Introduction to Statistics
Xiamen Academic Program

Dr. Jason A. Osborne

Friday, July 27, 2012
Topics we’ve covered

▶ Designed experiments vs observational studies
▶ Types of data (ordinal, continuous, etc), descriptive statistics
▶ Elementary probability, random variables, probability densities, mean ($E(Y)$) and variance ($V(Y)$).
▶ Sampling distributions of statistics, standard errors, central limit theorem.
▶ Hypothesis testing. $H_0 : \mu = \mu_0$, $H_0 : \mu_1 = \mu_2$, etc
▶ Analysis of variance $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t$
▶ Regression

Software we’ve seen

▶ SAS, JMP, R, Some applets

Software we haven’t seen

▶ SPSS, Systat, STATCRUNCH, S+
Simple linear regression of human height ($y$) on parent height ($x$)
Simple linear regression model

Let $x_1, \ldots, x_n$ denote the midparent heights. Let $Y_1, \ldots, Y_n$ denote the offspring’s heights (women multiplied by 1.08). Given $X = x_i,$

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \ldots, n(n = 928)$$

where $E_1, \ldots, E_n$ are

- independent,
- identically and
- normally distributed RVs with mean 0 and error variance $\sigma^2.$

(Write $E_i \overset{iid}{\sim} N(0, \sigma^2).$)

This implies

1. $\mu(x) = E(Y|X = x) = \beta_0 + \beta_1 x$
2. $\text{Var}(Y|X = x) = \sigma^2$ (Three unknown parameters $\beta_0, \beta_1, \sigma^2$ quantify the whole population of interest.)
Exercise: ignoring $x$, construct 95% c.i. for $\mu$. (Use output)

Many other questions to answer using regression analysis:

1. What is the meaning, in words, of $\beta_1$?

2. T/F: (a) $\beta_1$ is a statistic (b) $\beta_1$ is a parameter (c) $\beta_1$ is unknown.

3. What is the observed value of $\hat{\beta}_1$?

4. T/F: (a) $\hat{\beta}_1$ is a statistic (b) $\hat{\beta}_1$ is a parameter (c) $\hat{\beta}_1$ is unknown.

5. Is $\hat{\beta}_1 = \beta_1$?

6. How much does $\hat{\beta}_1$ vary about $\beta_1$ from sample to sample?

7. What values of $\beta_1$ are plausible in light of data?

8. What is the line that best fits these data, using the criterion that smallest sum of squared residuals is “best?”

9. How much of the observed variation in the heights of sons (the $y$-axis) is explained by this “best” line?

10. What is the estimated average height of sons whose midparent height is $x = 68$?
11. Is this the true average height in the whole population of sons whose midparent height is \( x = 68 \)?

12. Under the model, what is the true average height of sons with midparent height \( x = 68 \)?

13. What is the estimated std. dev. among the population of sons whose parents have midparent height \( x = 68 \)? Would you call this standard deviation a “standard error?”

14. What is the estimated standard deviation among the population of sons whose parents have midparent height \( x = 72 \)? Bigger, smaller, or the same as that for \( x = 68 \)? Is your answer obviously supported or refuted by inspection of the scatterplot?

15. What is the estimated standard error of the estimated average for sons with midparent height \( x = 68 \), \( \hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1 \)? Provide an expression for this standard error.

16. Is the estimated standard error of \( \hat{\mu}(72) \) bigger, smaller, or the same as that for \( \hat{\mu}(68) \)?
options ls=75 nodate;

data Galton;
  array cdata(14);
  if _n_ = 1 then input cdata1-cdata14 @ ;
  retain cdata1-cdata14; drop cdata1-cdata14 i;
  input parent @;
    do i = 1 to 14; input count @ ; son=cdata(i);
    output; end;
cards;
    61.7 62.2 63.2 64.2 65.2 66.2 67.2 68.2 69.2 70.2 71.2 72.2 73.2 73.7
73.0  0  0  0  0  0  0  0  0  0  1  3  0
72.5  0  0  0  0  0  1  2  1  2  7  2  4
71.5  0  0  0  1  3  4  3  5  10  4  9  2  2
70.5  1  0  1  1  3  12 18 14  7  4  3  3
69.5  0  0  1  16  4  17 27 20 33 25 20 11 4  5
68.5  1  0  7 11 16 25 31 34 48 21 18 4  3  0
67.5  0  3  5 14 15 36 38 28 38 19 11 4  0  0
66.5  0  3  3  5  2 17 17 14 13  4  0  0  0  0
65.5  1  0  9  5  7 11 11 7  7  5  2  1  0  0
64.5  1  1  4  4  1  5  5  0  2  0  0  0  0  0
64.0  1  0  2  4  1  2  2  1  1  0  0  0  0  0

;data big; set galton; drop j count;
  do j=1 to count;output; end;
proc print data=big(obs=20);

proc means; var son parent;
proc reg;
  model son=parent/clb;
  output out=out1 residual=r p=yhat ucl=pihigh lcl=pilow uclm=cihigh lclm=cilow
  stdp=stdmean;
The SAS System

The MEANS Procedure

Variable N Mean Std Dev Minimum Maximum
---------------------------------------------------------------------
son 928 68.0884698 2.5179414 61.7000000 73.7000000
parent 928 68.3081897 1.7873334 64.0000000 73.0000000
---------------------------------------------------------------------

The REG Procedure

Analysis of Variance

Sum of Mean
Source DF Squares Square F Value Pr > F
Model 1 1236.93401 1236.93401 246.84 <.0001
Error 926 4640.27261 5.01109
Corrected Total 927 5877.20663

Root MSE 2.23855 R-Square 0.2105
Dependent Mean 68.08847 Adj R-Sq 0.2096
Coeff Var 3.28770

Parameter Standard
Variable DF Estimate Error t Value Pr > |t|
Intercept 1 23.94153 2.81088 8.52 <.0001
parent 1 0.64629 0.04114 15.71 <.0001

Variable DF 95% Confidence Limits
Intercept 1 18.42510 29.45796
parent 1 0.56556 0.72702

questions regarding prediction, estimation when x=68, x=72

Obs parent son yhat stdmean cilow cihigh pilow pihigh r
1 68 . 67.8893 0.07457 67.7429 68.0356 63.4936 72.2849 .
2 72 . 70.4745 0.16871 70.1434 70.8056 66.0688 74.8801 .
Answers to Galton regression question

1. Change in average son’s height (inches) per one inch increase in midparent height (in the whole population.)

2. $\beta_1$ is an unknown parameter.

3. $\hat{\beta}_1 = 0.65$ son inches/midparent inch (from output.)

4. $\hat{\beta}_1 = 0.65$ is an observed value of a statistic.

5. $\beta_1$ is the slope of the population mean, $\hat{\beta}_1$ is the slope from the SLR of the observed data. $\hat{\beta}_1 = \beta_1$ is unlikely.

6. $\widehat{SE}(\hat{\beta}_1) = \sqrt{MS[E]/S_{xx}} = 0.04$ (from output.)

7. Add and subtract about 2 SE to get (0.57, 0.73)

8. $y = 23.9 + 0.65x$

9. $r^2 = 21\%$

10. $\mu(68) = \hat{\beta}_0 + 68\hat{\beta}_1 = 67.9$ (from output also.)
11. Not sure, as \( \mu(68) = \beta_0 + 68\beta_1 \) is unknown.
12. \( \mu(68) = \beta_0 + 68\beta_1 \).
13. \( \sqrt{MS[E]} = 2.24 \). Not a SE.
14. \( \sqrt{MS[E]} = 2.24 \). (Assume homoscedasticity.)
15. \( SE(\hat{\beta}_0 + 68\hat{\beta}_1) = 0.07 \).
16. \( SE(\hat{\beta}_0 + 72\hat{\beta}_1) = 0.17 \) (bigger). Expressions given by

\[
\hat{SE}(\hat{\mu}(68)) = \sqrt{MS[E]} \left( \frac{1}{n} + \frac{(68 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right)
\]

\[
= \sqrt{(1, 68)'MS[E](X'X)^{-1}(1, 68)}
\]