POPULATION REGRESSION FUNCTION (PRF)

2-1. A Hypothetical Example

- **Total population: 60 families**
- **Y=Weekly family consumption expenditure**
- **X=Weekly disposable family income**
- **60 families were divided into 10 groups of approximately the same income level**

  (80, 100, 120, 140, 160, 180, 200, 220, 240, 260)

2-1. A Hypothetical Example

- **Table 2-1 gives the *conditional distribution* of Y on the given values of X**
- **Table 2-2 gives the *conditional probabilities* of Y: $p(Y \mid X)$**
- ***Conditional Mean***

  (or Expectation): $E(Y \mid X=X_i)$

Table 2-2: Weekly family income X ($), and consumption Y ($)

Y \ X	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y ($)	**55**	**65**	**79**	**80**	**102**	**110**	**120**	**135**	**137**	**150**
	60	**70**	**84**	**93**	**107**	**115**	**136**	**137**	**145**	**152**
	65	**74**	**90**	**95**	**110**	**120**	**140**	**140**	**155**	**175**
	70	**80**	**94**	**103**	**116**	**130**	**144**	**152**	**165**	**178**
	75	**85**	**98**	**108**	**118**	**135**	**145**	**157**	**175**	**180**
	--	**88**	**--**	**113**	**125**	**140**	**--**	**160**	**189**	**185**
	--	**--**	**--**	**115**	**--**	**--**	**--**	**162**	**--**	**191**
Total	325	462	445	707	678	750	685	1043	966	1211
Mean	65	77	89	101	113	125	137	149	161	173

2-1. A Hypothetical Example

- **Figure 2-1 shows the population regression line (curve). It is the**

  regression of Y on X
- **Population regression curve is the**

  locus of the conditional means or expectations of the dependent variable

  for the fixed values of the explanatory variable X (Fig.2-2)

2-2. The concepts of population regression function (PRF)

- **$E(Y \mid X=X_i) = f(X_i)$ is Population Regression Function (PRF) or**

  Population Regression (PR)
- **In the case of linear function we have linear population regression function (or equation or model)**

  $$E(Y \mid X=X_i) = f(X_i) = ß_1 + ß_2 X_i$$

2-2. The concepts of population regression function (PRF)

$$E(Y \mid X=X_i) = f(X_i) = ß_1 + ß_2 X_i$$

- **$ß_1$ and $ß_2$ are regression coefficients, $ß_1$ is intercept and $ß_2$ is slope coefficient**
- **Linearity in the Variables**
- **Linearity in the Parameters**

2-4. Stochastic Specification of PRF

- **$U_i = Y - E(Y \mid X=X_i)$ or $Y_i = E(Y \mid X=X_i) + U_i$**
- **U_i = Stochastic disturbance or stochastic error term. It is nonsystematic component**
- **Component $E(Y \mid X=X_i)$ is systematic or deterministic. It is the mean consumption expenditure of all the families with the same level of income**
- **The assumption that the regression line passes through the conditional means of Y implies that $E(U_i \mid X_i) = 0$**

2-5. The Significance of the Stochastic Disturbance Term

- **U_i = Stochastic Disturbance Term is a surrogate for all variables that are omitted from the model but they collectively affect Y**
- **Many reasons why not include such variables into the model as follows:**

2-5. The Significance of the Stochastic Disturbance Term

Why not include as many as variable into the model (or the reasons for using u_i)

+ ***Vagueness of theory***
+ ***Unavailability of Data***
+ ***Core Variables vs. Peripheral Variables***
+ ***Intrinsic randomness in human behavior***
+ ***Poor proxy variables***
+ ***Principle of parsimony***
+ ***Wrong functional form***

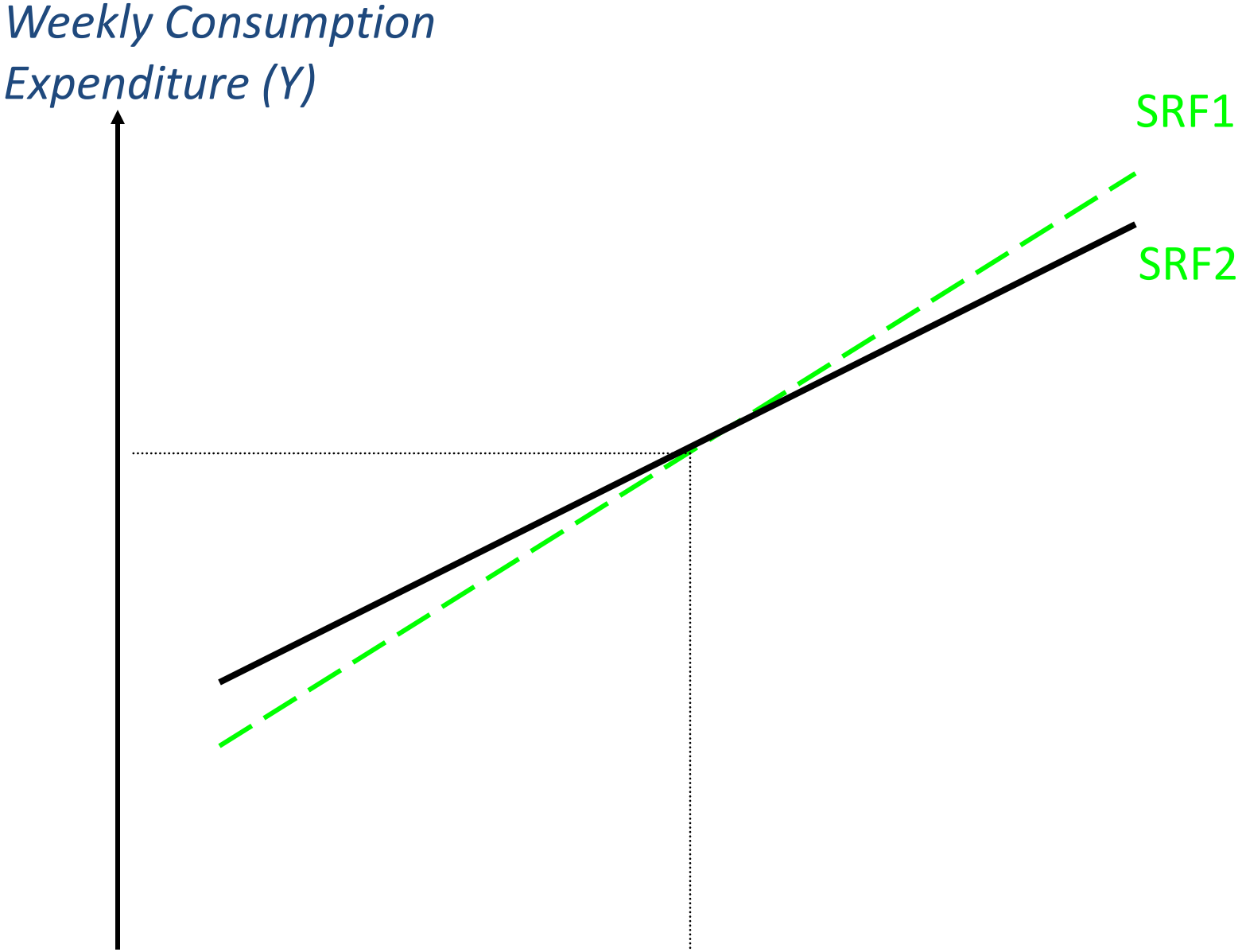
2-6. The Sample Regression Function (SRF)

Table 2-4: A random sample from the population

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

Table 2-5: Another random sample from the population

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260


Weekly Consumption Expenditure (Y)
SRF1
SRF2

2-6. The Sample Regression Function (SRF)

- **Fig.2-3: SRF1 and SRF 2**
- **$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$** (2.6.1)
- **$\hat{Y}_i$ = estimator of $E(Y \mid X_i)$**
- **$\hat{\beta}_1$ = estimator of β_1**
- **$\hat{\beta}_2$ = estimator of β_2**
- **Estimate = A particular numerical value obtained by the estimator in an application**
- **SRF in stochastic form: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$**
  or $Y_i = \hat{Y}_i + \hat{u}_i$ (2.6.3)

2-6. The Sample Regression Function (SRF)

- **Primary objective in regression analysis is to estimate the PRF $Y_i = \beta_1 + \beta_2 X_i + u_i$ on the basis of the SRF $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$ and how to construct SRF so that $\hat{\beta}_1$ close to β_1 and $\hat{\beta}_2$ close to β_2 as much as possible**

2-6. The Sample Regression Function (SRF)

- **Population Regression Function PRF**
- **Linearity in the parameters**
- **Stochastic PRF**
- **Stochastic Disturbance Term u_i plays a critical role in estimating the PRF**
- **Sample of observations from population**
- **Stochastic Sample Regression Function SRF used to estimate the PRF**

2-7. Summary and Conclusions

- **The key concept underlying regression analysis is the concept of the population regression function (PRF).**
- **This book deals with linear PRFs: linear in the unknown parameters. They may or may not linear in the variables.**

2-7. Summary and Conclusions

- **For empirical purposes, it is the stochastic PRF that matters. The stochastic disturbance term u_i plays a critical role in estimating the PRF.**
- **The PRF is an idealized concept, since in practice one rarely has access to the entire population of interest. Generally, one has a sample of observations from population and use the stochastic sample regression (SRF) to estimate the PRF.**

Regression Analysis

Linear regression

- Linear dependence: constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns).
- Regression analysis describes the relationship between two (or more) variables.
- Examples:
  - Income and educational level
  - Demand for electricity and the weather
  - Home sales and interest rates
- Our focus:
  - Gain some understanding of the mechanics.
    - the regression line
    - regression error
  - Learn how to interpret and use the results.
  - Learn how to setup a regression analysis.

Two main questions:

- Prediction and Forecasting
  - Predict home sales for December given the interest rate for this month.
  - Use time series data (e.g., sales vs. year) to forecast future performance (next year sales).
  - Predict the selling price of houses in some area.
    - Collect data on several houses (# of BR, #BA, sq.ft, lot size, property tax) and their selling price.
    - Can we use this data to predict the selling price of a specific house?
- Quantifying causality
  - Determine factors that relate to the variable to be predicted; e.g., predict growth for the economy in the next quarter: use past history on quarterly growth, index of leading economic indicators, and others.
  - Want to determine advertising expenditure and promotion for the 1999 Ford Explorer.
    - Sales over a quarter might be influenced by: ads in print, ads in radio, ads in TV, and other promotions.

Motivated Example

- Predict the selling prices of houses in the region.
  - Intuitively, we should compare the house for which we need a predicted selling price with houses that have sold recently in the same area, of roughly the same size, same style etc.
    - Idea: Treat it as a multiple sample problem.
    - Unfortunately, the list of houses meeting these criteria may be quite small, or there may not be a house of exactly the same characteristics.
    - Alternative approach: Consider the factors that determine the selling price of a house in this region.
- Collect recent historical data on selling prices, and a number of characteristics about each house sold (size, age, style, etc.).
  - Idea: one sample problem
    - To predict the selling price of a house without any particular knowledge of the house, we use the average selling price of all of the houses in the data set.
  - Better idea:
    - One of the factors that cause houses in the data set to sell for different amounts of money is the fact that houses come in various sizes.
    - A preliminary model might posit that the average value per square foot of a new house is $40 and that the average lot sells for $20,000. The predicted selling price of a house of size X (in square feet) would be: $20{,}000 + 40X$.
    - A house of 2,000 square feet would be estimated to sell for $20{,}000 + 40(2{,}000) = \$100{,}000$.

Motivated Example

- Probability Model:
  - We know, however, that this is just an approximation, and the selling price of this particular house of 2,000 square feet is not likely to be exactly $100,000.
  - Prices for houses of this size may actually range from $50,000 to $150,000.
  - In other words, the *deterministic* model is not really suitable. We should therefore consider a *probabilistic* model.
- Let Y be the actual selling price of the house. Then

$$Y = 20{,}000 + 40x + \varepsilon,$$

where ε (Greek letter epsilon) represents a random error term (which might be positive or negative).

  - If the error term ε is usually small, then we can say the model is a good one.
  - The random term, in theory, accounts for all the variables that are not part of the model (for instance, lot size, neighborhood, etc.).
  - The value of ε will vary from sale to sale, even if the house size remains constant. That is, houses of the exact same size may sell for different prices.

Regression Model

- The variable we are trying to predict (Y) is called the dependent (or response) variable.
- The variable x is called the independent (or predictor, or explanatory) variable.
- Our model assumes that

$$E(Y \mid X = x) = \beta_0 + \beta_1 x \qquad \text{(the "population line")} \qquad (1)$$

The interpretation is as follows:

  - When X (house size) is fixed at a level x, then we assume the mean of Y (selling price) to be linear around the level x, where β_0 is the (unknown) intercept and β_1 is the (unknown) slope or incremental change in Y per unit change in X.
  - β_0 and β_1 are not known exactly, but are estimated from sample data. Their estimates are denoted b_0 and b_1.

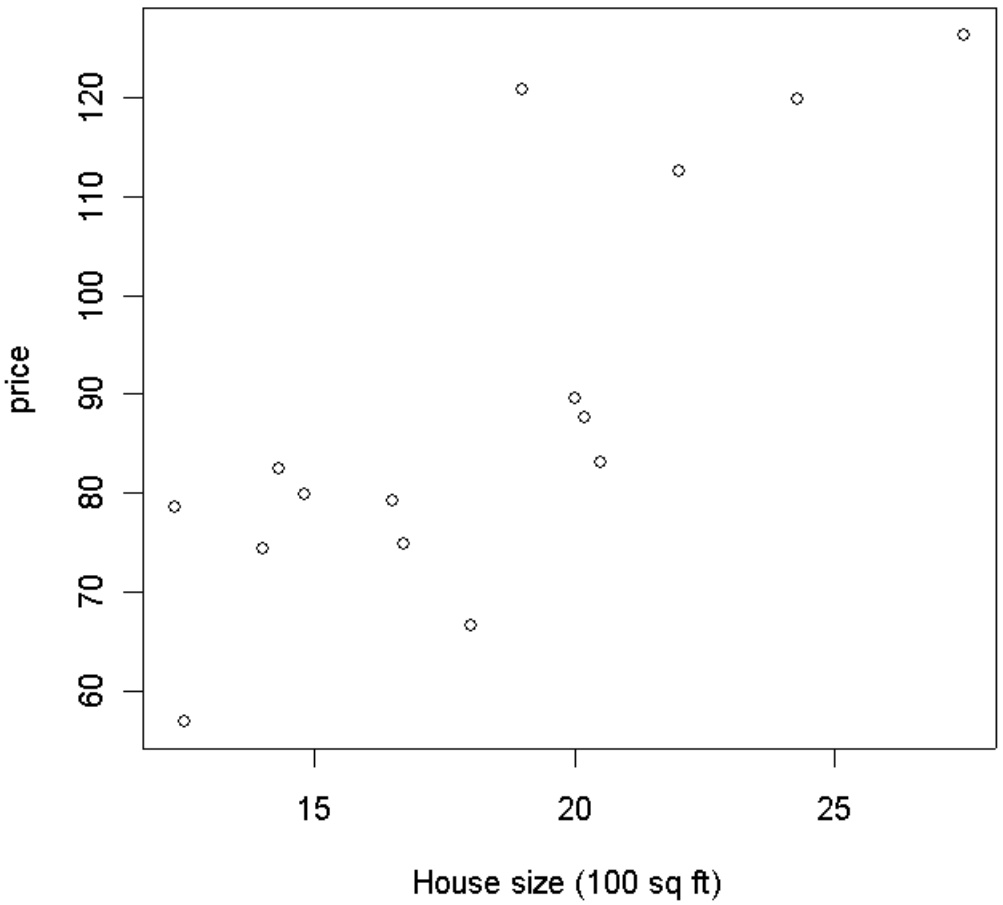
- A *simple regression* model: Consider a model with only one independent variable,.
- A *multiple regression* model: a model with multiple independent variables.

House Number	*Y:* Actual Selling Price ($1,000s)	*X*: House Size (100s ft2)
1	89.5	20.0
2	79.9	14.8
3	83.1	20.5
4	56.9	12.5
5	66.6	18.0
6	82.5	14.3
7	126.3	27.5
8	79.3	16.5
9	119.9	24.3
10	87.6	20.2
11	112.6	22.0
12	120.8	.019
13	78.5	12.3
14	74.3	14.0
15	74.8	16.7
Averages	88.84	18.17

Sample 15 houses from the region.

Least Squares Estimation

- price<- c(89.5,79.9,83.1,56.9,66.6,82.5,126.3,79.3,119.9,87.6,112.6,120.8,78.5,74.3,74.8)
- size<- c(20.0,14.8,20.5,12.5,18.0,14.3,27.5,16.5,24.3,20.2,22.0,19.0,12.3,14.0,16.7)
- plot(size,price,xlab= "House size (100 sq ft)",ylab="Selling price ($1,000)" main="House Size (X) vs Selling Price (Y)")



Assumptions

- These data do not form a perfect line. This is not surprising, considering that our data are random. In other words, if we assume equation (1) then our line predicts the mean for any given level x. However, when we actually take a measurement (i.e., observe the data), we observe:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \text{for } i = 1,2,\ldots, n = 15,$$

where ε_i is the random error associated with the ith observation.

  - Since we don't know the true values of β_0 and β_1, it is clear that we do not observe the actual errors (ε_i) precisely either.

- Assumptions about the Error
  - $E(\varepsilon_i) = 0$ for $i = 1, 2,...,n$.
  - $\sigma(\varepsilon_i) = \sigma_\varepsilon$ where σ_ε is unknown.
  - The errors are independent, that is, the error in the ith observation is independent of the error observed in the jth observation.
  - The ε_i are normally distributed (with mean 0 and standard deviation σ_ε).

Least Squares Estimation

- Recall β_0 and β_1 are (unknown) population parameters.
  - From the sample data, we will calculate numbers $\hat{\beta}_0$ and $\hat{\beta}_1$ that are estimates of the population parameters.
  - How should these numbers be chosen? For any choice of $\hat{\beta}_0$ and $\hat{\beta}_1$, we can write the following prediction equation

    $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

  - The "hat" is used to denote a value estimated from the model, as opposed to one that is actually observed.
  - For each house in our sample of 15 we could check to see how well this equation works at predicting the actual selling prices. Define e_i to be the error associated with the ith observation. That is:

    $e_i = y_i$ - (estimated selling price)

    These are sometimes called the *residuals* or simply *errors*.
- We will pick the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $\Sigma_i e_i^2$, the sum of the squares of the residuals. This method is often called Least Squares Regression.

Using the Equation

- Method of Least squares leads to that the intercept is 18.354 and the slope is 3.879.
  - How do we predict the selling price of a house of 1,650 square feet?
    - Plug in the value 16.50 (1,650 translated to 100s of square feet) in the regression equation and get *predicted selling price* = 18.354 + 3.879× (16.50) = 82.357.
    - Translate to a dollar amount, i.e., $82,357. This is the best estimate you have of the selling price of this house, that is, without any further information about the house (e.g., neighborhood, number of rooms, lot size, age, etc.).
- Analyzing a Regression
- Estimating the Standard Error
  - From the assumptions about the error, the magnitude of σ_ε should be a good guide to the accuracy of a prediction.
  - The number σ_ε is a population parameter, so we cannot know for certain what its value is.
  - We therefore use an estimate s_ε that is provided in the regression output under the name "standard error of the estimate" or just "standard error."

Making Predictions

- The estimate s_ε is calculated by $(SSE/(n-2))^{1/2}$.
  - The reason why we divide by n - 2 and not n - 1 has to do with the degrees of freedom issue.
  - The value of s_ε gives us some idea of the standard deviation of the errors if the model is used to estimate selling prices. In addition, we will make use of the normality assumption to help us make assessments of a prediction.
- Suppose a house occupies 2,000 square feet. How do we predict the selling price?
  - prediction interval: This is used if our goal is to determine a 95% confidence interval on the actual selling price of the house. A 95% prediction interval for the actual selling price is given by

    $(18.354 + 3.879 \times 20) \pm t_{(n-2,\,0.025)} s_\varepsilon = 95.94 \pm 28.07.$
  - confidence interval: This is used if our goal is to determine a 95% confidence interval on the mean selling price of all houses of this size (2,000 square feet). ($E[Y \mid X = x]$)

    It is $95{,}940 \pm t_{(n-2,\,0.025)} s_\varepsilon / \sqrt{n} = 95.94 \pm 7.25$.
  - In the above examples use the t distribution with n - 2 degrees of freedom. If $n - 2 \geq 30$ then the standard normal distribution can be used instead.

Making Inferences about Coefficients

- To assess the accuracy of the model, it involves determining whether a particular variable like house size has any effect on the selling price.
  - Suppose that when a regression line is drawn it produces a horizontal line. This means the selling price of the house is unaffected by the size of the house.
  - A horizontal line has a slope of 0, so when no linear relationship exists between an independent variable and the dependent variable we should expect to get $\beta_1 = 0$.
  - But of course, we only observe estimate of β_1, which might only be "close" to zero. To systematically determine when β_1 might in fact be zero, we will make inferences about it using our estimate , specifically, we will do hypothesis tests and build confidence intervals.
- **Testing β_1**, we can test any of the following:
  - $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$
  - $H_0 : \beta_1 \geq 0$ versus $H_A : \beta_1 < 0$
  - $H_0 : \beta_1 \leq 0$ versus $H_A : \beta_1 > 0$
- In each case, the null hypothesis can be reduced to H_0: $\beta_1 = 0$. The test statistic in each case is $\left(\hat{\beta}_1 - 0\right) / s_{\hat{\beta}_1}$

Example

- Can we conclude at the 1% level of significance that the size of a house is linearly related to its selling price? Test $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$
  - Note this is a two-sided test, we are interested in whether there is any relationship at all between price and size.
  - Calculate $T = (3.879 - 0) / 0.794 = 4.88$.
  - That is, we are 4.88 standard deviations from 0. So at the 1% level (corresponding to thresholds $\pm t_{(13,\ 0.005)} = \pm 3.012$), we reject H_0.
  - There is sufficient evidence to conclude that house size does linearly affect selling price.
- To get a *p*-value on this we would need to look up 4.88 inside the *t*-table.
  - It is 0.00024 or 0.024%; very small indeed.
- A 95% confidence interval for β_1 is given by $\hat{\beta}_1 \pm t_{(n-2,0.025)} s_{\hat{\beta}_1}$
  - For this example: It is $3.879 \pm (2.160)(0.794) = 3.879 \pm 1.715$.
  - Using the 15 data points, we are 95% confident that every extra square foot increases the price of the house by anywhere from \$21.64 to \$55.94.

Example

- Can we conclude at the 1% level of significance that the size of a house is linearly related to its selling price? Test $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$
  - Note this is a two-sided test, we are interested in whether there is any relationship at all between price and size.
  - Calculate $T = (3.879 - 0) / 0.794 = 4.88$.
  - That is, we are 4.88 standard deviations from 0. So at the 1% level (corresponding to thresholds $\pm t_{(13,\ 0.005)} = \pm 3.012$), we reject H_0.
  - There is sufficient evidence to conclude that house size does linearly affect selling price.
- To get a *p*-value on this we would need to look up 4.88 inside the *t*-table.
  - It is 0.00024 or 0.024%; very small indeed.
- A 95% confidence interval for β_1 is given by $\hat{\beta}_1 \pm t_{(n-2,0.025)} s_{\hat{\beta}_1}$
  - For this example: It is $3.879 \pm (2.160)(0.794) = 3.879 \pm 1.715$.
  - Using the 15 data points, we are 95% confident that every extra square foot increases the price of the house by anywhere from \$21.64 to \$55.94.

Method III: Measuring the Strength of the Linear Relationship

- Consider the following equation:

$$Y_i - \bar{Y} = (\hat{Y} - \bar{Y}) + e_i.$$

  - Squaring both sides and summing over all data points, and after a little algebra, we get:

    $\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y} - \bar{Y})^2 + \sum_i e_i^2$, which we usually rewrite as:

    $SST = SSR + SSE,$ (2)

    where $SST = \sum_i (Y_i - \bar{Y})^2$, $SSR = \sum_i (\hat{Y} - \bar{Y})^2$ and $SSE = \sum_i e_i^2$.

  - Interpretation:
    - SST stands for the "total sum of squares" - this is essentially the total variation in the data set, i.e., the total variation of selling prices.
    - SSR stands for "sum of squares due to regression" - this is the squared variation around the mean of the estimated selling prices. This is sometimes called the total variation explained by the regression.
    - SSE stands for "sum of squares due to error" - this is simply the sum of the squared residuals, and it is the variation in the Y variable that remains unexplained after taking into account the variable X.
  - The interpretation of equation (2) is that the total variation in Y (SST) is made up of two parts: the total variation explained by the regression (SSR) and the remaining unexplained variation (SSE).

Regression Statistics

- Define R^2 = SSR/SST = 1- SSE/SST
  - The fraction of the total variation explained by the regression.
  - R^2 is a measure of the explanatory power of the model.
  - Multiple-R = $(R^2)^{1/2}$ (in one variable case = $|\rho_{XY}|$)
- According to the definition of R^2, adding extraneous explanatory variables will artificially inflate the R^2.
  - We must be careful in interpreting this number.
  - Introducing extra variables can lead to spurious results and can interfere with the proper estimation of slopes for the important variables.
- In order to penalize an excess of variables, we consider the adjusted R^2, which is

  adjusted R^2 = 1- [SSE/(n-k-1)]/[SST/(n-1)] .

  Here n is the number of data and k is the number of explanatory variables.
  - The adjusted R^2 thus divides numerator and denominator by their DF.

How to determine the value of used cars that customers trade in when purchasing new cars?

- Car dealers across North America use the "Red Book" to help them determine the value of used cars that their customers trade in when purchasing new cars.
  - –The book, which is published monthly, lists average trade-in values for all basic models of North American, Japanese and European cars.
  - –These averages are determined on the basis of the amounts paid at recent used-car auctions.
  - –The book indicates alternative values of each car model according to its condition and optional features, but it does not inform dealers how the odometer reading affects the trade in value.
- Question: In an experiment to determine whether the odometer reading should be included in the Red Book, an interested buyer of used cars randomly selects ten 3-year-old cars of the same make, condition, and optional features.
  - –The trade-in value and mileage for each car are shown in the following table.

Data

Odometer Reading(1,000 miles) 59 92 61 72 52 67 88 62 95 83

Trade-in Value ($100s) 37 41 43 39 41 39 35 40 29 33

- Run the regression, with Trade-in Value as the dependent variable (*Y*) and Odometer Reading as the independent variable (*X*). The output appears on the following page.
- Regression Statistics
  - Multiple $R = 0.893$, $R^2 = 0.798$, Adjusted $R^2 = 0.773$ Standard Error = 2.178
  - Analysis of Variance

	df	SS	MS	F	Significance F
Regression	1	150.14	150.14	31.64	0.000
Residual	8	37.96	4.74		
Total	9	188.10			

  - Testing

	Coeff.	Stnd Error	t-Stat	P-value
Intercept	56.205	3.535	15.90	0.000
x	-0.26682	0.04743	-5.63	0.000

F and *F*-significance

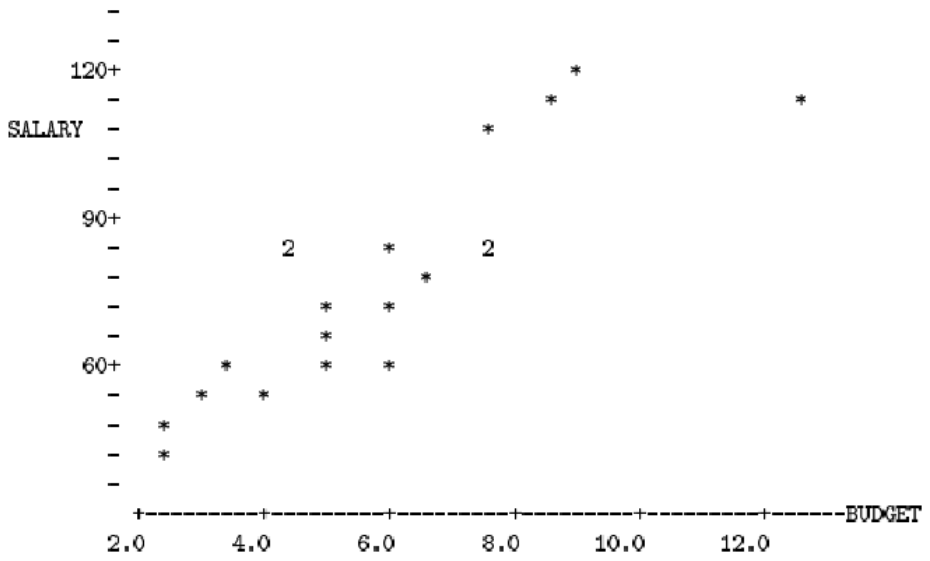
- *F* is a test statistic testing whether the estimated model is meaningful; i.e., statistically significant.
  - *F* =MSR/MSE
  - A large *F* or a small p-value (or *F*-significance) implies that the model is significant.
  - It is unusual not to reject this null hypothesis.

Questions

- What does the regression line tell us about the relationship between the two variables?
- Can we conclude at the 5% significance level that, for all cars of the type described in the experiment, higher mileage results in a lower trade-in value?
- Predict with 95% confidence the trade-in value of such a car that has been driven 60,000 miles.
- A large national courier company has a policy of selling its cars when the odometer reading reaches 75,000 miles. The company is about to sell a large number of 3-year-old cars, each equipped with the same optional features and in the same condition as the 10 cars described in the experiment. The company president would like to know the cars' mean trade-in price. Determine the 95% confidence interval estimate of the expected value of all cars that have been driven 75,000 miles.

Salary-budget Example

- A large corporation is concerned about maintaining parity in salary levels across different divisions.
  - –As a rough guide, it determines that managers responsible for comparable budgets in different divisions should have comparable compensation.
- Data Analysis: The following is a list of salary levels for 20 managers and the sizes of the budgets they manage.
  - –salary<- c(59.0,67.4,50.4,83.2,105.6,86.0,74.4,52.2,82.6,59.0,44.8,111.4,122.4, 82.6,57.0,70.8,54.6,111.0,86.2,79.0)
  - –budget<- c(3.5,5.0,2.5,6.0,7.5.4.5.6.0.4.0.4.5.5.0.2.5.12.5.9.0.7.5.6.0. 5.0,3.0, 8.5, 7.5, 6.5)
  - –Salary *Y* ($1000s)
  - –Budget *X* ($100,000s)



Salary-budget Example

- Want to fit a straight line to this data.
  - The slope of this line gives the marginal increase in salary with respect to increase in budget responsibility.
  - The regression equation is SALARY = 31.9 + 7.73 BUDGET
  - Each additional $100,000 of budget responsibility translates to an expected additional salary of $7,730.
  - If we wanted to know the average salary corresponding to a budget of 6.0, we get a salary of 31.9 + 7.73(6.0) = 78.28.
- Why is the least squares criterion the correct principle to follow?
- Assumptions Underlying Least Squares
  - The errors $\varepsilon_1,\ldots, \varepsilon_n$ are independent of the values of $X_1,\ldots,X_n$.
  - The errors have expected value zero; i.e., $E[\varepsilon_i] = 0$.
  - All the errors have the same variance: $Var[\varepsilon_i] = \sigma^2$, for all i = 1,…,*n*.
  - The errors are uncorrelated; i.e., $Corr[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$.
- The first two assumptions imply that $E[Y|X = x] = \beta_0 + \beta_1 x$.
  - Do we necessarily believe that the variability in salary levels among managers with large budgets is the same as the variability among managers with small budgets?

How do we evaluate and use the regression line?

- Evaluate the explanatory power of a model.
  - Without using *X*, how do we predict *Y*?
  - Determine how much of the variability in *Y* values is explained by the *X*.
- Measure variability using sums of squared quantities.
- The ANOVA table.
  - ANOVA is short for analysis of variance.
  - This table breaks down the total variability into the explained and unexplained parts.
  - Total SS (9535.8) measures the total variability in the salary levels.
    - Without using *x*, we will use sample mean to do prediction.
  - The Regression SS (6884.7) is the explained variation.
    - It measures how much variability is explained by differences in budgets.
  - Error SS (2651.1) is the unexplained variation.
    - This reflects differences in salary levels that cannot be attributed to differences in budget responsibilities.
  - The explained and unexplained variation sum to the Total SS.
- R-squared: R = SSR/SST = 6884.7/9538.8 = 72:2%