

Simple Linear Regression

Applied Regression and Other Multivariable Methods
Sections 5-7 - 5-11, 12-1 - 12-4

5

Inference About the Slope

- As with all estimates, $\hat{\beta}_1$ subject to sampling var
- Because $Y|X \sim \text{Normal}$, the estimate $\hat{\beta}_1 \sim \text{Normal}$
 - A linear combination of indep Normals is Normal
 - Thm: If $Y_i \sim N(\mu_i, \sigma_i)$, then

$$L = \sum c_i Y_i \sim N\left(\sum c_i \mu_i, \sqrt{\sum c_i^2 \sigma_i^2}\right)$$

- Can write $\hat{\beta}_1$ as linear combination of E 's
- Standard error of $\hat{\beta}_1$

$$S_{\hat{\beta}_1} = \frac{S_{Y|X}}{S_x \sqrt{n-1}}$$

- Use std error to form CI or test hypothesis
- Degrees of freedom $n - 2$

$$\text{Confidence Int} : \hat{\beta}_1 \pm t_{n-2, \alpha/2} S_{\hat{\beta}_1}$$

$$\text{Hypothesis Test: } T = (\hat{\beta}_1 - \beta_1^0) / S_{\hat{\beta}_1}$$

5-1

Inference About the Slope

- Want small $S_{\hat{\beta}_1}$ for inference
- In situation where X is under experimental control

If S_X made large \rightarrow small $S_{\hat{\beta}_1}$

Can increase S_X by increasing dispersion of X

Can also increase n to decrease $S_{\hat{\beta}_1}$, increase df

- Test $H_0 : \beta_1 = 0$ to see if linear association
- Does X help explain Y through a **linear** model?
 - Rejecting does not mean linear model is “best”
 - Not rejecting doesn't mean X unimportant
 - See page 56 for examples

5-2

Inference About the Intercept

- Sometimes interested in intercept β_0
- Standard error of $\hat{\beta}_0$

$$S_{\hat{\beta}_0} = S_{Y|X} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}$$

- In situation where X is under experimental control

If S_X made large \rightarrow small $S_{\hat{\beta}_0}$

Increase S_X by increasing dispersion

If \bar{X} close to zero \rightarrow small $S_{\hat{\beta}_0}$

- Can also increase n to decrease $S_{\hat{\beta}_1}$, increase df
- Test $H_0 : \beta_0 = 0$ to see if line goes through origin
 - Does not test linear model fit
 - Really only meaningful if X around zero

5-3

SAS Procedures

- model statement in proc reg has many options
- To construct confidence intervals use

– alpha= , clm, cli, clb

```
proc reg;
  model sbp=age / clb alpha=.01; /* Form 99% CI for parameters */
run;
```

Dependent Variable: sbp

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6394.02269	6394.02269	21.33	<.0001
Error	28	8393.44398	299.76586		
Corrected Total	29	14787			

Root MSE	17.31375	R-Square	0.4324
Dependent Mean	142.53333	Adj R-Sq	0.4121
Coef Var	12.14716		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	98.71472	10.00047	9.87	<.0001
age	1	0.97087	0.21022	4.62	<.0001

Parameter Estimates			
Variable	DF	99% Confidence Limits	
Intercept	1	71.08080 126.34863	
age	1	0.38999 1.55175	

5-4

Inference about the Line

- Line describes the mean population response for X
- Predicted mean at $X = X_0$ is

$$\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

- Standard error of $\hat{\mu}_{Y|X_0}$ is

$$S_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}$$

- New predicted observation at $X = X_0$ is

$$\hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

- Standard error of \hat{Y}_{X_0} is

$$S_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}$$

- New obs doesn't have to fall on line → bigger var

- Recall $Y_{X_0} = \mu_{Y|X_0} + E$ and $\text{Var}(E) = S_{Y|X}^2$

5-5

Interpolation vs Extrapolation

- Must use caution in interpretation of \hat{Y}_{X_0} , $\hat{\mu}_{Y|X_0}$
- If X_0 within range of observed X 's → interpolation
- If X_0 outside range of observed X 's → extrapolation
- Extrapolation should be avoided
 - No assurances still linear outside range of data
 - Example: Fish activity and Water Temp
- Can also construct confidence/prediction bands
- Prediction bands wider than confidence bands
- Most narrow at \bar{X}

5-6

SAS Procedures

```
proc reg;
  model sbp=age /cli clm; /* Confidence int for i=indiv m=mean */
  plot sbp*age / conf pred; /* Create plot with conf and pred bands */
run;
```

Dependent Variable: sbp

Output Statistics

Obs	Dep Var sbp	Predicted Value	Std Error Mean Predict	95% CL Mean
1	114.0000	115.2195	6.7058	101.4832 128.9558
2	124.0000	117.1613	6.3382	104.1781 130.1444
3	116.0000	118.1321	6.1568	105.5204 130.7439
4	120.0000	119.1030	5.9774	106.8588 131.3472
5	125.0000	122.9865	5.2825	112.1657 133.8072
6	130.0000	126.8700	4.6362	117.3731 136.3668
7	110.0000	131.7243	3.9332	123.6676 139.7810
8	136.0000	133.6661	3.6984	126.0901 141.2420
9	144.0000	136.5787	3.4139	129.5857 143.5717
10	120.0000	136.5787	3.4139	129.5857 143.5717
11	124.0000	139.4913	3.2289	132.8771 146.1055
12	128.0000	139.4913	3.2289	132.8771 146.1055
13	160.0000	141.4330	3.1700	134.9395 147.9265
14	138.0000	142.4039	3.1612	135.9285 148.8792
15	135.0000	142.4039	3.1612	135.9285 148.8792
16	142.0000	143.3748	3.1663	136.8889 149.8606
17	220.0000	144.3456	3.1853	137.8208 150.8704
18	145.0000	144.3456	3.1853	137.8208 150.8704
19	130.0000	145.3165	3.2180	138.7248 151.9082
20	142.0000	147.2582	3.3225	140.4525 154.0640
21	158.0000	150.1708	3.5675	142.8632 157.4785
22	154.0000	153.0835	3.9001	145.0946 161.0724
23	150.0000	153.0835	3.9001	145.0946 161.0724
24	140.0000	155.9961	4.2999	147.1881 164.8041

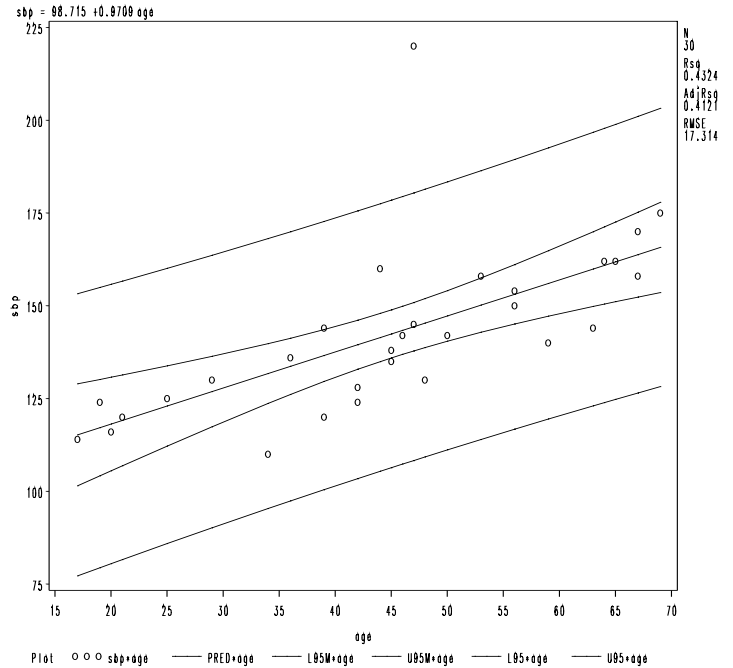
5-7

Output Statistics

Obs	95% CL Predict		Residual
1	77.1867	153.2523	-1.2195
2	79.3939	154.9286	6.8387
3	80.4909	155.7734	-2.1321
4	81.5833	156.6227	0.8970
5	85.9069	160.0661	2.0135
6	90.1549	163.5851	3.1300
7	95.3551	168.0935	-21.7243
8	97.4003	169.9318	2.3339
9	100.4302	172.7271	7.4213
10	100.4302	172.7271	-16.5787
11	103.4142	175.5684	-15.4913
12	103.4142	175.5684	-11.4913
13	105.3779	177.4882	18.5670
14	106.3520	178.4558	-4.4039
15	106.3520	178.4558	-7.4039
16	107.3210	179.4285	-1.3748
17	108.2848	180.4064	75.6544
18	108.2848	180.4064	0.6544
19	109.2435	181.3895	-15.3165
20	111.1455	183.3709	-5.2582
21	113.9602	186.3815	7.8292
22	116.7292	189.4377	0.9165
23	116.7292	189.4377	-3.0835
24	119.4531	192.5391	-15.9961
25	123.0159	196.7432	-15.8796
26	123.8945	197.8063	1.1496
27	124.7684	198.8742	0.1787
28	126.5018	201.0242	6.2370
29	126.5018	201.0242	-5.7630
30	128.2167	203.1929	9.2952

Sum of Residuals 0
 Sum of Squared Residuals 8393.44398
 Predicted Residual SS (PRESS) 9108.39568

Data for Chapter 5 Example



Regression Diagnostics

- Will study more procedures throughout semester
- These focus on simple linear regression
- Assumptions
 - 1 Model is correct (linearity)
 - 2 Independent observations
 - 3 Errors normally distributed
 - 4 Constant variance

$$\begin{aligned}
 Y_i &= \hat{\mu}_{Y_i|X_i} + (Y_i - \hat{\mu}_{Y_i|X_i}) \\
 Y_i &= \hat{Y}_i + \hat{E}_i \\
 \text{observed} &= \text{predicted} + \text{residual}
 \end{aligned}$$

- Diagnostics will use predicted and residual values

Diagnostics

- Normality
 - Histogram/Boxplot of residuals
 - Normal probability plot / QQ plot
 - Shapiro-Wilks/Kolmogorov-Smirnov Test
- Variance
 - Plot \hat{E}_i vs \hat{Y}_i (residual plot)
 - Bartlett's or Levene's Test (provided repeat X_i obs)
- Independence
 - Plot \hat{E}_i vs time/space
 - Runs test/Durbin-Watson Test
- Outliers
 - Is it influential? With and without analysis
 - Formal tests (e.g. standardized residuals)
 - Investigate why result may occur, don't try to eliminate

Normality Assumption

- Histogram/Boxplot
 - Is histogram of residuals bell-shaped?
 - Is boxplot/histogram symmetric?
- Normal Probability/QQ Plot
 - Ordered residuals vs cumulative normal probs
 - Is it approximately linear?

Constant Variance

- Often experiments with non-constant variance
- Size of residual associated with predicted value
- Residual plot
 - Plot \hat{E}_i vs \hat{Y}_i
 - Is the range constant for different levels of \hat{Y}_i
- Bartlett's and Levene's Test
 - More formal test
 - Compares pooled var with sample variances
 - Bartlett sensitive to Normality assumption

5-12

Independence

- Plot of the residuals over time
 - Is there a drift or pattern as trials proceed?
- Plot residuals versus relevant variables
 - Often variables omitted from analysis
 - Experimental conditions (e.g., temp)
 - May result in inclusion of factor in next exp
- Durbin-Watson or Runs Test
 - DW model statement option
 - Assumes observations presented in time order
 - Runs tests look at number of pos/neg residuals in a row

5-13

SAS Procedures

```
proc reg;
  model sbp=age;
  plot r.*nqq.;           /* Generate QQ Plot */
  plot r.*p.;             /* Generate residual Plot */
  output out=fit r=res p=pred;

proc gplot;                /* Generate Residual Plot */
  plot res*pred /vref=0 frame;

proc univariate normal pctdef=4; /* Check Normality of Residuals */
  var res;
  histogram res / normal kernel (L=2);
  qqplot res / normal (L=1 mu=est sigma=est);
run;
```

5-14

The UNIVARIATE Procedure
Variable: res (Residual)

Moments			
N	30	Sum Weights	30
Mean	0	Sum Observations	0
Std Deviation	17.012616	Variance	289.429103
Skewness	3.06973017	Kurtosis	13.6058151
Uncorrected SS	8393.44398	Corrected SS	8393.44398
Coeff Variation	.	Std Error Mean	3.10606451

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	17.01262
Median	-0.52040	Variance	289.42910
Mode	.	Range	97.37869
		Interquartile Range	12.33250

Tests for Normality			
Test	--Statistic--	-----p Value-----	
Shapiro-Wilk	W 0.699983	Pr < W	<0.0001
Kolmogorov-Smirnov	D 0.225738	Pr > D	<0.0100
Cramer-von Mises	W-Sq 0.338205	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq 2.140256	Pr > A-Sq	<0.0050

Quantiles (Definition 4)	
Quantile	Estimate
100% Max	75.654375
99%	75.654375
95%	44.256311
90%	9.148620
75% Q3	3.906773
50% Median	-0.520403
25% Q1	-8.425731
10%	-15.984417
5%	-18.894204
1%	-21.724310
0% Min	-21.724310

5-15

Extreme Observations

---Lowest---		---Highest---	
Value	Obs	Value	Obs
-21.7243	14	7.42134	1
-16.5787	23	7.82915	26
-15.9961	13	9.29523	30
-15.8796	27	18.56699	25
-15.4913	8	75.65438	2

Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic---	-----p Value-----
Kolmogorov-Smirnov	D 0.22573825	Pr > D <0.010
Cramer-von Mises	W-Sq 0.33820476	Pr > W-Sq <0.005
Anderson-Darling	A-Sq 2.14025586	Pr > A-Sq <0.005

Quantiles for Normal Distribution

Percent	Observed	Estimated
1.0	-21.72431	-39.577263
5.0	-18.89420	-27.983263
10.0	-15.98442	-21.802545
25.0	-8.42573	-11.474835
50.0	-0.52040	-0.000000
75.0	3.90677	11.474835
90.0	9.14862	21.802545
95.0	44.25631	27.983263
99.0	75.65438	39.577263

