Regression Analysis Using SAS and Stata

Hsueh-Sheng Wu

CFDR Workshop Series

Spring 2012

BGSU
Center for Family and Demographic Research

Outline

- What is regression analysis?
- Why is regression analysis popular?
- A primitive way of conducting regression analysis
- A better way of conducting regression analysis:
  Corrections for violations in regression assumptions for
  - Linearity
  - Mean independence
  - Homoscedasticity
  - Uncorrelated disturbances
  - Normal disturbance
- Conclusions

What Is Regression?

Regression is used to study the relation between a single dependent variable and one or more independent variables. In regression, the dependent variable *y* is a linear function of the *x*'s, plus a random disturbance ε.

$$y = a + b_1x_1 + b_2x_2 + \varepsilon$$

y is the dependent variable

a is the intercept

x_1 and x_2 *are independent variables*

b_1 *and* b_2 *are regression coefficients*

ε represents the combined effects of all the causes of y that are not included in the equation, but can influence the relations between x's and y

Five Assumptions of Regression

1. Linearity
   - y is a linear function of the x's
2. Mean independence
   - the mean of the disturbance term is always 0 and does not depend on the value of x's
3. Homoscedasticity
   - The variance of ε does not depend on the x's
4. Uncorrelated disturbances
   - The value of ε for any individual in the sample is not correlated with the value of ε for any other individuals
5. Normal disturbance
   - ε has a normal distribution

What Is Regression Analysis Popular?

- Statistical convenience. All statistic software provide regression analysis.
- Intuitive logic. Regression analysis fits our thinking style, that is, once we observed a phenomenon (i.e., dependent variable), what may contribute to this phenomenon.
- Various types of regression models
  - Based on the number of independent variables
    - Simple regression
    - Multiple Regression
  - Based on the type of the dependent variable
    - Ordinary least square regression
    - Logistic regression
    - Ordered logistic regression
    - Multinomial logistic regression
    - Poisson regression

A Primitive Way of Conducting Regression Analysis

- Decide a research question
  e.g., Whether the price of the car is determined by the weight, length, and the repair records of cars
- Decide dependent variable and independent variables
  Dependent variable: the price of the car
  Independent variables: the weight, length, and repair records
- Find a data set
  Data set: the information on prices, weights, lengths, and repair records of 74 cars
- Decide the regression model
  Ordinary Least Square (OLS) model is used because price is a continuous variable
- Run the regression analysis
- Interpret the results

Stata and SAS Commands for Regression Analysis

SAS commands:

```
proc reg data = auto;
MODEL price = weight length rep78;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: price Price

Number of Observations Read	74
Number of Observations Used	69
Number of Observations with Missing Values	5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	246375736	82125245	16.16	<.0001
Error	65	330421222	5083403		
Corrected Total	68	576796959			

Root MSE	2254.64042	R-Square	0.4271
Dependent Mean	6146.04348	Adj R-Sq	0.4007
Coeff Var	36.68442		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	6850.95187	4312.73825	1.59	0.1170
weight	Weight (lbs.)	1	5.25210	1.10343	4.76	<.0001
length	Length (in.)	1	-103.60163	37.78457	-2.74	0.0079
rep78	Repair Record 1978	1	844.94616	302.03629	2.80	0.0068

Stata commands:
webuse auto.dta, clear
reg price weight length rep78

Stata Output:

Source	SS	df	MS
Model	246375736	3	82125245.5
Residual	330421222	65	5083403.42
Total	576796959	68	8482308.22

Number of obs	= 69
F(3, 65)	= 16.16
Prob > F	= 0.0000
R-squared	= 0.4271
Adj R-squared	= 0.4007
Root MSE	= 2254.6

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	5.252098	1.103427	4.76	0.000	3.048401	7.455794
length	-103.6016	37.78457	-2.74	0.008	-179.0626	-28.14063
rep78	844.9462	302.0363	2.80	0.007	241.738	1448.154
_cons	6850.952	4312.738	1.59	0.117	-1762.181	15464.08

BGSU Center for Family and Demographic Research

A Better Way of Conducting Regression Analysis

- Decide a research question
- Decide dependent variable and independent variables
- Find a data set
- Decide the regression model
- Run the regression analysis
- Check the violations of the regression assumptions
- Interpret the results

Linearity Assumption

What does it mean?

- The dependent variable *y* is a linear function of the *x*'s
- Possible causes of violating this assumption:
  - Inaccurate specification of the regression models
  - Influential observations

What are the consequences?

- Biased estimates of intercept and regression coefficients
- Inaccurate prediction of y

Linearity Assumption (Cont.)

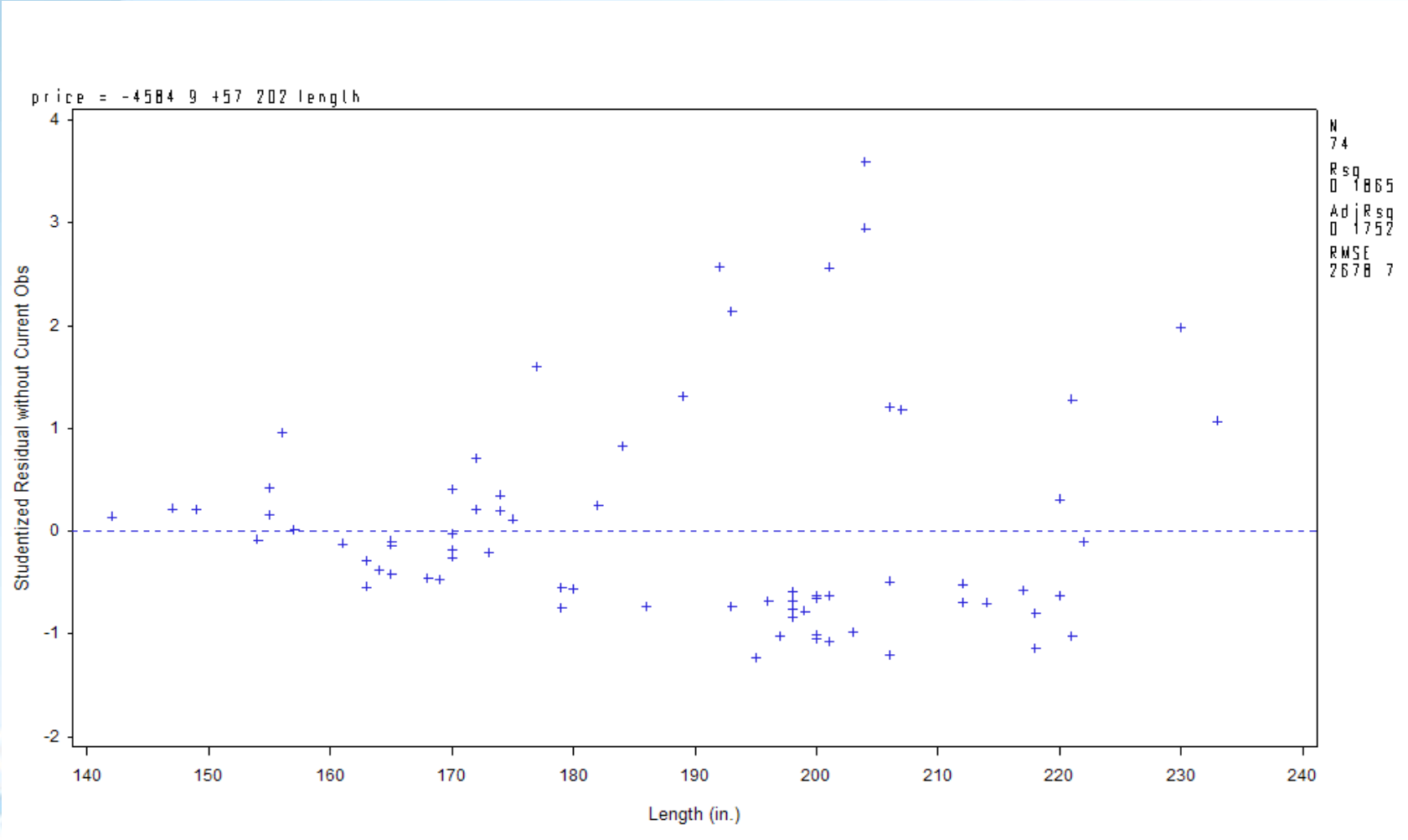
How to detect the inaccurate specification of the models?

- Plot y against x
- Plot residuals against x
- Plot residuals against y_{hat}

SAS commands:

```
proc reg data=auto;
model price = length;
plot price*length;
plot rstudent.*length;
plot rstudent.*p. / noline;
run;
```

Linearity Assumption (Cont.)

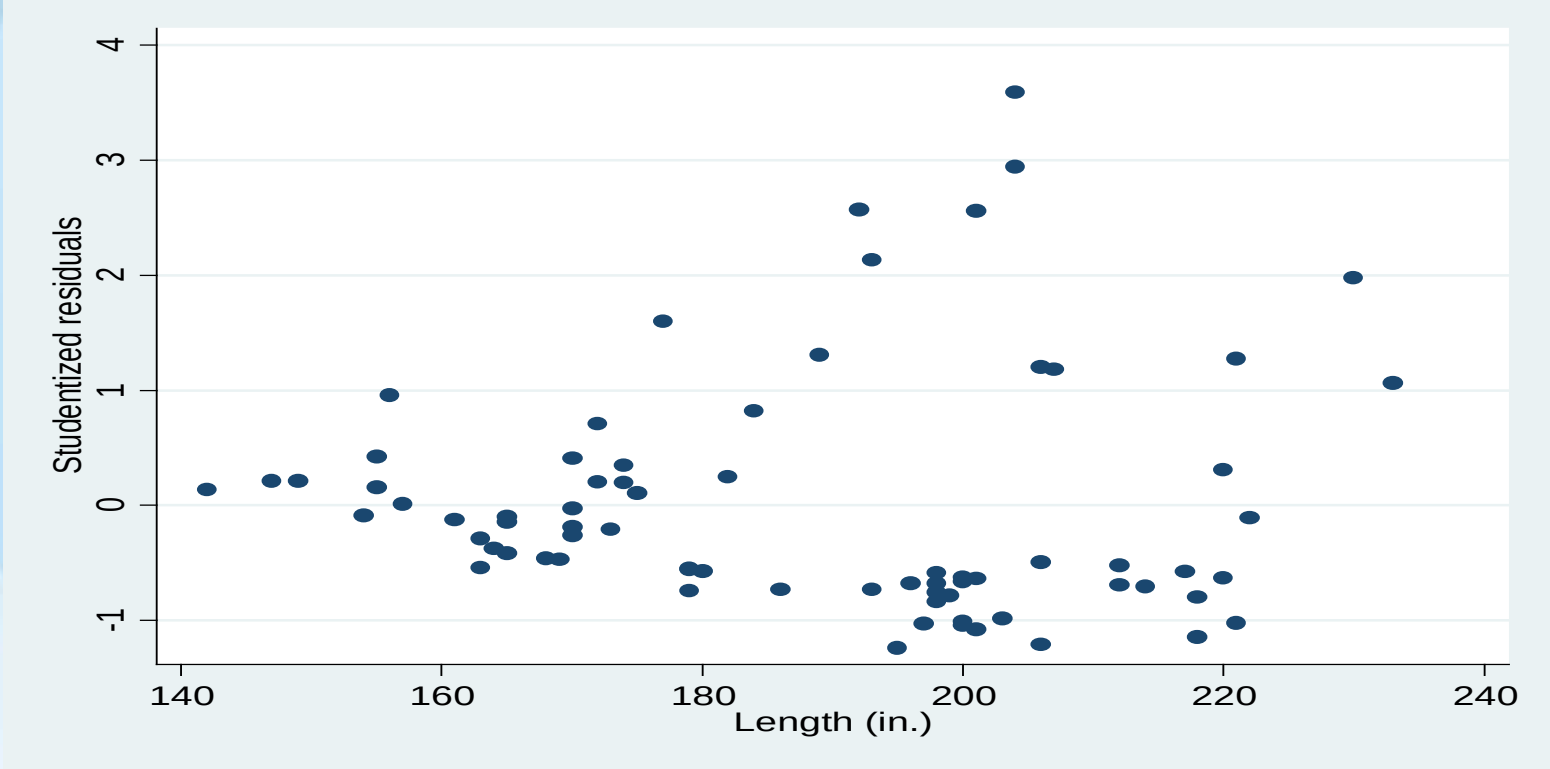


Linearity Assumption (Cont.)

Stata commands:

```
webuse auto.dta, clear
reg price length
predict r, rstudent
predict yhat, xb
scatter price length
scatter r length
scatter r yhat
```

Linearity Assumption (Cont.)



Linearity Assumption (Cont.)

Check for influential observations:

- Outliers:

If observations have standardized residuals that exceed =2 or -2, they may indeed outliers.

- Observations with high leverage:

If observation has leverage that is large than (2k+2)/n, where k is the number of predictors and n is the number of observations, these observations are said to have high leverage

- Observations with high impact on the regression coefficients:

Influential observations can be determined by either Cook's D statistics, DFITS, or DFBETA statistics.

  - –If observations have the value of Cook's D statistics larger than 4/n,
  - –If the DFITS statistics whose absolute values are larger than 2*sqrt(k/n),
  - –If the DFBETA statistics whose absolute value greater than 2/sqrt(n), they are influential observations.

Linearity Assumption (Cont.)

SAS commands:

```
proc reg data = in.auto;
model price = weight length rep78;
Output out=in.outlier(keep = make price weight length rep78 r lever cooked dffit)
rstudent = r h=lever cookd = cooked dffits = dffit;
run;
quit;

Proc print data = in.outlier;
Var make r;
Where abs(r)>2 & r ~=. ;
run;

Proc print data = in.outlier;
Var make lever;
Where lever  > (2*3+2)/69 & lever ~=.;
run;
```

Center for Family and Demographic Research

Linearity Assumption (Cont.)

```
proc reg data = in.auto;
model price = weight length rep78 / influence;
ods output OutputStatistics=in.dfbetas;
id make;
run;
quit;

proc print data=in.dfbetas;
var make DFFITS;
Where  abs(DFFITs) > (2*sqrt(3/69)) & DFFITS ~=. ;
Run;

proc print data=in.dfbetas;
var make DFB_Intercept DFB_weight DFB_length DFB_rep78 ;
Where  abs(DFB_weight) > (2/sqrt(69)) & DFB_weight ~=. ;
Run;
```

Linearity Assumption (Cont.)

Obs	make	DFFITS
2	Linc. Mark V	0.4797
4	Cad. Eldorado	0.8512
5	Linc. Versailles	0.5270
15	AMC Pacer	-1.0048
18	Volvo 260	0.5247
39	Cad. Seville	1.0777
49	Audi Fox	0.6182
66	Plym. Arrow	-1.0159

Obs	make	Intercept	weight	length	rep78
2	Linc. Mark V	-0.0130	0.2530	-0.1184	0.1010
4	Cad. Eldorado	0.4435	0.4704	-0.4156	-0.4082
5	Linc. Versailles	0.2646	0.4147	-0.3478	0.0204
15	AMC Pacer	-0.8790	-0.9209	0.9525	0.0170
39	Cad. Seville	0.6489	0.9956	-0.8688	0.1391
49	Audi Fox	-0.2089	-0.5201	0.4191	-0.2670
51	VW Dasher	-0.1254	-0.2461	0.1961	0.0210
66	Plym. Arrow	-0.9049	-0.9223	0.9670	0.0298

Linearity Assumption (Cont.)

Stata commands:

```
reg price weight length rep78
predict r, rstudent
predict lever, leverage
predict cooked, cooksd
predict dfit, dfits
list  make r if abs(r) > 2  &  r ~=.
list make lever if lever > (2*3+2)/69 & lever ~=.
list make cooked if cooked >4/69 & cooked ~=.
list make dfit if abs(dfit)>2*sqrt(3/69) & dfit ~=.

dfbeta
list make _dfbeta_1 _dfbeta_2 _dfbeta_3 if abs(_dfbeta_1) > (2/sqrt(69)) &
    _dfbeta_1 ~=.
```

Linearity Assumption (Cont.)

```
. list make dfit if abs(dfit)>2*sqrt(3/69) & dfit ~=.
```

	make	dfit
2.	AMC Pacer	-1.004767
12.	Cad. Eldorado	.8511783
13.	Cad. Seville	1.077664
27.	Linc. Mark V	.4797307
28.	Linc. Versailles	.5269713
42.	Plym. Arrow	-1.015867
54.	Audi Fox	.6182262
74.	Volvo 260	.5247175

```
. list make _dfbeta_1 _dfbeta_2 _dfbeta_3 if abs(_dfbeta_1) > (2/sqrt(69)) & _dfbeta_1 ~=.
```

	make	_dfbeta_1	_dfbeta_2	_dfbeta_3
2.	AMC Pacer	-.9209325	.9525123	.0170096
12.	Cad. Eldorado	.47041	-.4156323	-.4082073
13.	Cad. Seville	.9955547	-.8688278	.1390504
27.	Linc. Mark V	.2530411	-.118375	.1010498
28.	Linc. Versailles	.4147299	-.3477834	.0203597
42.	Plym. Arrow	-.9222513	.9670225	.0297615
54.	Audi Fox	-.5201173	.4191374	-.2670405
70.	VW Dasher	-.2461434	.1960774	.0209733

```
.
```

Linearity Assumption (Cont.)

Solutions:

- Re-specify the model by mathematically transforming x's. e.g., for a curvilinear relation, you can square the x's.
  - log transform
  - exponentiation is the use of the inverse of a logarithm, as in $x' = \varepsilon^x$
  - polynomial transformation is the use of powers of the variable, as in $x' = x^2$, $x' = x^3$, x' = SQRT(x). We use this approach often in multiple regression.
  - rescale the x variable into a dummy (dichotomous) variable
- Restrict the range of x
- Identify the influential cases and examine whether they should be included in the sample

Mean Independence

What does it mean?

- The mean of the disturbance term is always 0 and does not depend on the value of x's.
- Possible causes of violating this assumption:
  - omitted x variables: if any of the omitted variables is associated with the x's.
  - reverse causation: if y influence x's, then ε is associated with the x's.
  - measurement error in the x: x includes not only x but also something else. This something else will get into ε.

What are the consequences?

- Biased estimates of intercept and regression coefficients
- Inaccurate prediction of Y

Mean Independence (Con.)

How to detect the violation?

Link test: if the current model is a good model, no additional predictors have significant associations with the dependent variable.

Mean Independence (Cont.)

SAS commands for Link test:

```
proc reg data=auto;
model price = length;
output out=auto2 (keep= price length yhat) predicted=yhat;
run;
quit;
data auto3;
set auto2;
yhat2= yhat**2;
run;
proc reg data=auto3;
model price = yhat yhat2;
run;
```

SAS results:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	6200.15909	7068.00012	0.88	0.3833
yhat	Predicted Value of price	1	-1.10941	2.35895	-0.47	0.6396
yhat2		1	0.00017179	0.00019107	0.90	0.3716

Mean Independence (Cont.)

Stata commands:

webuse auto.dta, clear

reg price length

predict yhat, xb

gen yhat2 = yhat*yhat

reg price yhat yhat2

Stata results:

```
. reg price yhat  yhat2

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  2,    71) =    8.63
       Model |  124242430      2  62121214.9            Prob > F      =  0.0004
    Residual |  510822966     71  7194689.67            R-squared     =  0.1956
-------------+------------------------------           Adj R-squared =  0.1730
       Total |  635065396     73  8699525.97            Root MSE      =  2682.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yhat |  -1.109417   2.358951    -0.47   0.640    -5.813032    3.594197
       yhat2 |   .0001718   .0001911     0.90   0.372    -.0002092    .0005528
       _cons |   6200.167       7068     0.88   0.383    -7893.024    20293.36
------------------------------------------------------------------------------
```

Mean Independence (Cont.)

Solutions:

- Use of past literatures to justify your model
- Use experimental design to collect your data, which not only support the mean independence assumption, but also avoid reverse causation
- If you use survey design and have measures of relevant variables that have not been included in the model, you can include these variables in the model to reduce the possibility of violating this assumption
- Use simultaneous equations to model reciprocal relations between x's and y
- Choose measures with high reliability or include measurement models in regression analysis

Homoscedasticity

What does it mean?

- Homoscedasticity means that the variance of ε is the same across all levels of x's.
- Possible causes of violating this assumption.
  - Improvement in data collection techniques: During the course of data collection, the interviewers are getting better and less likely to commit error in collecting data.
  - Learning: Respondents are less likely to have errors in answering the same questions when being interviewed in the follow-up survey than in the baseline survey.
  - Outliers

What are the consequences?

- Inefficiency: observations with larger disturbance variance contain less information than observations with smaller disturbance variance. but OLS weights them equally.
- Bias in standard errors can leads to incorrect conclusions.

Homoscedasticity (Cont.)

How to detect the violation?

- Plot residuals against X
- Plot residuals against Y_{hat}
- White test

SAS commands:

```
proc reg data=auto;
model price = length weight rep78/ spec;
run; quit;
```

Homoscedasticity (Cont.)

Solutions:

- Re-specify the model or transform the dependent variable
- Use robust standard errors
- Use weighted least squares only if you know what weights to use

Uncorrelated Disturbances

What it means?

- The disturbance variables for any two individuals must be uncorrelated.
- Possible causes of violating this assumption
  - Sample design: simple random sampling is not likely to cause this problem, but a cluster sampling is.
  - The selection of unit of analysis, e.g., the couple
  - The use of panel data

What are the consequences?

- Inefficient estimates
- Downward bias in estimated standard errors, which means that there will be a tendency to conclude that relations exist when they really don't.

Uncorrelated Disturbances (Cont.)

How to detect the violation?

- Calculate the residuals for all respondents and then examine correlations between the residuals of suspected groups of respondents
- Intra-class correlation

Solutions:

- Include the cluster variables into the models as a control
- Use regression with robust standard errors
- Use generalized least squares

Uncorrelated Disturbances (Cont.)

Solutions:

- Including the correlations among respondents into the regression models

SAS commands:

```
proc genmod data=auto;
class foreign;
Model price = price weight rep78;
repeated subject=foreign / type=ind ;
run;
```

Stata commands:

```
reg price length rep78 weight, cluster(foreign)
```

Normality

What does it mean:

- The disturbance term ε need to be normally distributed, but x's and y do not.
  - Positive Skewness
  - Negative Skewness
  - Positive Kurtosis
  - Negative Kurtosis
- Possible causes of violation of this assumption
  - The true distribution of the variable, e.g., some variables follow a binomial or poisson distribution.
  - Measurement artifacts
  - Inadequate sample

What are the consequences?

- When the sample is extremely small (e.g., below 100), the violation of this assumption leads to inaccurate estimates of confidence intervals and p-values. As the sample gets larger, the central limit theorem suggested that we can get pretty accurate confidence intervals and p-values.

Normality (Cont.)

How to detect the violation:

- Graphic methods: Stem-and-leaf plot, (skeletal) box plot, dot plot, histogram
- Shapiro-Wilk W test for normality

Normality (Cont.)

SAS Commands:

```
proc reg data=auto;
model price= length weight rep78;
output out=auto2 (keep= price length weight rep78 res yhat)
residual=res predicted=yhat;
run;

proc univariate data=auto2 normal;
var res;
qqplot res / normal(mu=est sigma=est);
run;
```

Stata commands:

```
reg price length rep78 weight
swilk r
```

Normality (Cont.)

Solutions:

- Using larger samples
- Using conservative p-values (e.g., using 0.01 rather than 0.05)

Conclusions

- Regression analysis is the most commonly used technique in social sciences
- To accurately use regression analysis, you need to check for possible violations of the regression analysis
- Other useful resources for learning conducting regression
  - http://www.ats.ucla.edu/stat/sas/webbooks/reg/default.htm
  - http://www.ats.ucla.edu/stat/stata/webbooks/reg/
  - http://www.indiana.edu/~statmath/stat/all/panel/
  - http://dss.princeton.edu/online_help/analysis/regression_intro.htm
- If you have any questions about running regression analysis, CFDR provides programming support. Please feel free to contact Hsueh-Sheng Wu @ 372-3119 or wuh@bgsu.edu