

STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 3: Multiple Linear Regression

Multiple Linear Regression

- generalizes the simple linear regression model by allowing more terms than just the intercept and slope
- a simple example

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\text{Var}(Y|X_1 = x_1, X_2 = x_2) = \sigma^2$$

- main question: will the adding of X_2 help $E(Y|X_1 = x_1)$?
- if yes, how much?

Multiple Regression - Con't

- United Nations data in Section 3.1 of text
 - Fertility: birth rate per 1000 females in 2000
 - PPgdp: per person gross domestic product in 2001
 - Purban: percentage of urban population
 - Locality: 193 regions

- Y : $\log(\text{Fertility})$ X_1 : $\log(\text{PPgdp})$, X_2 : Purban

$$R^2 = 46\% \quad \text{for} \quad \widehat{E(Y|X_1)} = 2.703 - 0.153x_1$$

$$R^2 = 35\% \quad \text{for} \quad \widehat{E(Y|X_2)} = 1.750 - 0.013x_2$$

- higher PPgdp or Purban leads to lower birth rate
- what if we consider both X_1 and X_2 in regression?

Multiple Linear Regression - Con't

- for $E(Y|X_1, X_2)$: $R^2 = 46\% + 35\%$ only if X_1 and X_2 are completely unrelated and measure different things.

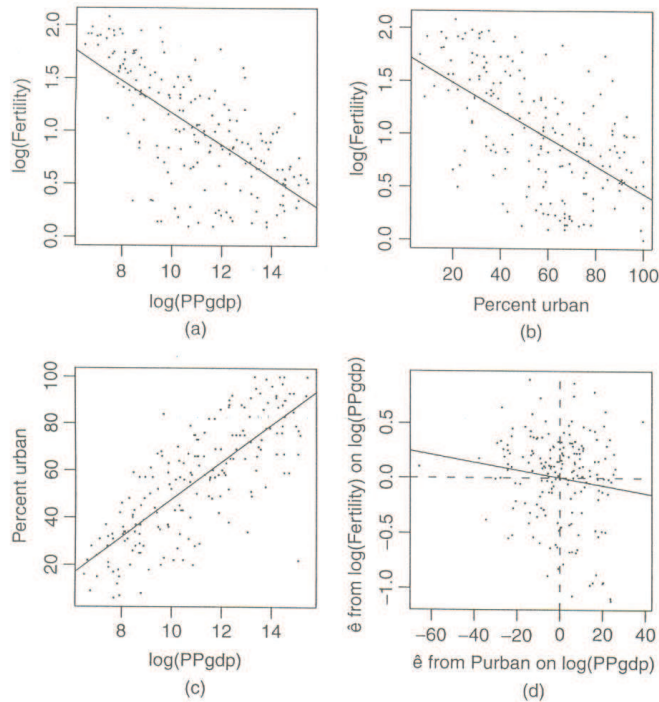
Q: will this be the case for UN data?

- more often situation: $46\% \leq R^2 \leq 46\% + 35\%$
- how much additional explanation was offered by X_2 ?
- let $\hat{e}_{Y|X_1}$ be the residuals of regressing Y on X_1 :
variability of Y not explained by X_1 , or variability of Y after the effect of X_1 is removed
- let $\hat{e}_{X_2|X_1}$ be the residuals of regressing X_2 on X_1 :
variability of X_2 not explained by X_1 , or variability of X_2 after the effect of X_1 is removed

Added-Variable Plot

● regression and residual plots: (b) v.s. (d)

$\hat{\beta}_2 = -0.013$ ignoring X_1 , $\hat{\beta}_2 = -0.004$ adjusting for X_1



Multiple Linear Regression (MLR)

- in general, multiple linear model:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\text{Var}(Y|X) = \sigma^2$$

- a linear function of the parameters $\{\beta_0, \dots, \beta_p\}$
- $p = 1 \Rightarrow$ simple linear regression
- when $p = 2$, fit a 2D plane in a 3D space

Regression Surface for $p = 2$

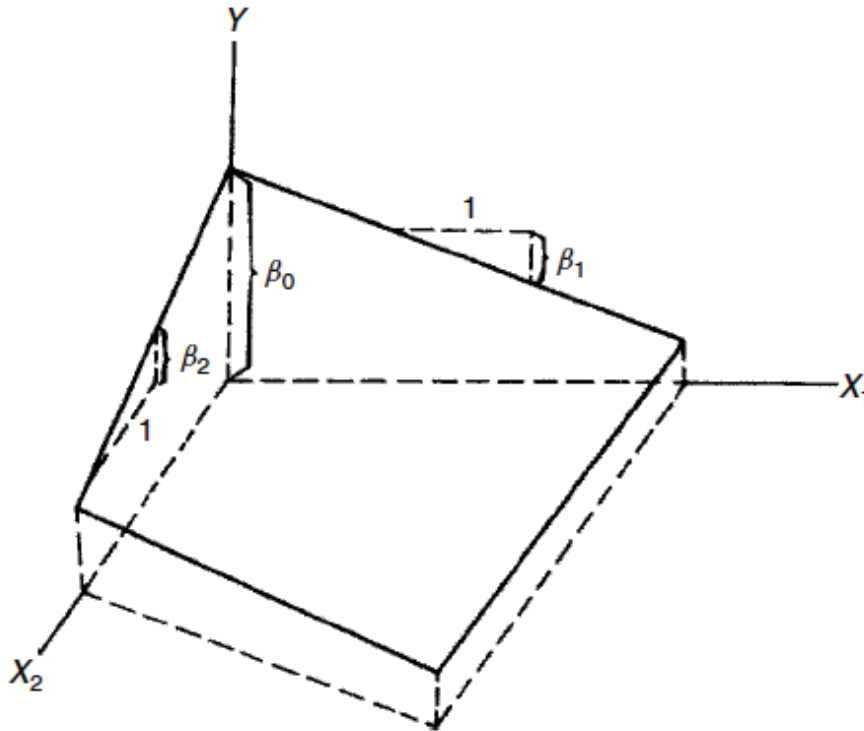


FIG. 3.2 A linear regression surface with $p = 2$ predictors.

Terms and Predictors

- the textbook distinguishes “predictors” and “terms”, only for convenience and to avoid confusion
- predictors: the "original data that you collect"
- e.g., height, weight, color, smoking or not
- terms: created from predictors, the X -variable in our multiple regression models
- e.g., height^2 , $\log(\text{weight})$, $\text{height} \times \text{weight}$, color, ...
- an important question in multiple linear regression:
the selection of a "good" set of terms

Matrix Notation for MLR

● Recall: $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

$$\text{Var}(Y|X) = \sigma^2$$

● observed values: $(x_{11}, x_{12}, \cdots, x_{1p}, y_1)$

$$(x_{21}, x_{22}, \cdots, x_{2p}, y_2)$$

$$\vdots$$
$$\vdots$$

$$(x_{n1}, x_{n2}, \cdots, x_{np}, y_n)$$

● $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Matrix Notation for MLR - Con't

- the i th row of \mathbf{X} will be denoted as \mathbf{x}_i'

- $$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

- there are $(p + 1)$ parameters, including the intercept β_0

Matrix Notation for MLR - Con't

- multiple linear regression in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- the i th row is $y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i$
 $\Rightarrow y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$
- about the vector of errors \mathbf{e} :

$$\mathbf{E}(\mathbf{e}) = \mathbf{0}, \quad \text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

- if we add normality assumption:

$$\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

OLS for Multiple Linear Regression

- $RSS(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$

- if $(\mathbf{X}'\mathbf{X})^{-1}$ exists, $RSS(\beta)$ is minimized by

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

- $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{Y}$: similar to SXX , SXY

- Residuals: $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$

- $RSS = \hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$

- $\sigma^2 = \text{Var}(Y|X)$ is estimated with $\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}$

- with the **normality** assumption we have

$$(n - (p + 1))\hat{\sigma}^2 / \sigma^2 \sim \chi^2(n - (p + 1))$$

OLS using Matrices

- model: $E(Y|X = \mathbf{x}) = \beta' \mathbf{x}$ and $\text{Var}(Y|X = \mathbf{x}) = \sigma^2$

- OLS estimates $\hat{\beta}$ of β minimize

$$\begin{aligned} RSS(\beta) &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y} + \beta'(\mathbf{X}'\mathbf{X})\beta - 2\mathbf{Y}'\mathbf{X}\beta \end{aligned}$$

- some matrix differentiation results to proceed:

$$\frac{\partial c\beta}{\partial \beta} = c' \quad \text{and} \quad \frac{\partial \beta' V \beta}{\partial \beta} = (V + V')\beta$$

- by setting the derivative of $RSS(\beta)$ to zero, we have

normal equation: $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$

- thus the OLS estimate is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

Properties of OLS Estimates

- assume $E(e) = 0$ and $\text{Var}(e) = \sigma^2 \mathbf{I}_n$, $\hat{\beta}$ is unbiased

$$\begin{aligned} E(\hat{\beta}|\mathbf{X}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \beta \end{aligned}$$

- for variance, we will need $\text{Var}(\mathbf{B}'\mathbf{Z}) = \mathbf{B}'\text{Var}(\mathbf{Z})\mathbf{B}$

$$\begin{aligned} \text{Var}(\hat{\beta}|\mathbf{X}) &= \text{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{Y}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\sigma^2\mathbf{I}_n]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Residual Sum of Squares

- $RSS = RSS(\hat{\beta}) = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$
 $= \mathbf{Y}'\mathbf{Y} + \hat{\beta}'(\mathbf{X}'\mathbf{X})\hat{\beta} - 2\mathbf{Y}'\mathbf{X}\hat{\beta}$
- $\hat{\beta}'(\mathbf{X}'\mathbf{X})\hat{\beta} = \hat{\beta}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \hat{\beta}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}\hat{\beta}$
- $RSS = \mathbf{Y}'\mathbf{Y} - \hat{\beta}\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}}, \text{ with } \hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$
- why does this satisfy the pythagorean property?
- $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\beta}$
 $= \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{0}$
- $\hat{\mathbf{e}}$ is orthogonal to column space of \mathbf{X} , denoted by $S(\mathbf{X})$
- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ is the projection of \mathbf{Y} onto $S(\mathbf{X}) \implies \hat{\mathbf{e}} \perp \hat{\mathbf{Y}}$

Geometric Interpretation of OLS

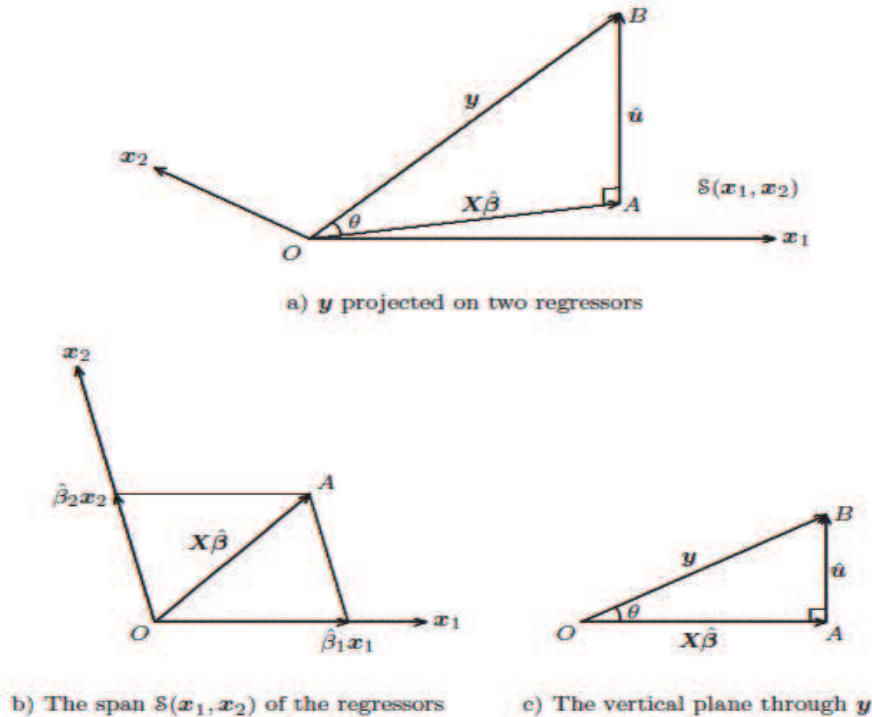


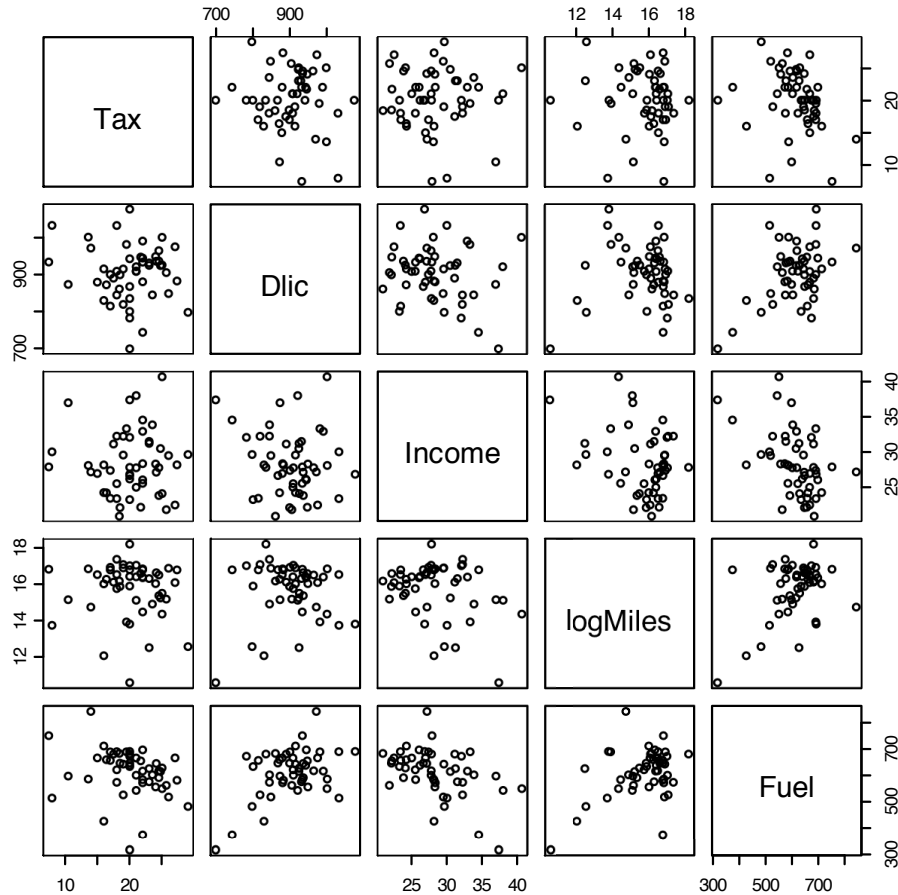
Figure 2.11 Linear regression in three dimensions

● express the simple linear regression in matrix form

Fuel Consumption Data

- goal: effect of gasoline tax on fuel consumption over U.S. states, including Washington D.C. (how many?)
- Y : Fuel, fuel consumption averaged over state population
- 4 terms:
 1. Tax: tax on gasoline in each state
 2. Dlic: number of driver licenses averaged over state population
 3. Income: personal income in each state
 4. logMiles: total length of highway in each state, in log miles (base two)

Fuel Consumption Data - Con't



Fuel Consumption Data - Con't

TABLE 3.2 Sample Correlations for the Fuel Data

Sample Correlations

	Tax	Dlic	Income	logMiles	Fuel
Tax	1.0000	-0.0858	-0.0107	-0.0437	-0.2594
Dlic	-0.0858	1.0000	-0.1760	0.0306	0.4685
Income	-0.0107	-0.1760	1.0000	-0.2959	-0.4644
logMiles	-0.0437	0.0306	-0.2959	1.0000	0.4220
Fuel	-0.2594	0.4685	-0.4644	0.4220	1.0000

Fuel Consumption Data - Con't

- the linear regression model to be fitted

$$E(\text{Fuel}|X) = \beta_0 + \beta_1 \text{Tax} + \beta_2 \text{Dlic} + \beta_3 \text{Income} + \beta_4 \log(\text{Miles})$$

TABLE 3.3 Multiple Linear Regression for the Fuel Data

Coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	154.1928	194.9062	0.791	0.432938
Tax	-4.2280	2.0301	-2.083	0.042873
Dlic	0.4719	0.1285	3.672	0.000626
Income	-6.1353	2.1936	-2.797	0.007508
logMiles	18.5453	6.4722	2.865	0.006259

Residual standard error: 64.89 on 46 degrees of freedom

Multiple R-Squared: 0.5105

F-statistic: 11.99 on 4 and 46 DF, p-value: 9.33e-07

ANOVA for Multiple Linear Regression

● Comparing

$$E(Y|X = \mathbf{x}) = \beta_0 + \sum_{j=1}^p \beta_j x_j \text{ with } E(Y|X = \mathbf{x}) = \beta_0$$

● similar to simple linear regression

The Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	p	SS_{reg}	SS_{reg}/p	$MS_{reg}/\hat{\sigma}^2$	
Residual	$n - (p + 1)$	RSS	$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}$		
Total	$n - 1$	$SY Y$			

ANOVA for Fuel Consumption

- The test is:

$$NH: E(Y|X = \mathbf{x}) = \beta_0 \quad \text{vs} \quad AH: E(Y|X = \mathbf{x}) = \beta' \mathbf{x}$$

Fuel Consumption Data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	4	201994	50499	11.992	9.33e-07
Residuals	46	193700	4211		
Total	50	395694			

- Coefficient of Determination

$$R^2 = \frac{SS_{reg}}{SYY} = \frac{SYY - RSS}{SYY} = 1 - \frac{RSS}{SYY}$$

- $R^2 = \frac{201994}{395694} = 0.5105$ for fuel consumption

Hypothesis Test for One Term

- Fuel Consumption: what will happen if we delete Tax from the model?
- NH: $\beta_1 = 0$, $\beta_0, \beta_2, \beta_3, \beta_4$ arbitrary
AH: $\beta_1 \neq 0$, $\beta_0, \beta_2, \beta_3, \beta_4$ arbitrary
- fit a model with all terms: Model B (Bigger)
- fit a model with all terms but tax - Model S (smaller)
- Let RSS_B be the RSS from Model B
- let RSS_S be the RSS from Model S
- which is bigger? $RSS_S \geq RSS_B$ (why?)

Hypothesis Test for One Term

- $RSS_S - RSS_B$ gives the "contribution" of Tax after adjusting all other terms

	df	SS	MS	F	Pr(>F)
Excluding Tax	47	211964			
Including Tax	46	193700	4211		
Difference	1	18264	18264	4.34	0.043
			(SS/df)	(MS/ $\hat{\sigma}^2$)	

- so for this problem, Tax is statistically significant
- this F -test for one term is the same as t -test ($t^2 = F$)

Sequential Analysis of Variance Tables

- in Model B with Tax, Dlic, Income, logMiles, $SS_{reg} = 201994$ tells how much variation is explained
- for the smaller model (without Tax), $SS_{reg} = 201994$ from Model B is decomposed into two parts:
 - 1: "Dlic, Income, logMiles", $SS_{others} = 183730$
 - 2: "Tax, given Dlic, Income, logMiles" $SS_{tax} = 18264$
- we could continue this decomposition
 - e.g., 1: "logMiles"
 - 2: "Income, given logMiles"
 - 3: "Dlic, given Income, logMiles"
 - 4: "Tax, given Dlic, Income, logMiles"

Sequential Analysis of Variance Tables - Con't

- for the fuel example, 4 parts
- in general, the order of decomposition matters

a) First analysis

	Df	Sum Sq	Mean Sq
Dlic	1	86854	86854
Tax	1	19159	19159
Income	1	61408	61408
logMiles	1	34573	34573
Model B	4	201994	50499
Residuals	46	193700	4211

(b) Second analysis

	Df	Sum Sq	Mean Sq
logMiles	1	70478	70478
Income	1	49996	49996
Dlic	1	63256	63256
Tax	1	18264	18264
Model B	4	201994	50499
Residuals	46	193700	4211

(terms entering the model from top to bottom)

Predictions and Fitted Values

- similar to simple linear regression
- prediction: given a new \mathbf{x}_* , predict y_* with
- $\hat{y}_* = \mathbf{x}_*' \hat{\boldsymbol{\beta}}$
- $\text{sepred}(\hat{y}_* | \mathbf{x}_*) = \hat{\sigma} \sqrt{1 + \mathbf{x}_*' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_*}$
- fitted value: given a value \mathbf{x} , want to estimate the mean function at \mathbf{x}
- $\hat{E}(Y | X = \mathbf{x}) = \hat{y} = \mathbf{x}' \hat{\boldsymbol{\beta}}$
- $\text{sefit}(\hat{y} | \mathbf{x}) = \hat{\sigma} \sqrt{\mathbf{x}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}}$
- $\text{sepred}(\hat{y}_* | \mathbf{x}) = \sqrt{\hat{\sigma}^2 + \text{sefit}(\hat{y}_* | \mathbf{x})^2}$