Chapter 10
Multivariate Regression

Halima Bensmail

CS502

Monday 4-7pm

LAS Hall C

Regression Using Many Independent Variables

- Identifying and Summarizing Data
- Linear Regression Model
- Basic Checks of the Model
- Added Variable Plots
- Some Special Independent Variables
- Is a Group of Independent Variables Important?
- Matrix Notation

Summarizing the Data

- The data consists of:
  - $(X_1, Y_1)=(x_{11}, x_{12}, \ldots , x_{1k}, y_1)$
  - $(X_2, Y_2)=(x_{21}, x_{22}, \ldots , x_{2k}, y_2)$
  - . . .
  - . . .
  - $(X_n,Y_n)=(x_{n1}, x_{n2}, \ldots , x_{nk}, y_n)$
- Begin the analysis of the data by examining each variable in isolation of the others.

The next step

- is to measure the effect of each x on y.
- Scatter plots
- Correlations
- Regression Lines
- A scatterplot matrix
- Method of Least Squares
  - $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k$.

Inference for Multiple Regression

Multiple regression model (matrix notation)

$$Y = X\,b + \varepsilon$$

where

Y	n dimensional vector
X	$n \times (1+p)$ dimensional matrix
b	$1+p$ dimensional vector
ε	n dimensional vector

Thus the model can be written as

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Least squares approach: Minimize

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

The Linear Regression Model

- The model is
- response = nonrandom regression plane + random error,
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i, \quad i = 1, \ldots, n.$
- The expected response is a linear combination of the explanatory variables, that is,
  - $E\, y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.
- The observed response is the expected response plus a random error term.
- The quantities β_0 , ..., β_k are unknown, yet nonrandom, parameters. These quantities determine a plane in k+1 dimensions.

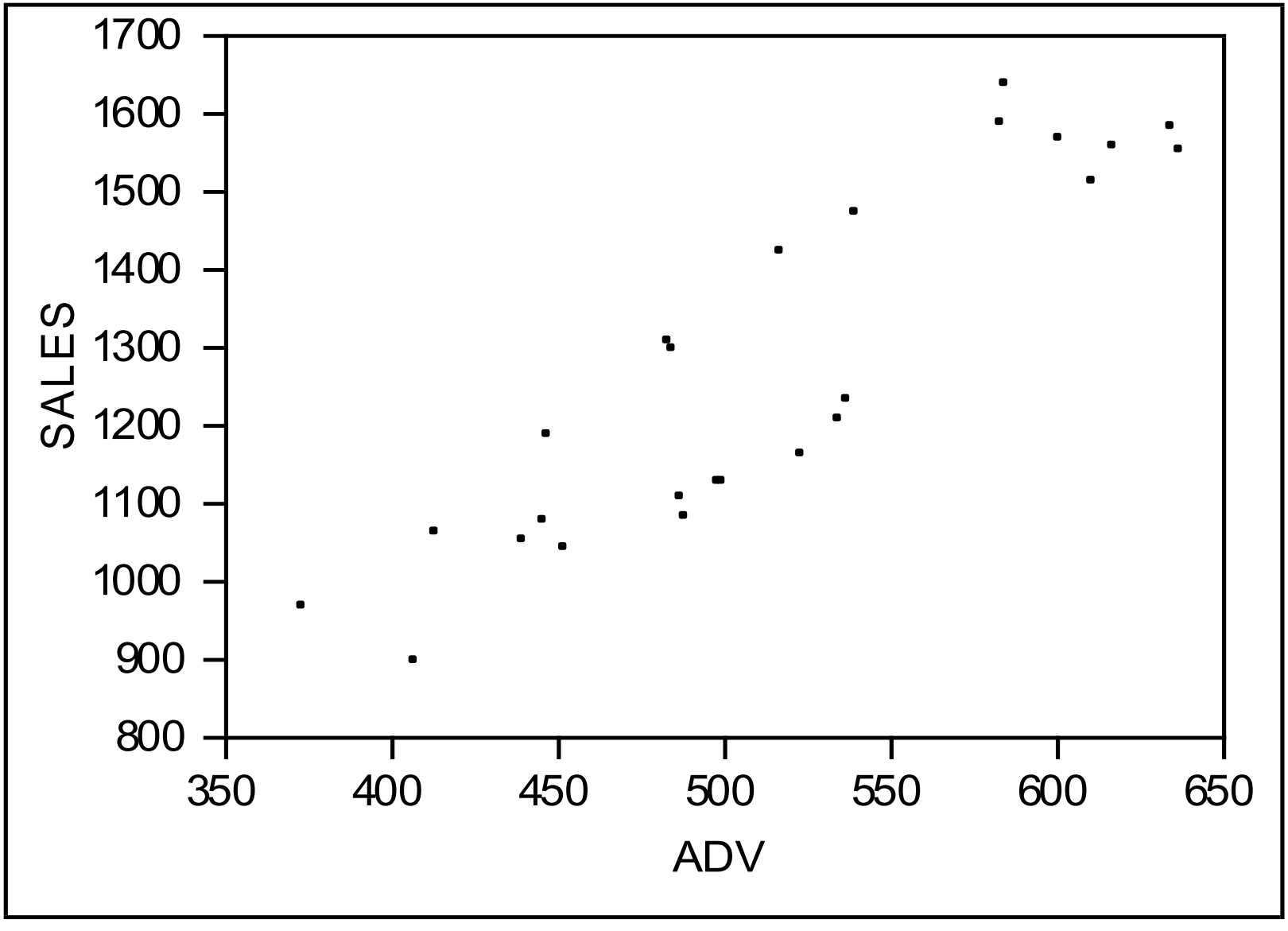
Random Errors

- The quantity **e** represents the random deviation, or error, of an individual response from the plane.
- The random errors $\{e_1, e_2, ., e_n\}$ are assumed to be randomly selected from an unknown population of errors.
- We assume that the expected value of each error is 0 so that the expected response is given by the regression plane, that is,
  - $E\, y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$.
- The regression plane is nonrandom. Thus,
- $Var\ (y) = Var\ (e) = \sigma^2$.
- If the ***jth*** variable is continuous, we interpret β_j as the **expected change** in y per unit change in x_j assuming all the other variables are held fixed.

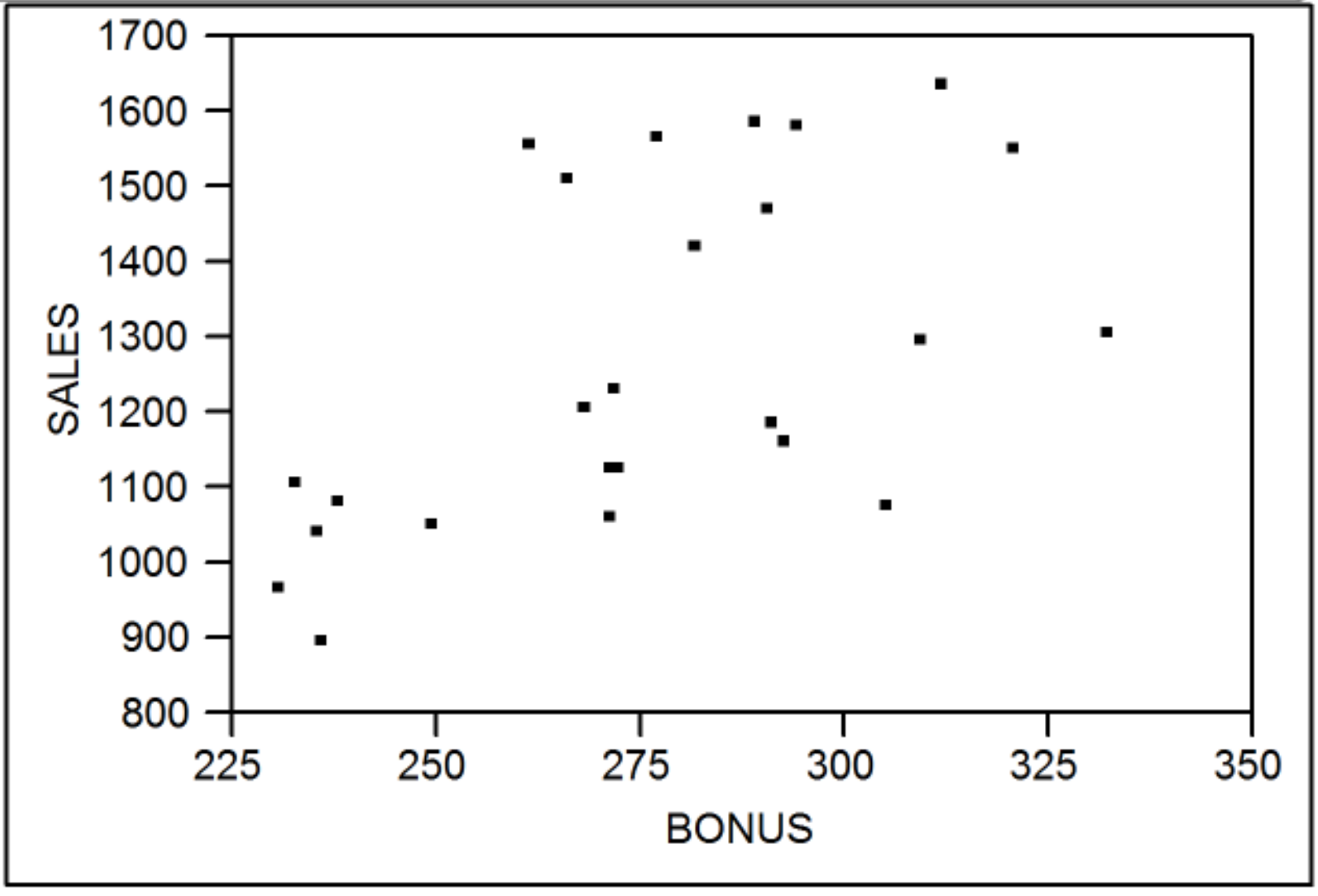
Meddicorp Example

- Data on Meddicorp company that sells medical supplies to hospitals.
- Y= Meddicorp's sales (in thousand of dollars)
- X_1: Amount meddicorp spent on advertising
- X_2: Total amount of bonuses paid (in thousand)

Bivariate Fit of SALES By ADV



Bivariate Fit of SALES By BONUS



Summary of Fit

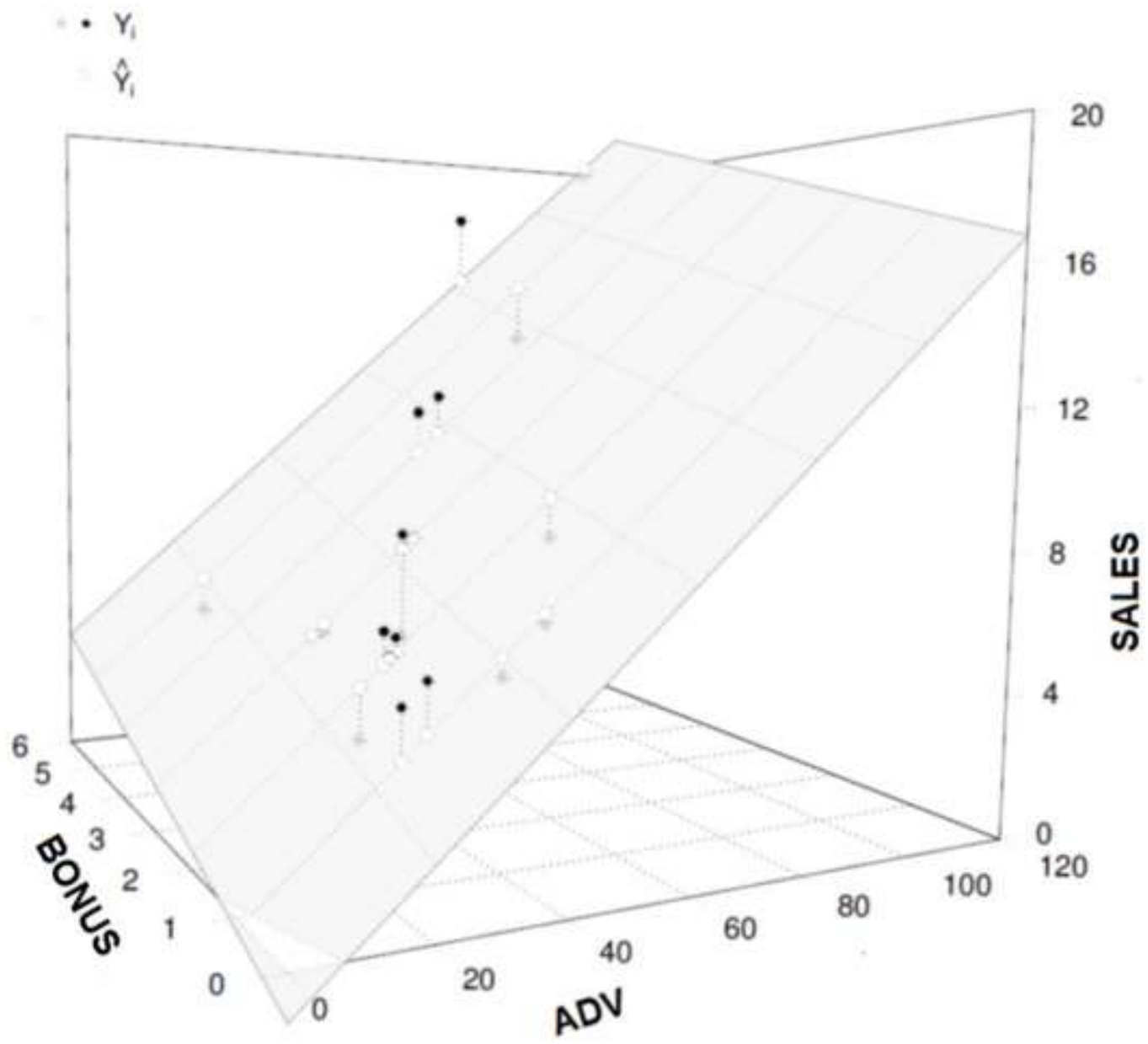
RSquare	0.854953
RSquare Adj	0.841767
Root Mean Square Error	90.74432
Mean of Response	1269.02
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1067814.1	533907	64.8376
Error	22	181159.7	8235	Prob > F
C. Total	24	1248973.7		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-516.497	189.8665	-2.72	0.0125
ADV	2.4732589	0.275278	8.98	<.0001
BONUS	1.8561942	0.71559	2.59	0.0166

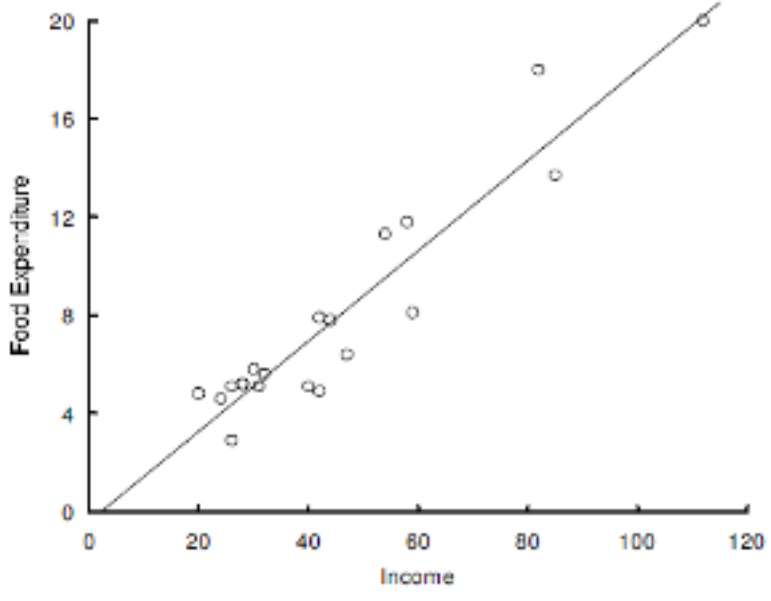

Y_i
$\hat{Y}_i$
20
16
12
8
4
0
SALES
6
5
4
3
2
1
0
BONUS
0
20
40
60
80
100
120
ADV

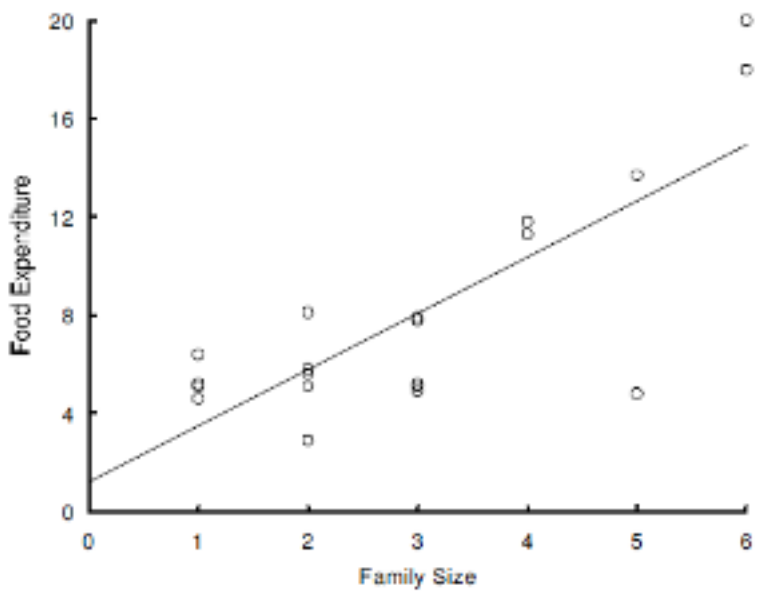
Multiple regression
Example: Food expenditure and family income

Expenditure ($)	Income ($)	Exp. ($)	Inc. ($)
2400	41200	1450	37500
2650	50100	2020	36900
2350	52000	3750	48200
4950	66000	1675	34400
3100	44500	2400	29900
2500	37700	2550	44750
5106	73500	3880	60550
3100	37500	3330	52000
2900	56700	4050	67700
1750	35600	1150	20600

Example: Food expenditure and family income

Data:
- Sample of 20 households
- Food expenditure (response variable)
- Family income and family size





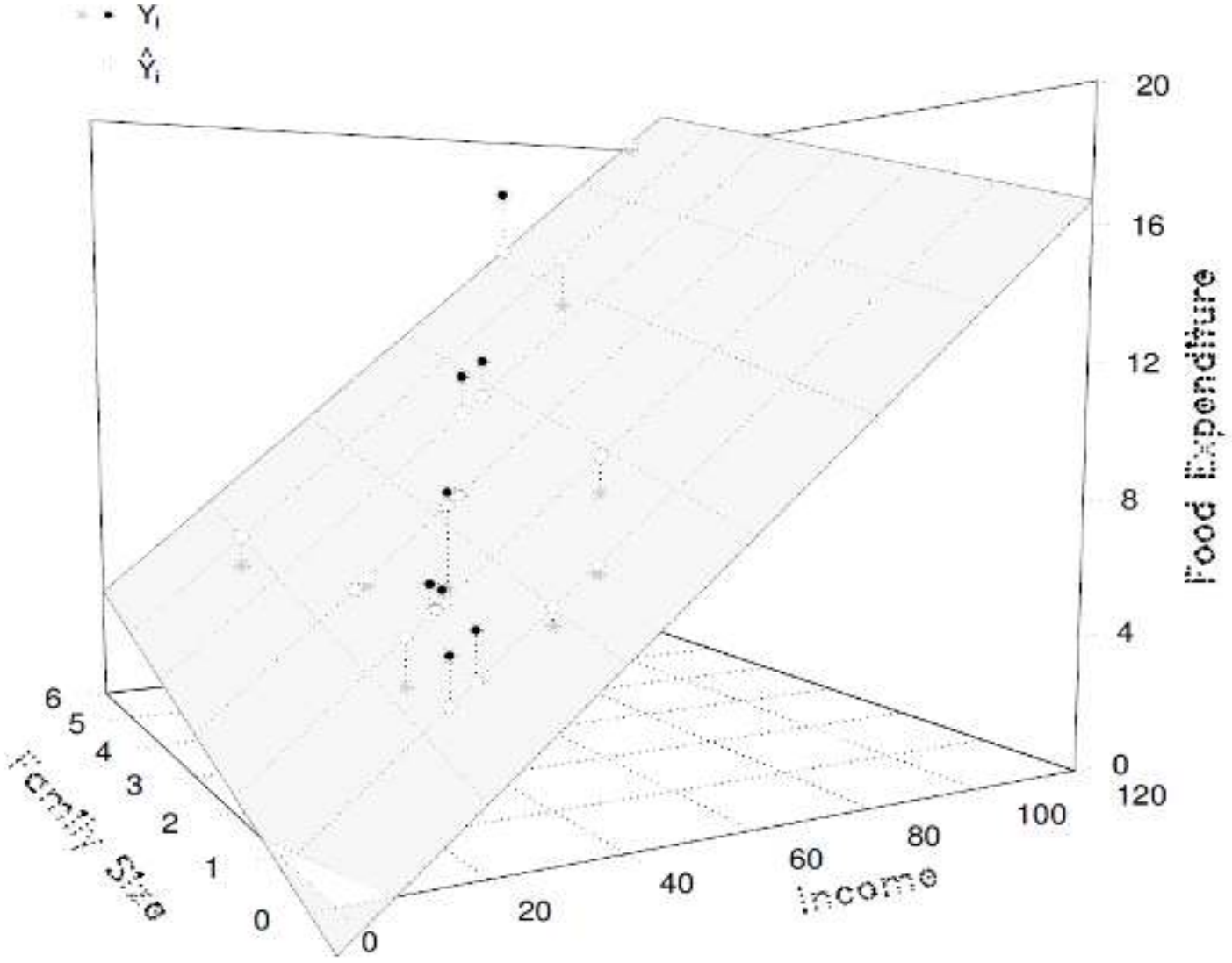
```
. regress food income
------------------------------------------------------------------------------
        food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .1841099   .0149345    12.33   0.000     .1527336    .2154862
       _cons |  -.4119994   .7637666    -0.54   0.596    -2.016613    1.192615
------------------------------------------------------------------------------

. regress food number
------------------------------------------------------------------------------
        food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      number |   2.287334   .4224493     5.41   0.000     1.399801    3.174867
       _cons |   1.217365   1.410627     0.86   0.399    -1.746252    4.180981
------------------------------------------------------------------------------
```

Example: Food expenditure and family income

Data: $(\text{Food}_i, \text{Income}_i, \text{Number}_i)$, $i = 1, \ldots, 20$

Fitted regression model:

$$\widehat{\text{Food}} = \hat{b}_0 + \hat{b}_1 \, \text{Income} + \hat{b}_2 \, \text{Number}$$



Inference for Multiple Regression

Example: Food expenditure and family income

Interpretation of regression coefficients

```
. quietly regress food income
. predict e_food1, residuals
. quietly regress number income
. predict e_num, residuals
. regress e_food1 e_num
```

e_food1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e_num	.7931055	.2375541	3.34	0.004	.2940229	1.292188

```
. quietly regress food number
. predict e_food2, residuals
. quietly regress income number
. predict e_inc, residuals
. regress e_food2 e_inc
```

e_food2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e_inc	.1482117	.0159172	9.31	0.000	.114771	.1816525

The Variability

- Interpret the Total Sum of Squares, to be the total variation in the data set.
- Total SS = $\Sigma\ (y_i - \bar{y}\ )^2$.
- Now compute the fitted value.
- $\hat{y}_i = b_0 + b_1\, x_{i1} + b_2\, x_{i2} + \ldots + b_k\, x_{ik}$.
- We now have two "estimates" of y_i , $\hat{y}_i$ and $\bar{y}$.

$$y_i - \bar{y} = y_i - \hat{y}_i \quad + \quad \hat{y}_i - \bar{y}$$

- "the deviation without knowledge of the regression plane"
- = "the deviation with knowledge of the regression plane"
- + "the deviation explained by the regression plane."
- As before,

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 \quad + \quad \sum_i (\hat{y}_i - \bar{y})^2$$

- Total SS = Error SS + Regression SS

Residuals

- The residual,$\hat{e}_i$ should be close to the true error, e_i.
- $\hat{e}_i = y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik})$
- is close to
- $y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} .) = e_i$.
- With the residuals, we define the estimator of σ^2 to be
- $s^2 = \Sigma \hat{e}_i^2 / (n-(k+1)) = SSE/ (n-(k+1))$
- Again, there is a dependency among residuals. For example, the average of residuals is 0. This reasoning leads us to divide by n-(k+1) in lieu of n-1.
- We may also express s^2 in terms of the sum of squares quantities in the ANOVA (analysis of variance) table. That is,
- $s^2 = (n-(k+1))^{-1} SSE = MSE$

The ANOVA
Analysis of Variance

- This leads us to the ANOVA table:

Source	SS	df	MS
Model	Model SS	k	Model MS
Error	Error SS	n-(k+1)	Error MS
Total	Total SS	n-1	

  - The ANOVA table is merely a bookkeeping device used to keep track of the sources of variability.
  - Recall, R^2, is the proportion of variability explained by the regression plane. **R^2 = SSR / SST.** $R^2 = \rho_{XY}^2$.
- A coefficient of determination adjusted for degrees of freedom is
- $R_a^2 = 1 - (SSE/(n-(k+1)) / (SST/(n-1)) = 1 - s^2 / s_y^2$.
  - Algebra - whenever an explanatory variable is added to the model, R^2 never decreases. (not true for R_a^2.)
  - As the model fit improves, as measured through s^2 , the adjusted R^2 becomes larger and vice versa.

Is the Model Adequate?

- The nonrandom portion of our model is
- $E\ y = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + ... + \beta_k\ x_k$.
- We translate the question, "Is the model adequate?" into
- $H_0: \beta_1 = ... = \beta_k = 0$.
  - Thus, we can use the tests of hypothesis machinery to aid our decision making process.
  - The alternative hypothesis is that **at least one of the slope** parameters **does not equal to zero.**
  - The **larger the ratio** of regression sum of squares to the error sum of squares, the better is the model fit. If we standardize this ratio by the respective degrees of freedom, we get the so-called "F-ratio."
  - **F-ratio = (Regression SS / k) / (Error SS / (n-(k+1))**
- **= Regression MS / Error MS = Regression MS / s^2.**

  - Both **R^2** and the **F-ratio** are useful for summarizing model adequacy. The sampling distribution of the F-ratio is known, at least under the null hypothesis.

F-Distribution

- Both the statistic and the theoretical curve are named for R. A. Fisher.

- Like the normal and the t-distribution, the F-distribution is a continuous idealized histogram.
- The F-distribution is indexed by two degree of freedom parameters: one for the numerator, $\mathbf{df_1}$, and one for the denominator, $\mathbf{df_2}$.

- Declare H_0 to be invalid if **F-ratio exceeds an F-value**. The F-value is computed using a significance level with $\mathbf{df_1 = k}$ and $\mathbf{df_2 = n\text{-}k\text{-}1}$ degrees of freedom.

Is an Independent Variable Important?

- "Is x_j important?" - $H_0 : \beta_j = 0$ valid?
- We respond to this question by looking at the t-ratio
  - $test(b_j) = b_j / SE(b_j)$

- 1. Declare H_0 invalid in favor of H_a: β_j NE 0 if:
  - $|test(b_j)|$ exceeds a t-value
- with **n-(k+1)** degrees of freedom. Use a significance level divided by 2.
- 2. Declare H_0 invalid in favor of $H_a : \beta_j > 0$ if:
- $test(b_j)$ exceeds a t-value with **n-(k+1)** degrees of freedom.

The t-ratio [Data: Rent]

- Alternatively, one can construct p-values.
- A useful convention:
- Rent/sft = 1.14 - .112 Miles - .000281 Footage.
- (.064) (.0183) (.0000775)
- The parameter estimates are: $b_0 = 1.14$, $b_1 = -.112$ and $b_2 = .000281$.
- The corresponding standard errors are: $se(b_0)=.064$, $se(b_1)=.0183$ and $se(b_2)=.0000775$.
- **For regression with 1 explanatory variable, F-ratio = (t-ratio)^2 and F-value = (t-value)^2** .
- The F-test has the advantage that it works for **more than one** explanatory variable.
- The t-test has the advantage that one can consider **1-sided alternatives.**

Meddicorp example

- Sales = -516.49 +2.47 ADV +1.85 BONUS.
- (189.86) (.2175) (.716)
- The parameter estimates are: b_0 =189.86, b_1 = 2.47 and b_2 = 1.85.
- The corresponding standard errors are: se(b_0)=189.86, se(b_1)=.2175 and se(b_2)=.716.
- R^2=85% and R^2_a =84% are good so we have a good fit
- F-test=64.83
- P_{value}=P($F_{(2,22)}$ > 64.83) =0.0001 which is smaller than 5%
- So the model is adequate

Relationships between Correlation and Regression

- 1. $R^2 = r^2_{y,\hat{y}}$

- Because it can be interpreted as the correlation between the response and the fitted values, sometimes R (the positive root square of R^2) is referred to as the **multiple correlation coefficient.**

- 2. Both F-ratio and R^2 are measures of model fit. Because of the following algebraic relationship, we know that as R^2 increases, so does the F-ratio.
- $$F\text{-}ratio = ((1/ R^2 - 1))^{-1} (n-(k+1))/k.$$
  $$= R^2 /(1- R^2) \ . \ (n-(k+1))/k$$

Visualizing Multivariate Regression Data

- The **Added Variable plot** is a plot of the response versus an explanatory variable after "controlling for" the effects of additional explanatory variables. It is also called: Partial regression plot.
- 1. Regress y on $\{x_2, ..., x_k\}$ to get residuals $\hat{e}_1$.
- 2. Regress x_1 on $\{x_2, ..., x_k\}$ to get residuals $\hat{e}_2$
- 3. A plot of $\hat{e}_1$ versus $\hat{e}_2$.
- Summarize this plot via a correlation coefficient. Denote this correlation by $r(y, x_1 \mid x_2, ..., x_k)$.
- Idea: The residual
- $\hat{e} = y - (b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k)$ is the response *controlled for* values of the explanatory variables.

Partial Correlations and t-ratios

- Quicker way: run a regression of y on x_1 , x_2 , ..., x_k.
- Denote the t-ratio for β_1 by $t(b_1)$. We have
- $$r(y, x_1 \mid x_2, \ldots, x_k) = \frac{t(b_1)}{\sqrt{t(b_1)^2 + n - (k+1)}}$$

- Larger t-ratios can be interpreted as having a higher correlation between the dependent variable and the predictor, after controlling for the effects of other predictors.

Partial correlation Example(fridge)

- When we add a new variable to the explanatory variable, to summarize the effect of this variable to the dependent variable given the other predictors, we calculate the partial correlation coefficient given by the previous formula.

- **Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-810.3293	396.319	-2.04	0.0489
R_CU_FT	59.43786	26.98895	2.20	0.0347
F_CU_FT	104.37307	16.62632	6.28	<.0001
SHELVES	39.453118	14.51731	2.72	0.0104

- R^2=62% is still small, can we do better if we add the Energy cost variable?

Partial correlation

- R_CU_FT, F_CU_FT and SHELVES are used to predict the Price of a fridge.
- BUT **……..> We want to add E-cost?**
-

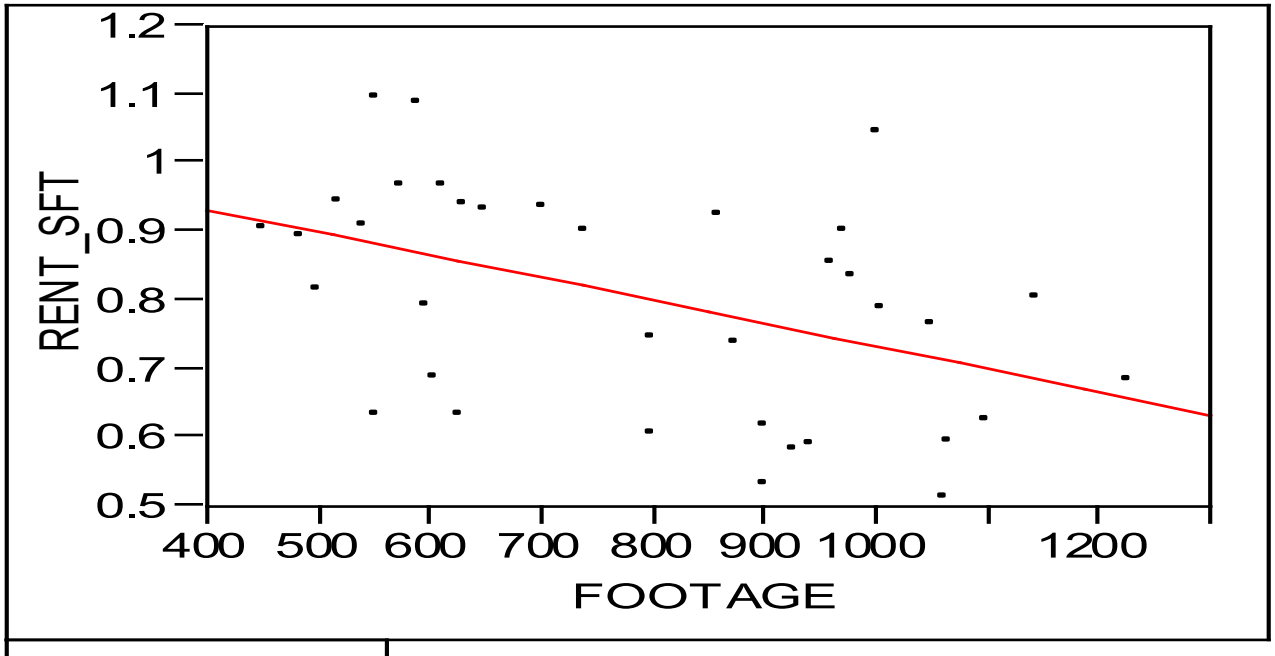
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-919.2204	366.3991	-2.51	0.0174
R_CU_FT	82.729809	26.29214	3.15	0.0036
F_CU_FT	175.26836	30.68579	5.71	<.0001
SHELVES	41.588347	13.36161	3.11	0.0039
E_COST	-8.169977	3.06696	-2.66	0.0120

- **Corr(Price, E-cost| R-CU-FT, F-Cu-FT, Shelves)** is interpreted to be the correlation between price and E-Cost in the presence of the other variables and is equal to:
- $-2.66/(\sqrt{[(-2.66)^2+37-(4+1)]})=-2.66/6.25=-0.42.$

Indicator/Dummy Variables and Interaction

Bivariate Fit of RENT_SFT By FOOTAGE



Linear Fit

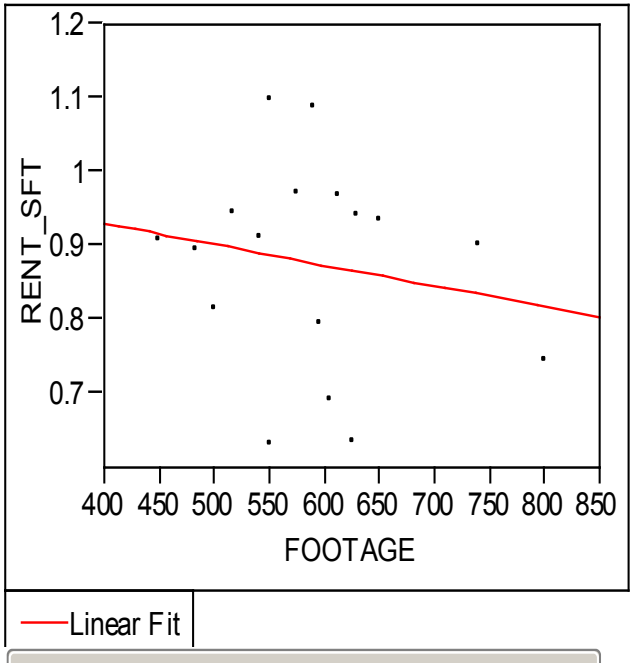
RENT_SFT = 1.0679446 - 0.0003322 FOOTAGE

Summary of Fit

RSquare	0.207491
RSquare Adj	0.184182
Root Mean Square Error	0.145654
Mean of Response	0.805157
Observations (or Sum Wgts)	36

TWOBED=0 One bedroom

Bivariate Fit of RENT_SFT By FOOTAGE

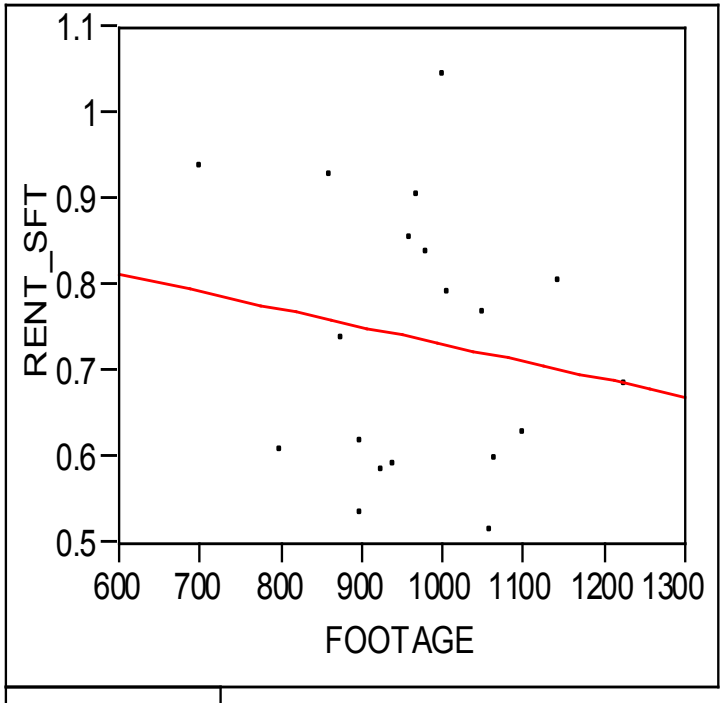


Linear Fit

RENT_SFT = 1.0451369 - 0.0002825 FOOTAGE

TWOBED=1 Two Bedrooms

Bivariate Fit of RENT_SFT By FOOTAGE



Linear Fit

RENT_SFT = 0.9353345 - 0.0002017 FOOTAGE

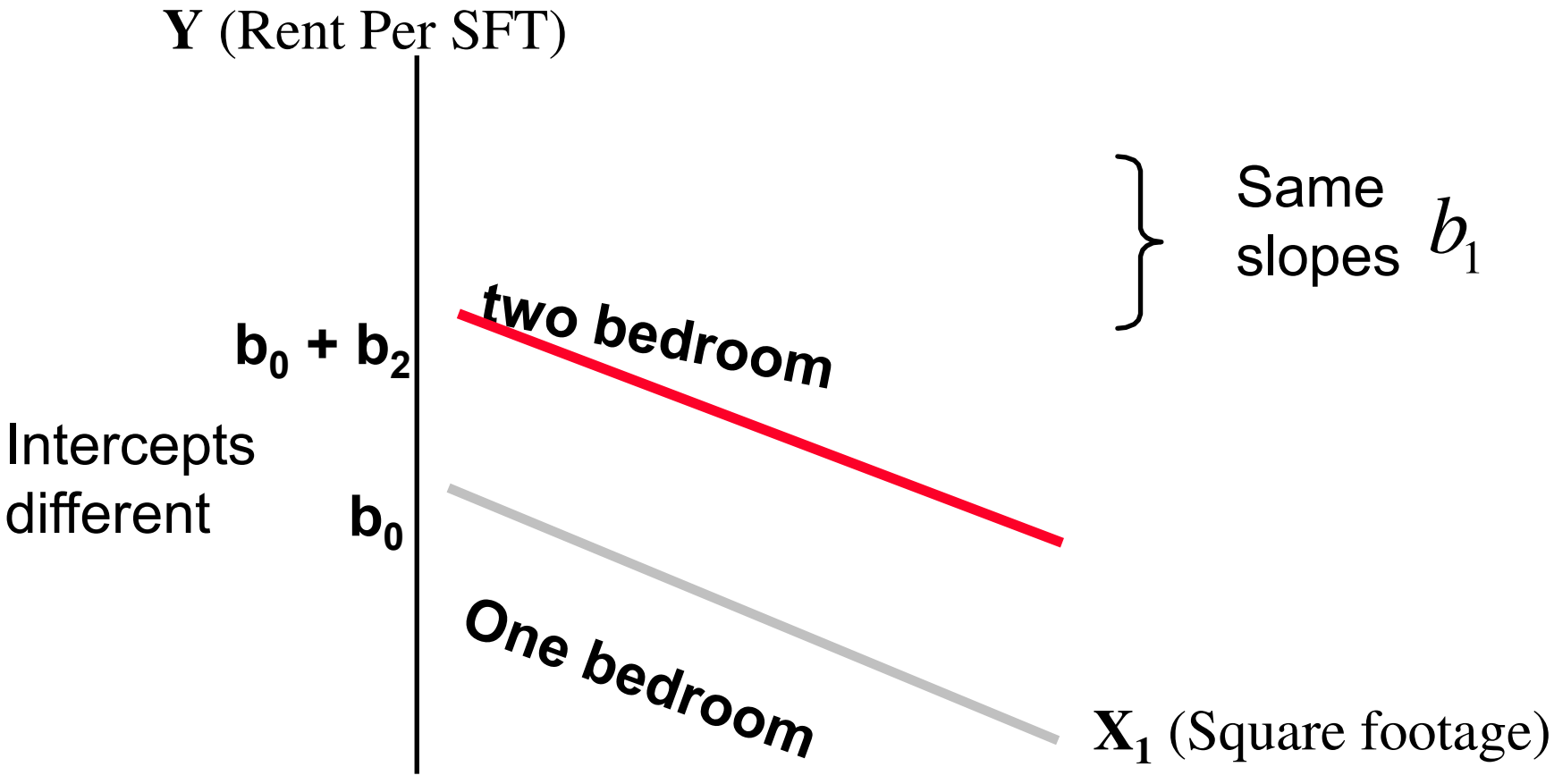
Two separate regression equations

Dummy Variable

- Define $D = 0$ if an apartment has one bedroom and = 1 if it has two bedrooms.
- The variable D is said to be an *indicator* (dummy) variable, in that it indicates the presence, or absence, of two bedrooms.
- To interpret βs, we now consider the model
- $y = \beta_0 + \beta_1 x_1 + \beta_2 D + e.$
- Taking expectations, we have $E\, y = \beta_0 + \beta_1 x_1 + \beta_2 D$
- $E\, y = (\beta_0 + \beta_2) + \beta_1 x_1$ for two bedroom(D=1)
- $= \beta_0 + \beta_1 x_1$ for one bedroom(D=0)
  - The least squares method of calculating the estimators, and the resulting theoretical properties, are the still valid when using categorical variables.

Dummy-Variable Models

Two separate regression equations



- What happened if the Dummy variable is a Nominal variable?

Response RENT_SFT

Whole Model

FOOTAGE

TWOBED

Summary of Fit

RSquare	0.213818
RSquare Adj	0.16617
Root Mean Square Error	0.147253
Mean of Response	0.805157
Observations (or Sum Wgts)	36

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	0.19460859	0.097304	4.4875
Error	33	0.71555262	0.021683	Prob > F
C. Total	35	0.91016121		0.0189

Parameter Estimates

	Term	Estimate	Std Error	t Ratio	Prob>|t|
	Intercept	1.0123207	0.142065	7.13	<.0001
X1	FOOTAGE	-0.000227	0.000233	-0.97	0.3382
D	TWOBED	-0.052527	0.101931	-0.52	0.6098

Interpreting b_2

Interpretation

- $= (\beta_0 + \beta_2) + \beta_1 x_1$ for two bedroom(D=1)
- $= \beta_0 + \beta_1 x_1$ for one bedroom(D=0)

- We have same slope and different intercept.
- It looks like we are fitting two different but parallel line to the data.
- This process allows to answer the question:

whether there is a difference in the average value of y variable for the two groups after adjusting for the effect of the quantitative variable (x_1)?

Also how much the average difference on y is*?*

Interpretation of β

- For indicator variable such *D*, we interpret β_2 as the expected increase of y when going from the base level of (D=0) to the alternative level (D=1).
- Here it is the expected increase of Rent_SFT when going from two bedroom to one bedroom .
- Example:
- **$y=1.0123 - 0.00022\, x_1 - 0.05\, D$**
- using the least squares method as we have seen before.
- We have also **s= ……..**and **R^2=…….%**
- We expect the rent per square foot to be smaller by 0.05 $ for a two bedroom as compared to one bedroom apartment.
- Then test whether β_2 is statistically significant or could this difference have occurred purely by chance?

Question

- **Does the coding of the two groups matter?NO**
- **Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.9597939	0.229298	4.19	0.0002
FOOTAGE	-0.000227	0.000233	-0.97	0.3382
TWOBED	0.0525268	0.101931	0.52	0.6098

Regression model when one explanatory variable is categorical

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8677955	0.071997	12.05	<.0001
FOOTAGE	-0.000423	0.000079	-5.33	<.0001
AGE	0.1271177	0.021049	6.04	<.0001

The Coefficient 0.127 indicates that as the value assigned to Age increases, so does the amount of Rent-Sft.
On average there is a difference of 0.127 units on Rent_sft Between different apartment age.

Age=1 if old
=2 if intermediate
=3 if new

-So we pay 0.127 ($1000)more on average for a new apartment than for an intermediate
-We pay 0.127 ($1000)more on average for an intermediate Apartment than for an old one

Better option= yes
Create dummy variables

If old is used as the base-level

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.9943763	0.065795	15.11	<.0001
FOOTAGE	-0.000412	0.000083	-4.96	<.0001
New	0.2506243	0.043152	5.81	<.0001
Inter	0.1053645	0.047396	2.22	0.0334

Difference in the intercept between new and old

Difference in the intercept between intermediate and old

Interaction

- Definition:

- An interaction term is a variable that is created as a nonlinear function of two or more explanatory variables.

- This is usually a special case of linear regression because we can create the nonlinear term as a new explanatory variable and run a linear regression.

- We can always use t-test to check if the new variable is important or not…..

Modeling Interaction

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

x_1x_2 is a **cross-product** or **interaction term**

$$y = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2$$

The slope of x_1 depends on x_2 value

$$y = \beta_0 + (\beta_2 + \beta_3 x_1) x_2 + \beta_1 x_1$$

The slope of x_2 depends on x_1 value

Testing H_0: $b_3=0$ will determine the existence of interaction

Interaction Terms

- Why if the change in the expected **y** per unit change in x_1 depend on x_2 ?
- Start with E $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. (called additive)

- Add an *interaction variable* $x_3 = x_1 x_2$ to get
- E $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- To interpret β_3, as x_1 moves from x_1 to $x_1 + 1$, we get
- *change* = E y_{new} - E y_{old} =
- $(\beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_3 (x_1 + 1) x_2)$-
- $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$
- $= \beta_1 + \beta_3 x_2$.

Interpretation

- Here we say that the **partial change in Expected y** due to **movement of x_1 depend on the value of x_2.**

- We say also that the partial changes due to each variable are not unrelated but rather “move together”.

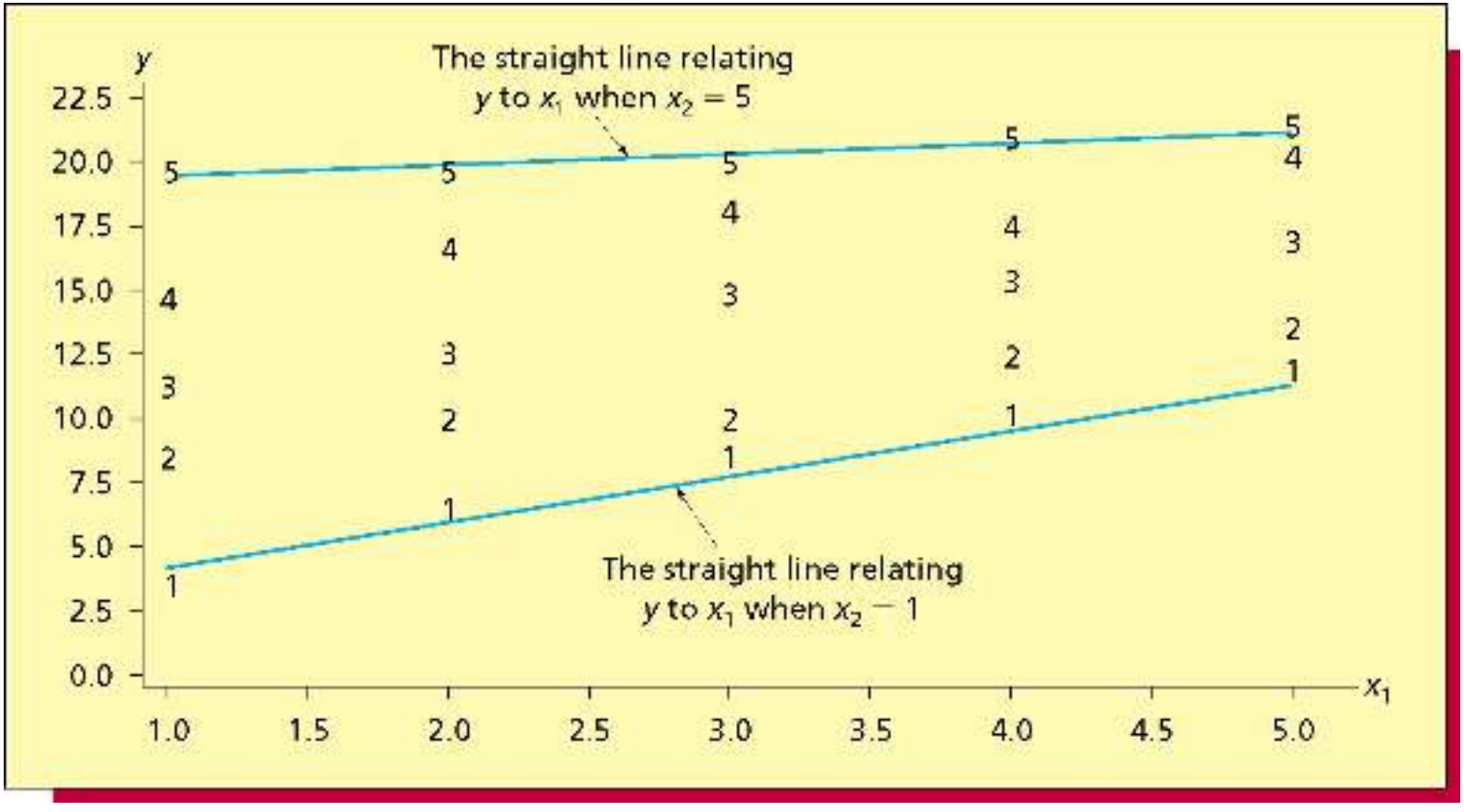
Combining a continuous and an indicator Interaction Terms-Indicators

- y - RENT_SFT, x_1 - MILES, D - TWOBED
- D = 0 if the apartment is a 1 bedroom and
- = 1 if the apartment is a 2 bedroom.
- Then, using an interaction term,
  - $\mathbf{E}\, y = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 x_1 D$
- $E\, y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_1$ for 2 bedrooms
- $E\, y = \beta_0 + \beta_1\, x_1$ for 1 bedroom.
- So here we have the choice for two possibilities:
- 1- fitting one regression model to both kind of bedrooms assuming **<u>one variability</u>** parameter or
- 2-fitting two non-parallel regression models, one for one bedroom and another to two bedrooms and thus we assume **<u>different variability</u>** parameters.

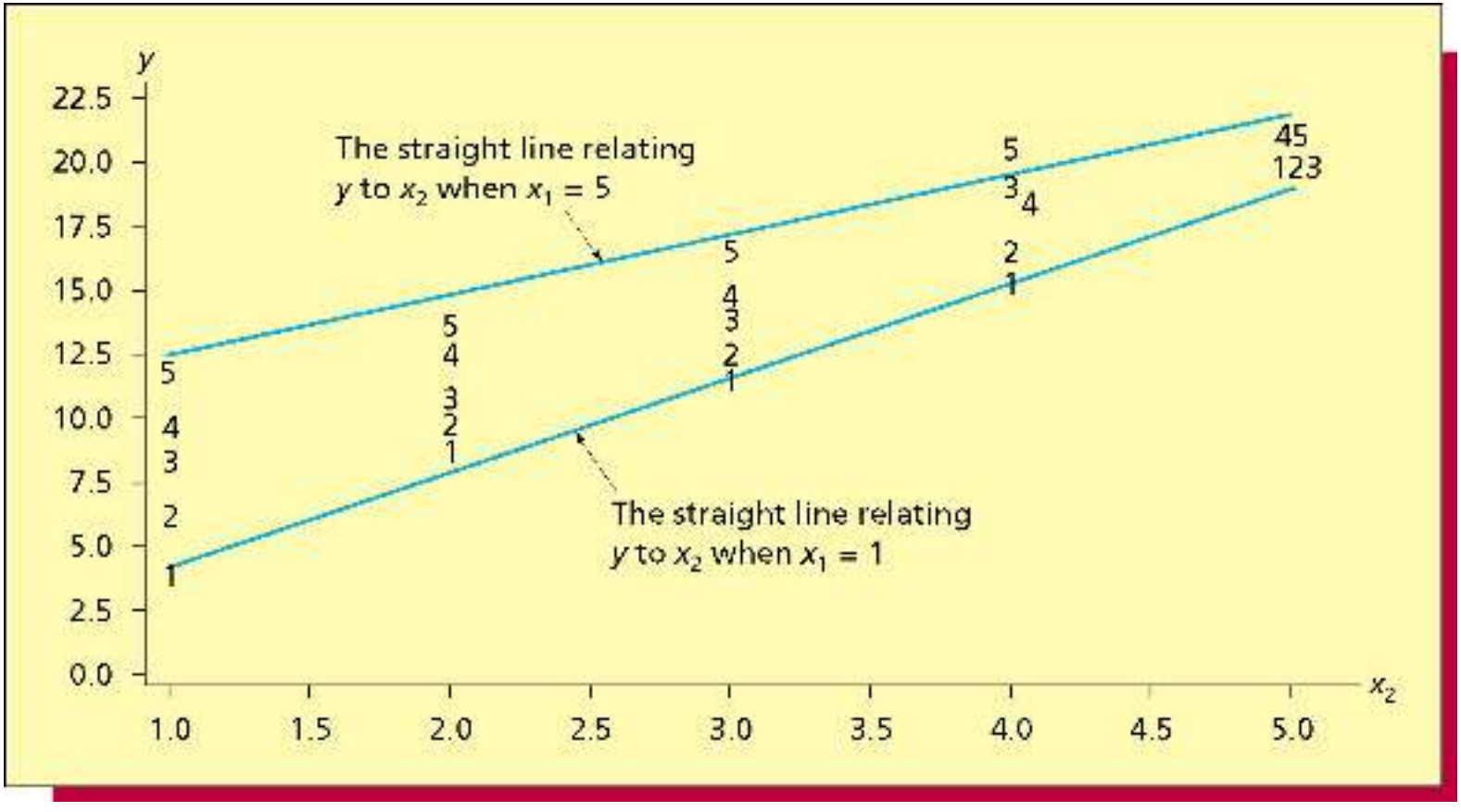
Interaction Variables

Sales Region	Radio and TV Expenditures x1	Print Expenditures x2	Sales Volume y
1	1	1	3.27
2	1	2	8.38
3	1	3	11.28
4	1	4	14.5
5	1	5	19.63
6	2	1	5.84
7	2	2	10.01
8	2	3	12.46
9	2	4	16.67
10	2	5	19.83
11	3	1	8.51
12	3	2	10.14
13	3	3	14.75
14	3	4	17.99
15	3	5	19.85
16	4	1	9.46
17	4	2	12.61
18	4	3	15.5
19	4	4	17.68
20	4	5	21.02
21	5	1	12.23
22	5	2	13.58
23	5	3	16.77
24	5	4	20.56
25	5	5	21.05

Interaction Variables



Interaction Variables



Response Sales

Whole Model

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.3497	0.688281	-3.41	0.0026
Radio Tv	2.3611	0.207524	11.38	<.0001
Paper	4.1831	0.207524	20.16	<.0001
Radio*Paper	-0.3489	0.062571	-5.58	<.0001

$$\hat{y} = -2.35 + (2.36 - 0.35x_2)x_1 + 4.18x_2$$

Interaction exists, the slope of x_1 decreases as x_2 increases

Radio advertisement effect on sales diminished as the paper advertisement increases.

Indicators and Several Continuous Variables

- y - total tax paid as a percent of total income (TAXPERCT)
- x_1 - total income (TOTALINC),
- x_2 - earned income (EARNDINC),
- x_3 - federal itemized or standard deductions (DEDUCTS),
- x_4 - marital status (MARRIED, =1 if married, =0 if single).

- We can combine the indicator variable, x_4 , with each of the other explanatory variables to get the model

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- $+ \beta_{14} x_1 x_4 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + e$ *(6 explanatory variable: long)*
- The deterministic portion of this model can be written as:

- $E\, y = (\beta_0 + \beta_4) + (\beta_1 + \beta_{14}) x_1 + (\beta_2 + \beta_{24}) x_2 + (\beta_3 + \beta_{34}) x_3$ for married filers
- $E\, y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ for single filers.

- (are 2 three-explanatory variable regression model: simpler)