Topic 3 - Statistical Inference

STAT 525 - Fall 2013

Outline

- Normal error regression model
- Inference concerning β_1
- Inference concerning β_0
- Inference concerning prediction

Normal Error 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
  - $\varepsilon_i \sim \mathrm{N}(0, \sigma^2) \longleftarrow$ **NEW**
  - Uncorrelated $\longrightarrow$ independent error terms
- Defines distribution of random variable Y

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

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

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
                            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     Standard
Variable   DF      Estimate        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
```


Testing for Linear Relationship

- Term $\beta_1 X_i$ defines linear relationship
- Will then test $H_0 : \beta_1 = 0$
- Test requires
  - Test statistic
  - Sampling distribution of the test statistic

Note: form of test statistic is often

$$\frac{\text{point estimate} - E(\text{point estimate}|H_0)}{s(\text{point estimate})}$$

Distribution of b_1

- Rewrite $b_1 = \sum k_i Y_i$ where

$$k_i = \frac{X_i - \overline{X}}{\sum (X_i - \overline{X})^2}$$

Note: $\sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum (X_i - \overline{X})Y_i$
$\sum k_i = 0$ and $\sum k_i X_i = 1$

- Can now describe distribution of b_1

Distribution of b_1

- Normal since linear combination of *i.i.d.* Y_i's

$$
\begin{aligned}
E(b_1) &= E(\sum k_i Y_i) \\
&= \sum k_i E(Y_i) \\
&= \sum k_i \beta_0 + \sum k_i \beta_1 X_i \\
&= 0 + \beta_1 \\
\mathrm{Var(b_1)} &= \mathrm{Var}(\sum \mathrm{k_i Y_i}) \\
&= \sum k_i^2 \mathrm{var(Y_i)} \\
&= \sigma^2 / \sum (X_i - \overline{X})^2
\end{aligned}
$$

Distribution of $\frac{b_1 - \beta_1}{s(b_1)}$

- Rewrite as

$$\frac{b_1 - \beta_1}{\sigma(b_1)} \div \frac{s(b_1)}{\sigma(b_1)}$$

- Since Y_i's are *i.i.d.* normal
  - b_1 is normal $\longrightarrow$ 1st term is standard normal
  - The quantity $\sum (Y_i - \hat{Y}_i)^2/\sigma^2 \sim \chi^2_{n-2}$
  - The variable $s^2(b_1)/\sigma^2(b_1) \sim \chi^2_{n-2}/(n-2)$
  - This variable is independent of b_0 and b_1

$$\downarrow$$

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$$

Steps of Hypothesis Test

- $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$
- Compute the test statistic

$$t^\star = \frac{b_1 - 0}{s(b_1)} = \frac{9.32 - 0}{0.31} = 30.07$$

- Compute P-value using sampling dist

$$P(|t_{n-2}| \geq |t^\star|) = 6.5 \times 10^{-12} (< .0001)$$

- Compare to α and draw conclusion

  Reject H_0 at $\alpha = .05$ level, evidence suggests a positive linear relationship

Distribution of b_0

- Rewrite $b_0 = \sum k_i Y_i$ where

$$k_i = \frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$$

Note: $\sum k_i = 1$ and $\sum k_i X_i = 0$

- Can now describe distribution of b_0

Distribution of b_0

- Normal since linear combination of *i.i.d.* Y_i's

$$\begin{aligned}
E(b_0) &= E(\sum k_i Y_i) \\
&= \sum k_i E(Y_i) \\
&= \sum k_i \beta_0 + \sum k_i \beta_1 X_i \\
&= \beta_0 + 0
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var(b_0)} &= \mathrm{Var}(\sum \mathrm{k_i Y_i}) \\
&= \sum k_i^2 \mathrm{var(Y_i)} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right)
\end{aligned}$$

Steps of Hypothesis Test

- $H_0 : \beta_0 = 0$ and $H_a : \beta_0 \neq 0$
- Compute the test statistic

$$t^\star = \frac{b_0 - 0}{s(b_0)} = \frac{-61.12 - 0}{25.13} = -2.43$$

- Compute P-value using sampling dist

$$P(|t_{n-2}| \geq |t^\star|) = 0.0333$$

- Compare to α and draw conclusion

  Reject H_0 at $\alpha = .05$ level, evidence suggests the intercept is different from zero

Confidence Intervals

- Could also form confidence intervals
- General form for parameter β_l

$$b_l \pm t(1 - \alpha/2, n-2) s(b_l)$$

- Reject $H_0 : \beta_l = \beta_{l0}$ if β_{l0} is not in CI
- These CIs generated in SAS with `clb` option

Comments

- Test of intercept usually not of interest
- When errors not normal, procedures are generally reasonable approximations
  - Bootstrapping as alternative approach
- Procedures can be modified for one-sided test / confidence bound
- At design stage
  - $\mathrm{Var}(b_1)$ smaller when $\sum (X_i - \overline{X})^2$ large
  - $\mathrm{Var}(b_0)$ smallest when $\overline{X} = 0$

Interval Estimation of $E(Y_h)$

- Often interested in estimating the mean response for particular X_h

$$\hat{Y}_h = b_0 + b_1 X_h$$

- Need sampling dist of $\hat{Y}_h$ to form CI
  - Rewrite $\hat{Y}_h = \sum k_i Y_i$ where
  $$k_i = \frac{1}{n} + \frac{(X_h - \overline{X})(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}$$
  - Similar construction as b_0 (i.e., $X_h = 0$)
  - $E(\hat{Y}_h) = E(Y_h)$
  - $\mathrm{Var}(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}\right)$
  - CI: $\hat{Y}_h \pm t(1-\alpha/2, n-2)s(\hat{Y}_h)$

Interval Estimation of $Y_{h(new)}$

- Consider predicting future observation
- Unlike the expected value, a new observation does not fall directly on the regression line. Must account for added variability.
- In other words, $\hat{Y}_{h(new)} = E(\hat{Y}_h) + \varepsilon_h$
- Consider variability of $Y|X_h \longrightarrow \sigma^2$
- $\mathrm{Var}(\hat{Y}_{h(new)}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}\right)$

SAS Commands

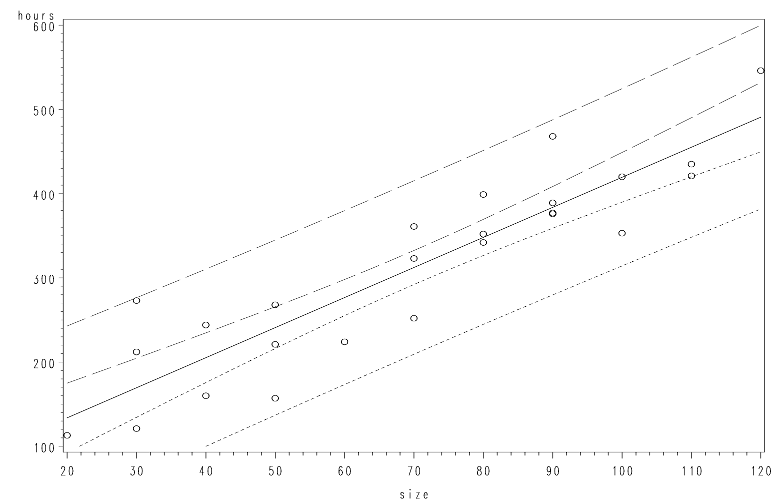
```
data a1;
   infile 'C:\Textdata\CH01TA01.txt';
   input size hours;

data a2; size=65;  output;
         size=100; output;
data a3; set a1 a2;

symbol1 v=circle i=rlclm95;
symbol2 v=circle i=rlcli95;
proc gplot data=a1;
   plot hours*size=1 hours*size=2 /overlay;
run;

proc reg data=a3;
   model hours=size/clm cli alpha=.10;
   id size;
run;
```


Scatterplot



Dependent Variable: hours

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	252378	252378	105.88	<.0001
Error	23	54825	2383.71562		
Cor Total	24	307203			

Root MSE	48.82331	R-Square	0.8215
Dependent Mean	312.28000	Adj R-Sq	0.8138
Coeff Var	15.63447		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	62.36586	26.17743	2.38	0.0259
size	1	3.57020	0.34697	10.29	<.0001

Output Statistics

Obs	size	Dep Var hours	Predicted Value	Std Error Mean Predict	90% CL Mean	
1	80	399.0000	347.9820	10.3628	330.2215	365.7425
2	30	121.0000	169.4719	16.9697	140.3880	198.5559
3	50	221.0000	240.8760	11.9793	220.3449	261.4070
4	90	376.0000	383.6840	11.9793	363.1530	404.2151
5	70	361.0000	312.2800	9.7647	295.5446	329.0154
6	60	224.0000	276.5780	10.3628	258.8175	294.3385
7	120	546.0000	490.7901	19.9079	456.6706	524.9096
8	80	352.0000	347.9820	10.3628	330.2215	365.7425
9	100	353.0000	419.3861	14.2723	394.9251	443.8470
10	50	157.0000	240.8760	11.9793	220.3449	261.4070
11	40	160.0000	205.1739	14.2723	180.7130	229.6349
12	70	252.0000	312.2800	9.7647	295.5446	329.0154
22	90	468.0000	383.6840	11.9793	363.1530	404.2151
23	40	244.0000	205.1739	14.2723	180.7130	229.6349
24	80	342.0000	347.9820	10.3628	330.2215	365.7425
25	70	323.0000	312.2800	9.7647	295.5446	329.0154
26	65	.	294.4290	9.9176	277.4315	311.4264
27	100	.	419.3861	14.2723	394.9251	443.8470

Output Statistics

Obs	size	Dep Var hours	Predicted Value	Std Error Mean Predict	90% CL Predict	
1	80	399.0000	347.9820	10.3628	262.4411	433.5230
2	30	121.0000	169.4719	16.9697	80.8847	258.0591
3	50	221.0000	240.8760	11.9793	154.7171	327.0348
4	90	376.0000	383.6840	11.9793	297.5252	469.8429
5	70	361.0000	312.2800	9.7647	226.9460	397.6140
6	60	224.0000	276.5780	10.3628	191.0370	362.1189
7	120	546.0000	490.7901	19.9079	400.4244	581.1558
8	80	352.0000	347.9820	10.3628	262.4411	433.5230
9	100	353.0000	419.3861	14.2723	332.2072	506.5649
10	50	157.0000	240.8760	11.9793	154.7171	327.0348
11	40	160.0000	205.1739	14.2723	117.9951	292.3528
12	70	252.0000	312.2800	9.7647	226.9460	397.6140
22	90	468.0000	383.6840	11.9793	297.5252	469.8429
23	40	244.0000	205.1739	14.2723	117.9951	292.3528
24	80	342.0000	347.9820	10.3628	262.4411	433.5230
25	70	323.0000	312.2800	9.7647	226.9460	397.6140
26	65	.	294.4290	9.9176	209.0432	379.8148
27	100	.	419.3861	14.2723	332.2072	506.5649

Confidence Band

- Consider looking at entire regression line
- Want to define likely region where line lies
- Replace $t(1-\alpha/2, n-2)$ with Working-Hotelling value in each confidence interval

$$W^2 = 2F(1-\alpha; 2, n-2)$$

- Boundary values define a hyperbola
- Confidence level α covers *all* X_h

$$\Pr(|\hat{Y}_h - Y_h| \leq W s(\hat{Y}_h), \forall X_h) \geq 1-\alpha$$

- Will be discussed more in Chapter 4

Confidence Bands

- Prediction wider than confidence interval
- Band narrowest at $\overline{X}$
- Theory comes from fact that (b_0, b_1) is multivariate normal
  - Joint confidence region for (β_0, β_1) is an ellipse
  - Cov(b_0, b_1)=Cov$(\sum k_{i0}Y_i, \sum k_{i1}Y_i)$=$-\overline{X}Var(b_1)$
- Band width for $X_h >$ individual CI width
- Can find α for individual CIs that gives same results

Background Reading

- Appendix A
- KNNL Chapters 2 and 3
- SAS template file `knnl054.sas`