Final review!!!

Week 14

There will be a quiz next week

Timestamper!

Fill out course evaluations

- Florida attorney general suspected contractors were illegally setting bids for government contracts by fixing them higher than they would be if truly competitive. They collected several measures to predict the price of a contract bid by the lowest bidder (in dollars):
  - Engineer's estimate of a fair contract price (in dollars)
  - Ratio of low (winning) bid price to DOT engineer's estimate of fair price
  - Status of contract (1 if fixed)
  - District of construction project (1,2,3,4, or 5)
  - Number of bidders on contract
  - Estimated number of days to complete work
  - Length of road project
  - Percentage of costs allocated to asphalt
  - Percentage of costs allocated to base material
  - Percentage of costs allocated to excavation
  - Percentage of costs allocated to mobilization
  - Percentage of costs allocated to structures
  - Percentage of costs allocated to traffic control
  - Subcontractor utilization (1 if yes)

- Found a nice model last time, where we predicated the winning bid using the number of bidders and the number of days to complete the work. Now lets test if anything related to the allocation of money on the project can help predict the winning bid (all the percentage of costs variables).
  - Engineer's estimate of a fair contract price (in dollars)
  - Ratio of low (winning) bid price to DOT engineer's estimate of fair price
  - Status of contract (1 if fixed)
  - District of construction project (1,2,3,4, or 5)
  - Number of bidders on contract
  - Estimated number of days to complete work
  - Length of road project
  - Percentage of costs allocated to asphalt
  - Percentage of costs allocated to base material
  - Percentage of costs allocated to excavation
  - Percentage of costs allocated to mobilization
  - Percentage of costs allocated to traffic control
  - Subcontractor utilization (1 if yes)
- What model do they want to test?

  $Y= \beta_0 + \beta_1 Numbids + \beta_2 Daysest + \beta_3 PctAsph + \beta_4 PctBase + \beta_5 PctExcav + \beta_6 PctMobil + \beta_7 PctTraff + \varepsilon$
- What procedure should you use to compare this new model to the original ($Y= \beta_0 + \beta_1 Numbids + \beta_2 Daysest + \varepsilon$) ?
  - Nested F

- Doesn't look like any significant, but?
- Can we just take them all out?
  - 81% correct
  - Can't just remove them all because they could be competing with each other, so don't look significant when they are important
  - What if just 1 nonsig predictor? 55%

Nested F

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-741784.83	139936.259		-5.301	.000
	DAYSEST	7835.210	376.721	.778	20.798	.000
	NUMBIDS	54624.411	25315.921	.081	2.158	.032

a. Dependent Variable: LOWBID

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-767821.053	212392.127		-3.615	.000
	NUMBIDS	55003.502	28088.628	.081	1.958	.051
	DAYSEST	7870.876	411.490	.781	19.128	.000
	PCTASPH	145771.142	260004.489	.023	.561	.576
	PCTBASE	-342571.622	693348.486	-.021	-.494	.622
	PCTEXCAV	249527.248	614700.375	.017	.406	.685
	PCTMOBIL	117893.596	1048586.350	.004	.112	.911
	PCTTRAFF	-459713.286	544140.596	-.031	-.845	.399

a. Dependent Variable: LOWBID

Nested F

$$F = \frac{(SSModel_{full} - SSModel_{subset}) / \#\ predictors\ tested}{SSE_{full}/(n-k-1)}$$

$$F = \frac{(602{,}570{,}309{,}493{,}757 - 600{,}661{,}479{,}096{,}113)/ 5}{321{,}941{,}908{,}903{,}964/(279-7-1)} = .321$$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	600661479096113.500	2	300330739548056.750	255.955	.000[b]
	Residual	323850739301607.200	276	1173372243846.403		
	Total	924512218397720.800	278			

Reduced

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	602570309493757.000	7	86081472784822.420	72.461	.000[b]
	Residual	321941908903964.000	271	1187977523630.864		
	Total	924512218397721.000	278			

Full

a. Dependent Variable: LOWBID

b. Predictors: (Constant), PCTTRAFF, PCTMOBIL, PCTBASE, DAYSEST, PCTASPH, NUMBIDS, PCTEXCAV

$$F = \frac{(SSModel_{full} - SSModel_{subset})/ \#\ predictors\ tested}{SSE_{full}/(n-k-1)}$$

Nested F

$$F = \frac{(602{,}570{,}309{,}493{,}757 - 600{,}661{,}479{,}096{,}113)/ 5}{321{,}941{,}908{,}903{,}964/(279-7-1)} = .321$$

- With numerator degrees of freedom equal to the number of predictors being tested
  - 5
- denominator degrees of freedom equal to the error degrees of freedom for the full model
  - 271

duced

Full

M						
1						
	Total	720.800				

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	602570309493757.000	7	86081472784822.420	72.461	.000[b]
	Residual	321941908903964.000	271	1187977523630.864		
	Total	924512218397721.000	278			

a. Dependent Variable: LOWBID

b. Predictors: (Constant), PCTTRAFF, PCTMOBIL, PCTBASE, DAYSEST, PCTASPH, NUMBIDS, PCTEXCAV

Nested F

- Critical value of F for an alpha of .05, for our degrees of freedom = 2.21, our F-statistic: .321

H_0: $\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=0$

H_a: not all betas are zero

Since our test statistic is less than the critical value for an alpha of .05, we fail to reject the null hypothesis. This suggests that variables about the allocation of money on the project are not useful in predicting the average winning bid, after accounting for the number of bids and the estimated days to complete the project.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	600661479096113.500	2	300330739548056.750	255.955	.000[b]
	Residual	323850739301607.200	276	1173372243846.403		
	Total	924512218397720.800	278			

a. Dependent Variable: LOWBID

b. Predictors: (Constant), DAYSEST, NUMBIDS

Reduced

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	602570309493757.000	7	86081472784822.420	72.461	.000[b]
	Residual	321941908903964.000	271	1187977523630.864		
	Total	924512218397721.000	278			

a. Dependent Variable: LOWBID

b. Predictors: (Constant), PCTTRAFF, PCTMOBIL, PCTBASE, DAYSEST, PCTASPH, NUMBIDS, PCTEXCAV

Full

Logistic regression

- When a response variable is binary
  - Can we predict when a contract was fixed based on the lowest bid and the number of bidders?
- Assumptions
  - Linearity
  - Independence
    - No pairing or clustering of the data in space or time
    - Each project has many different aspects to it, not unreasonable to think they are independent.
  - Randomness
    - Want random sample from population or random assignment within an experiment
    - Doesn't really say, probably took all bids within a time period, or randomly selected. Both of which would be ok….
    - Spinner model appropriate for deciding if fixed? –no, but still useful for looking at relationships

Linearity: Box-Tidwell

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

Lowest bid: Do we need to scale?

Since p>.05, we fail to reject the null hypothesis. This suggests that we meet the assumption of linearity for LowBid and the log odds of whether a contract is fixed or not.

- Do the same thing for numbids

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LOWBID	.000	.000	.398	1	.528	1.000
	LOWBID by logLowbid	.000	.000	.390	1	.532	1.000
	Constant	-.589	.224	6.937	1	.008	.555

a. Variable(s) entered on step 1: LOWBID, LOWBID * logLowbid .

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	64.781	2	.000
	Block	64.781	2	.000
	Model	64.781	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	288.991[a]	.207	.288

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			STATUS		Percentage Correct
	Observed		0	1	
Step 1	STATUS	0	167	20	89.3
		1	50	42	45.7
	Overall Percentage				74.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LOWBID	.000	.000	8.351	1	.004	1.000
	NUMBIDS	-.619	.100	38.066	1	.000	.538
	Constant	1.591	.366	18.861	1	.000	4.907

- Assess model

H_0: $\beta_1 = \beta_2 = 0$

H_a: at least one beta is not equal to zero

Since $p<.001$, we reject the null hypothesis, this suggests that the lowest bid and number of bids are useful in predicting the odds of whether a contract is fixed or not.

The model correctly predicts 74.9% of contracts, and does better (89%) for non-fixed contracts than fixed (45.7)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	64.781	2	.000
	Block	64.781	2	.000
	Model	64.781	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	288.991[a]	.207	.288

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			STATUS		Percentage Correct
	Observed		0	1	
Step 1	STATUS	0	167	20	89.3
		1	50	42	45.7
	Overall Percentage				74.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LOWBID	.000	.000	8.351	1	.004	1.000
	NUMBIDS	-.619	.100	38.066	1	.000	.538
	Constant	1.591	.366	18.861	1	.000	4.907

- Test the betas

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Since p<.01, we reject the null hypothesis, this suggests that there is a significant log-linear relationship between the lowest bid and the odds of whether a contract is fixed or not, after accounting for the number of bids.

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

Since p<.001, we reject the null hypothesis, this suggests that there is a significant linear relationship between the number of bids and the log odds of whether a contract is fixed or not, after accounting for the lowest bid.

Beta for lowbid very small because the numbers are so large. We can scale it into $10,000

Follow up

- What is the regression equation?
- Log(odds^)=1.591 +.003Lowbidscaled -.619Numbids
- What is the probability of a contract being fixed if the lowest bid was $200,000 and the number of bidders was 5.

$$\pi = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \ldots + \beta_k X_k)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \ldots + \beta_k X_k)}}$$

Lets figure out what the equation comes to:

1.591 + .003(200,000/10,000) -.619(5) =-1.44

π= e^-1.44 / 1+e^-1.44

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lowbidscaled	.003	.001	8.351	1	.004	1.003
	NUMBIDS	-.619	.100	38.066	1	.000	.538
	Constant	1.591	.366	18.861	1	.000	4.907

a. Variable(s) entered on step 1: lowbidscaled, NUMBIDS.

Follow up

- Interpret the beta for numbids:
- For every additional bidder, the odds of a contract being fixed decreases by a factor of .538, after accounting for the lowest bid.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lowbidscaled	.003	.001	8.351	1	.004	1.003
	NUMBIDS	-.619	.100	38.066	1	.000	.538
	Constant	1.591	.366	18.861	1	.000	4.907

a. Variable(s) entered on step 1: lowbidscaled, NUMBIDS.

Can we do better?

- Hypothesize the DOT estimate of a fair bid might be important in predicting whether a contract is fixed or not, and that the relationship between the lowest bid and the status of a contract might depend on the DOT estimate of the fair bid.
- What equation do they want to test?
- $\log(odds)=\beta_0 + \beta_1 Lowbid + \beta_2 Numbids + \beta_3 DOTest + \beta_4 LowbidDOTest$

Is this a better model?

- Significant betas, but is it making the model significantly better? How do we answer?
- LRT test – likelihood ratio test

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LOWBID	.000	.000	10.000	1	.002	1.000
	NUMBIDS	-.601	.107	31.793	1	.000	.549
	DOTEST	.000	.000	10.120	1	.001	1.000
	DOTEST by LOWBID	.000	.000	3.412	1	.065	1.000
	Constant	1.725	.401	18.490	1	.000	5.612

a. Variable(s) entered on step 1: LOWBID, NUMBIDS, DOTEST, DOTEST * LOWBID .

Nested likelihood ratio test

H_0: $\beta_{DOTest}=\beta_{LowbidDOTest}=0$

H_a: not all betas are zero

Test statistic: -2LLNested - -2LLFull: 288.991-272.112 =16.879

Degrees of freedom = # of predictors tested = 2

What if significant?

We reject the null, suggests that the DOT's estimate of a fair bid, and the interaction of this and the lowest bid are useful in predicting the odds of a contract being fixed, after accounting for the number of bids and the lowest bid, and that we should use the full model

What if nonsignificant?

We fail to reject the null hypothesis, this suggests that the DOT's estimate of fair bid and the interaction of this and the lowest bid are not useful in predicting the odds of a contract being fixed, after accounting for the number of bids and the lowest bid. This suggests that we should use the nested model.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	272.112[a]	.254	.353

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Full

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	288.991[a]	.207	.288

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Nested

Non-parametric

- Randomization/Permutation:
  - Disrupt the pairing but use all the data
- Bootstrap:
  - Keep the pairing, but choose a ‘new’ sample
- Wilcoxon-Mann-Whitney and Kruskal-Wallis
  - Based on ranking of data, and medians, not means
- Commonly used when:
  - Don’t meet assumptions like normality or equal variance
  - Small sample sizes
  - Outliers
- Would you ever run if the assumptions are met? 88%

Bootstrap

- Can we predict the length of the project from how much of the budget goes to different things?
  - Initial testing settled on three variables
  - Percentage of costs allocated to excavation, mobilization, and structures

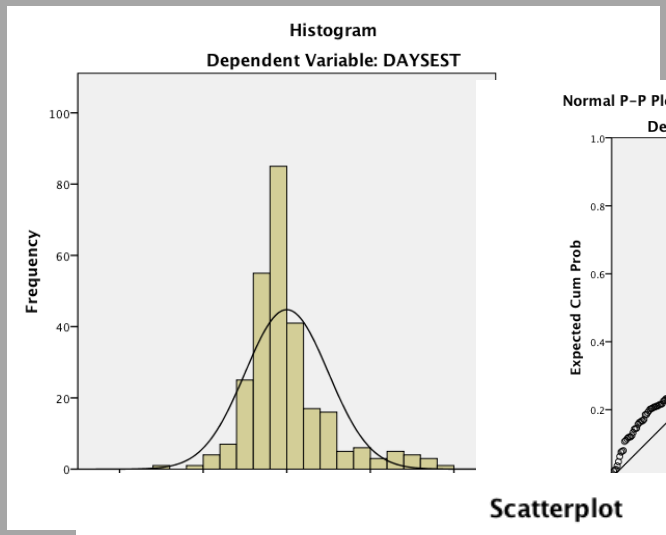
Coefficients[a]

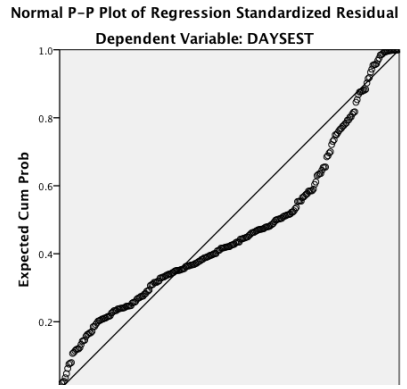
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	70.642	16.845		4.194	.000
	PCTEXCAV	388.156	79.787	.266	4.865	.000
	PCTMOBIL	628.951	143.752	.228	4.375	.000
	PCTSTRUC	478.512	73.659	.353	6.496	.000

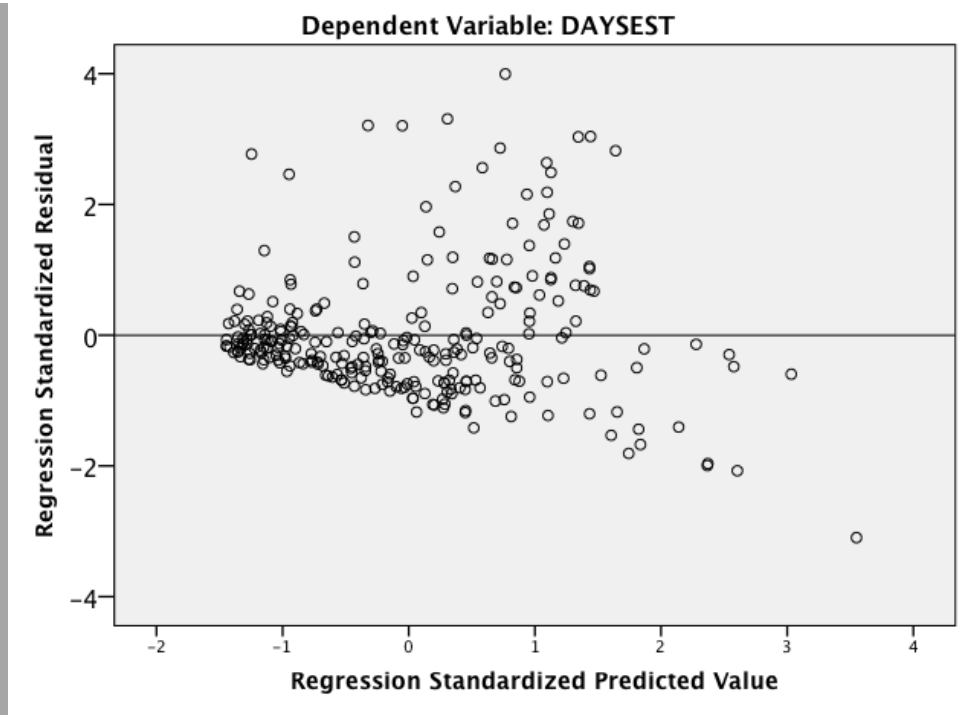
a. Dependent Variable: DAYSEST

Let's check assumptions

- Anything look weird?
- Not normally distributed
- 'the plot thickens'
- Several outliers







Bootstrap to the rescue!

- p-values all slightly less significant, but still very significant. Percent spent on structures is most biased (10.735).
- What is bias?
  - How much the average bootstrapped test statistic differs from the original value.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	70.642	16.845		4.194	.000	37.481	103.803
	PCTEXCAV	388.156	79.787	.266	4.865	.000	231.085	545.227
	PCTMOBIL	628.951	143.752	.228	4.375	.000	345.956	911.946
	PCTSTRUC	478.512	73.659	.353	6.496	.000	333.504	623.520

a. Dependent Variable: DAYSEST

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	70.642	-1.613	13.429	.001	45.993	92.611
	PCTEXCAV	388.156	3.676	110.252	.002	194.246	622.534
	PCTMOBIL	628.951	-2.261	184.526	.001	278.213	980.708
	PCTSTRUC	478.512	10.735	118.787	.001	267.704	757.822

- Test and interpret beta for PctStruc

H_0: $\beta_3 = 0$

H_a: $\beta_3 \neq 0$

Since p<.001, we can reject the null hypothesis. This suggests that there is a linear relationship between the percent spent on structures and the average estimated length of a project, after accounting for the percent spent on excavation and mobilization. Given that the assumptions of normality and equal variance were not met, we ran a bootstrap analysis. This confirmed the above results (p<.01).

For every additional percentage point spent on structures, the average estimated length of the project is expected to increase by 478.512. Using the bootstrap analysis with biased corrected confidence intervals, we can be 95% confident that the actual average increase will be between 267.704 and 757.822, after accounting for the percent spent on excavation and mobilization.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	70.642	16.845		4.194	.000	37.481	103.803
	PCTEXCAV	388.156	79.787	.266	4.865	.000	231.085	545.227
	PCTMOBIL	628.951	143.752	.228	4.375	.000	345.956	911.946
	PCTSTRUC	478.512	73.659	.353	6.496	.000	333.504	623.520

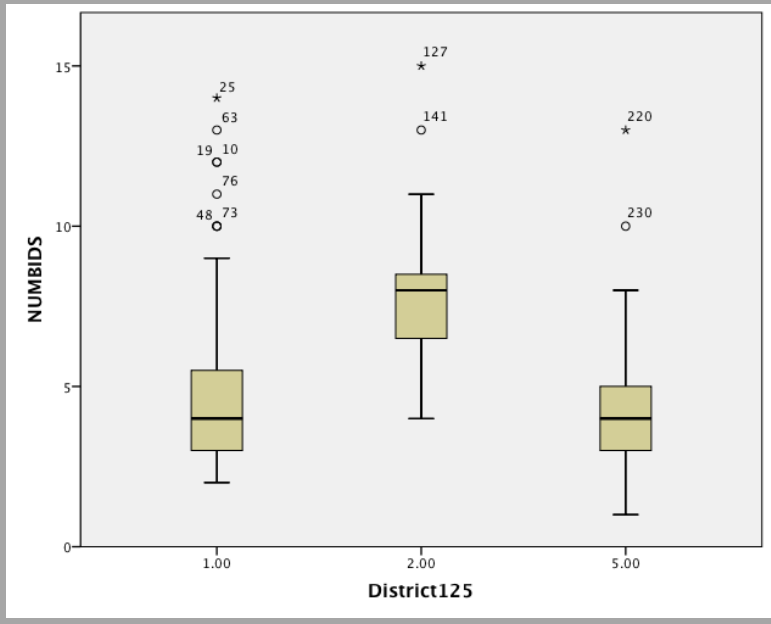
a. Dependent Variable: DAYSEST

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	70.642	-1.613	13.429	.001	45.993	92.611
	PCTEXCAV	388.156	3.676	110.252	.002	194.246	622.534
	PCTMOBIL	628.951	-2.261	184.526	.001	278.213	980.708
	PCTSTRUC	478.512	10.735	118.787	.001	267.704	757.822

Kruskal-Wallis

- We might hypothesize there is more corruption in certain districts. Since we have shown that the number of bidders correlates with the status of a contract, lets see if the number of bidders differs by district.
- What kind of test?
  - One-way ANOVA



Check assumptions

- Normality?
  - Normal P-P plot looks ok
- Equal variance?
  - Since p<.05, we reject the null for Levene's test (H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, H_a: Not all variances are equal), suggesting that we fail the equal variance assumption. However, the residual plot seems to have similar spread for the tree districts, and the rule of thumb (2.678/1.95 =1.37) is less than 2, so suggests we meet assumption. Taken together, 2 out of 3 tests suggest we meet the assumption.
- Outliers?
  - Yes, several standardized residuals above 2

Descriptive Statistics

Dependent Variable: NUMBIDS

District125	Mean	Std. Deviation	N
1.00	4.61	2.678	119
2.00	7.71	2.532	31
5.00	4.13	1.950	122
Total	4.75	2.591	272

Levene's Test of Equality of Error Variances[a]

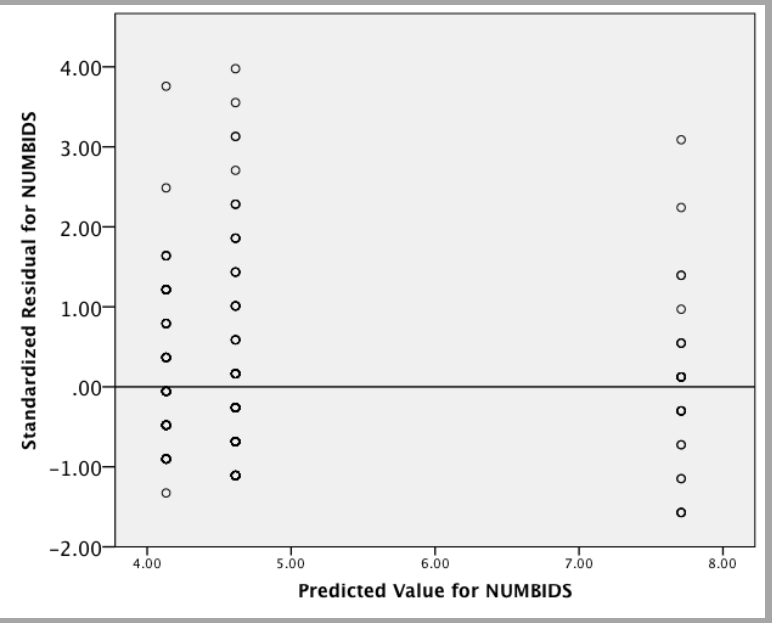
Dependent Variable: NUMBIDS

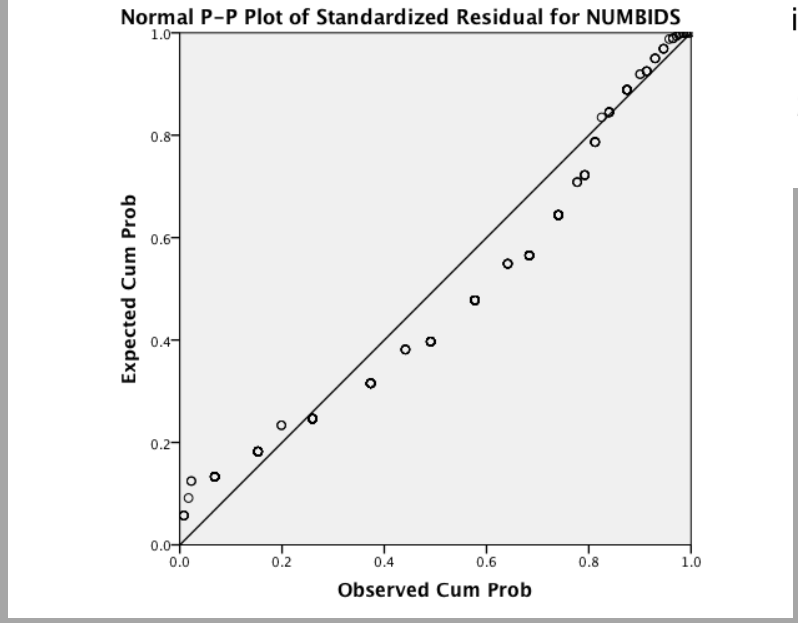
F	df1	df2	Sig.
4.428	2	269	.013

Tests the null hypothesis that the error

iable is equal

25





Kruskal-Wallis

H_0: $\theta_1 = \theta_2 = \theta_3$

H_a: medians are not all equal

- Since p<.001, we reject the null hypothesis, this suggests that the median number of bidders differs for the three districts.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of NUMBIDS is the same across categories of District125.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

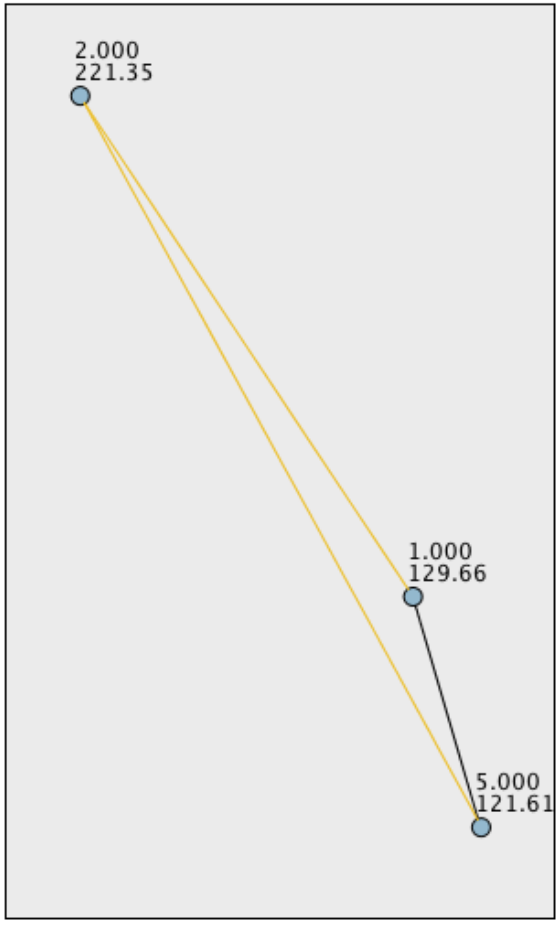
Follow-up: posthoc tests

$H_0: \theta_i = \theta_j$

$H_a: \theta_i \neq \theta_j$

Since p<.001 for district 2 versus 5 and 1, we can reject the null hypothesis. This suggests that the median number of bidders is larger for district two than the other districts. For district 1 versus 5, p>.05, and we fail to reject the null hypothesis. This suggests that the median number of bidders is not statistically different for these two districts.

Pairwise Comparisons of District125



ch node shows the sample average rank of District125.

Each node shows the sample average rank of District125.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
5.000-1.000	8.057	10.001	.806	.420	1.000
5.000-2.000	99.748	15.613	6.389	.000	.000
1.000-2.000	-91.691	15.653	-5.858	.000	.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Good Luck!!