**Topic 2 : Simple Linear Regression**

**Outline**
- Description of linear regression model
- Least Squares
- Fitted regression line
- Residuals

**Leaning Tower of Pisa Example**
- Dependent (response) variable: lean ($Y$)
- Independent (predictor) variable: year ($X$)
- Have $i = 1, 2, \ldots, n = 13$ pairs of $(X_i, Y_i)$
- $Y_i = i^{\text{th}}$ dependent variable
- $X_i = i^{\text{th}}$ independent variable
- Will build a model such that $E(Y_i) = f(X_i)$

**Is Linear Trend Reasonable?**
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 \( \iff \) \( E(\varepsilon_i) = 0 \)
  - Variance \( \sigma^2 \) \( \iff \) \( \text{Var}(\varepsilon_i) = \sigma^2 \)
  - Uncorrelated \( \iff \) \( \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \)

Features of the Model

- \( Y_i = \) constant term + random term
  - constant 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 \( X_i \)
  - \( \text{Cov}(Y_i, Y_j) = \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \)

Estimation of Model Parameters

- Consider deviations 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
  - Squared deviations only contribute positively
  - 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 \( X \)
  \[ \beta_1 = \frac{\Delta E(Y)}{\Delta X} \rightarrow \Delta E(Y) = \beta_1 \Delta X \]
- \( b_1 \) is the least squares estimate of \( \beta_1 \)
  \[ b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \]
- 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 $\beta_0$
  \[ b_0 = \bar{Y} - b_1 \bar{X} \]
- Fitted line goes through $(\bar{X}, \bar{Y})$

**Properties of Estimates**

- **Gauss-Markov** theorem states that in a linear regression these least squares estimators
  - Are unbiased $\implies E(b_i) = \beta_i$
  - Have minimum variance among all unbiased linear estimators
  - BLUE = best linear unbiased estimators
- In other words, these estimates are the most precise of any estimator where
  - $b_i$ is of the form $\sum k_i Y_i$
  - $E(b_i) = \beta_i$
- Note: No distribution for the $\epsilon_i$ has been specified

**Estimated Regression Line**

- The estimated regression line is
  \[ \hat{Y}_i = b_0 + b_1 X_i \]
- $\hat{Y}_i$ is known as the fitted value
- Each fitted value also equals the mean response for that $X_i$ (recall $Y|X_i$ a random variable)
- Extension of the Gauss-Markov theorem
  - $E(\hat{Y}_i) = E(Y_i)$
  - $\hat{Y}_i$ minimum variance among linear estimators

**Example**

The Graduate Chair of Department Z administered a newly designed entrance test to the 30 incoming Master’s students as part of a study to determine whether a student’s grade point average (GPA) at the end of the first year ($Y$) can be predicted from the entrance test score ($X$). The results of the study follow. Assume that the linear regression model is appropriate.

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 an entrance test score of 5.0
4. State the expected change in grade point if the entrance test score were 0.5 units higher
### Answers

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

2. State the estimated regression function

3. Obtain a point estimate for an entrance test score of 5.0

4. State the expected change in grade point if the entrance test score were 0.5 units higher

### Properties of Residuals

- The residuals are the differences between the observed and fitted values
  
  $e_i = Y_i - \hat{Y}_i$

- 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

\[ \begin{array}{cccccc}
X & Y & X - \bar{X} & Y - \bar{Y} & (X - \bar{X})(Y - \bar{Y}) & (X - \bar{X})^2 \\
5.5 & 3.1 & 0.5 & 0.6 & 0.30 & 0.25 \\
4.8 & 2.3 & -0.2 & -0.2 & 0.04 & 0.04 \\
4.7 & 3.0 & -0.3 & 0.5 & -0.15 & 0.09 \\
3.9 & 1.9 & -1.1 & -0.6 & 0.66 & 1.21 \\
4.5 & 2.5 & -0.5 & 0.0 & 0.00 & 0.25 \\
6.2 & 3.7 & 1.2 & 1.2 & 1.44 & 1.44 \\
6.0 & 3.4 & 1.0 & 0.9 & 0.90 & 1.00 \\
5.2 & 2.6 & 0.2 & 0.1 & 0.02 & 0.04 \\
4.7 & 2.8 & -0.3 & 0.3 & -0.09 & 0.09 \\
4.3 & 1.6 & -0.7 & -0.9 & 0.63 & 0.49 \\
4.9 & 2.0 & -0.1 & -0.5 & 0.05 & 0.01 \\
5.4 & 2.9 & 0.4 & 0.4 & 0.16 & 0.16 \\
5.0 & 2.3 & 0.0 & -0.2 & 0.00 & 0.00 \\
6.3 & 3.2 & 1.3 & 0.7 & 0.91 & 1.69 \\
4.6 & 1.8 & -0.4 & -0.7 & 0.28 & 0.16 \\
4.3 & 1.4 & -0.7 & -1.1 & 0.77 & 0.49 \\
5.0 & 2.0 & 0.0 & -0.5 & 0.00 & 0.00 \\
5.9 & 3.8 & 0.9 & 1.3 & 1.17 & 0.81 \\
4.1 & 2.2 & -0.9 & -0.3 & 0.27 & 0.81 \\
4.7 & 1.5 & -0.3 & -1.0 & 0.30 & 0.09 \\
100.0 & 50.0 & 0.0 & 0.0 & 7.66 & 9.12
\end{array} \]
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)

### SAS Proc Reg

```sas
proc reg data=a1;
  model lean=year/clb p r;
  output out=a2 p=pred r=resid;
  id year;
run;
```

```sas
proc gplot data=a2;
  plot resid*year/vref=0;
  where lean ne .;
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</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 693.6923  Adj R-Sq 0.9869
Coeff Var 0.60271

### Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>year</th>
<th>Dep Var</th>
<th>Predicted</th>
<th>Std Error</th>
<th>Value Mean</th>
<th>Predict 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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></td>
</tr>
<tr>
<td>14</td>
<td>113</td>
<td>991.8901</td>
<td>9.9848</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Variable | DF | Estimate | Error | t Value | Pr > |t| |
|----------|----|----------|-------|---------|------|---|
| Intercept| 1  | -61.12088| 25.12982| -2.43   | 0.0333 |
| year     | 1  | 9.31888  | 0.30991| 30.07   | <.0001 |

| Variable | DF | 95% Confidence Limits | |
|----------|----|------------------------|
| Intercept| 1  | -116.43124 -5.81052    |
| year     | 1  | 8.63656 10.00080       |
Normal Error Regression Model

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

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

Maximum Likelihood 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 \( L = f_1 \times f_2 \times \cdots \times f_n \)
- Find \( \beta_0, \beta_1 \) and \( \sigma^2 \) which maximizes \( L \)
- Obtain similar estimators \( b_0 \) and \( b_1 \)
- Estimate of \( \sigma^2 \) is different

Normal Error Model

- Normal error assumption greatly simplifies the theory of analysis
- Sampling distributions used to construct confidence intervals / perform hypothesis tests follow known distributions (e.g., \( t, F \))
- While not always true in practice, most inference only sensitive to large departures from normality
- See pages 31-32 for more details

Background Reading

- Appendix A
- KNNL Chapters 1 and 2
- SAS template file \texttt{pisa.sas}