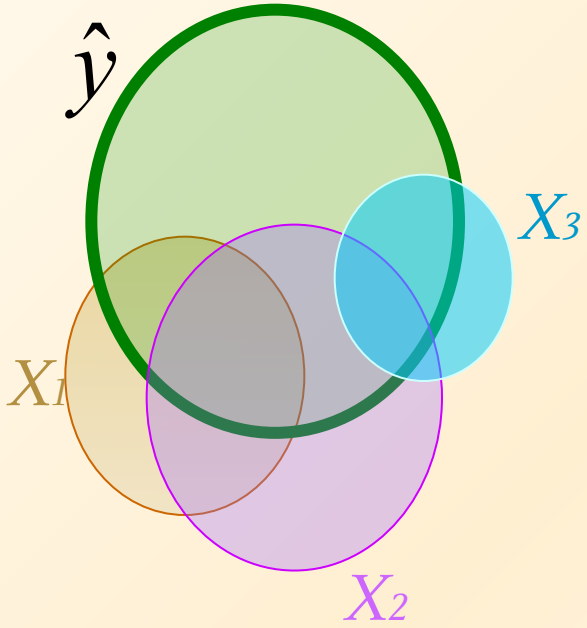
Linear Regression Analysis

Regression Analysis

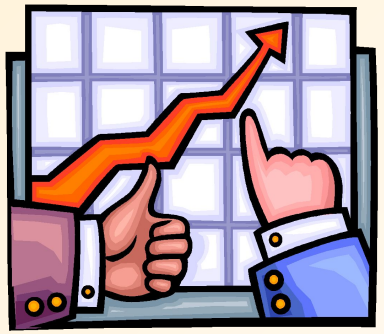
$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k + \varepsilon$$



Simple and Multiple Regression Analysis

What does regression analysis do?

- Examines whether changes/differences in values of one variable (dependent variable Y) are linked to changes/differences in values of one or more other variables (independent variables X_1, X_2, etc.), **while controlling** for the changes in values of all other Xs.
  - E.g., Relationship between salary and gender for people who have the same levels of education, work experience, position level, seniority, etc.



- The DV (Y) must be metric.
- The IVs (Xs) must be either metric or dummy var.

- Central Question Addressed:
  - Is Y a function of X_1, X_2, etc.? How ?
  - Is there a relationship between Y and X_1, X_2 , etc., (in each case, after controlling for the effects of all other Xs)? In what way?
  - What is the relative impact of each X on Y, holding all other Xs constant (that is, all other Xs being equal)?

Simple and Multiple Regression Analysis

More specifically,

- Do values of Y tend to increase/decrease as values of X_1, X_2, etc. increase/decrease?

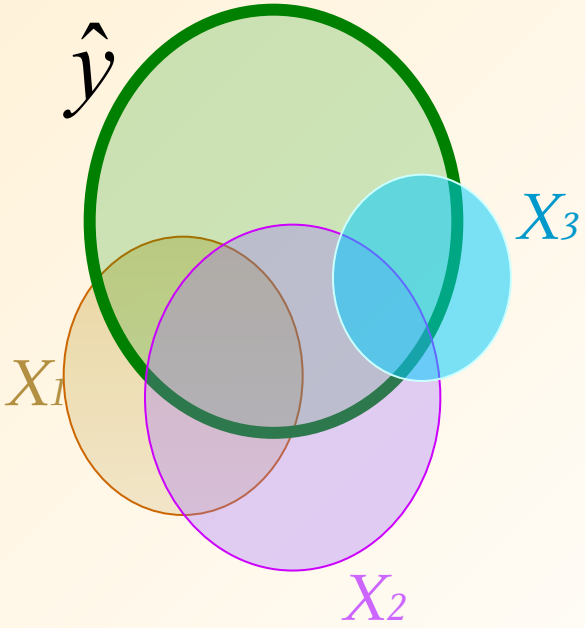
  <u>If so,</u>

  - By how much?

    And

  - How strong is the connection/relationship between Xs and Y?

    - <u>what % of differences/variations in Y values</u> (e.g., income) among study subjects <u>can be explained by</u> (or attributed to) differences in <u>X values</u> (e.g. years of education, years of experience, etc.)?

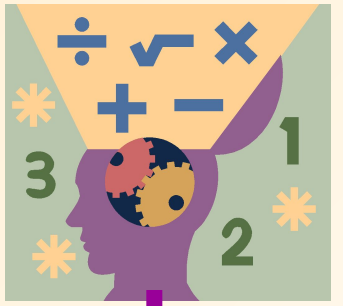


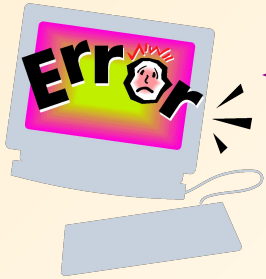
Simple and Multiple Regression Analysis

- NOTE: Once we can determine how values of Y change as a function of values of X_1, X_2, etc., we will also be able to **predict/estimate** the value of Y from specific values of X_1, X_2, etc.

$$\hat{Y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k + \epsilon$$

- Therefore, regression analysis, in a sense, is about ESTIMATING values of Y, using information about values of Xs:





- Estimation, by definition, involves?

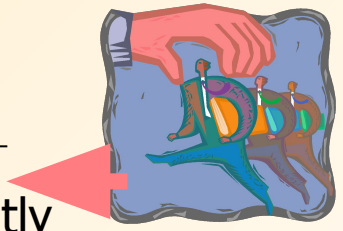
- The objective?
  - To minimize error in estimation.
  - Or, to compute estimates that are as close to the true/actual values as possible.

Simple and Multiple Regression Analysis

QUESTION: What is the simplest way to obtain an estimate for some population characteristic (e.g., number of credit cards per U.S. household)?

ANSWER:

1. Select a representative sample from the population and
2. Compute the mean for that sample (e.g., compute the average number of CCs for the sample households).



Regression analysis can be viewed as a technique that often significantly improves the accuracy of estimation results relative to using the mean value.

So, suppose we were to estimate the number of credit cards for U.S. households, based on information from a random sample of, say, n = 8 families.

Simple and Multiple Regression Analysis

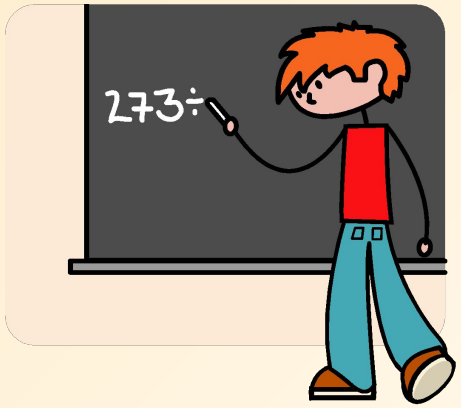
*Estimating Number of Credit Cards**

i Family Number	y_i **Actual** # of Credit Cards
1	4
2	6
3	6
4	7
5	8
6	7
7	8
8	10

$$\sum Y_i = 56$$

$\hat{y} =$ Estimate?

$$\hat{y} = \bar{y} = \frac{56}{8} = 7$$



QUESTION: Can we determine how much error in estimation we are committing by using as our estimate, for each of these households?

$$\overline{Y} = 7$$

* This example was adopted from Hair, Black, Babin, Anderson, & Tatham, (2006). *Multivariate Data Analysis*, 6th ed., Prentice Hall.

Simple and Multiple Regression Analysis

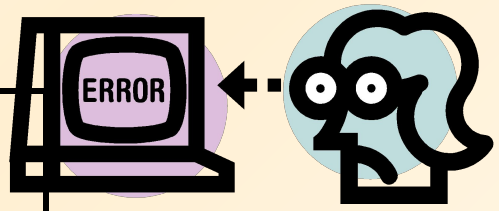
Estimating Number of Credit Cards

i Family Number	y_i **Actual** # of Credit Cards	$\hat{y} = \bar{y}$ **Estimate** for # of Credit Cards	Error in Estimation
1	4	7	?
2	6	7	?
3	6	7	?
4	7	7	?
5	8	7	?
6	7	7	?
7	8	7	?
8	10	7	?

$$\sum y_i = 56 \qquad \hat{y} = \bar{y} = \frac{56}{8} = 7$$

Simple and Multiple Regression Analysis

Estimating Number of Credit Cards

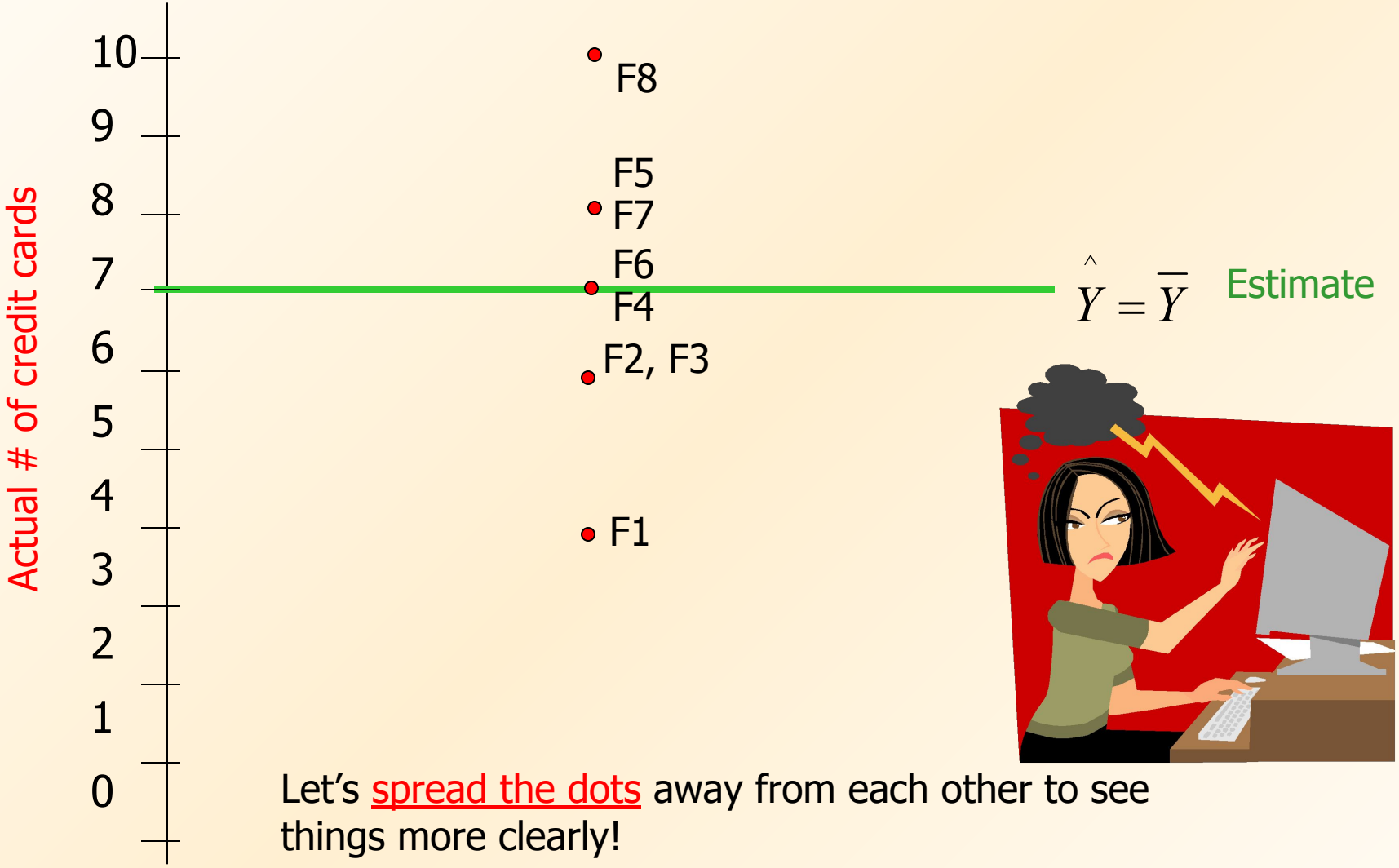


i Family Number	y_i **Actual** # of Credit Cards	$\hat{y} = \bar{y}$ **Estimate** for # of Credit Cards	$y_i - \bar{y}$ Error in Estimation
1	4	7	-3
2	6	7	-1
3	6	7	-1
4	7	7	0
5	8	7	+1
6	7	7	0
7	8	7	+1
8	10	7	+3

$$\sum y_i = 56 \qquad \hat{y} = \bar{y} = \frac{56}{8} = 7$$

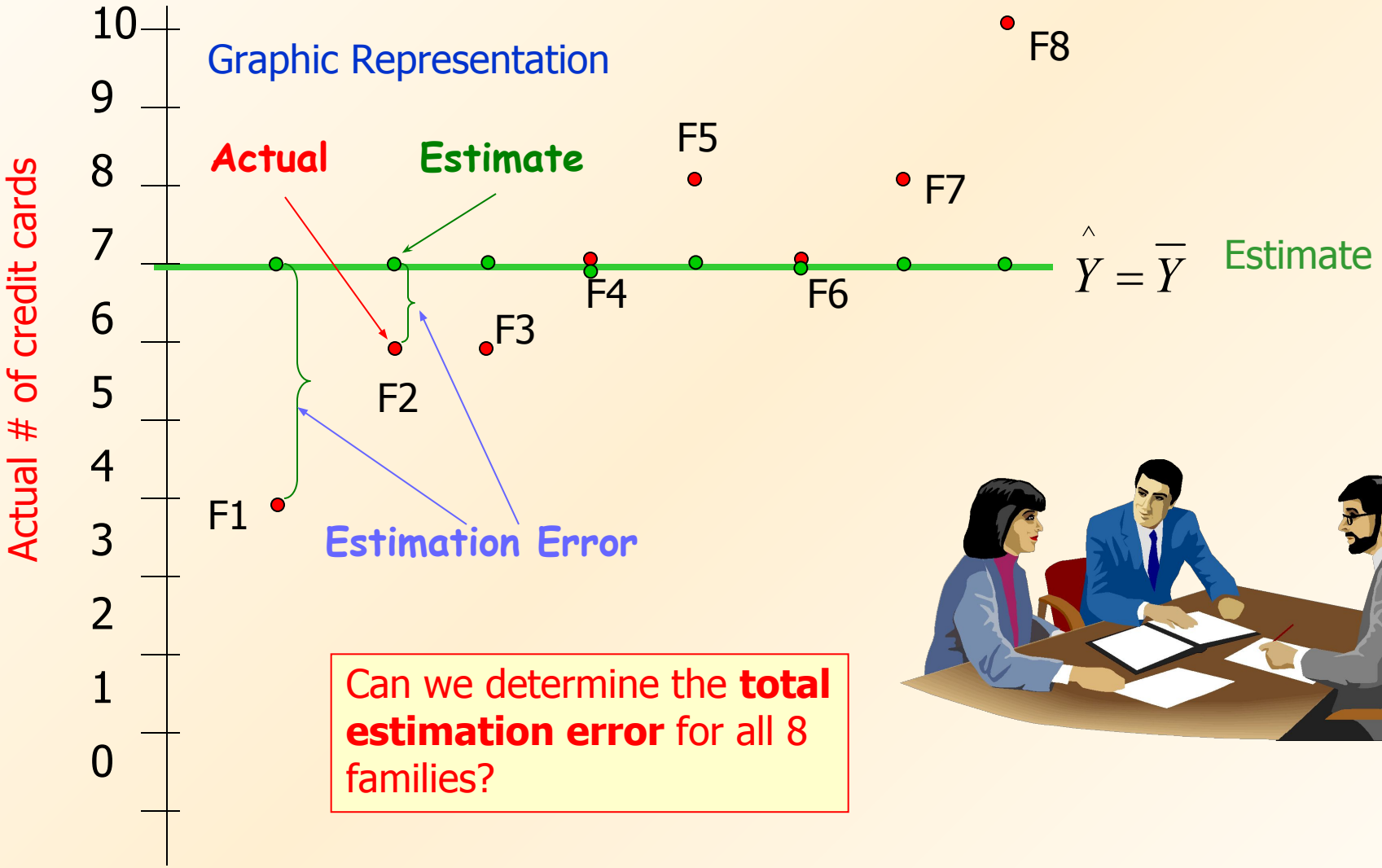
Lets now see all this graphically

Simple and Multiple Regression Analysis



Let's spread the dots away from each other to see things more clearly!

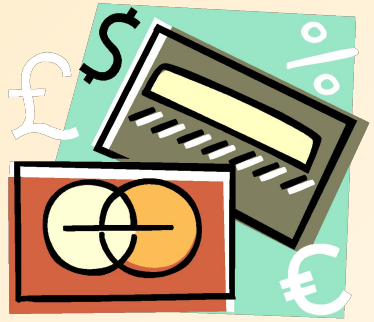
Simple and Multiple Regression Analysis



Can we determine the **total estimation error** for all 8 families?

Simple and Multiple Regression Analysis

i Family Number	y_i **Actual** # of Credit Cards	$\hat{y} = \bar{y}$ **Estimate** for # of Credit Cards	$y_i - \bar{y}$ Error in Estimation
1	4	7	-3
2	6	7	-1
3	6	7	-1
4	7	7	0
5	8	7	+1
6	7	7	0
7	8	7	+1
8	10	7	+3
	$\sum y_i = 56$	$\hat{y} = \bar{y} = \frac{56}{8} = 7$	$\sum (y_i - \bar{y})$ **= 0**



What would be the total estimation error for all 8 families combined?

Solution?

Simple and Multiple Regression Analysis

Estimating Number of Credit Cards

i Family Number	y_i **Actual** # of Credit Cards	$\hat{y} = \bar{y}$ **Estimate** for # of Credit Cards	$y_i - \bar{y}$ Error in Estimation	$(y_i - \bar{y})^2$ Errors Squared
1	4	7	-3	9
2	6	7	-1	1
3	6	7	-1	1
4	7	7	0	0
5	8	7	+1	1
6	7	7	0	0
7	8	7	+1	1
8	10	7	+3	9

$\sum y_i = 56$ $\quad \hat{y} = \bar{y} = \frac{56}{8} = 7$ $\quad \sum (y_i - \bar{y}) = 0$ $\quad \sum (y_i - \bar{y})^2 = \mathbf{22}$

SST = Sum of Squares Total

Simple and Multiple Regression Analysis

22 = SST = Index for total (combined) amount of estimation error
for all families (observations) in the sample when using the mean as the estimate.

✔ SST is also the sum of squared deviations from the mean.

- Remember the formula for computing **Variance**?

- Objective in Estimation?

  Minimize error, maximize precision.

- Can we cut down the amount of estimation error (SST)? How?

  Yes, we can, by using information about other variables suspected to be strong predictors (strongly related to) # of credit cards possessed by families (e.g., family size, family income, etc.)..

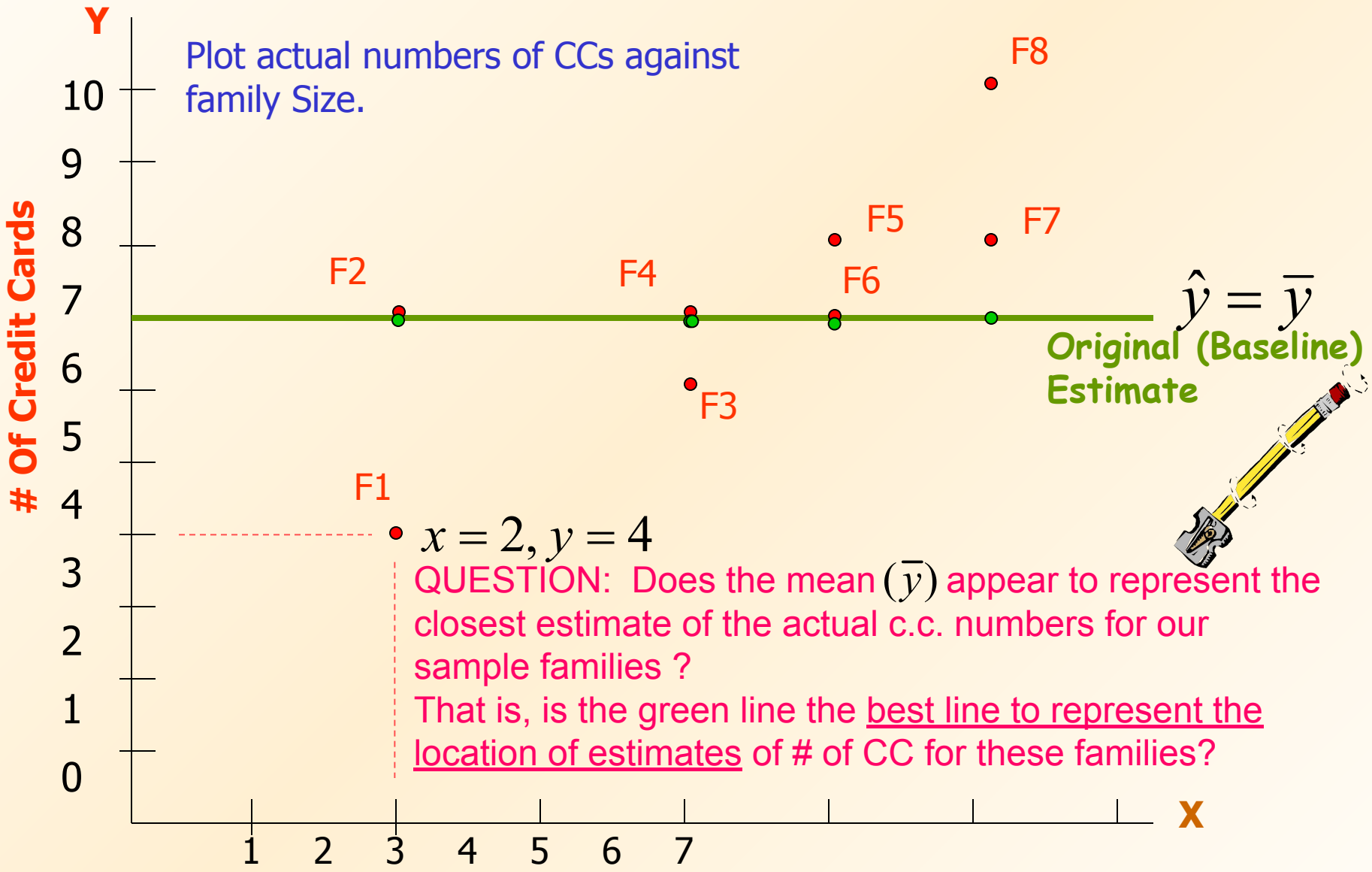
Simple and Multiple Regression Analysis

i Family Number	*y* **Actual** # of Credit Cards	*x* **Family Size**
1	4	2
2	6	2
3	6	4
4	7	4
5	8	5
6	7	5
7	8	6
8	10	6

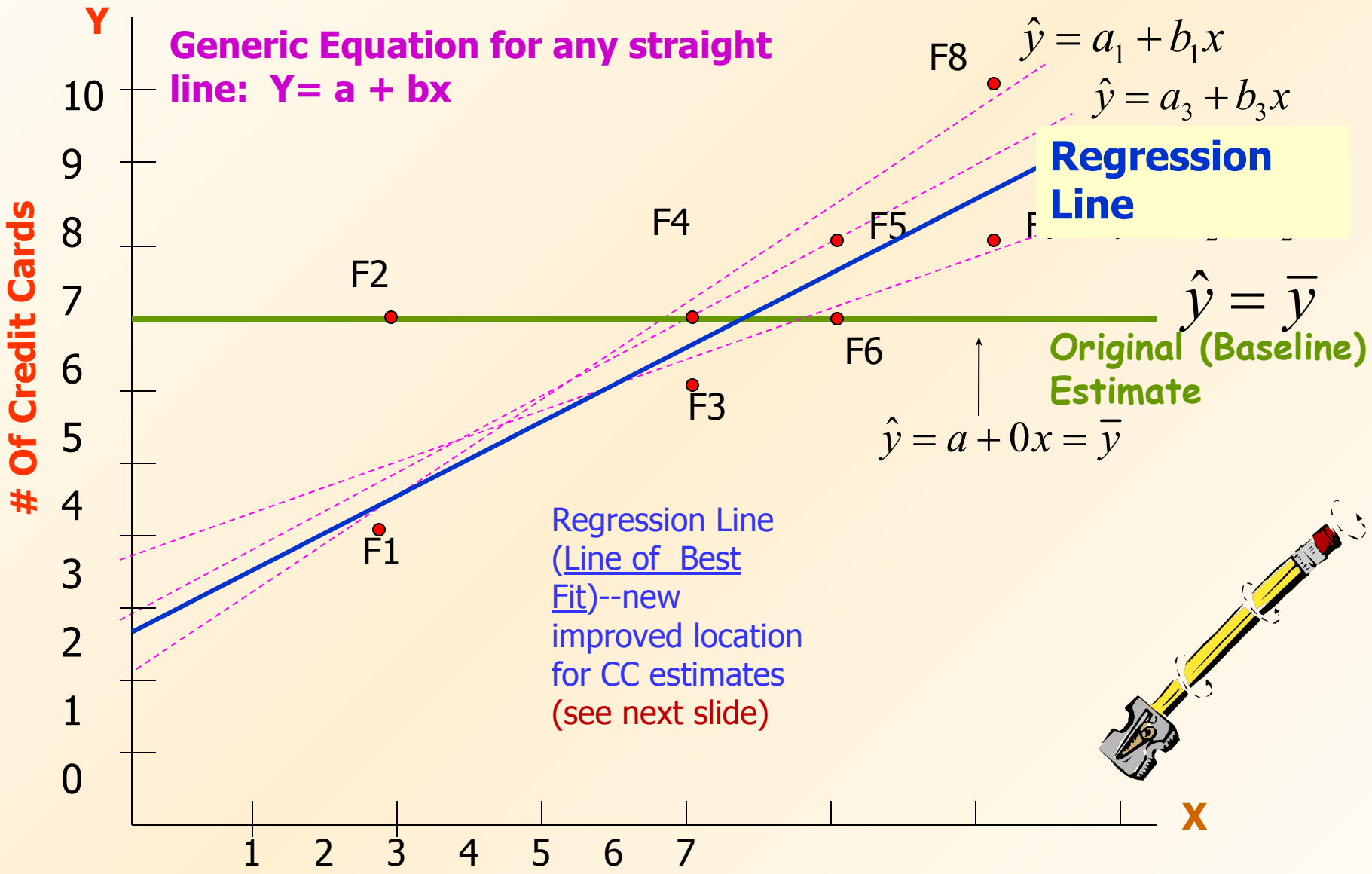
We now can attempt to estimate # of credit cards from the information on family size, rather than from its own mean.

Let's first see this graphically!

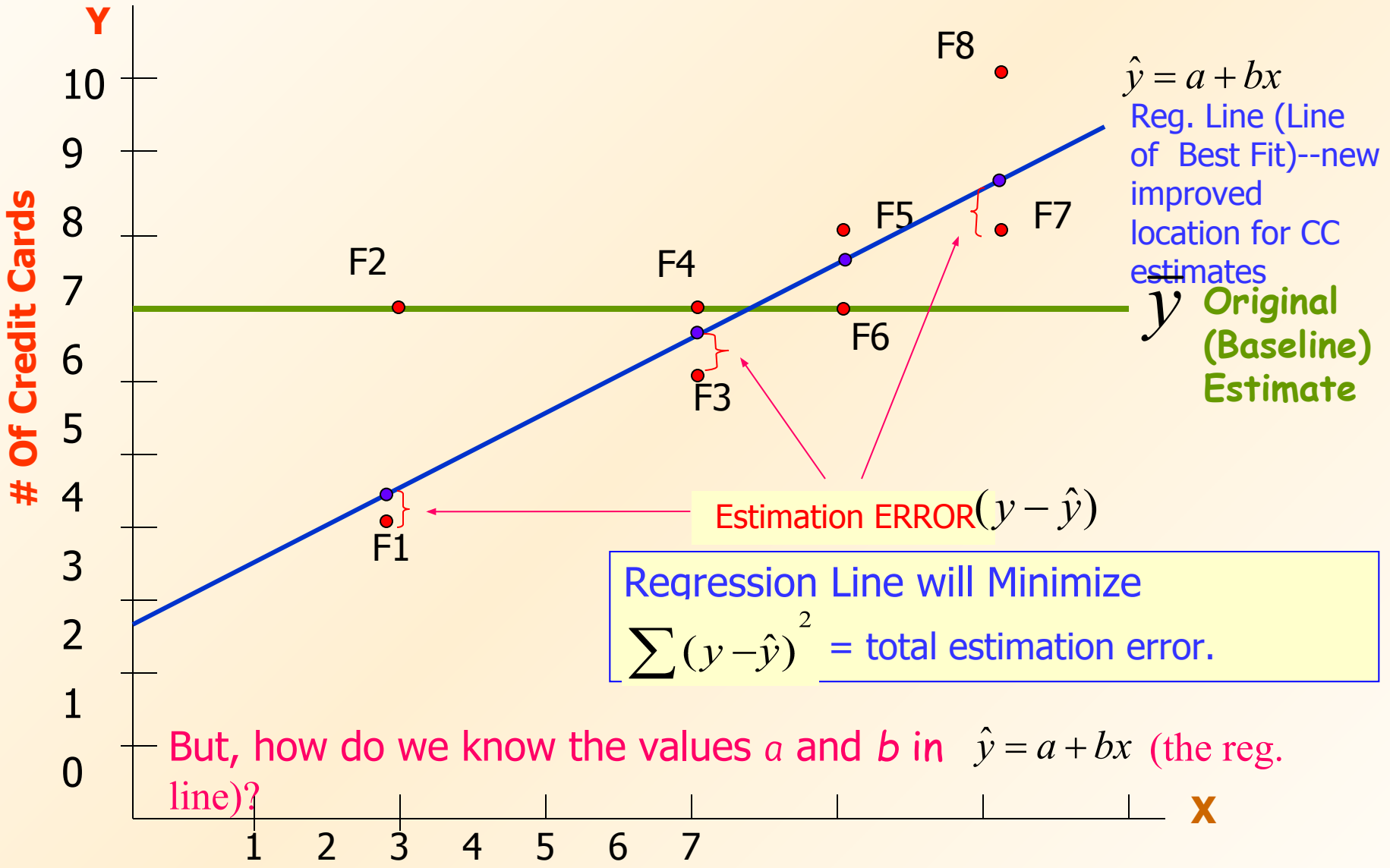
Simple and Multiple Regression Analysis



Simple and Multiple Regression Analysis



Simple and Multiple Regression Analysis



Actual # of credit cards

EQUATION FOR REGRESSION LINE (LINE OF BEST FIT)--

Values of *a* and *b for the regression line*:

$$\hat{y} = a + bx$$

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Let's use above formulas to compute the values of "*a*" and "*b*" for the regression line in our example.

We will need: $\bar{y}$, $\bar{x}$, $\sum (x - \bar{x})(y - \bar{y})$, and $\sum (x - \bar{x})^2$

Simple and Multiple Regression Analysis

We need: $\bar{y},\ \bar{x}, \sum(x-\bar{x})(y-\bar{y}),$ and $\sum(x-\bar{x})^2$

i Family Number	y **Actual #** of Credit Cards	x **Family Size**	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	4	2	?	?	?	?
2	6	2	?	?	?	?
3	6	4	?	?	?	?
4	7	4	?	?	?	?
5	8	5	?	?	?	?
6	7	5	?	?	?	?
7	8	6	?	?	?	?
8	10	6	?	?	?	?

$\bar{Y} = \frac{56}{8} = 7 \quad \bar{x} = \frac{34}{8} = 4.25 \quad \sum(x-\bar{x})(y-\bar{y}) = ? \quad \sum(x-\bar{x})^2 = ?$

Simple and Multiple Regression Analysis

We need: $\bar{y},\ \bar{x}, \sum(x-\bar{x})(y-\bar{y}),$ and $\sum(x-\bar{x})^2$

i Family Number	y **Actual #** of Credit Cards	x **Family Size**	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	4	2	-2.25	-3	6.75	5.0625
2	6	2	-2.25	-1	2.25	5.0625
3	6	4	-.25	-1	.25	.0625
4	7	4	-.25	0	0	.0625
5	8	5	.75	1	.75	.5625
6	7	5	.75	0	0	.5625
7	8	6	1.75	1	1.75	3.0625
8	10	6	1.75	3	5.25	3.0625

$\bar{Y} = \frac{56}{8} = 7 \quad \bar{x} = \frac{34}{8} = 4.25$

$\sum(x-\bar{x})(y-\bar{y}) = 17 \quad \sum(x-\bar{x})^2 = 17.5$

Simple and Multiple Regression Analysis

REGRESSION LINE (LINE OF BEST FIT):

$$\hat{y} = a + bx$$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{17}{17.5} = .971$$

$$a = \bar{y} - b\bar{x} = 7 - .971(4.25) = 2.87$$

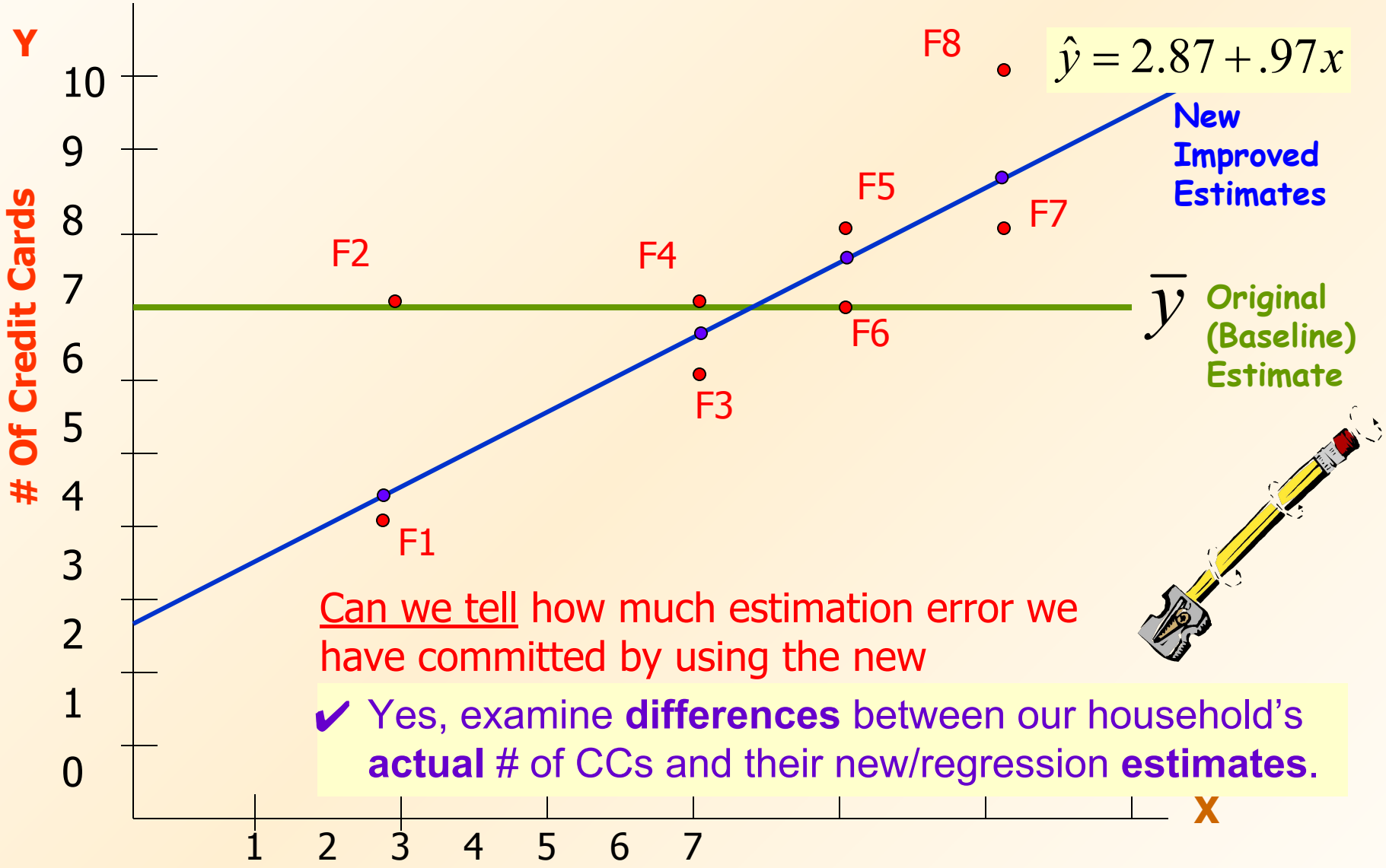
a =2.87 b = .97

$$\hat{y} = 2.87 + .97x$$

? Y-Intercept

? Regression Coefficient

Simple and Multiple Regression Analysis



Can we tell how much estimation error we have committed by using the new

✔ Yes, examine **differences** between our household's **actual** # of CCs and their new/regression **estimates**.

Simple and Multiple Regression Analysis

$\hat{y} = 2.87 + .97x$ $\hat{y}$

i Family Number	y **Actual #** of Credit Cards	x **Family Size**	$\hat{y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	4	2	?	?	?
2	6	2	?	?	?
3	6	4	?	?	?
4	7	4	?	?	?
5	8	5	?	?	?
6	7	5	?	?	?
7	8	6	?	?	?
8	10	6	?	?	?

$\Sigma(y - \hat{y})^2$

Simple and Multiple Regression Analysis

$\hat{y} = 2.87 + .97x$

$\hat{y} = 2.87 + .97(2) = 4.81$

i Family Number	y **Actual #** of Credit Cards	x Family Size	$\hat{y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	4	2	4.81	-.81	.66
2	6	2	4.81	1.19	1.42
3	6	4	6.76	-.76	.58
4	7	4	6.76	.24	.06
5	8	5	7.73	.27	.07
6	7	5	7.73	-.73	.53
7	8	6	8.7	-.7	.49
8	10	6	8.7	1.3	1.69

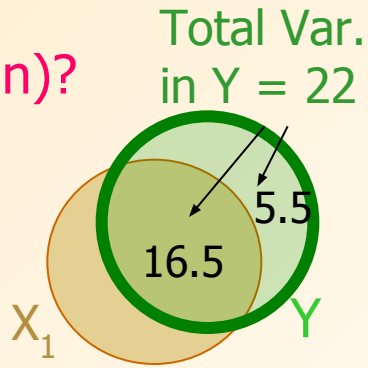
5.486 $= \sum (y - \hat{y})^2$

SSE = Sum of Squares Error (SS Residual)

Simple and Multiple Regression Analysis

Total Baseline Error using the mean (**SS Total**) **22.0**
New or Remaining Error (**SS Error** or **SS Residual**) **5.486 ~ 5.5**

QUESTION: How much of the original estimation error have we explained
away (eliminated) by using the regression model (instead of the mean)?



22 – 5.486 = 16.514 (SS Regression or SS Explained)

QUESTION: **What %** of estimation error have we explained
(eliminated by using the regression model?

R^2 = 16.514 / 22 = .751 or 75% What is this called?
% of differences in # of CCs among households that is explained by differences in their family size.
What does the remaining 25% represent?

Percent of variation (differences) in number of credit cards owned by families that can be accounted for by: (a) all other potential predictors not included in the model, beyond family size, and (b) unexplainable random/chance variations.

Simple and Multiple Regression Analysis

R^2 = SS Regression / SS Total = 16.5/22 = 75%

R^2 is a measure of our success regarding accuracy of our estimation effort.

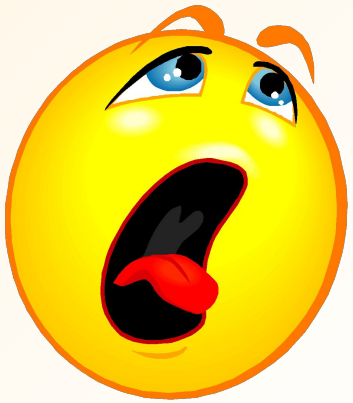
- ✔ R^2 = % of estimation error that we have been able to explain away by using the regression model, instead of using the mean.
- ✔ R^2 indicates how much better we can predict Y from information about Xs, rather than from using its own mean.
- ✔ R^2 = % of differences (variations) in Y values that is explained by (attributable to) differences in X values.

Note: When dealing with only two variables (a single X and Y):

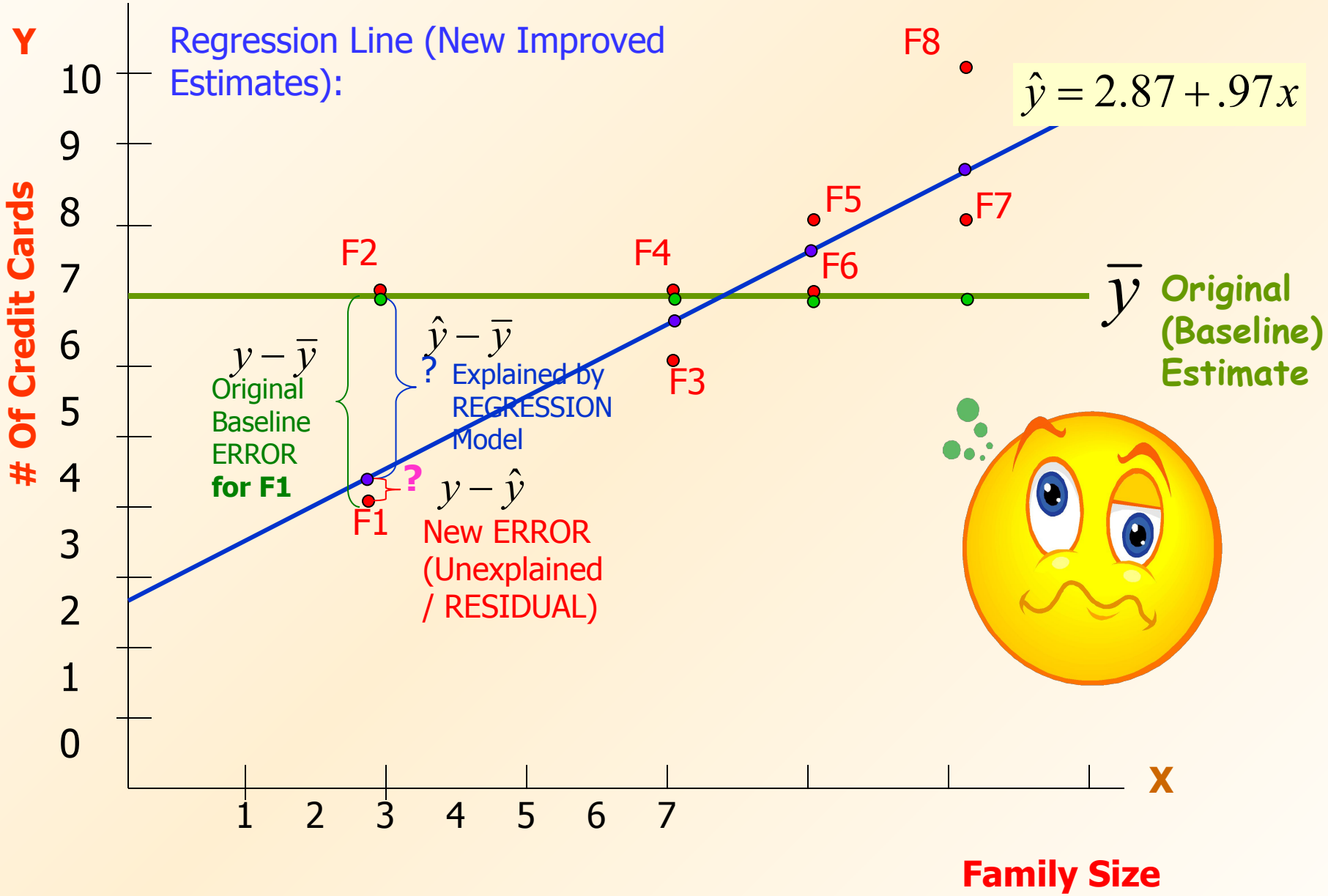
$$r = \sqrt{R^2} = \sqrt{\frac{16.514}{22}} = \sqrt{.75} = .866$$

Pearson Correlation of Y with X_1 (NOT controlling for any other var.)

Let's now examine all this graphically!



Simple and Multiple Regression Analysis



Simple and Multiple Regression Analysis

5.5 = SSE = The amount of estimation error for the 8 sample families when using simple regression (i.e., a regression model that includes only information about family size).

Can we reduce the amount of estimation error (SSE) to an even lower level and, thus, improving the estimation process? How?

Yes, by adding information on a second variables suspected strongly related to # of credit cards (e.g., family income--X_2).

.

Simple and Multiple Regression Analysis

i Family Number	y_i **Actual** # of Credit Cards	x_1 **Family Size**	x_2 **Family Income**
1	4	2	14
2	6	2	16
3	6	4	14
4	7	4	17
5	8	5	18
6	7	5	21
7	8	6	17
8	10	6	25

We now can attempt to estimate # of CCs from our information on family size and family income!

Our regression model will now be a linear plane, rather than a straight line!

Generic Equation for a linear plane:

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

Let's examine the regression plane for our example graphically.

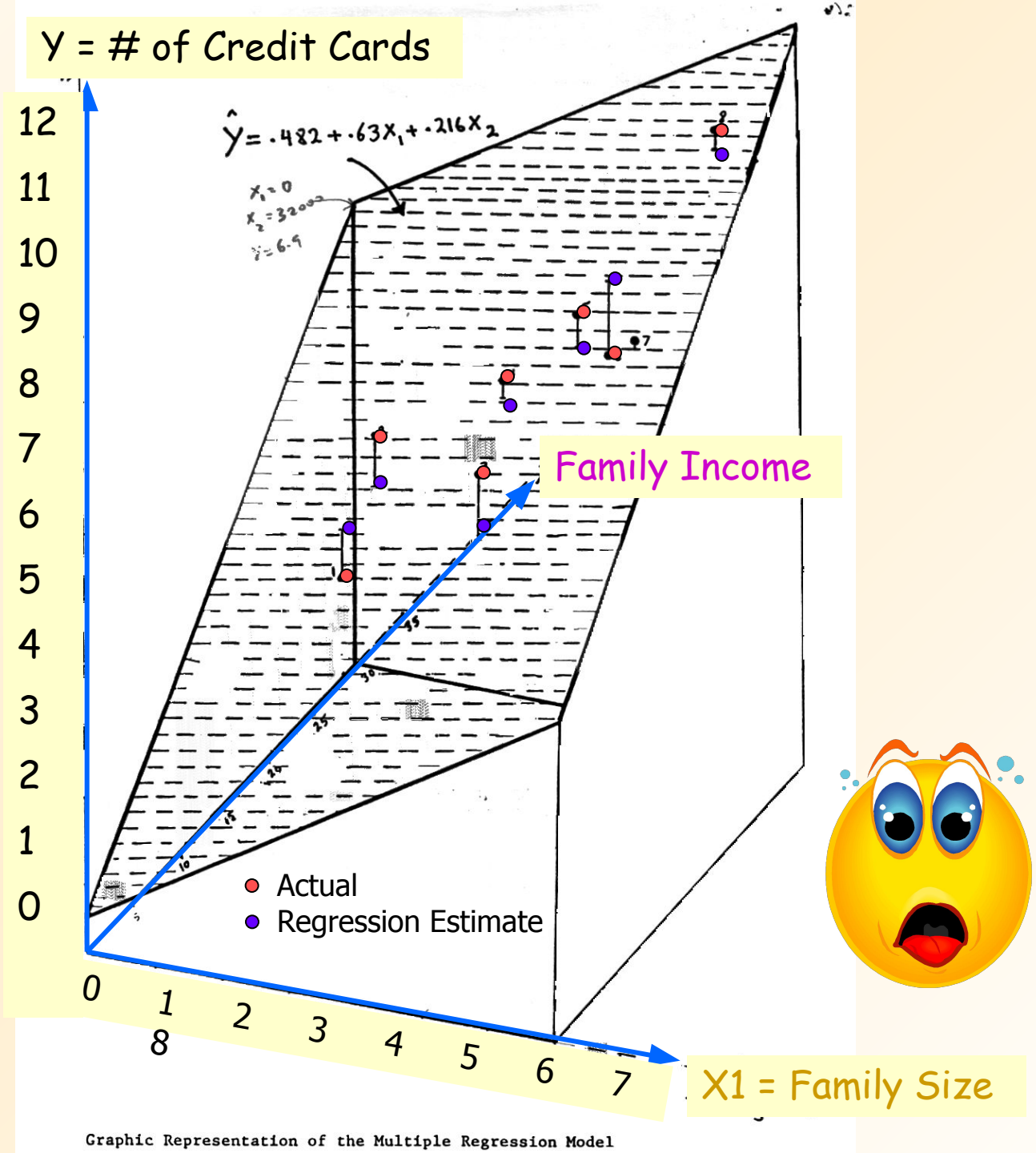
$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

Formulas are available for computing values of a, b_1 and b_2

MULTIPLE REGRESSION MODEL FOR OUR EXAMPLE:

$$\hat{y} = .482 + .63 x_1 + .216 x_2$$

Let's now see how much error in estimation we are committing by using this multiple regression model.



Graphic Representation of the Multiple Regression Model

Simple and Multiple Regression Analysis

$$\hat{y} = .482 + .63x_1 + .216x_2$$

i Family Number	y **Actual #** of Credit Cards	x_1 **Family Size**	x_2 **Family Income** ($000)	$\hat{Y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	4	2	14	?	?	?
2	6	2	16	?	?	?
3	6	4	14	?	?	?
4	7	4	17	?	?	?
5	8	5	18	?	?	?
6	7	5	21	?	?	?
7	8	6	17	?	?	?
8	10	6	25	?	?	?

$\hat{y}$

$$\Sigma(y - \hat{y})^2$$

Simple and Multiple Regression Analysis

$\hat{y} = .482 + .63x_1 + .216x_2$ $\quad \hat{y} = .482 + .63(2) + .216(14) = 4.77$

i Family Number	y **Actual #** of Credit Cards	x_1 **Family Size**	x_2 **Family Income** ($000)	$\hat{Y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	4	2	14	4.77	-.77	.59
2	6	2	16	5.20	.80	.64
3	6	4	14	6.03	-.03	.00
4	7	4	17	6.68	.32	.10
5	8	5	18	7.53	.47	.22
6	7	5	21	8.18	-1.18	1.39
7	8	6	17	7.95	.05	.00
8	10	6	25	9.67	.33	.11

SSE = Sum of Squares Error (Residual)

→ **3.05** $= \Sigma(y - \hat{y})^2$

Unique (additional) contribution of X_2 (family income) beyond X_1 = ?

5.5 – 3.05 = 2.45

Simple and Multiple Regression Analysis

The MULTIPLE REGRESSION MODEL FOR OUR EXAMPLE:

$$\hat{y} = .482 + .63x_1 + .216x_2$$

? ?

Y-Intercept, "*a*"

(**NOTE:** Only when all Xs can meaningfully take on value of zero, the intercept will have a meaningful/direct/ practical interpretation. Otherwise, it is simply an aid in increasing accuracy of estimation.

b_1 and ***b_2*** **= Regression Coefficients**

0.63: Among families of the same income, an increase in family size by one person would, on average, result in .63 more credit cards.

0.21: Among families of the same size, an income increase of $1,000, results in an average increase of 0.2 credit cards .

"*b*"s represent effect of each X on Y when all other Xs are controlled for/held constant/taken into account

- **i.e., after impacts of all other variables are accounted for (remember the high blood pressure-hearing problem connection?)**

Simple and Multiple Regression Analysis

The MULTIPLE REGRESSION MODEL FOR OUR EXAMPLE:

$$\hat{y} = .482 + .63x_1 + .216x_2$$

SST = 22 SSE = 3.05

What is our new R^2?

SS Regression = 22 – 3.05 = 18.95

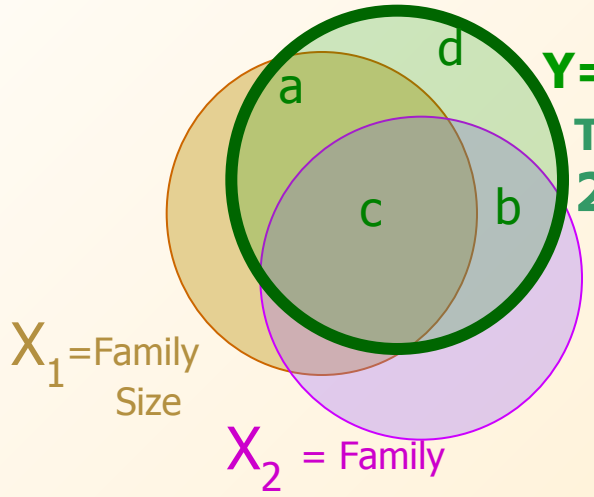
R^2 = 18.95 / 22 = .861 or 86%

→ Percent of differences in households' number of CCs that is explained by differences in family size and family income.

The Remaining 14%?

(3.05 / 22 = .14)

→ Percent of variation in number of credit cards that can be accounted for by (a) all other relevant factors not included in the model, beyond family size and income, and (b) unexplainable random/chance variations.



Y= # of CC

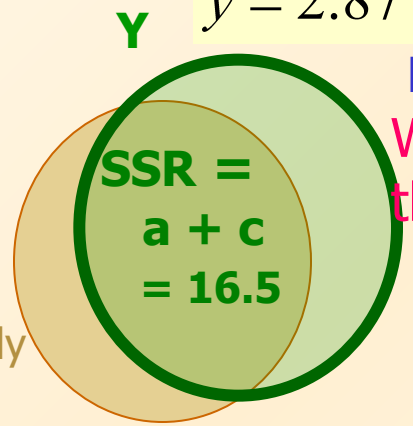
Total Variation/Error in Y = SS Total = a + b + c + d = 22

$\hat{y} = 2.87 + .97X_1$

r^2 = ? R^2 = (a+c) / (a+b+c+d)

R^2 = 16.5 / 22 = 0.75

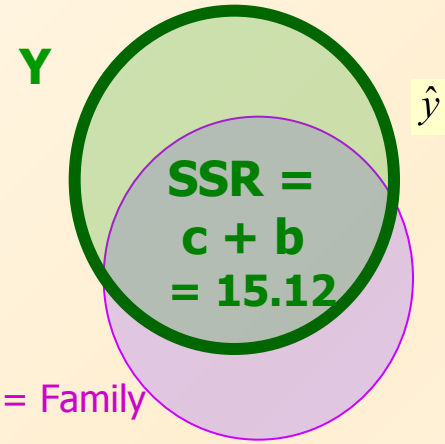
What do we call the square root of this?



Pearson/simple Correlation of Y with X_1 (not controlling for X_2)

$$r_{yx_1} = \sqrt{\frac{16.5}{22}} = \sqrt{0.75} = 0.867$$

$$r_{yx_1} = \sqrt{\frac{a+c}{a+b+c+d}}$$



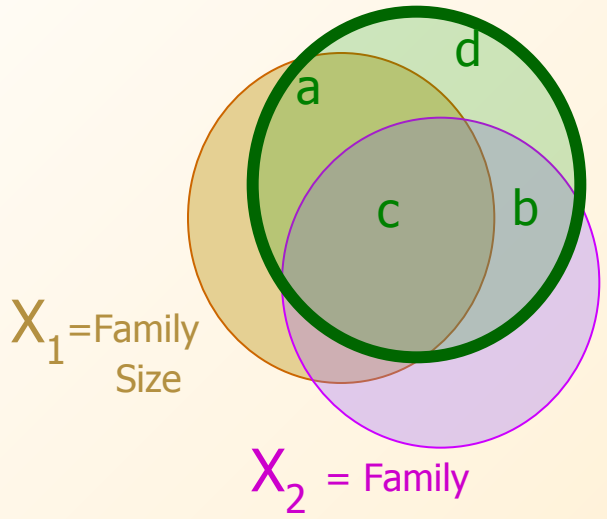
$\hat{y} = -0.063 + .398X_2$

Pearson/simple Correlation of Y with X_2 (not controlling for X_1) **?**

r^2 = (b+c) / (a+b+c+d) = 15.12 / 22 = 0.687

$$r_{yx_2} = \sqrt{\frac{b+c}{a+b+c+d}}$$

$$r_{yx_2} = \sqrt{\frac{15.11}{22}} = 0.829$$



$$\hat{y} = .482 + .63x_1 + .216x_2$$

R^2 Graphically = ?

NOTE: **c** is explained by both X_1 and X_2

SSR = a + b +c = 18.95

SST = a + b + c + d = 22

R^2 = SSR / SST = (a + b + c) / (a + b + c + d) = 18.95 / 22 = 86%

SSE = ?

SSE = d = 22 – 18.95 = 3.05

Simple and Multiple Regression Analysis

$\hat{y} = .482 + .63x_1 + .216x_2$ $\quad \hat{y} = .482 + .63(2) + .216(14) = 4.77$

i Family Number	y Actual # of Credit Cards	x_1 Family Size	x_2 Family Income ($000)	$\hat{Y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	4	2	14	4.77	-.77	.59
2	6	2	16	5.20	.80	.64
3	6	4	14	6.03	-.03	.00
4	7	4	17	6.68	.32	.10
5	8	5	18	7.53	.47	.22
6	7	5	21	8.18	-1.18	1.39
7	8	6	17	7.95	.05	.00
8	10	6	25	9.67	.33	.11

SSE = Sum of Squares Error (Residual) → $\mathbf{3.05} = \sum (y - \hat{y})^2$

Remember:

Unique (additional) contribution of X_2 = 5.5 – 3.05 = 2.45

Interpreting Regression Results

H_o: $R^2 = 0$. That is, There is <u>NO RELATIONSHIP</u> between the DV and <u>ANY OF</u> the IVs included in the regression model.

1. Is overall F significant? (i.e., $\alpha \leq 0.05$)
   - No → Don't reject Ho; No indep. Variable has a sig. relationship with dep. Variable. → **Stop**.
   - Yes → Reject Ho; One or more independent variables are significantly related to the dep. Variable.

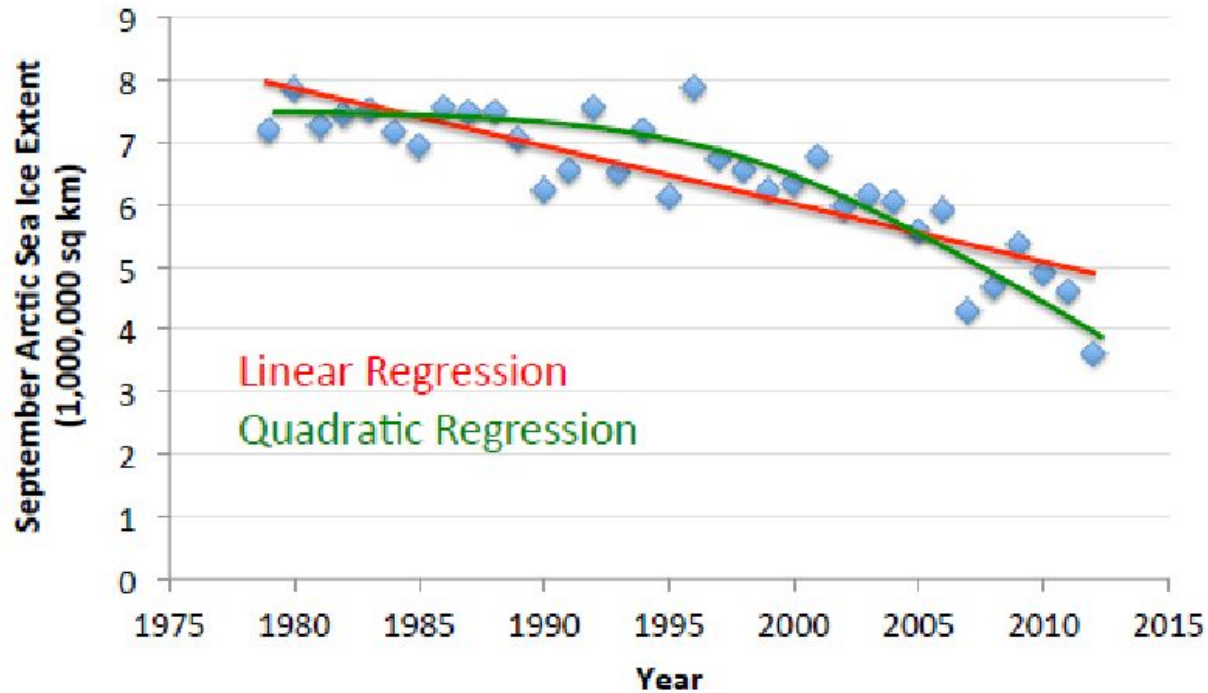
2. <u>Which independent variable(s)</u> have significant relationships with the dep. Var.? In the "Coefficients" table, look up the result of the t-test for each indep. variable's regression coefficient (b). H_o for t-test of a given variable hypothesizes that the coefficient b = 0. That is, there is no relationship between the corresponding independent variable and the dep. Variable. If a t-test's $\alpha \leq 0.05$, reject the null and conclude that the corresponding variable has a significant relationship with the dep. Variable.

3. Look up the sign of the regression coefficient (b) <u>ONLY FOR those</u> indep. variables that are found to have a significant relationship with the dependent variable (i.e., those with $\alpha \leq 0.05$), and state your conclusions accordingly.

Regression

Given:

- Data $\boldsymbol{X} = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right\}$ where $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$
- Corresponding labels $\boldsymbol{y} = \left\{y^{(1)}, \ldots, y^{(n)}\right\}$ where $y^{(i)} \in \mathbb{R}$



Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)

Prostate Cancer Dataset

- 97 samples, partitioned into 67 train / 30 test
- Eight predictors (features):
  - 6 continuous (4 log transforms), 1 binary, 1 ordinal
- Continuous outcome variable:
  - lpsa: log(prostate specific antigen level)

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

Based on slide by Jeff Howbert

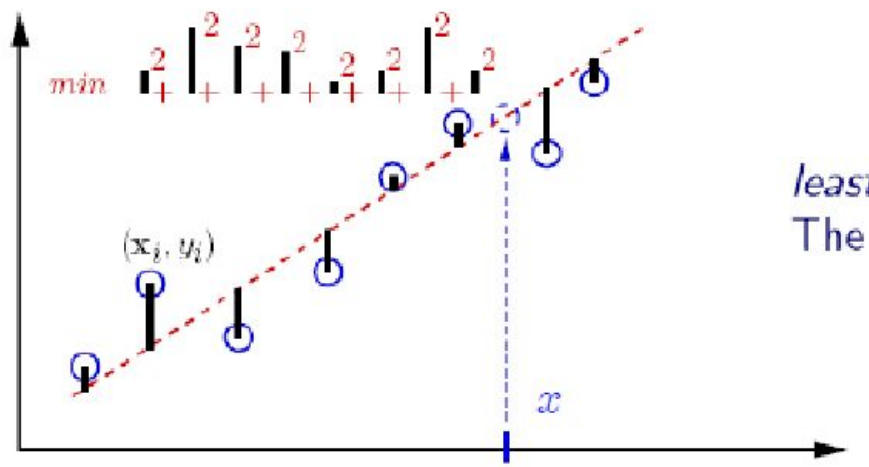
Linear Regression

- Hypothesis:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$$

Assume $x_0 = 1$

- Fit model by minimizing sum of squared errors



least squares (LSQ)
The fitted line is used as a predictor

Figures are courtesy of Greg Shakhnarovich

Least Squares Linear Regression

- Cost Function

$$J(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) - y^{(i)} \right)^2$$

- Fit by solving $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$