Week6: Review!

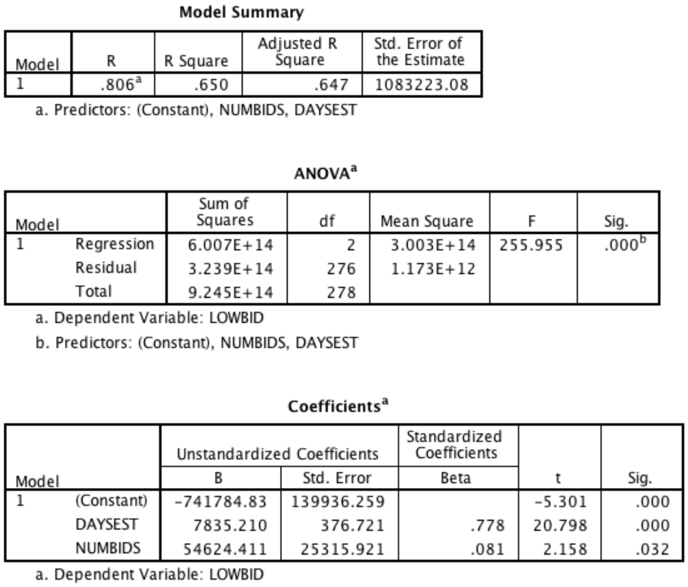
Exam and review session:

Science Center Hall D

Timestamper!!

- 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)

- You want to investigate how the winning bid varies with the number of bidders and the number of days to complete the work.
  - 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)
- What kind of test should you use?
  - Multiple regression



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.806[a]	.650	.647	1083223.08

a. Predictors: (Constant), NUMBIDS, DAYSEST

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.007E+14	2	3.003E+14	255.955	.000[b]
	Residual	3.239E+14	276	1.173E+12		
	Total	9.245E+14	278			

a. Dependent Variable: LOWBID

b. Predictors: (Constant), NUMBIDS, DAYSEST

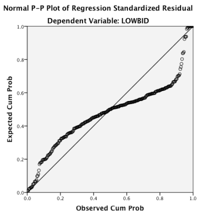
Coefficients[a]

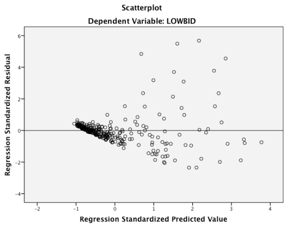
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-741784.83	139936.259		-5.301	.000
	DAYSEST	7835.210	376.721	.778	20.798	.000
	NUMBIDS	54624.411	25315.921	.081	2.158	.032

a. Dependent Variable: LOWBID

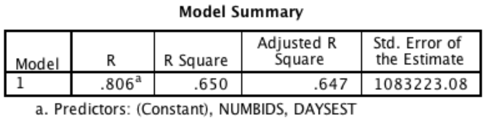
Check assumptions

- Conditions for multiple regression and how to check them?
  - **Zero Mean** - the distribution of the errors is centered at zero – always true with least squares regression!
  - **Constant Variance** - the variability of the errors is the same for all values of the predictor variable
    - Residual plot
  - **Independence** - the errors are independent of each other
    - Study design
  - **Normality** – In order to use standard distributions for confidence intervals and hypothesis tests, we often need to assume the random errors follow a normal distribution.
    - NPP plot, histogram
- **Random** – the data are obtained using a random process, like random sampling from a population of interest.
  - Study design





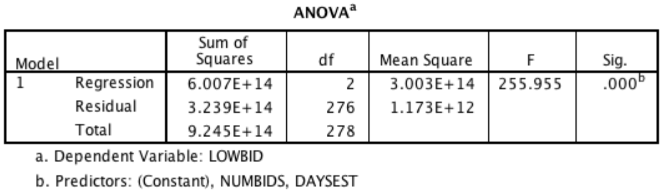
- Constant variance?
  - No, looks like the plot thickens, might want to transform by taking the log or square root of one or both of predictors or the response
- Normality?
  - A lot of dots far from the line…. Probably not
- Independence, random selection?
  - Didn't randomly select, so can't extent beyond our sample.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.806[a]	.650	.647	1083223.08

a. Predictors: (Constant), NUMBIDS, DAYSEST



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.007E+14	2	3.003E+14	255.955	.000[b]
	Residual	3.239E+14	276	1.173E+12		
	Total	9.245E+14	278			

a. Dependent Variable: LOWBID

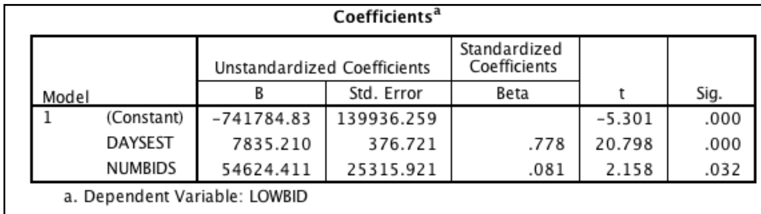
b. Predictors: (Constant), NUMBIDS, DAYSEST

- Assess the model
- What do you need to answer this question?
  - Hypotheses, decision, conclusion
- Asses the model

H_0: $\beta_1=\beta_2=0$, H_a: At least one beta is not equal to 0

Since the p-value is small ($p<.001$), we can reject the null hypothesis, this suggests that the number of bids and estimated length of the project combine to predict a significant amount of variability in the lowest bid.

- What predictors are useful?



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

- Length of project

H_0: $\beta_{daysest} = 0$, H_a: $\beta_{daysest} \neq 0$

Since the p-value is small ($p<.001$) we can reject the null hypothesis. This suggests that the length of the project has a significant positive linear relationship with the average winning bid, after accounting for the number of bids.

- Number of bids

H_0: $\beta_{numbids} = 0$, H_a: $\beta_{numbids} \neq 0$

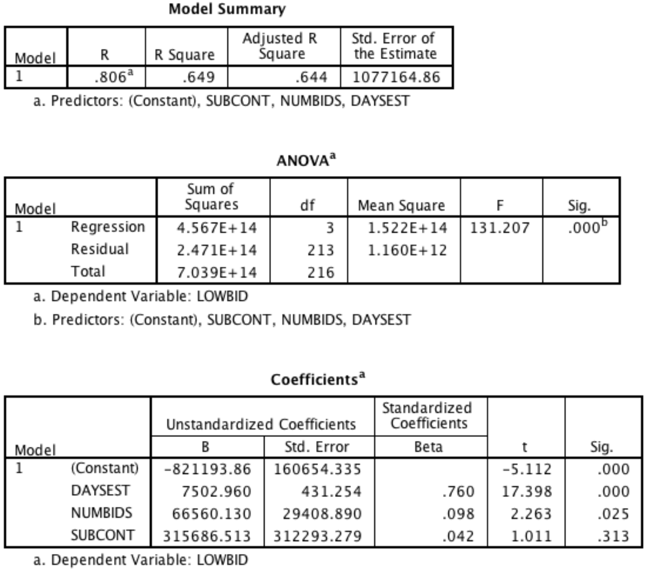
Since the p-value is small ($p<.05$), we can reject the null hypothesis, suggesting that the number of bids is positively and linearly related to the average winning bid, after accounting for the length of the project.

- Interpret $\beta_{numbids}$
  - For each additional bidder on the contract, the price of the winning bid increases on average by $54,624.41, after accounting for the length of the project.
- What if we wanted to investigate whether we needed two lines (or more accurately, two planes as there are two quantitative predictors) to represent the relationship between Daysest and Numbids and the winning bid, one for when a subcontractor is used, one for when it isn't. What model should we test?

$Y= \beta_0 + \beta_1$Daysest $+ \beta_2$Numbids $+ \beta_3$Subcont$+ \varepsilon$

- What would we need to test to answer the above question?

β_3



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.806[a]	.649	.644	1077164.86

a. Predictors: (Constant), SUBCONT, NUMBIDS, DAYSEST

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.567E+14	3	1.522E+14	131.207	.000[b]
	Residual	2.471E+14	213	1.160E+12		
	Total	7.039E+14	216			

a. Dependent Variable: LOWBID

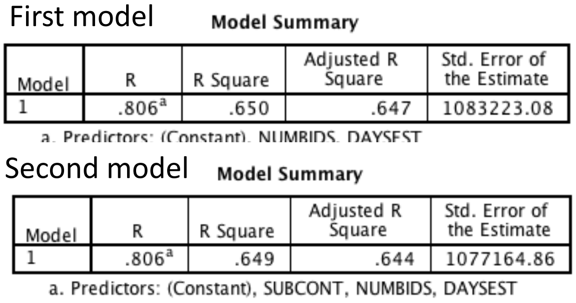
b. Predictors: (Constant), SUBCONT, NUMBIDS, DAYSEST

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-821193.86	160654.335		-5.112	.000
	DAYSEST	7502.960	431.254	.760	17.398	.000
	NUMBIDS	66560.130	29408.890	.098	2.263	.025
	SUBCONT	315686.513	312293.279	.042	1.011	.313

a. Dependent Variable: LOWBID

- If you had to choose between these two models, which would you choose?



First model **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.806[a]	.650	.647	1083223.08

a. Predictors: (Constant), NUMBIDS, DAYSEST

Second model **Model Summary**

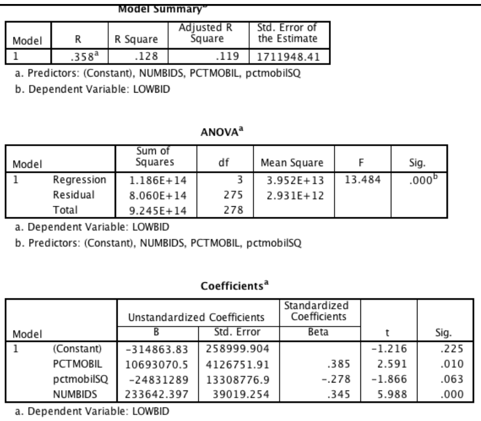
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.806[a]	.649	.644	1077164.86

a. Predictors: (Constant), SUBCONT, NUMBIDS, DAYSEST

- If you had to choose between these two models, which would you choose?
  - The adjusted R^2 is larger for the first model (64.7%) versus the second one (64.4%), suggesting that it accounts for more variability in the winning bid, after accounting for the number of predictors and sample size. However, the standard error of the estimate is smaller for the second model (1,077,164.86 versus 1,083,223.08), suggesting the typical error for the second model is smaller. To help make the final decision, I would take into consideration the fact that the second model has a non-significant term (H_0: $\beta_{subcontract} = 0$, H_a: $\beta_{subcontract} \neq 0$, $p>.05$), suggesting the model is more complicated than necessary. Therefore, I would chose the first model.
- What if we thought that the relationship between the number of bidders and the winning bid might vary depending on the length of the project, what model would we want to test?

$Y= \beta_0 + \beta_1 Daysest + \beta_2 Numbids + \beta_3 DayestNumbids + \varepsilon$

- You are interested in how the winning bid varies with the the number of bids and the percentage allocated to mobilization. After looking at the data, they thought they should include a square term for percentage of mobilization.
  - 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)
- What test would you use?
  - Multiple regression



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.358[a]	.128	.119	1711948.41

a. Predictors: (Constant), NUMBIDS, PCTMOBIL, pctmobilSQ

b. Dependent Variable: LOWBID

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.186E+14	3	3.952E+13	13.484	.000[b]
	Residual	8.060E+14	275	2.931E+12		
	Total	9.245E+14	278			

a. Dependent Variable: LOWBID

b. Predictors: (Constant), NUMBIDS, PCTMOBIL, pctmobilSQ

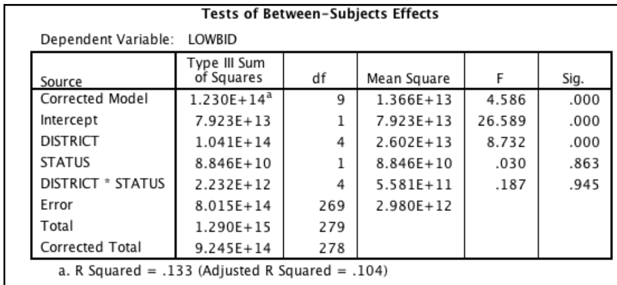
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-314863.83	258999.904		-1.216	.225
	PCTMOBIL	10693070.5	4126751.91	.385	2.591	.010
	pctmobilSQ	-24831289	13308776.9	-.278	-1.866	.063
	NUMBIDS	233642.397	39019.254	.345	5.988	.000

a. Dependent Variable: LOWBID

- Do you need the square term?
  - H_0: $\beta_{pctmobilsq}$ = 0, H_a: $\beta_{pctmobilsq} \neq 0$
  - Since p>05, we fail to reject the null hypothesis. This suggests that the square term is not useful in predicting average winning bid, after accounting for the number of bids and the linear relationship between pctmobil and the winning bid.
- Would you use this model?
  - The overall model is signficant (H_0: $\beta_1=\beta_2=\beta_3=0$, H_a: At least one beta is not equal to 0,p<.001), though it is predicting a very small amount of the variability in the winning bid (adjR^2=.119). In addition, the higher order square term is non-significant. Given this, I would not use the model. I might try taking out the higher order term and seeing what happens.
- What is the direction of the relationship between the percentage allocated to mobilization expenses and the winning bid?
  - With the square term in there, we can't just look at the slope for PctMobile. However, the square term is negative, suggesting that the relationship is some part of an inverted U shape.

- You are interested in how the winning bid varies with the District of the construction project and whether the status of a contract is fixed or not.
  - 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)
- What test would you use?
  - Two-way ANOVA



Tests of Between-Subjects Effects

Dependent Variable: LOWBID

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1.230E+14[a]	9	1.366E+13	4.586	.000
Intercept	7.923E+13	1	7.923E+13	26.589	.000
DISTRICT	1.041E+14	4	2.602E+13	8.732	.000
STATUS	8.846E+10	1	8.846E+10	.030	.863
DISTRICT * STATUS	2.232E+12	4	5.581E+11	.187	.945
Error	8.015E+14	269	2.980E+12		
Total	1.290E+15	279			
Corrected Total	9.245E+14	278			

a. R Squared = .133 (Adjusted R Squared = .104)

Posthoc tests?

- Test the main effects
- District

H_0: $\mu_{distr1} = \mu_{distr2} = \mu_{distr3} = \mu_{distr4} = \mu_{distr5}$

H_a: the means are not all equal

Since the p-value is small (p<.001), we can reject the null hypothesis, suggesting that there is a significant main effect of district, and that the average winning bid differs for the districts.

- Status

H_0: $\mu_{fixed} = \mu_{notfixed}$ H_a: $\mu_{fixed} \neq \mu_{notfixed}$

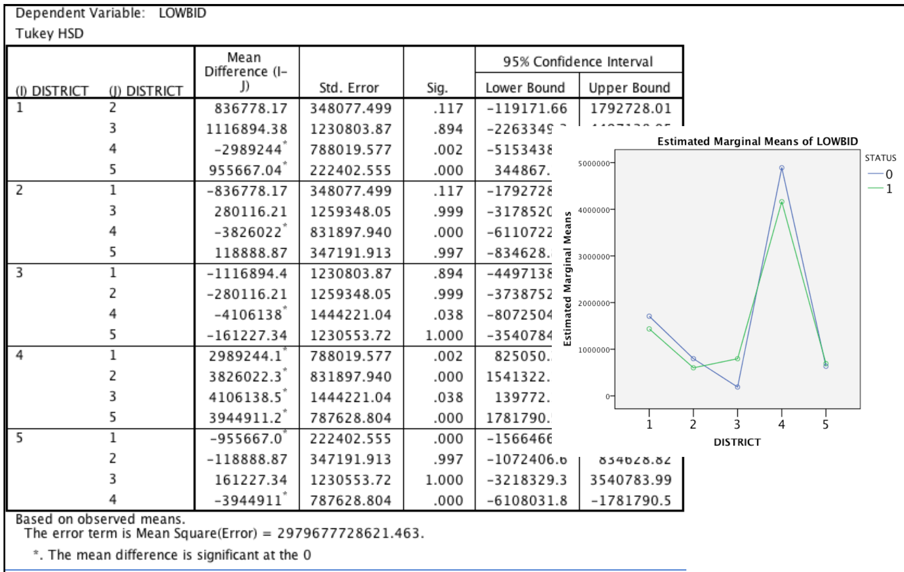
Since the p-value is large (p=.863), we fail to reject the null hypothesis. This suggests that there is not a significant difference in the average winning bid for whether it is fixed or not.

- Is there a significant interaction? And what implication does that have for interpreting the main effects?

H_0: the main effect of each factor is the same for each level of the other factor

H_a: the two factors interact

Since the p-value is large (p=.945), we fail to reject the null hypothesis. This suggest that the affect of district is the same for all levels of status. This means that we can interpret the main effects like normal. If it had been significant, we could not interpret the main effects, we would have to pull the two factors apart and test them separately.



Dependent Variable: LOWBID
Tukey HSD

(I) DISTRICT	(J) DISTRICT	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	836778.17	348077.499	.117	-119171.66	1792728.01
	3	1116894.38	1230803.87	.894	-2263349	[illegible]
	4	-2989244*	788019.577	.002	-5153438	[illegible]
	5	955667.04*	222402.555	.000	344867.	[illegible]
2	1	-836778.17	348077.499	.117	-1792728	[illegible]
	3	280116.21	1259348.05	.999	-3178520	[illegible]
	4	-3826022*	831897.940	.000	-6110722	[illegible]
	5	118888.87	347191.913	.997	-834628.	[illegible]
3	1	-1116894.4	1230803.87	.894	-4497138	[illegible]
	2	-280116.21	1259348.05	.999	-3738752	[illegible]
	4	-4106138*	1444221.04	.038	-8072504	[illegible]
	5	-161227.34	1230553.72	1.000	-3540784	[illegible]
4	1	2989244.1*	788019.577	.002	825050.	[illegible]
	2	3826022.3*	831897.940	.000	1541322.	[illegible]
	3	4106138.5*	1444221.04	.038	139772.	[illegible]
	5	3944911.2*	787628.804	.000	1781790.	[illegible]
5	1	-955667.0*	222402.555	.000	-1566466	[illegible]
	2	-118888.87	347191.913	.997	-1072406.6	834628.82
	3	161227.34	1230553.72	1.000	-3218329.3	3540783.99
	4	-3944911*	787628.804	.000	-6108031.8	-1781790.5

Based on observed means.
The error term is Mean Square(Error) = 2979677728621.463.
*. The mean difference is significant at the 0

H_0: $\mu_i = \mu_j$ H_a: $\mu_i \neq \mu_j$

Bids from district4 are significantly higher on average than bids from all other districts (4 vs 1 p<.01, 4 vs 2 p<.001, 4 vs 3 p<.05, 4 vs 5 p<.001). In addition, district1 is higher on average than district5 (p<.001).

- You are interested in how the winning bid varies by district and the percentage of costs allocated to traffic control.
  - 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)
- What test would you use?
  - Multiple regression

- What model would you use to test this?

I would need to change the variable district into 4 dummy variables, using district1 as the base category:

Ifdistrict2: 1 if from district 2, 0 otherwise

Ifdistrict3: 1 if from district 3, 0 otherwise

Ifdistrict4: 1 if from district 4, 0 otherwise

Ifdistrict5: 1 if from district 5, 0 otherwise

Y= β_0 + β_1Ifdistrict2 + β_2Ifdistrict3 + β_3Ifdistrict4 + β_4Ifdistrict5 + β_5Trafficcosts + ε

- What if we thought that not only the intercept would change for different districts, but also that the slopes would change – or that the relationship between Trafficcosts and the average winning bid depends on which district the bid is from. What model do we want to test?

Y= β_0 + β_1Ifdistrict2 + β_2Ifdistrict3 + β_3Ifdistrict4 + β_4Ifdistrict5 + β_5Trafficcosts + β_6Ifdistrict2Trafficcosts + β_7Ifdistrict3Trafficcosts + β_8Ifdistrict4Trafficcosts + β_9Ifdistrict5Trafficcosts+ ε

- A researcher ran a regression analysis that used mother's age (years) and income (dollars) to predict the weight of newborns (ounces). The prediction interval for a mother of 19 and an income of 20,000 is: 5 to 7.5). What does this mean?
  - We are 95% confident that a mother of 19 years and an income of $20,000 will have a baby between 5 and 7.5 ounces.
- What generally does a confidence interval mean:
  - This means that if we collected a new sample from our population and calculated a new CI, and did this repeatedly, 95% of the time, the interval will actually contain the true population parameter.
- Versus what is a pvalue?
  - The probability that you would find a difference as large as you found in your sample, if the null hypothesis was true.

- Conditions for ANOVA, and how to check?
  - Have a mean of zero
    - again automatically true due to analysis
  - Equal variance: have the same standard deviation **for each group**
    - Residual plot
    - Rule of thumb
    - Levene's test of equal variance
  - Follow a normal distribution
    - Normal probability plot
- Be independent: usually achieved by random assignment
  - Study design
- Randomization: if want to infer about population

Quiz!

- Imagine you are interested in buying a house and want to predict the cost of the house. You collect data on the following predictor variables:
  - Location (Cambridge, Summerville, Jamaica Plain)
  - Size, in square feet of living space
  - Year the house was built
  - Whether it is gas or electric heat
  - Whether the kitchen has been recently updated or not.
- Choose what type of analysis is best to answer the following questions.

- You are interested in how the cost of a house is related to the size of the house and year the house was built. (91%)
- You are interested in whether the cost of a house differs depending on what neighborhood it is in. (86%)
- You are interested in whether the cost of house differs depending on what kind of heat it has and whether the kitchen has been updated. (90%)
- You are interested in whether the relationship between the year the house was built and where it is located. (20%)

Quiz!

- You hypothesize that the cost of a house is related to the age of the house, the size of the house, and where it is located, such that the relationship between the age of the house and its cost might depend on location. What model are you interested in testing? (53%)

- Y^= β_0+ β_1Age + β_2Size + β_3Location + β_4AgeLocation (28%)

- Y^= β_0+ β_1Age + β_2Size + β_3IsCambridge + β_4IsSummerville + β_5IsCambridgeAge + β_6IsSummervilleAge

- Y^= β_0+ β_1Age + β_2Size + β_3IsCambridge + β_4IsSummerville + β_5IsCambridgeAge + β_6IsSummervilleAge + β_7IsCambridgeSize + β_8IsSummervilleSize (19%)

Grad projects

- Step 1: find a group! Use the board:
  - Subject title/topic
  - Project Idea: my idea
- Step 2: In class presentation on your idea – get feedback, work out the specifics
  - Second half of class starting March 25th
  - pass/fail
  - 4min recorded video – through collaborate or anything you want
- Step 3: Write an official project proposal with your group
  - Due April 15
  - 2 pages!!!
  - 10 points
- Step 4: Write final project with group
  - May 6
  - Around 10 pages
  - 10 points