22S:105
Statistical Methods and Computing

Introduction to Inference for Regression

Lecture 22
Apr. 14, 2006

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Simple Linear Regression

- If a scatterplot suggests a linear relationship between 2 variables, we want to summarize the relationship by drawing a straight line on the plot.
- A *regression line* summarizes the relationship between a response variable and an explanatory variable.
  - Both variables must be quantitative.
- definition: A *regression line* is a straight line that describes how a response variable Y changes as an explanatory variable X changes.
  - often used to predict the value of Y that corresponds to a given value of X.

Idea of linear regression

- We are considering a population for which a response variable and an explanatory variable are of interest.
- Example
  - population: adult Americans
  - response variable: systolic blood pressure (sbp)
  - explanatory variable: age
- Each value of the explanatory variable defines a *subpopulation* of the whole population.
  - example: subpopulations are all 21-yr-olds, all 22-yr-olds, etc.

- Each of the subpopulations has its own mean of the response variable, $\mu_{Y|X=x*}$
  - example: population mean sbp in 21-yr-old Americans is some fixed but unknown number $\mu_{Y|X=21}$
- The means for all these subpopulations lie on a straight line.

Other ideas of linear regression

- The distribution of the response variable in each subpopulation is normal.
  - example: sbp in 21-yr-old Americans has a normal distribution

    sbp in 61-yr-old Americans also follows a normal distribution, but with a different mean ($\mu_{Y|X=61}$)
- The standard deviation of the response variable is the same in all the subpopulations.

What's so great about all this?

We can describe the means of *all* the subpopulations by describing one straight line!

- It takes only 2 numbers to specify a straight line.
- We can use sample data to estimate these 2 numbers.
- The estimated line summarizes the relationship between the two variables in our sample data.
  - similar to how $\bar{x}$ summarizes sample values of a single variable
- We can use the estimated line to predict future values of the response variable based on the explanatory variable.

The population regression line

- We can write the population regression line as

$$\mu_{Y|X=x} = \alpha + \beta x$$

- α and β are unknown population parameters
- β is the *slope* of the line
  - For a 1-unit increase in X, we would expect a change of β units in Y
  - slope is "rise over run"

- α is the *intercept* of the line
  - This is $\mu_{Y|X=0}$
  - Often the notion of a subpopulation for which $X = 0$ is not meaningful.

    Example: There are no adults of age 0!
  - In these cases, consider the intercept to be the number that makes the line fit correctly in the range of observed X values.

Example: Powerboats and manatees in Florida

Data on powerboat registrations (in 1000's) in Florida and the number of manatees killed by boats in Florida.

OBS	YEAR	POWERBT	KILLED
1	1977	447	13
2	1978	460	21
3	1979	481	24
4	1980	498	16
5	1981	513	24
6	1982	512	20
7	1983	526	15
8	1984	559	34
9	1985	585	33
10	1986	614	33
11	1987	645	39
12	1988	675	43
13	1989	711	50
14	1990	719	47

Scatterplot

```
  Plot of KILLED*POWERBT.  Symbol used is '.'.

KILLED |
    60 +
       |
       |
       |                             .
       |                              .
       |                          .
    40 +                       .
       |
       |               .  . .
       |
       |
       |        .  .
    20 +      .    .
       |          . .
       |     .
       |
       |
       |
     0 +
       --+--------+--------+--------+--------+--
        400      500      600      700      800

                        POWERBT
```


Using sample data to estimate the intercept and slope

- We will write an estimated regression line based on sample data as

$$\hat{y} = a + bx$$

- a is the estimated intercept, and b is the estimated slope
- Example: the estimated regression line for the manatees-and-powerboats problem is

$$\hat{y} = -41.4 + 0.125x$$

- This means that for a 1-unit increase in powerboat registrations we would expect 0.125 more manatees to be killed.
  - Since we are measuring powerboat registrations in 1000's, this means for every additional 1000 powerboat registrations, we expect 0.125 more manatees to be killed.

- Note that it makes no sense in this problem to say that the intercept (-41.4) is the number of manatees that we would expect to be killed in a year when there were no powerboat registrations.
- **An estimated regression line is meaningful only for the range of X values actually observed.**
  - In the manatee problem, this is 450 to 725 (thousands). The estimated intercept makes the linear relationship come out right over this range of X values.

Prediction using an estimated regression line

Example: What is the predicted number of manatees killed in a year when there are 600 thousand powerboat registrations?

$$\begin{aligned} \hat{y} &= -41.4 + 0.125(600) \\ &= 33.6 \end{aligned}$$

Least squares: choosing the "best" estimated line

a and b are estimated by choosing a line as follows:

- for each observed value y_i in the sample data, compute the distance from y_i to the line
- square each of the distances
- add up all the squared distances
- choose the line that makes the sum of these squared distances the smallest

Notation

Recall:

- y_i is the observed value of the response variable for subject i
- $\hat{y}_i$ is the value predicted by the regression line for subject i

$$\hat{y}_i = a + bx_i$$

- A **residual** is the difference between an observed value and a predicted value of the response variable.

$$r_i = y_i - \hat{y}_i$$

How well does the regression line predict the response variable

- The **coefficient of determination** or R^2
  - the square of the correlation coefficient between the response variable and the explanatory variable
  - the proportion of the variability among the observed values of the response variable that is explained by the linear regression
- Example: in the manatee data, $R^2 = 0.8864$
  - 88.6% of the variability in number of manatee deaths is explained by number of powerboat registrations

Inference about the slope and intercept

- The least squares estimates of the intercept and slope based on our data are the point estimates of the population intercept and slope.
  - a is the point estimate of the population intercept α
  - b is the point estimate of the population slope β
- As usual, we also need to estimate the variability in our point estimates in order to compute confidence intervals and carry out hypothesis tests.
  - i.e., we need the *standard errors* of a and b
  - These depend on the sample standard deviation of the data

$s_{y|x}$ — the sample standard deviation from regression

- This is the estimate of the common $\sigma_{y|x}$ in all the subpopulations.
- $$\begin{aligned} s &= \sqrt{\frac{1}{n-2}\sum_i residual_i^2} \\ &= \sqrt{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2} \end{aligned}$$
- $n-2$ is the *degrees of freedom*
  - Recall that $\hat{y}_i = a + bx_i$. That is, there are two *estimated* quantities, a and b, involved in calculating the $\hat{y}_i$s.
  - The degrees of freedom is the sample size n minus the number of estimated quantitiees that are involved in calculating the sample standard deviation.

Confidence intervals for the regression slope

- The population slope β usually is the parameter in which we are most interested in regression.
- We need not only the point estimate b but also an interval that expresses the amount of uncertainty in the estimate.
- As usual, the form of the confidence interval is

$$\begin{aligned} estimate &\pm t^* SE_{estimate} \\ b &\pm t^* SE_b \end{aligned}$$

- The standard error of the least-squares slope b is
$$SE_b = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$
- For a two-sided, level C confidence interval, t^* is the upper $\frac{1-C}{2}$ cutoff for a t distribution with $n-2$ degrees of freedom.

Example: the manatee data

```
proc reg data = manatee ;
model killed = powerbt / clb ; /* clb option prints confidence intervals
                                  for regression coefficients */
run ;
```

```
Model: MODEL1
Dependent Variable: KILLED

                          Analysis of Variance

                            Sum of         Mean
 Source           DF       Squares       Square      F Value       Prob>F

 Model             1    1711.97866   1711.97866       93.615       0.0001
 Error            12     219.44991     18.28749
 C Total          13    1931.42857

     Root MSE       4.27639     R-square       0.8864
     Dep Mean      29.42857     Adj R-sq       0.8769
     C.V.          14.53141
```



```
                          Parameter Estimates

                   Parameter      Standard    T for H0:
 Variable  DF       Estimate         Error   Parameter=0    Prob > |T|

 INTERCEP   1     -41.430439    7.41221723        -5.589        0.0001
 POWERBT    1       0.124862    0.01290497         9.675        0.0001

                          Parameter Estimates

            Variable      DF       95% Confidence Limits

            Intercept      1      -57.58027      -25.28060
            powerbt        1        0.09674        0.15298
```

- $s_{y|x} = 4.276$
- The estimated slope $b = 0.1249$.
- $SE_b = 0.0129$
- To construct a 95% confidence interval for the unknown population slope β, we need the upper .025 cutoff for a t distribution with $n - 2 = 12$ degrees of freedom.

  – From Table C, this is 2.179.
- So our 95% confidence interval is

$$\begin{aligned} &0.1249 \pm 2.179(0.0129) \\ &0.1249 \pm 0.02811 \\ (&0.0968, 0.1530) \end{aligned}$$

- We are 95% confident that the unknown population slope β lies in this interval.

Testing the hypothesis of no linear relationship

- We often want to test the null hypothesis that there is no linear relationship between the explanatory variable and the response variable.

$$H_0 : \beta = 0$$

- If the slope is 0, the regression line is horizontal. This says that the means of all the subpopulations are the same! That is, there is no linear relationship (no correlation) between the two variables.
- Usually the alternative hypothesis of interest is two-sided.

$$H_a : \beta \neq 0$$

- The test statistic is a t statistic:

$$t = \frac{b}{SE_b}$$

- The p-value is obtained by comparing the observed value of the t statistic to a t distribution with $n-2$ degrees of freedom.

Example: the manatee data

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-41.430439	7.41221723	-5.589	0.0001
POWERBT	1	0.124862	0.01290497	9.675	0.0001

- Let's carry out the hypothesis test at the $\alpha = .05$ significance level.
- The t statistic value is 9.675, and the p-value is less than 0.0001.
- Therefore, we would have had less than 1 chance in 10,000 of obtaining sample data that produced a t statistic this far away from 0 or farther if the true population slope was 0.