STAT 525

Chapter 1
Linear Regression with One Predictor

Dr. Qifan Song

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

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

Data taken from Exercise 10.8, p698 in Moore and McCabe, *Intro to the Practice of Statistics*, 3rd ed.

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;
```

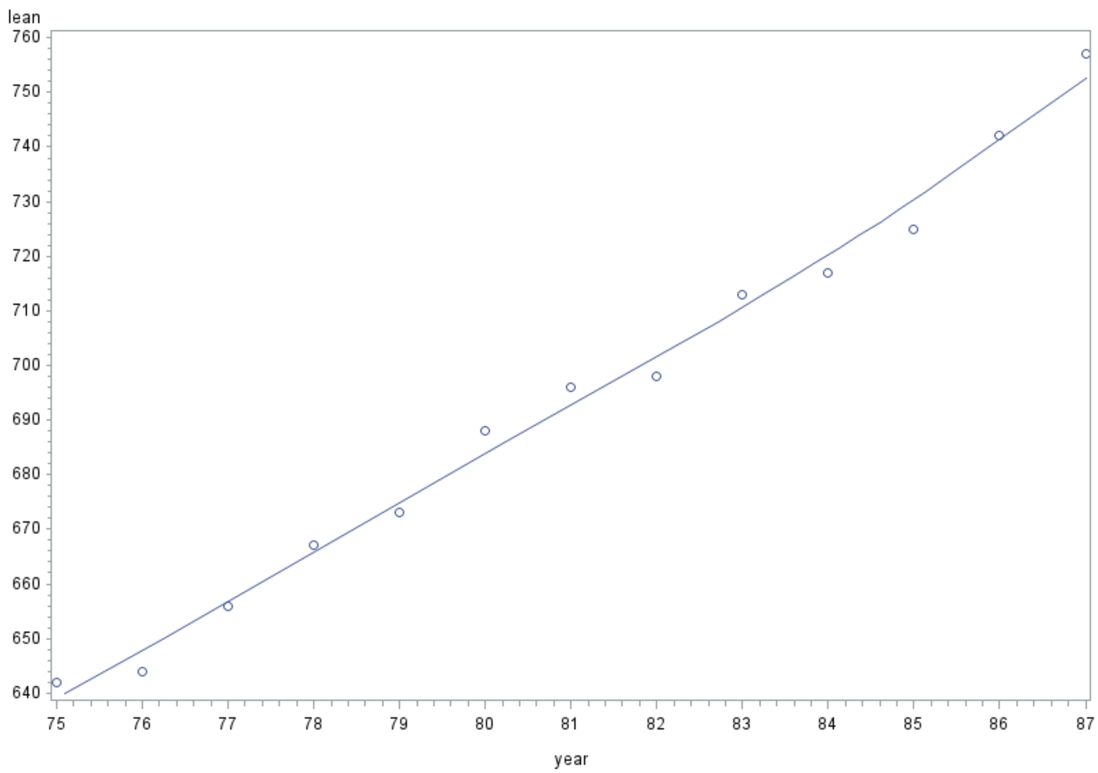
Comments:

`@@`: double trailing @ allows creating multiple observations from a single record.

`SM<70><S>`: plot smooth curve with smoothness level 70 (0-99).

What is the Trend?

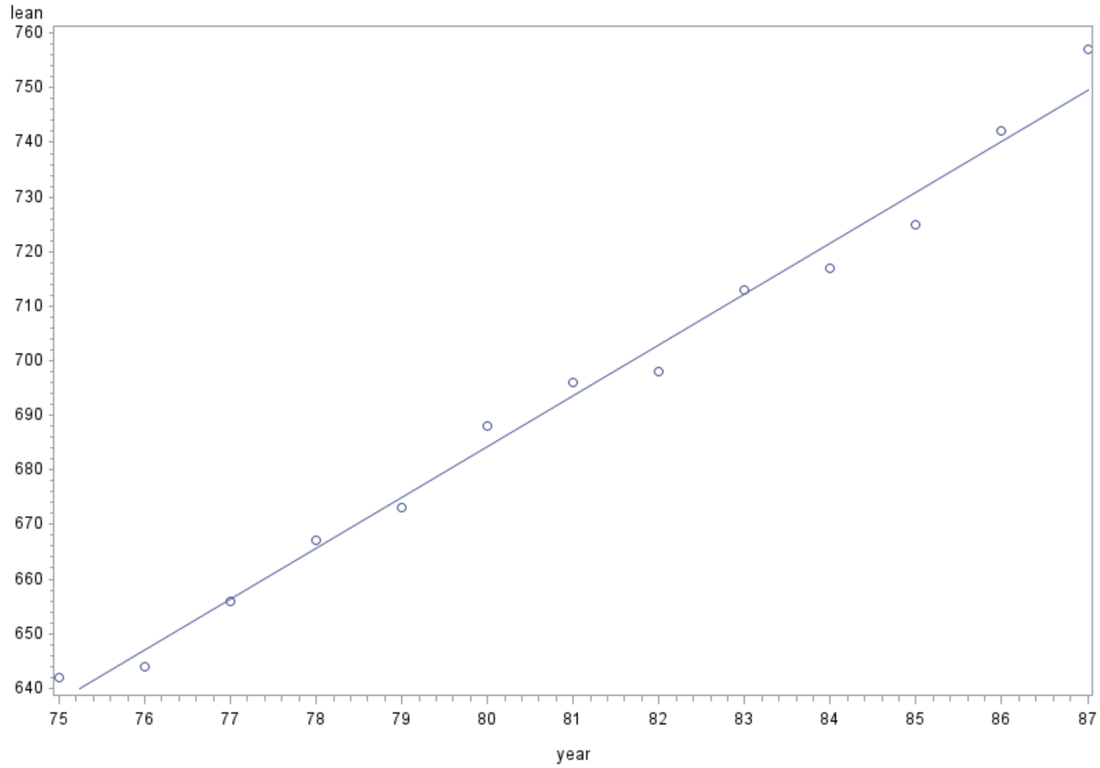
- Should always plot the data first!
- A first look at how Y changes as X is varied is available from a scatterplot.
- Helps to visually detect abnormalities in data set.



Linear Trend?

```
SYMBOL1 V=CIRCLE I=rl;
PROC GPLOT DATA=a1;
     PLOT lean*year / FRAME; WHERE lean NE .;
RUN;
```

`R<L/Q/C>`: plot a Linear Regression result.



Straight Line Equation

- Straight line describes “curve” well
- Formula for a straight line

$$E(Y_i) = \beta_0 + \beta_1 X_i, \text{ or } E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

  - β_0 is the intercept
  - β_1 is the slope

- Need to **estimate** β_0 and β_1
  i.e. determine their plausible values from the data
- Will use method of **least squares** (OLS estimator).

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_0 is the intercept
- β_1 is the slope
- ε_i is the i^{th} random error term
  - Mean 0, i.e. $E(\varepsilon_i) = 0$
  - Constant Variance σ^2, i.e. $Var(\varepsilon_i) = \sigma^2$
  - Uncorrelated, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$
  - Independent to X_i if X_i is random

Features of the Model

- Y_i = deterministic term ($\beta_0 + \beta_1 X_i$) + random term (ε_i)
- Implies Y_i is a random variable (for both fixed or random X_i cases)
  - $E(Y_i) = \beta_0 + \beta_1 X_i + 0$, or
    $E(Y_i | X_i = x) = \beta_0 + \beta_1 x + 0$ (underlying relationship)
  - $Var(Y_i) = 0 + \sigma^2$, or
    $Var(Y_i | X_i = x) \equiv \sigma^2$ (constant variance)
  - $Cov(Y_i, Y_j) = Cov(\varepsilon_i, \varepsilon_j) = 0$, or
    $Cov(Y_i, Y_j | X_i = x, X_j = x') \equiv 0$.
- For simplicity, we assume X_i's are fixed unless specified otherwise.

Estimation of Regression Function

- Consider the deviation of observed data Y_i from a straight line with slope a and intercept b,

$$Y_i - (aX_i + b)$$

  it measures how good the line $ax + b$ fits the data (X_i, Y_i) in terms of vertical distance

- Method of least squares (smallest sum of squared derivation)

  – Find the value of a and b which minimize

$$Q = \sum_{i=1}^{n} [Y_i - (aX_i + b)]^2$$

  – Motivated by $E(Y) = \arg\min_b E(Y - b)^2 \approx \arg\min_b \sum (Y_i - b)^2/n$.

Estimating/Interpreting the Slope

- β_1 is the true unknown slope
  - Defines change in $E(Y)$ for change in X, i.e.,

$$\beta_1 = \frac{\Delta E(Y)}{\Delta X}$$

- b_1 is the least squares estimate of β_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 or positive?

Estimating/Interpreting the Intercept

- β_0 is the true unknown intercept
  - β_0 is the expected value of Y under $X = 0$

$$E(Y) = \beta_1 X + \beta_0 = \beta_1 \times X + \beta_0 = \beta_0$$

  - Sometimes not of interest (scope of model)

- b_0 is the least squares estimate of β_0

$$b_0 = \overline{Y} - b_1 \overline{X}$$

  that is, the fitted line goes through $(\overline{X}, \overline{Y})$.

Properties of Estimates

- Under the *Gauss-Markov* theorem, these least squares estimators
  - are **unbiased**, $E(b_i) = \beta_i$
  - Have **minimum variance** among all unbiased linear estimators
- In other words, these estimates are the most precise among all estimators which satisfy
  - b_i is of form $\sum k_i Y_i$
  - $E(b_i) = \beta_i$

Estimated Regression Line

- Using the estimated parameters, the fitted regression line is

$$\widehat{Y}_i = b_0 + b_1 X_i$$

  where $\widehat{Y}_i$ is the estimated value at X_i (Fitted value).

- Fitted value $\widehat{Y}_i$ is also an estimate of the mean response $E(Y_i)$
- Extension of the Gauss-Markov theorem
  - $E(\widehat{Y}_i) = E(Y_i)$
  - $\widehat{Y}_i$ has minimum variance among all unbiased linear estimators

Example: Leaning Tower of Pisa

Based on the following table

1. Obtain the least squares estimate of β_0 and β_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

	X	Y	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})(Y-\overline{Y})$	$(X-\overline{X})^2$
	75	642	-6	-51.6923	310.1538	36
	76	644	-5	-49.6923	248.4615	25
	77	656	-4	-37.6923	150.7692	16
	78	667	-3	-26.6923	80.0769	9
	79	673	-2	-20.6923	41.3846	4
	80	688	-1	-5.6923	5.6923	1
	81	696	0	2.3077	0	0
	82	698	1	4.3077	4.3077	1
	83	713	2	19.3077	38.6154	4
	84	717	3	23.3077	69.9231	9
	85	725	4	31.3077	125.2308	16
	86	742	5	48.3077	241.5385	25
	87	757	6	63.3077	379.8462	36
$\sum$	1053	9018	0	0	1696	182

Answer

1. Obtain the least squares estimate of β_0 and β_1.

   $$b_1 = \frac{1696}{182} = 9.3187, \quad b_0 = \frac{9018}{13} - 9.3187\frac{1053}{13} = -61.1224$$

2. State the regression function

   $$\widehat{Y}_i = -61.1224 + 9.3187 X_i$$

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

   $$(\widehat{Y}|X = 102) = -61.1224 + 9.3187(102) = 889.3850$$

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 values

$$e_i = Y_i - \widehat{Y}_i$$

- This is not the error term $\varepsilon_i = Y_i - E(Y_i)$
- The e_i is observable while ε_i is not
- Residuals are highly useful in assessing the appropriateness of the model

Properties of Residuals

- $\sum e_i = 0$
- $\sum e_i^2$ are minimized
- $\sum Y_i = \sum \widehat{Y}_i$
- $\sum X_i e_i = 0$
- $\sum \widehat{Y}_i e_i = 0$

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

ReML Estimation of Error Variance

- In single population (i.e., ignoring X)

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

  - unbiased estimation
  - One df lost by using $\overline{Y}$ in place of μ_Y

- In regression model

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

  - unbiased estimation
  - Two df lost by using (b_0, b_1) in place of (β_0, β_1)
  - Also known as the *mean square error* (MSE)

PROC REG in SAS: Leaning Tower of Pisa

```
PROC REG DATA=a1;
   MODEL lean=year / CLB P R;
   OUTPUT OUT=a2 P=pred R=resid;
   ID year;
RUN;
```

```
               Analysis of Variance
                          Sum of         Mean
Source              DF       Squares       Square     F Value     Pr > F
Model                1         15804        15804      904.12     <.0001
Error               11     192.28571     17.48052
Corrected Total     12         15997

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

              Parameter Estimates
                Parameter     Standard
Variable    DF   Estimate        Error    t Value    Pr > |t| 95% Confidence Limits
Intercept    1  -61.12088     25.12982      -2.43      0.0333  -116.43124   -5.81052
year         1    9.31868      0.30991      30.07      <.0001     8.63656   10.00080
```

Output Statistics

Obs	year	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual
1	75	642.0000	637.7802	2.1914	4.2198	3.561
2	76	644.0000	647.0989	1.9354	-3.0989	3.706
3	77	656.0000	656.4176	1.6975	-0.4176	3.821
4	78	667.0000	665.7363	1.4863	1.2637	3.908
5	79	673.0000	675.0549	1.3149	-2.0549	3.969
6	80	688.0000	684.3736	1.2003	3.6264	4.005
7	81	696.0000	693.6923	1.1596	2.3077	4.017
8	82	698.0000	703.0110	1.2003	-5.0110	4.005
9	83	713.0000	712.3297	1.3149	0.6703	3.969
10	84	717.0000	721.6484	1.4863	-4.6484	3.908
11	85	725.0000	730.9670	1.6975	-5.9670	3.821
12	86	742.0000	740.2857	1.9354	1.7143	3.706
13	87	757.0000	749.6044	2.1914	7.3956	3.561
14	102	.	889.3846	6.6107	.	.

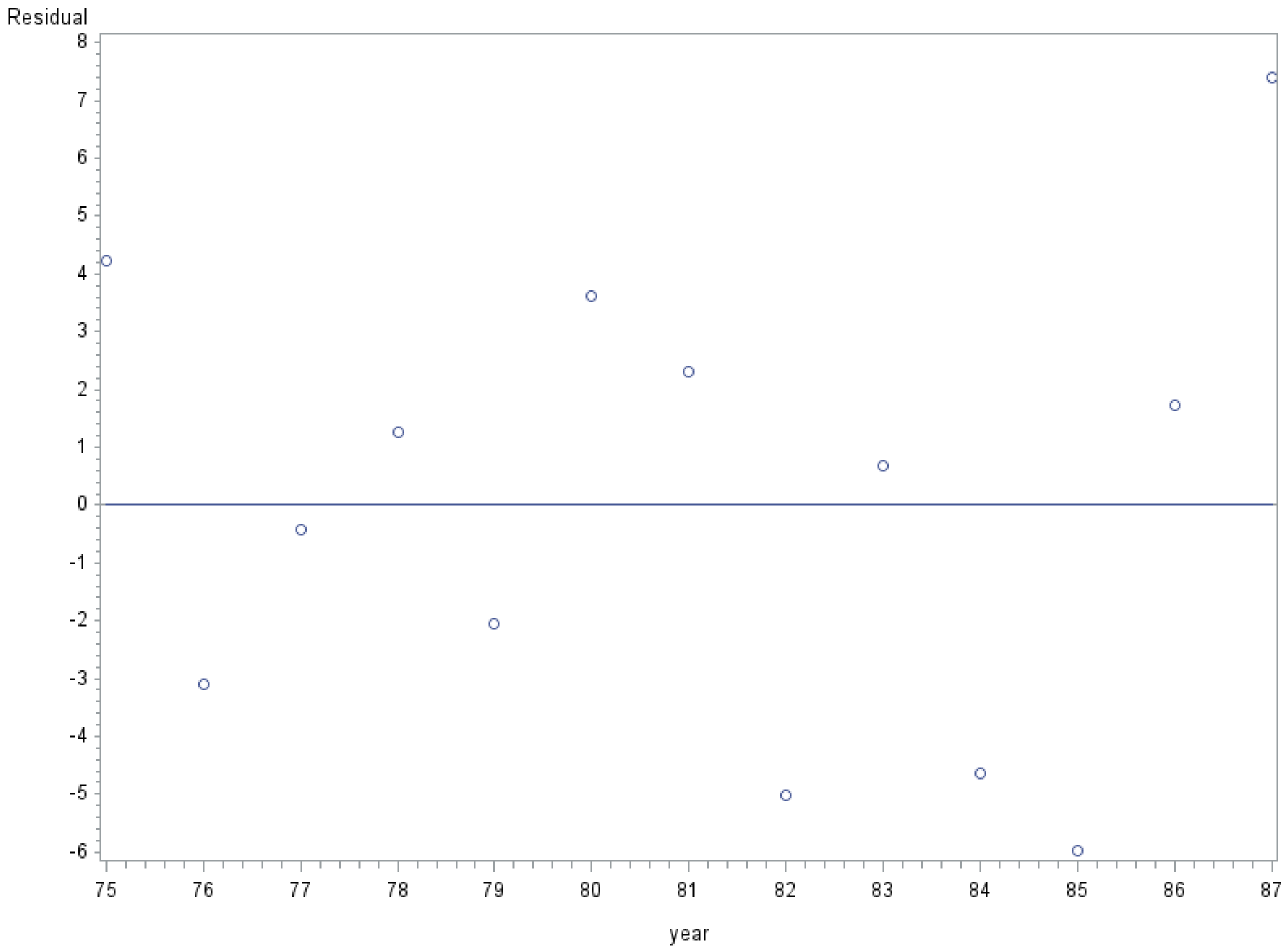
Comments:

`CLB`: confidence intervals for the β estimations

`P R`: display predicted/fitted values and residual values

`output`: Output statistics into a new data set

```
PROC GPLOT DATA=a2;
PLOT resid*year / FRAME VREF=0;
WHERE lean NE .;
RUN;
```



Normal Error Regression Model

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

- β_0 is the intercept
- β_1 is the slope
- the random error term is assumed to be **independent normally** distributed
- Defines distribution of random variable Y_i

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

Comments

- **The normality assumption doesn't affect the validity and unbiasedness of the least square estimators**
- The normality assumption will greatly simplifies the theory of analysis beyond estimations
- 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

Comments

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

$$
\begin{aligned}
f_i = & \quad \text{the likelihood of } Y_i \\
= & \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right\}
\end{aligned}
$$

- Likelihood function $L = \prod f_i$ (i.e. the joint probability distribution of the observations, viewed as function of parameters)
- Find β_0, β_1 and σ^2 which maximizes L.
- Obtain similar estimators b_0 and b_1 for β_0 and β_1 , but slightly different estimators for σ^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