Chapter 1
Linear Regression with One Predictor

Professor Dabao Zhang
**Goals of Regression Analysis**

- **Serve three purposes**
  - Describes an association between $X$ and $Y$
    - In some applications, the choice of which variable is $X$ and which is $Y$ can be arbitrary
    - Association generally does not imply causality
  - In experimental settings, helps select $X$ to control $Y$ at the desired level
  - Predict a future value of $Y$ at a specific value of $X$

- **Always** need to consider scope of the model
Example: Leaning Tower of Pisa

- Annual measurements of its lean available
- Measured in tenths of a mm > 2.9 meters
- Prior to recent repairs, its lean was increasing over time
- Goals:
  - To characterize lean over time
  - To predict future observations
## The Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th>year</th>
<th>lean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>642</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>644</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>656</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>667</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td>673</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>688</td>
</tr>
<tr>
<td>7</td>
<td>81</td>
<td>696</td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>698</td>
</tr>
<tr>
<td>9</td>
<td>83</td>
<td>713</td>
</tr>
<tr>
<td>10</td>
<td>84</td>
<td>717</td>
</tr>
<tr>
<td>11</td>
<td>85</td>
<td>725</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
<td>742</td>
</tr>
<tr>
<td>13</td>
<td>87</td>
<td>757</td>
</tr>
</tbody>
</table>

The Data and Relationship

• Response/Dependent variable: lean (Y)
• Explanatory/Independent variable: year (X)
• Observe lean from 1975 - 1987
• Is there a relationship between Y and X?
To Generate a Scatterplot in SAS

DATA a1; INPUT year lean @@;
CARDS;
75 642 76 644 77 656 78 667 79 673 80 688 81 696 82 698 83 713 84 717 85 725 86 742 87 757 102 .;
;
PROC PRINT DATA=a1; WHERE lean NE .; RUN;

SYMBOL1 V=CIRCLE I=SM70;
PROC GPLOT DATA=a1;
    PLOT lean*year / FRAME; WHERE lean NE .;
RUN;
What is the Trend?

- Should always plot the data first!!!!!
Linear Trend?

SYMBOL1 V=CIRCLE I=r1;
PROC GPLOT DATA=a1;
   PLOT lean*year / FRAME; WHERE lean NE .;
RUN; QUIT;
Straight Line Equation

- Straight line describes “curve” well
- Formula for a straight line
  \[ E[Y] = \beta_0 + \beta_1 X \]
  - \( \beta_0 \) is the intercept
  - \( \beta_1 \) in the slope
- Need to estimate \( \beta_0 \) and \( \beta_1 \)
  i.e. determine their plausible values from the data
- Will use method of least squares
Simple Linear Regression Model

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

- \( \beta_0 \) is the intercept
- \( \beta_1 \) is the slope
- \( \varepsilon_i \) is the \( i^{th} \) random error term
  - Mean 0 \( \iff \mathbb{E}(\varepsilon_i) = 0 \)
  - Variance \( \sigma^2 \) \( \iff \text{Var}(\varepsilon_i) = \sigma^2 \)
  - Uncorrelated \( \iff \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j \)
Features of the Model

• \( Y_i = \) deterministic term + random term
  – deterministic term is \( \beta_0 + \beta_1 X_i \)
  – random term is \( \varepsilon_i \)

• Implies \( Y_i \) is a random variable
  – \( E(Y_i) = \beta_0 + \beta_1 X_i + 0 \)
    \( \rightarrow E(Y) = \beta_0 + \beta_1 X \) (underlying relationship)
  – \( \text{Var}(Y_i) = 0 + \sigma^2 \)
    \( \rightarrow \) variance the same regardless of \( X_i \)
  – \( \text{Cov}(Y_i, Y_j) = \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j \)
Estimation of Regression Function

- Consider deviation of $Y_i$ from $E(Y_i)$
  \[ Y_i - (\beta_0 + \beta_1 X_i) \]

- Method of **least squares**
  - Find estimators of $\beta_0, \beta_1$ which minimize
  \[ Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \]
  - Deviations can be positive or negative
  - Square deviations so contribution positive
  - Calculus of solutions shown on pages 17-18
Estimating the Slope

- $\beta_1$ is the true unknown slope
  - Defines change in $E(Y)$ for change in $X$
    \[
    \beta_1 = \frac{\Delta E(Y)}{\Delta X} \implies \Delta E(Y) = \beta_1 \Delta X
    \]

- $b_1$ is the least squares estimate of $\beta_1$
  \[
  b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
  \]

- When will $b_1$ be negative?
Estimating the Intercept

• $\beta_0$ is the true unknown intercept
  - Defines $E(Y)$ when $X = 0$
    \[ E(Y) = \beta_0 + \beta_1 \times 0 = \beta_0 \]
  - Usually not of interest (scope of model)

• $b_0$ is the least squares estimate of $\beta_0$
  \[ b_0 = \bar{Y} - b_1 \bar{X} \]
  \[ \downarrow \]
  Fitted line goes through $(\bar{X}, \bar{Y})$
Properties of Estimates

• Under the Gauss-Markov theorem, these least squares estimators
  – Are unbiased $\iff E(b_l) = \beta_l, \ l = 0, 1$
  – Have minimum variance among all unbiased linear estimators

• In other words, these estimates are the most precise of any estimator where
  – $b_l$ is of the form $\sum k_i Y_i$
  – $E(b_l) = \beta_l$
Estimated Regression Line

- Using the estimated parameters, the fitted regression line is
  \[ \hat{Y}_i = b_0 + b_1 X_i \]

  where \( \hat{Y}_i \) is the estimated value at \( X_i \)

- Fitted value \( \hat{Y}_i \) is also an estimate of the mean response \( E[Y_i] \)

- Extension of the Gauss-Markov theorem
  - \( E(\hat{Y}_i) = E(Y_i) \)
  - \( \hat{Y}_i \) minimum variance among linear estimators
Example: Leaning Tower of Pisa

Based on the following table

1. Obtain the least squares estimate of $\beta_0$ and $\beta_1$.

2. State the regression function

3. Obtain a point estimate for the year 2002 ($X = 102$)

4. State the expected change in lean over two years
<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>X - X</th>
<th>Y - Y</th>
<th>(X - X)(Y - Y)</th>
<th>(X - X)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>642</td>
<td>-6</td>
<td>-51.6923</td>
<td>310.1538</td>
<td>36</td>
</tr>
<tr>
<td>76</td>
<td>644</td>
<td>-5</td>
<td>-49.6923</td>
<td>248.4615</td>
<td>25</td>
</tr>
<tr>
<td>77</td>
<td>656</td>
<td>-4</td>
<td>-37.6923</td>
<td>150.7692</td>
<td>16</td>
</tr>
<tr>
<td>78</td>
<td>667</td>
<td>-3</td>
<td>-26.6923</td>
<td>80.0769</td>
<td>9</td>
</tr>
<tr>
<td>79</td>
<td>673</td>
<td>-2</td>
<td>-20.6923</td>
<td>41.3846</td>
<td>4</td>
</tr>
<tr>
<td>80</td>
<td>688</td>
<td>-1</td>
<td>-5.6923</td>
<td>5.6923</td>
<td>1</td>
</tr>
<tr>
<td>81</td>
<td>696</td>
<td>0</td>
<td>2.3077</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>82</td>
<td>698</td>
<td>1</td>
<td>4.3077</td>
<td>4.3077</td>
<td>1</td>
</tr>
<tr>
<td>83</td>
<td>713</td>
<td>2</td>
<td>19.3077</td>
<td>38.6154</td>
<td>4</td>
</tr>
<tr>
<td>84</td>
<td>717</td>
<td>3</td>
<td>23.3077</td>
<td>69.9231</td>
<td>9</td>
</tr>
<tr>
<td>85</td>
<td>725</td>
<td>4</td>
<td>31.3077</td>
<td>125.2308</td>
<td>16</td>
</tr>
<tr>
<td>86</td>
<td>742</td>
<td>5</td>
<td>48.3077</td>
<td>241.5385</td>
<td>25</td>
</tr>
<tr>
<td>87</td>
<td>757</td>
<td>6</td>
<td>63.3077</td>
<td>379.8462</td>
<td>36</td>
</tr>
</tbody>
</table>

\[ \sum 1053 \quad 9018 \quad 0 \quad 0 \quad 1696 \quad 182 \]
**Answers**

1. Obtain the least squares estimate of $\beta_0$ and $\beta_1$.

   \[
   b_1 = \frac{1696}{182} = 9.3187 \quad \rightarrow \quad b_0 = \frac{9018}{13} - 9.3187 \frac{1053}{13} = -61.1224
   \]

2. State the regression function

   \[
   \hat{Y}_i = -61.1224 + 9.3187X_i
   \]

3. Obtain a point estimate for the year 2002 ($X = 102$)

   \[
   (\hat{Y}|X = 102) = -61.1224 + 9.3187(102) = 889.3850
   \]

4. State the expected change in lean over two years

   Since the slope is 9.3187, a two unit increase in $X$ results in a $2 \times 9.3187 = 18.6374$ increase in lean
Residuals

- The *residual* is the difference between the observed and fitted value

\[ e_i = Y_i - \hat{Y}_i \]

- This is **not** the error term \( \varepsilon_i = Y_i - E(Y_i) \)

- The \( e_i \) is observable while \( \varepsilon_i \) is not

- Residuals are highly useful in assessing the appropriateness of the model
Properties of Residuals

(1) $\sum e_i = 0$
(2) $\sum e_i^2$ is minimized
(3) $\sum Y_i = \sum \hat{Y}_i$
(4) $\sum X_i e_i = 0$
(5) $\sum \hat{Y}_i e_i = 0$

These properties follow directly from the least squares criterion and normal equations (pg 23-24)
Estimation of Error Variance

• In single population (i.e., ignoring $X$)

$$s^2 = \frac{\sum(Y_i - \bar{Y})^2}{n - 1}$$

– Unbiased estimate of $\sigma^2$
– One df lost by using $\bar{Y}$ in place of $\mu$

• In regression model

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2}$$

– Unbiased estimate of $\sigma^2$
– Two df lost by using $(b_0, b_1)$ in place of $(\beta_0, \beta_1)$
– Also known as the mean square error (MSE)
PROC REG DATA=a1;
    MODEL lean=year / CLB P R;
    OUTPUT OUT=a2  P=pred R=resid;
    ID year;
RUN;

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>15804</td>
<td>15804</td>
<td>904.12</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>11</td>
<td>192.28571</td>
<td>17.48052</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>12</td>
<td>15997</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 4.18097  R-Square 0.9880
Dependent Mean 693.69231  Adj R-Sq 0.9869
Coeff Var 0.60271

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| | 95% Confidence Limits |
|----------|----|----------|----------------|---------|-------|   |                        |
| Intercept| 1  | -61.12088| 25.12982       | -2.43   | 0.0333| 95% Confidence Limits |
| year     | 1  | 9.31868  | 0.30991        | 30.07   | <.0001| 8.63656 | 10.00080          |
### Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>year</th>
<th>lean Value</th>
<th>Mean Predicted</th>
<th>Std Error</th>
<th>Residual</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>642.0000</td>
<td>637.7802</td>
<td>2.1914</td>
<td>4.2198</td>
<td>3.561</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>644.0000</td>
<td>647.0989</td>
<td>1.9354</td>
<td>-3.0989</td>
<td>3.706</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>656.0000</td>
<td>656.4176</td>
<td>1.6975</td>
<td>-0.4176</td>
<td>3.821</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>667.0000</td>
<td>665.7363</td>
<td>1.4863</td>
<td>1.2637</td>
<td>3.908</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td>673.0000</td>
<td>675.0549</td>
<td>1.3149</td>
<td>-2.0549</td>
<td>3.969</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>688.0000</td>
<td>684.3736</td>
<td>1.2003</td>
<td>3.6264</td>
<td>4.005</td>
</tr>
<tr>
<td>7</td>
<td>81</td>
<td>696.0000</td>
<td>693.6923</td>
<td>1.1596</td>
<td>2.3077</td>
<td>4.017</td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>698.0000</td>
<td>703.0110</td>
<td>1.2003</td>
<td>-5.0110</td>
<td>4.005</td>
</tr>
<tr>
<td>9</td>
<td>83</td>
<td>713.0000</td>
<td>712.3297</td>
<td>1.3149</td>
<td>0.6703</td>
<td>3.969</td>
</tr>
<tr>
<td>10</td>
<td>84</td>
<td>717.0000</td>
<td>721.6484</td>
<td>1.4863</td>
<td>-4.6484</td>
<td>3.908</td>
</tr>
<tr>
<td>11</td>
<td>85</td>
<td>725.0000</td>
<td>730.9670</td>
<td>1.6975</td>
<td>-5.9670</td>
<td>3.821</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
<td>742.0000</td>
<td>740.2857</td>
<td>1.9354</td>
<td>1.7143</td>
<td>3.706</td>
</tr>
<tr>
<td>13</td>
<td>87</td>
<td>757.0000</td>
<td>749.6044</td>
<td>2.1914</td>
<td>7.3956</td>
<td>3.561</td>
</tr>
<tr>
<td>14</td>
<td>102</td>
<td>889.3846</td>
<td>6.6107</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PROC GPLOT DATA=a2;
   PLOT resid*year / FRAME VREF=0;
   WHERE lean NE .;
RUN; QUIT;
Normal Error Regression Model

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \]

- \( \beta_0 \) is the intercept
- \( \beta_1 \) in the slope
- \( \varepsilon_i \) is the \( i^{th} \) random error term
  - \( \varepsilon_i \sim N(0, \sigma^2) \) ← NEW
  - Uncorrelated → independent error terms
- Defines distribution of random variable \( Y \)
  \[ Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2) \]
Comments

• The least square estimates are unbiased without the normality assumption

• The normality assumption greatly simplifies the theory of analysis

• The normality assumption makes it easy to construct confidence intervals / perform hypothesis tests

• Most inferences are only sensitive to large departures from normality

• See pages 26-27 for more details
Maximum Likelihood Estimation

- Assumption of Normality gives us more choices of methods for parameter estimation

\[ Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \]

\[ f_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \right\} \]

- Likelihood function \( L = f_1 \times f_2 \times \cdots \times f_n \) (i.e. the joint probability distribution of the observations, viewed as function of parameters)

- Find \( \beta_0, \beta_1 \) and \( \sigma^2 \) which maximizes \( L \)

- Obtain similar estimators \( b_0 \) and \( b_1 \) for \( \beta_0 \) and \( \beta_1 \), but slightly different estimators for \( \sigma^2 \) (see HW#1)
Chapter Review

- Description of Linear Regression Model
- Least Squares & Parameter Estimation
- Fitted Regression Line
- Normality Assumption
- PROC REG in SAS: First Touch