + 0 =	3.1
5	-2.3		3.1 - 2.3 =	.8
6	-1.9		.8 - 1.9 =	-1.1
7	.2		-1.1 + .2 =	-.9
8	-.3		-.9 - .3 =	-1.2
9	.2		-1.2 + .2 =	-1.0
10	-.1		-1.0 - .1 =	-1.1

- The issue we begin to observe is that adjacent error terms tend to be both of the same size and magnitude
- This will result in a very poorly fit regression function, which subsequently results in large variances of estimated regression coefficients when OLS methods are used
- Although one can use residual plots in an attempt to visually identify trends, quantitative statistical methods have also been developed that can detect such patterns

First-Order Autoregressive Error Model

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

where

- ρ is the autocorrelation parameter, $|\rho| < 1$
- u_t are independent $\sim \mathcal{N}(0, \sigma^2)$

n.b. This model is identical to the SLR model except for the structure of the error term, where each error term in the model consists of a fraction of the previous error term (when $\rho > 0$) plus a new disturbance term u_t

The Durbin-Watson Test for Autocorrelation

- Assumes a first-order autoregressive error model with fixed level of the predictor variable
- The test consists of determining whether or not the autocorrelation parameter ρ is zero, the implication being that if it is equal to zero, the autoregressive structure of the error terms collapses, and we return to a SLR model since the disturbance terms u_t are independent
- The hypothesis test for the Durbin-Watson test is

$$H_0 : \rho = 0$$

$$H_a : \rho > 0$$

The Durbin-Watson Test for Autocorrelation cont'd

- The Durbin-Watson test statistic is obtained by using ordinary least squares to fit the regression function, calculating the ordinary residuals

$$e_t = Y_t - \hat{Y}_t$$

and subsequently calculating the statistic

$$D = \frac{\sum_{i=2}^n (e_t - e_{t-1})^2}{\sum_{i=1}^n e_i^2}$$

The Durbin-Watson Test for Autocorrelation cont'd

- The Durbin-Watson test statistic is obtained by using ordinary least squares to fit the regression function, calculating the ordinary residuals

$$e_t = Y_t - \hat{Y}_t$$

and subsequently calculating the statistic

$$D = \frac{\sum_{i=2}^n (e_t - e_{t-1})^2}{\sum_{i=1}^n e_i^2}$$

- Exact critical values are difficult to obtain, but Durbin and Watson have been able to obtain lower and upper bounds d_L and d_U such that a value of D outside these bounds leads to a definite decision

The Durbin-Watson Test for Autocorrelation cont'd

The decision rule for the Durbin-Watson test is

- If $D > d_U$, do not reject H_0
- If $D < d_L$, reject H_0
- If $d_L \leq D \leq d_U$, the test is inconclusive

$$H_0 : \rho = 0$$

$$H_a : \rho > 0$$

The Durbin-Watson Test for Autocorrelation cont'd

The decision rule for the Durbin-Watson test is

- If $D > d_U$, do not reject H_0
- If $D < d_L$, reject H_0
- If $d_L \leq D \leq d_U$, the test is inconclusive

$$H_0 : \rho = 0$$

$$H_a : \rho > 0$$

What autocorrelation pattern(s) will generate small or large statistics?

The Durbin-Watson Test for Autocorrelation cont'd

The decision rule for the Durbin-Watson test is

- If $D > d_U$, do not reject H_0
- If $D < d_L$, reject H_0
- If $d_L \leq D \leq d_U$, the test is inconclusive

$$H_0 : \rho = 0$$

$$H_a : \rho > 0$$

What autocorrelation pattern(s) will generate small or large statistics?

```
# R Code  
> dwtest(V1~V2,data=myData)
```

Durbin-Watson test

```
data: V1 ~ V2  
DW = 0.7347, p-value = 0.0001748  
alternative hypothesis: true autocorrelation is greater than 0
```

Durbin-Watson Statistical Table (page 1 of 2)

Level of Significance $\alpha = .05$										
n	$p - 1 = 1$		$p - 1 = 2$		$p - 1 = 3$		$p - 1 = 4$		$p - 1 = 5$	
	d_l	d_u								
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Final Notes on the Durbin-Watson Test

- If a test for negative autocorrelation is required, the test statistic to be used is $4 - D$, where D is defined as above; the test is conducted in the same manner, i.e., if $4 - D$ falls below d_L we conclude $\rho < 0$, that negative autocorrelation exists
- A two-sided test for $H_0 : \rho = 0$ versus $H_0 : \rho \neq 0$ can be conducted by employing two one-sided tests with the Type I risk in the two sided test equal to 2α
- The Durbin-Watson test is effective in detecting first-order autoregressive errors, but fail in detecting second-order or n^{th} -order autoregressive errors

Acknowledgements

Some of the figures in this presentation are taken from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani