Topic 2 : Simple Linear Regression

STAT 525 - Fall 2013

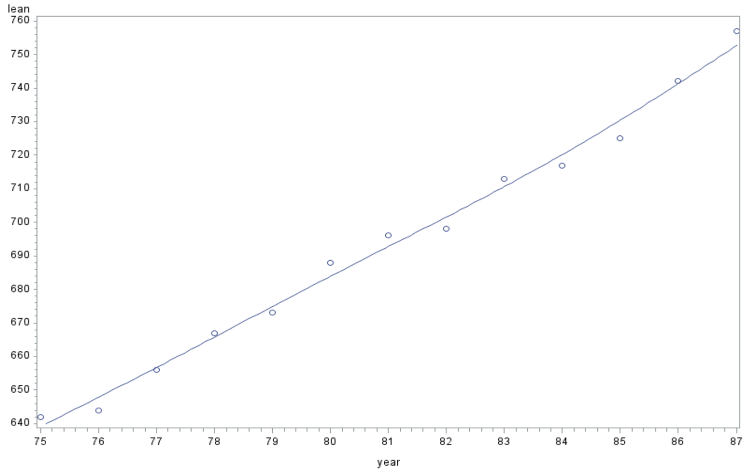
Outline

- 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

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

- β_0 is the intercept
- β_1 is the slope
- ε_i is the i^{th} random error term
  - Mean 0 $\longleftrightarrow E(\varepsilon_i) = 0$
  - Variance $\sigma^2 \longleftrightarrow \mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$
  - Uncorrelated $\longleftrightarrow \mathrm{Cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = 0$

Features of the Model

- Y_i = constant term + random term
  - constant term is $\beta_0 + \beta_1 X_i$
  - random term is ε_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)
  - $\mathrm{Var}(Y_i) = 0 + \sigma^2$

    $\rightarrow$ variance the same regardless of X_i
  - $\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = 0$

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 β_0, β_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

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

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

- b_1 is the least squares estimate of β_1

$$b_1 = \frac{\sum\limits_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}$$

- When will b_1 be negative?

Estimating the Intercept

- β_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 β_0

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

$$\downarrow$$

Fitted line goes through $(\overline{X}, \overline{Y})$

Properties of Estimates

- Gauss-Markov theorem states that in a linear regression these least squares estimators
  - Are **unbiased** $\longleftrightarrow E(b_l) = \beta_l$
  - 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_l is of the form $\sum k_i Y_i$
  - $E(b_l) = \beta_l$
- Note: No distribution for the ε_i has been specified

Estimated Regression Line

- The estimated regression line is

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

where $\hat{Y}_i$ 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 β_0 and β_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

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

Answers

1. Obtain the least squares estimates of β_0 and β_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

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 ε_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)

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

  - Unbiased estimate of σ^2
  - One df lost by using $\overline{Y}$ in place of μ

- In regression model

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

  - Unbiased estimate of σ^2
  - Two df lost by using (b_0, b_1) in place of (β_0, β_1)
  - Also known as the *mean square error* (MSE)

SAS Proc Reg

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

proc gplot data=a2;
   plot resid*year/vref=0;
   where lean ne .;
run;
```


Analysis of Variance

Source	DF	Sum of Squares	Mean 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		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-61.12088	25.12982	-2.43	0.0333
year	1	9.31868	0.30991	30.07	<.0001

Variable	DF	95% Confidence Limits	
Intercept	1	-116.43124	-5.81052
year	1	8.63656	10.00080

Output Statistics

Obs	year	Dep Var lean	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	113	.	991.8901	9.9848	.	.

Normal Error Regression Model

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

- β_0 is the intercept
- β_1 in the slope
- ε_i is the i^{th} random error term
  - $\varepsilon_i \sim \mathrm{N}(0, \sigma^2) \longleftarrow$ **NEW**
  - Uncorrelated $\longrightarrow$ independent error terms
- Defines distribution of random variable Y

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

Maximum Likelihood Estimation

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 \mathrm{X_i}, \sigma^2)$$
$$\downarrow$$
$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right\}$$

- Likelihood function $L = f_1 \times f_2 \times \cdots \times f_n$
- Find β_0, β_1 and σ^2 which maximizes L
- Obtain similar estimators b_0 and b_1
- Estimate of σ^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 pisa.sas