Chap 11 Notes: Inference About Regression

J. Harner A. Billings

Department of Statistics
West Virginia University

Stat 211 Fall 2007

Outline

Hypothesis Test on the Slope of the Regression line

- The sample regression coefficient b_1 estimates the population regression coefficient β_1.
- We can perform tests of hypothesis on β_1.

Example: (Cont. from Sect. 3.3, T.V. - Cholesterol)

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \\ \alpha &= 0.05 \end{aligned}$$

Test Statistic:

$$T = \frac{b_1 - 0}{SE(b_1)} = \frac{b_1}{\left(\frac{S_E}{\sqrt{SS(X)}}\right)}$$

where,

$$S_E = \sqrt{\frac{SS(Y) - b_1^2 SS(X)}{n-2}}.$$

- Note: S_E is called the "residual standard deviation."

- Recall: (from Sec. 3.3)

$$
\begin{aligned}
SS(X) &= \sum X^2 - \frac{(\sum X)^2}{n} \\
SS(Y) &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\
SS(XY) &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\
b_1 = \hat{\beta}_1 &= \frac{SS(XY)}{SS(X)} \\
b_0 = \hat{\beta}_0 &= \bar{Y} - b_1 \bar{X}
\end{aligned}
$$

So,

$$
\begin{aligned}
S_E &= \sqrt{\frac{1917.88 - (18.2^2)(4.97)}{8 - 2}} \\
&= \sqrt{\frac{1917.88 - 1646.26}{6}} \\
&= \sqrt{\frac{271.62}{6}} \\
&= \sqrt{45.27} \\
&= 6.73
\end{aligned}
$$

Note: S_E can also be computed as

$$S_E = \sqrt{\frac{(Y - \hat{Y})}{n - 2}}$$

Next, we compute the test statistic:

$$\begin{aligned} T &= \frac{b_1}{\left(\frac{S_E}{\sqrt{SS(X)}}\right)} \\ &= \frac{18.2}{\left(\frac{6.73}{\sqrt{4.97}}\right)} \\ &= \frac{18.2}{3.02} \\ &= 6.03 \end{aligned}$$

Use t-table (Table F) with $\mathrm{d.f.} = n - 2 = 8 - 2 = 6$ to compute the P-value.

Since

$$H_A : \beta \neq 0$$

$$\begin{aligned} P - \text{value} &= 2 \times P(T > |\text{test statistic value}|) \\ &= 2 \times P(T > |6.03|) \\ &= 2 \times P(T > 6.03) \end{aligned}$$

Looking in Table F, for d.f. = 6 , we see that

$$P(T > 5.959) = 0.0005$$

So,

$$P(T > 6.03) < 0.0005$$

and the P-value $< 2 \times 0.0005 = 0.001$.

Decision: Reject H_0 if

$$P - \text{value} \leq \alpha.$$

Since

$$P - \text{value} < 0.001 < 0.05,$$

we reject H_0.

Conclusion: The slope of the regression equation is significantly different from 0. This implies that the X's "contain" information about the corresponding Y's. Hence the X's can be used to predict the Y's.

Assumptions:

1. The Y-values are independent of each other.
2. The relation between X and Y is linear:

$$Y = \beta_0 + \beta_1 X$$

3. For each value X, the standard deviation of the Y's are all equal, i.e., the standard deviation of the Y's does not change when the X-value is changed.
4. For each X-value, the corresponding Y's follow a Normal Distribution.

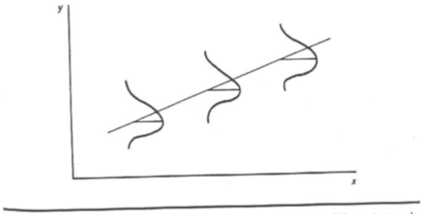

Figure 15.2 Assumptions Required in Using the Formula for Confidence Intervals for Predicted y

Note: If n is large, i.e., $n \geq 30$, we can relax the normality assumption of Y for each X.
Note: If n is large, i.e., ($n \geq 30$), the test statistic will follow an approximate normal distribution.

$$Z = \frac{b_1}{\left(\frac{S_E}{\sqrt{SS(X)}}\right)}$$

and we can find the P- value of this test statistic by using the standard normal table (Table E).

Confidence Interval on the Slope of the Population Regression Line (β_1)

A $(1-\alpha) \times 100\%$ C.I. on β_1 is given by

$$b_1 \pm t \frac{S_E}{\sqrt{SS(X)}}$$

Example: Construct a 90% C.I. on β_1 for the T.V. - cholesterol data.

In Sect. 3.3 we computed

$$SS(X) = 4.97.$$

In Sect. 11.1 we computed

$$S_E = 6.73.$$

Next find t (Table: F)

For a 90% C.I.,

$$\alpha = 1 - 0.90 = 0.10$$

but we use $\frac{\alpha}{2} = 0.05$ to index the table.

$$\text{d.f.} = n - 2 = 8 - 2 = 6.$$

So,

$$t = 1.943.$$

Next compute the limits of the 90% C.I. for β_1.

$$\begin{aligned} b_1 &\pm t\frac{S_E}{\sqrt{SS(X)}} \\ 18.2 &\pm 1.943\frac{6.73}{\sqrt{4.97}} \\ &\vdots \\ 18.2 &\pm 5.866 \end{aligned}$$

Our 90% C.I. on β_1 goes from 12.33 to 24.066.

Assumptions:
Same a for hypothesis test on β_1.

Hypothesis Test on the Population Correlation Coefficient ρ

The sample correlation coefficient, r, estimates the population correlation coefficient, ρ.

We can perform a test of hypothesis on ρ.

Example: (Cont.)

$$
\begin{aligned}
H_0 : \rho &= 0 \qquad \text{(no correlation between X and Y)} \\
H_A : \rho &\neq 0 \qquad \text{(no correlation between X and Y)} \\
\alpha &= 0.05
\end{aligned}
$$

Text Statistic:

$$\begin{aligned} T &= r \times \sqrt{\frac{n-2}{1-r^2}} \\ &= 0.926 \times \sqrt{\frac{8-2}{1-0.926^2}} \\ &= 0.926 \times \sqrt{\frac{8-2}{0.1425}} \\ &= 0.926 \times \sqrt{42.0982} \\ &= 6.01 \end{aligned}$$

P-value of test statistic:
Since

$$H_A : \rho \neq 0$$

is 2-sided,

$$
\begin{aligned}
P - \text{value} &= 2 \times P(T > |\text{test statistic value}|) \\
&= 2 \times P(T > 6.01).
\end{aligned}
$$

Using Table F with d.f. $= n - 2 = 6$, we see

$$P(T > 5.959) = 0.0005$$

So,

$$P(T > 6.01) < 0.0005$$

and

$$P - \text{value} < 2 \times 0.0005 = 0.001$$

Decision: Reject H_0 if,

$$P - \text{value} \leq \alpha$$

Since

$$P - \text{value} < 0.001 < 0.05,$$

we reject H_0.

Conclude: There does seem to be a correlation between "number of hours of T.V. per day" and cholesterol level, at the 5% significance level.

Assumptions: Each variable is normally distributed.

The Connection Between Correlation And Regression

Consider a test

$$H_0 : \beta = 0$$

$$H_A : \beta \neq 0$$

If we fail to reject H_0, this indicates that our regression line is (approximately) horizontal.

This means that the X's provide little, if any, value in predicting the Y's , i.e., The predicted Y would be (nearly) the same for all values of X.

This corresponds to saying that variables X and Y are uncorrelated.

In fact, were we to conduct the test:

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

we would fail to reject H_0.

Similarly, if we test:

$$H_0 : \beta = 0$$

$$H_A : \beta \neq 0$$

and reject H_0, we would reject H_0 in the test:

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

Look at the “T.V. hours” and “cholesterol level” examples.

In Sect. 11.1, we tested:

$$
\begin{aligned}
H_0 : \rho &= 0 \\
H_A : \rho &\neq 0 \\
\alpha &= 0.05 \\
T &= 6.01 \qquad \text{(test statistics)} \\
P - \text{value} &= 2 \times P(T \geq |\text{test statistics}|) \\
&= 2 \times P(T \geq 6.01) < 0.001
\end{aligned}
$$

and we reject H_0.
In Sect. 11.1, we tested:

$$
\begin{aligned}
H_0 : \beta &= 0 \\
H_A : \beta &\neq 0 \\
\alpha &= 0.05.
\end{aligned}
$$

(Cont.)

$$
\begin{aligned}
T &= 6.03 \qquad \text{(test statistic)} \\
P - \text{value} &= 2 \times \mathrm{P}(\mathrm{T} \geq |\text{test statistic}|) \\
&= 2 \times \mathrm{P}(\mathrm{T} \geq 6.03) < 0.001
\end{aligned}
$$

and we reject H_0.

Note: In both tests,

1. The same P-values;
2. The test-statistic values are identical (except for round-off error);
3. Same decision (Reject H_0).

Confidence Interval for Predicted Values of Y and the Mean Of Y

Rather than use a single value for $\hat{Y}$, it may be better to specify a range of values in which we expect Y to be, i.e., use a Confidence interval.

A (1-α)100% C.I. for predicted mean value of Y at some value $X = X_0$ is

$$(b_0 + b_1 X_0) \pm tS_E \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS(X)}}$$

where

$$
\begin{aligned}
SS(X) &= \sum X^2 - \frac{(\sum X)^2}{n} \\
SS(Y) &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\
S_E &= \sqrt{\frac{SS(Y) - b^2.SS(X)}{n-2}}
\end{aligned}
$$

and t is from a t-distribution with

$$\text{d.f.} = n - 2$$

for $\alpha/2$.

Example (Cont.)

Construct a 95% C.I. for the mean cholesterol level of a persons who watch 2 hours of T.V. per day.

First, compute $\bar{X}$:

$$\bar{X} = \frac{\sum X}{n} = \frac{18.5}{8} = 2.3125$$

Next, compute S_E.
From Sect. 3.3, we saw that

$$\begin{aligned} SS(X) &= 4.97 \\ SS(Y) &= 1917.88 \end{aligned}$$

Example (Cont.)

So,

$$
\begin{aligned}
S_E &= \sqrt{\frac{SS(Y) - b_1^2 SS(X)}{n-2}} \\
&= \sqrt{\frac{1917.88 - (18.2)^2 4.97}{8-2}} \\
&\vdots \\
&= 6.73
\end{aligned}
$$

Next, Find t (Table F)

$$
\begin{aligned}
\text{d.f.} &= n - 2 = 8 - 2 = 6 \\
\frac{\alpha}{2} &= \frac{0.05}{2} = 0.025 \\
t &= 2.447
\end{aligned}
$$

Example (Cont.)

For $X = 2$ hours T.V. per day the error term of the 95% C.I. on mean cholesterol level is

$$
\begin{aligned}
tS_r\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS(X)}} &= 2.447 \times 6.73\sqrt{\frac{1}{8} + \frac{(2 - 2.231)^2}{4.97}} \\
&= 16.464\sqrt{0.14434} \\
&= 16.464(0.3799) \\
&= 6.25
\end{aligned}
$$

For $X = 2$,

$$
\begin{aligned}
\hat{Y} &= b_0 + b_1.X_0 \\
&= 157.6 + 18.2X_0 \\
&= 157.6 + 18.2(2) \\
&= 194
\end{aligned}
$$

Example (Cont.)

Thus a 95% C.I. for the predicted value of the mean of Y when $X = 2$ is given by

$$194 \pm 6.25$$

So our C.I. goes from

$$187.75 \text{ to } 200.25.$$

Interpretation:

With 95% confidence, the mean cholesterol level of persons who watch 2 hours of T.V. per day is between 187.75 and 200.25.

Assumptions:

1. the data points are normally distributed about the regression line (in the Y direction), i.e., the Y-values have a normal distribution for each particular value of X.
2. these normal distributions of the Y's are the same for each value of X.

Figure 15.2 Assumptions Required in Using the Formula for Confidence Intervals for Predicted y

Prediction Interval For Y

For a specific value of $X = X_0$, we may desire to predict an individual value $\hat{Y}$, rather than the predicted mean value of Y.

A (1-α)100% prediction interval for single Y at some value $X = X_0$

$$(b_0 + b_1 X_0) \pm t S_E \times \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS(X)}},$$

where S_E is the residual standard deviation and t is from a t-distribution with d.f. $= n - 2$ using $\alpha/2$.

Note: This looks similar to a C.I. on the mean value of Y, except for the additional term under the square root.

Example (Cont.)

Construct a 95% prediction interval for the cholesterol level who watches 2 hours of T.V. per day.

From previous examples in Sect. 11.1 and Sect. 11.2,

$$\begin{aligned}
\bar{X} &= 2.3125 \\
SS(X) &= 4.97 \\
SS(Y) &= 1917.88 \\
S_E &= 6.73
\end{aligned}$$

Example (Cont.)

To find the appropriate value t use Table F with

$$\begin{aligned} \text{d.f.} &= n-2 = 8-2 = 6 \\ \frac{\alpha}{2} &= \frac{0.05}{2} = 0.025 \\ \text{So,} \quad t &= 2.45 \end{aligned}$$

For $X = 2$ hours T.V. per day, the error term of the 95% prediction interval is

$$\begin{aligned} & tS_E\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{SS(X)}} \\ = \; & 2.45 \times 6.73\sqrt{1+\frac{1}{8}+\frac{(2-2.231)^2}{4.97}} \\ = \; & 16.464\sqrt{1.4434} \\ = \; & 16.464(1.06974) \\ = \; & 17.6122 \end{aligned}$$

Example (Cont.)

For $X = 2$,

$$\begin{aligned}\hat{Y} &= b_0 + b_1.X_0 \\ &= 157.6 + 18.2X_0 \\ &= 157.6 + 18.2(2) \\ &= 194\end{aligned}$$

Thus a 95% prediction interval for a single predicted value of Y when $X = 2$ is given by

$$194 \pm 17.6122$$

So our 95% prediction interval goes from

$$176.39 \text{ to } 211.61$$

Example (Cont.)

Interpretation: With 95% confidence, the predicted cholesterol level of an individual who watches 2 hours of T.V. per day is between 176.39 to 211.61.

Assumptions: Same as for C.I.

Note: The 95% prediction interval for an individual (from 176.4 to 211.6) is wider than the 95% confidence interval for the mean cholesterol level of people who watch 2 hours T.V per day (from 187.75 to 200.25).