Lab 12: 2-sample t-test and 1-way ANOVA
STAT 301

Group Members: Type all names below:
Student 1 (will type answers to questions 1a, 2b, 2e, 2i, 2m): Syrena Hilgendorf
Student 2 (will type answers to questions 1b, 2a, 2f, 2l, 2n): Marie Klapacz
Student 3 (will type answers to questions 1b, 2c, 2g, 2k, 2o): Lorpu Kokoi
Student 4 (will type answers to questions 1c, 2d, 2h, 2j): Neha Maddali
If you have less than four students in your group, split up the remaining questions among your group.

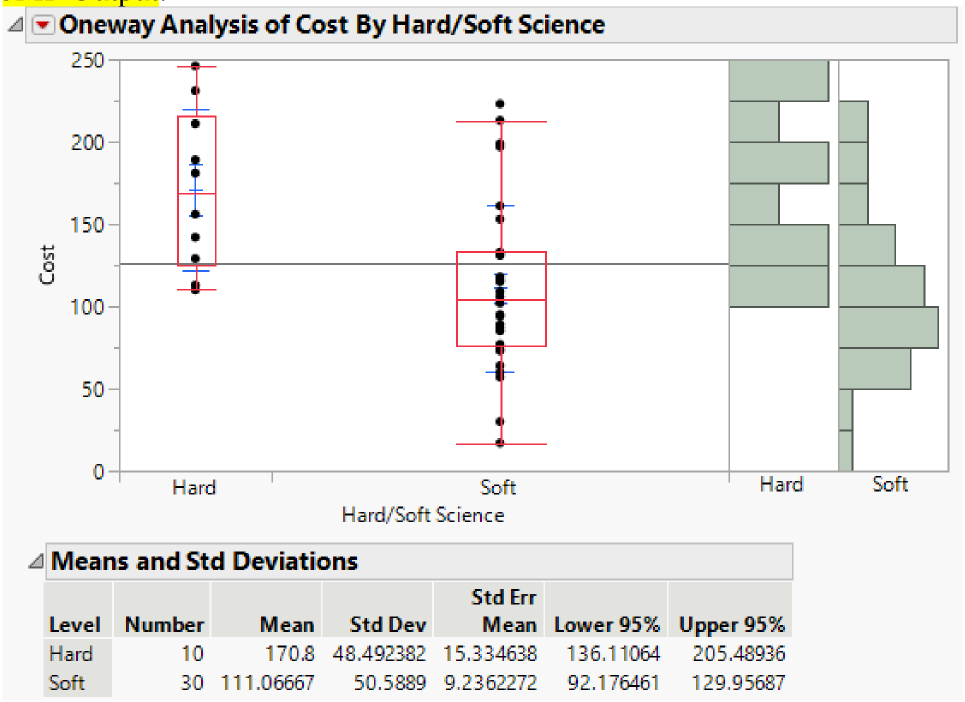
In this lab, you will work with the TextbookCosts Data set.[1] Textbooks were randomly selected, and their price was recorded.

1. **Hard and Soft Sciences.** The data includes the costs of textbooks split by hard and soft sciences.

   a. Create graphs and descriptive statistics to compare the hard and soft sciences textbooks. Follow the "*Descriptive Statistics and Graphs for Comparing Multiple Means using Fit Y by Z*" in the JMP Guide to get histograms and boxplots. Based on these plots, is there a visual indication of differences in the average costs of textbooks for the hard and soft sciences? Explain.

   Type answer here: Yes, there is a visual indication of differences in the average costs of textbooks for the hard and soft sciences. The box plot shows that the mean is different for the hard and soft sciences. The box plot also shows that there is a difference in the IQRs for the hard and soft sciences. There are also fewer observations for the hard sciences compared to the soft sciences.

   JMP Output:



Oneway Analysis of Cost By Hard/Soft Science

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Hard	10	170.8	48.492382	15.334638	136.11064	205.48936
Soft	30	111.06667	50.5889	9.2362272	92.176461	129.95687

Research Question 1: Is there evidence of a difference in the population mean price of textbooks for the hard and soft sciences?

   b. Conduct a *t*-test assuming equal variances to answer Research Question 1. Show all steps of

[1] Data from http://www.lock5stat.com/datapage.html

your hypothesis test. Follow the "*Hypothesis Test for Difference in Two Independent Means*" in the JMP Guide to get the test statistic and *p*-value.

Type answer here:

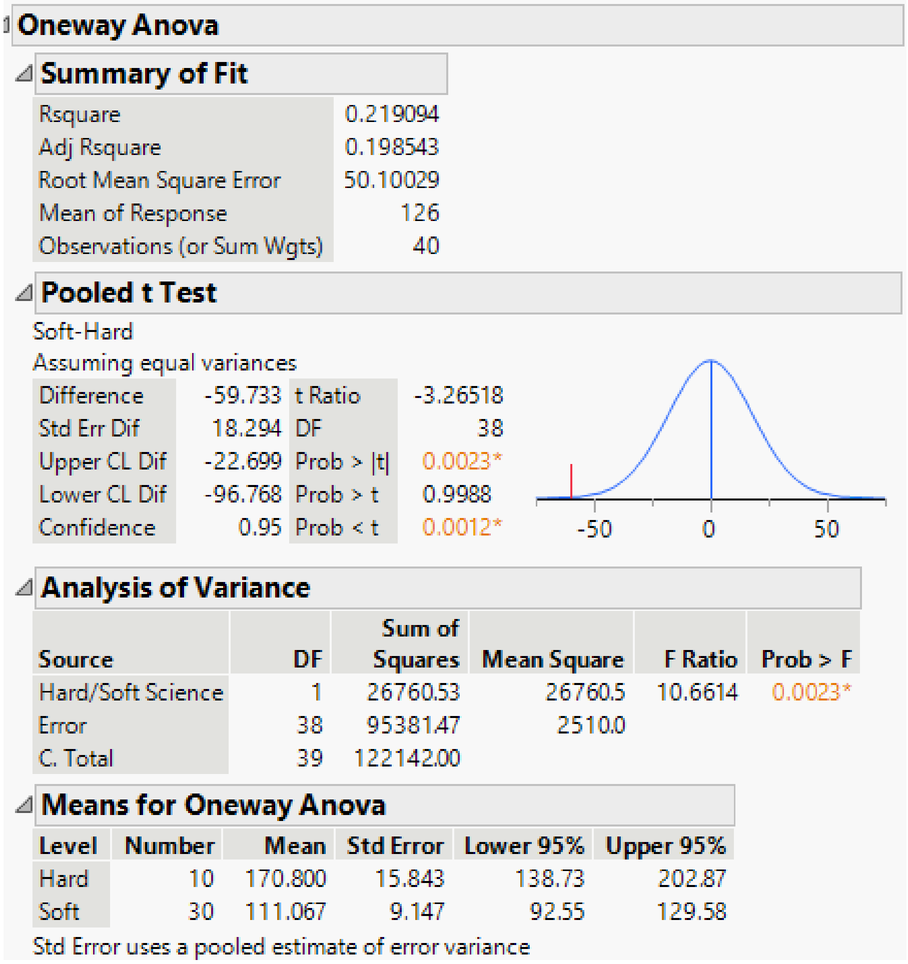
Hypothesis statements:
H_0: $\mu_1 = \mu_2$ where μ_1 = the mean of soft science textbooks and μ_2 = hard science textbooks
H_A: μ_1 not equal to μ_2
Check conditions: Independence is met because Textbooks were randomly selected.

Test statistic = (-3.26518)^2 = 10.6614004324

p-value = 0.0023

Conclusion: There is overwhelming evidence to suggest that the population mean price of textbooks for the soft sciences are different from the hard sciences.

JMP Output:



Oneway Anova

Summary of Fit

Rsquare	0.219094
Adj Rsquare	0.198543
Root Mean Square Error	50.10029
Mean of Response	126
Observations (or Sum Wgts)	40

Pooled t Test

Soft-Hard
Assuming equal variances

Difference	-59.733	t Ratio	-3.26518
Std Err Dif	18.294	DF	38
Upper CL Dif	-22.699	Prob > |t|	0.0023*
Lower CL Dif	-96.768	Prob > t	0.9988
Confidence	0.95	Prob < t	0.0012*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Hard/Soft Science	1	26760.53	26760.5	10.6614	0.0023*
Error	38	95381.47	2510.0		
C. Total	39	122142.00			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Hard	10	170.800	15.843	138.73	202.87
Soft	30	111.067	9.147	92.55	129.58

Std Error uses a pooled estimate of error variance

Research Question 2: How different is the population mean price of textbooks for the hard and soft sciences?

c. Create and interpret a 95% confidence interval using the *t*-distribution assuming equal variances to answer Research Question 2. Note that the interval should show up in your *t*-test output in JMP.

Type answer here:

Confidence interval = [-22.699, -96.768]

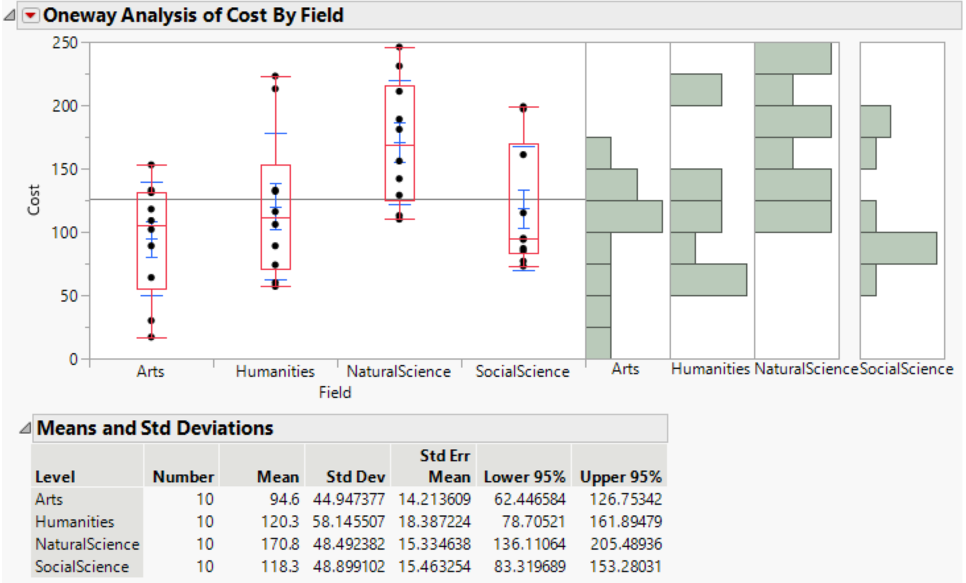
Interpretation: We are 95% confident that the population mean price of textbooks for soft sciences is between 22.70 and 96.77 dollars lower than textbooks for hard sciences.

2. **Field.** The data also includes the costs of textbooks split by different fields; Arts, Humanities, Natural Science, and Social Science.

    a. Create graphs and descriptive statistics to compare textbooks for the four fields. Based on these plots, is there a visual indication of differences in the average costs of textbooks for the four fields? Explain.

Type answer here: Yes, there is visual indication of differences in the average costs of textbooks for the four fields. The box plot shows that the mean is different for each of the four categories. The box plot also shows that there is a difference in IQRs for the categories.

JMP Output:



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Arts	10	94.6	44.947377	14.213609	62.446584	126.75342
Humanities	10	120.3	58.145507	18.387224	78.70521	161.89479
NaturalScience	10	170.8	48.492382	15.334638	136.11064	205.48936
SocialScience	10	118.3	48.899102	15.463254	83.319689	153.28031

Now we will create a model to predict the cost of textbooks using a multiple regression model with indicator variables for field. We will refer to this multiple regression model as Model 1.

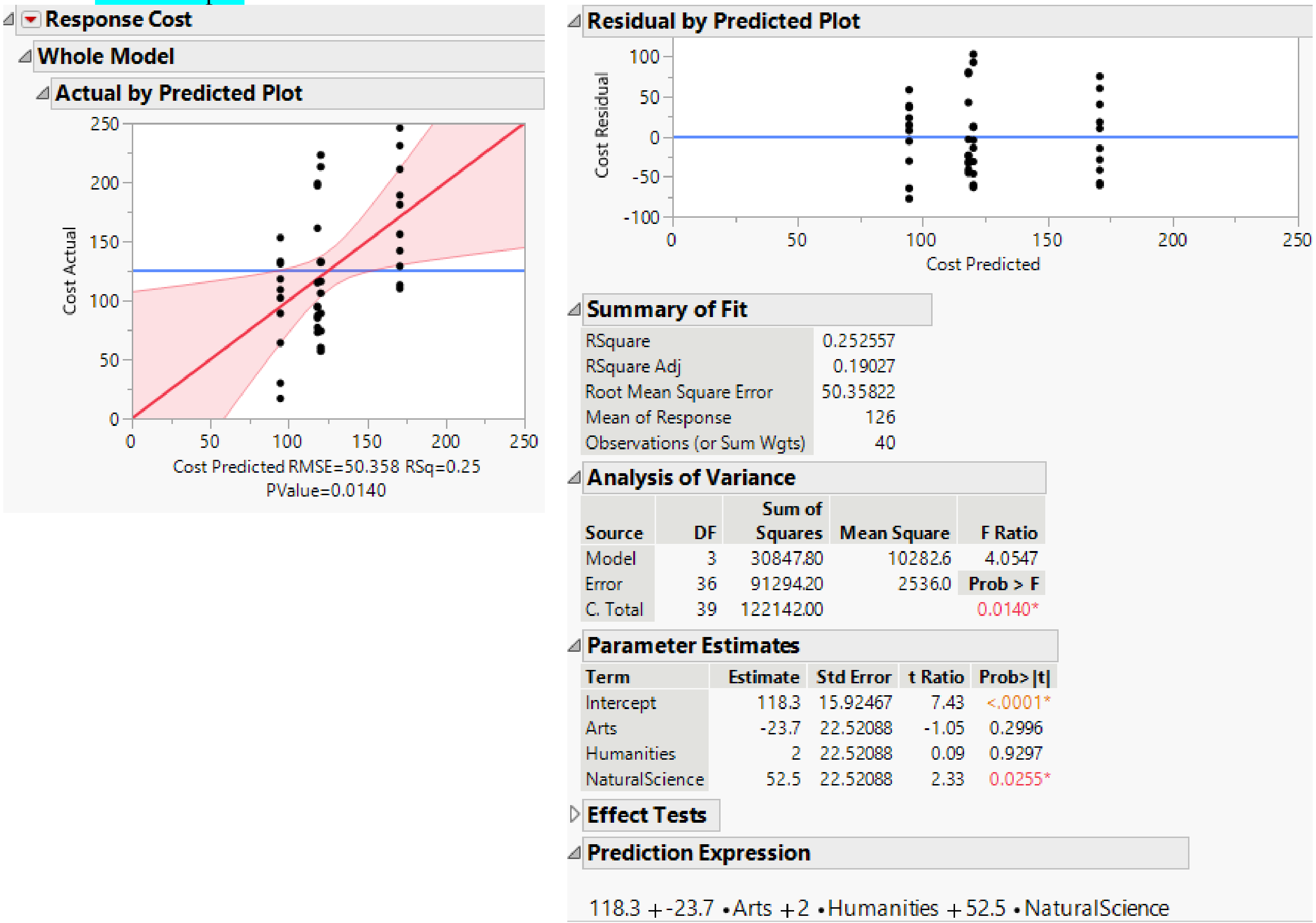
b. (Model 1) In this example, we have 4 groups. Choose a base-level/reference group, define all indicator variables, and write out a multiple regression population model with a categorical variable of field and a response variable of textbook costs. (Hint: There are many correct answers, and you should not be doing any analysis in JMP yet.)

Type answer here:
Reference group: Social science
Indicator variables: Arts, Humanities, Natural Science
Multiple regression population model: μ_y = B0 + B1(x1) + B2(x2) + B3(x3)
x1 = arts if 1, or 0 otherwise
x2 = humanities if 1, or 0 otherwise
x3 = natural science if 1, or 0 otherwise

c. Use JMP to create indicator variables and estimate the multiple regression model. What is the estimated model (i.e., multiple regression equation)?

Type answer here: predicted cost = 118.3 - 23.7*Arts + 2*Humanities + 52.5*NaturalScience

JMP Output:



Summary of Fit

RSquare	0.252557
RSquare Adj	0.19027
Root Mean Square Error	50.35822
Mean of Response	126
Observations (or Sum Wgts)	40

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	3	30847.80	10282.6	4.0547
Error	36	91294.20	2536.0	Prob > F
C. Total	39	122142.00		0.0140*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	118.3	15.92467	7.43	<.0001*
Arts	-23.7	22.52088	-1.05	0.2996
Humanities	2	22.52088	0.09	0.9297
NaturalScience	52.5	22.52088	2.33	0.0255*

Effect Tests

Prediction Expression

118.3 + -23.7 • Arts + 2 • Humanities + 52.5 • NaturalScience

d. Based on your estimated model (Model 1), compute the predicted cost of textbooks for each field.

Type answer here:
Arts: The predicted price for Art is 118.3 - 23.7 = 94.6 dollars
Humanities: The predicted price for Humanities is 118.3 +2 = 120.3 dollars
Natural Sciences: The predicted price for Natural Science is 118.3 + 52.5 = 170.8 dollars
Social Science: The predicted price for Social Science is 118.3 dollars

e. How do these predictions compare to the sample means from your descriptive statistics analysis?

Type answer here: All of the predictions we made match the sample means from the descriptive statistics analysis included in part 2a.

Research Question 3: Is there evidence that fields predict textbook costs?

f. (Model 1) Use the multiple regression notation and write out the following steps for a hypothesis test to test the overall model answering Research Question 3.

Type answer here:

Hypothesis statements: H0: B1 = B2 = B3 where B1 = arts, B2 = humanities, B3 = natural science

Ha: at least one B_i doesn't equal 0

Test statistic = 4.0547
p-value = 0.0140
Conclusion: There is moderate evidence to suggest that fields predict textbook costs.

Now we look at the same dataset while fitting an ANOVA model (instead of a multiple regression model with indicator variables). We will refer to this ANOVA model as Model 2.

g. (Model 2) Write out the population ANOVA model appropriate for this example. (Hint: You should not be doing any analysis in JMP yet.)

Type answer here: $y_{ik} = \mu_{ik} + \alpha_k + \epsilon$ where i = 1,2,3 and k = 1,2,3 and ϵ~N(0, σ)

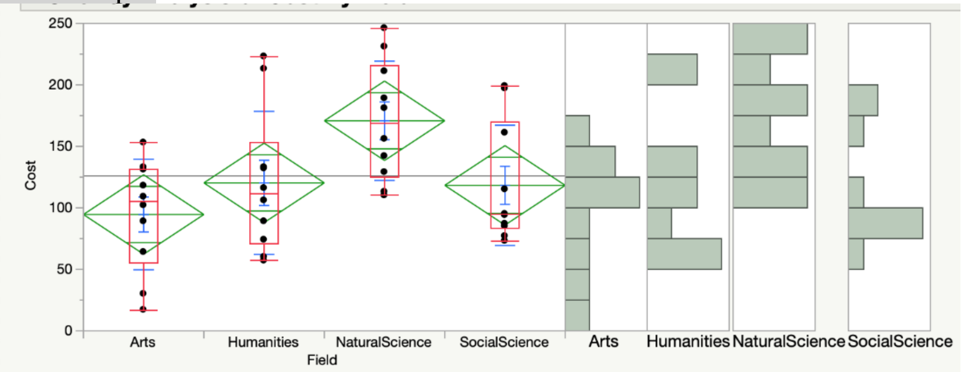
where $\mu1$ is the population mean textbook price for arts, $\mu2$ is the population mean textbook price for humanities, and $\mu3$ is the population mean textbook price for natural sciences. $\alpha1$ is the population effect for arts, $\alpha2$ is the population effect for humanities, and $\alpha3$ is the population effect for natural sciences.

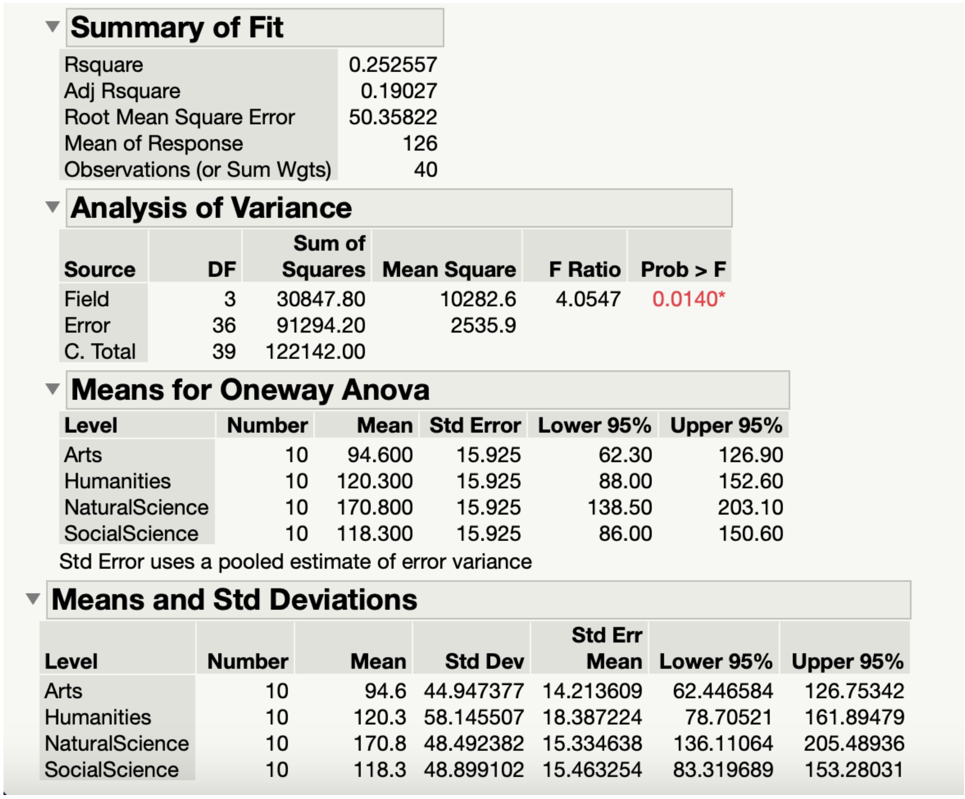
Yik= μi + epsilon ik for i=1,2,3,4 and k=1,2,3,4..10 10 observations for each field)
OR
Yik= μ+alpha k + epsilon ik

Research Question 4: Does the population mean textbook cost differ for any of the 4 fields?

h. Follow the JMP instructions "*One-way ANOVA using Fit Y by X*" in the JMP Guide to get your ANOVA output.

JMP Output:





▼ **Summary of Fit**

Rsquare	0.252557
Adj Rsquare	0.19027
Root Mean Square Error	50.35822
Mean of Response	126
Observations (or Sum Wgts)	40

▼ **Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Field	3	30847.80	10282.6	4.0547	0.0140*
Error	36	91294.20	2535.9		
C. Total	39	122142.00			

▼ **Means for Oneway Anova**

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Arts	10	94.600	15.925	62.30	126.90
Humanities	10	120.300	15.925	88.00	152.60
NaturalScience	10	170.800	15.925	138.50	203.10
SocialScience	10	118.300	15.925	86.00	150.60

Std Error uses a pooled estimate of error variance

▼ **Means and Std Deviations**

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Arts	10	94.6	44.947377	14.213609	62.446584	126.75342
Humanities	10	120.3	58.145507	18.387224	78.70521	161.89479
NaturalScience	10	170.8	48.492382	15.334638	136.11064	205.48936
SocialScience	10	118.3	48.899102	15.463254	83.319689	153.28031

i. (Model 1 vs. 2) Compare your predictions from Model 1 to the means reported in the ANOVA output. Comment on your observations. Explain why this makes sense.

Type answer here: All of the predictions we made from Model 1 match the means reported in the ANOVA output. This makes sense because in ANOVA tests we compare the means for each group to the overall mean. In regression we compare the means for each group to the reference group which in our case was SocialScience.

j. (Model 2) Using the ANOVA output, is there statistical evidence that the population mean textbook cost differs for any of the 4 fields (i.e., answer Research Question 4)?

Type answer here:

Hypothesis statements:
H_0: $\mu_1 = \mu_2 = \mu_3$ where μ_1 = the mean of Arts textbooks, μ_2 = the mean of Humanities textbooks, μ_3 = mean of NaturalScience textbooks
H_A: at least one equality doesn't hold

F ratio: 4.0547

P−value: 0.0140

Conclusion: There is moderate evidence to suggest that the population mean textbook cost differs depending on the field.

k. (Model 1 vs. 2) Compare your hypotheses statements for Models 1 and 2. Carefully explain why these two differently written hypotheses imply the same thing.

Type answer here:

Model 1 hypothesis statements:

- H0: B1 = B2 = B3 where B1 = arts, B2 = humanities, B3 = natural sciences
- Ha: at least one B_i doesn't equal 0

Model 2 hypothesis statements:

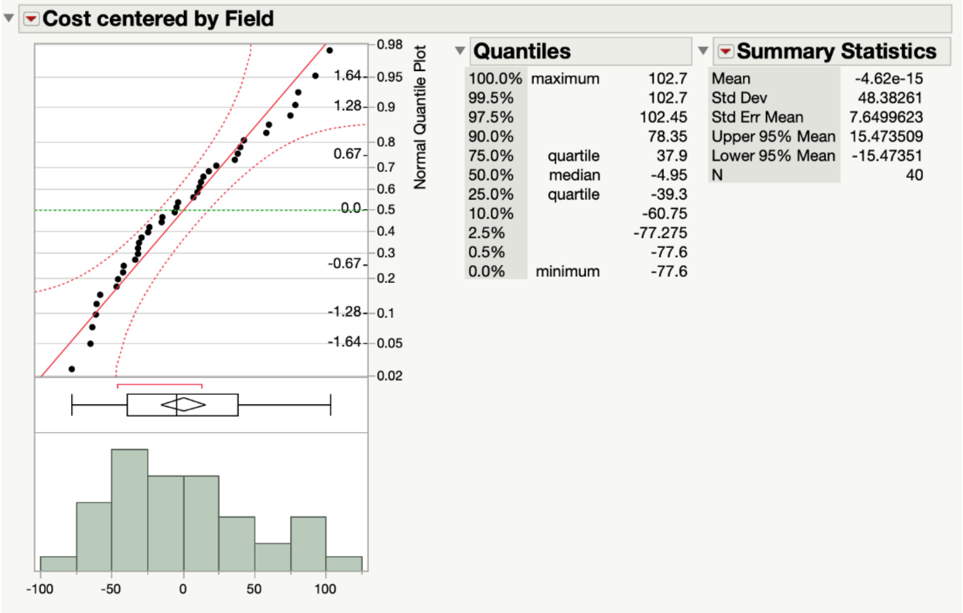
- H_0: $\mu_1 = \mu_2 = \mu_3$ where μ_1 = the mean of Arts textbooks, μ_2 = the mean of Humanities textbooks, μ_3 = mean of NaturalScience textbooks
- H_A: at least one equality doesn't hold

These hypothesis statements imply the same thing because a regression model (Model 1) compares the means for each group against a baseline group, while ANOVA (Model 2) compares the means for each group against the grand mean. In both cases, we are using means to conduct an analysis.

l. Continue to follow the JMP instructions "*One-way ANOVA using Fit Y by X*" in the JMP Guide to get your Normal Quantile Plot and Residual Plot.

JMP Output:

Normal:



Cost centered by Field

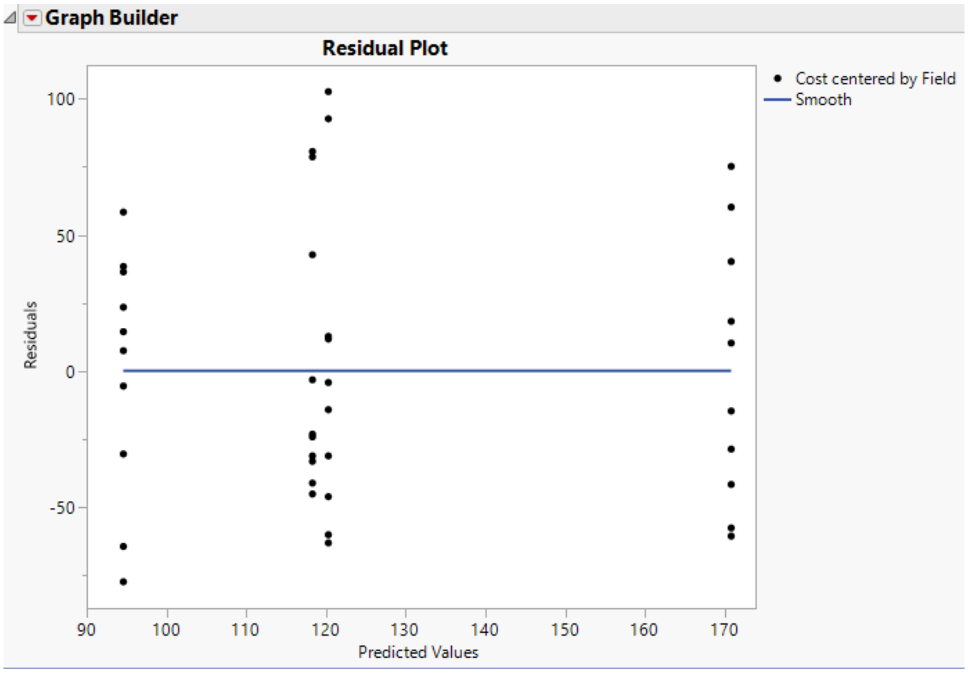
Quantiles

100.0%	maximum	102.7
99.5%		102.7
97.5%		102.45
90.0%		78.35
75.0%	quartile	37.9
50.0%	median	-4.95
25.0%	quartile	-39.3
10.0%		-60.75
2.5%		-77.275
0.5%		-77.6
0.0%	minimum	-77.6

Summary Statistics

Mean	-4.62e-15
Std Dev	48.38261
Std Err Mean	7.6499623
Upper 95% Mean	15.473509
Lower 95% Mean	-15.47351
N	40

Residual:



Graph Builder

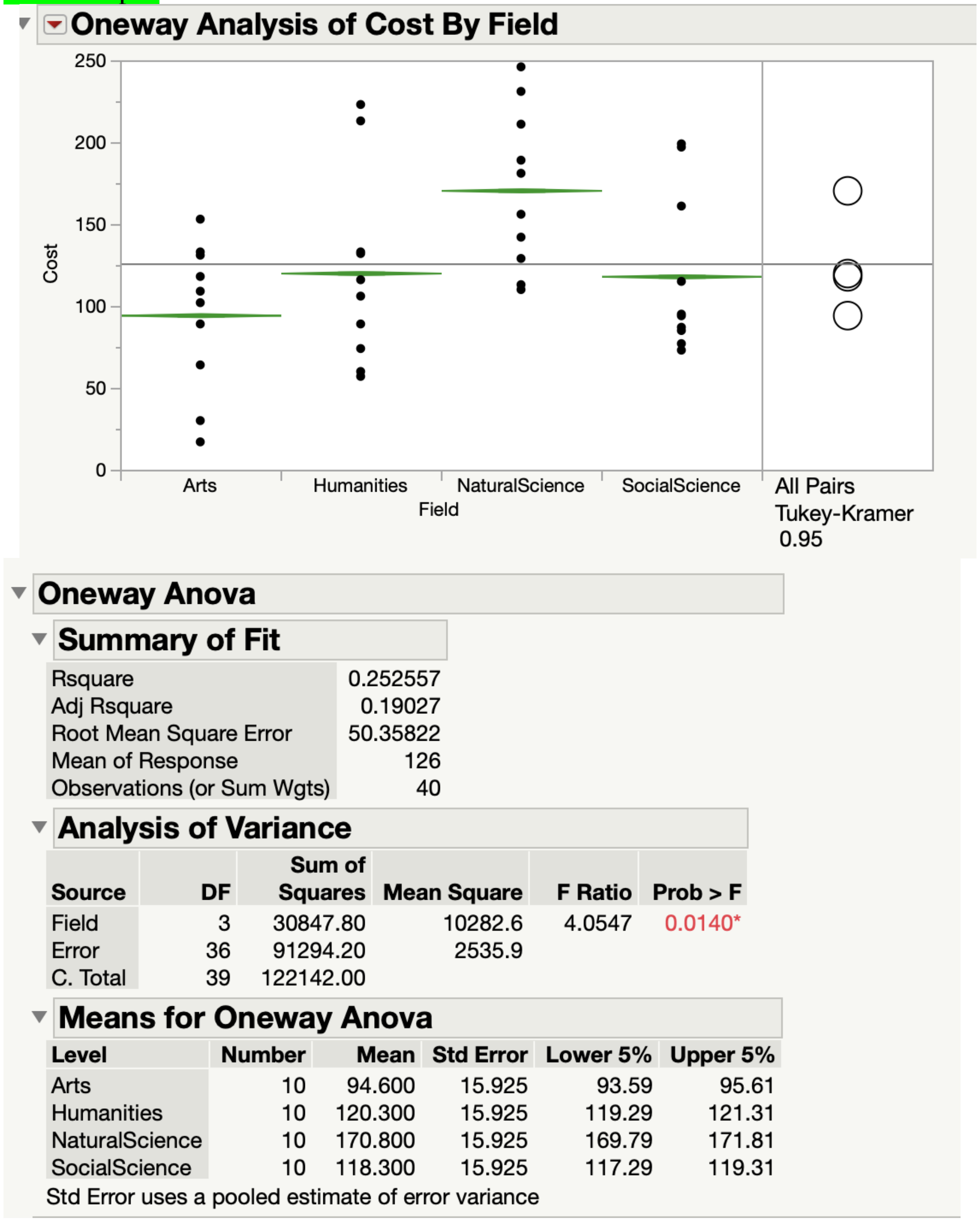
m. Are the assumptions/conditions met for the ANOVA to be used? List all appropriate conditions and describe why they are or are not met.

Type answer here: Samples are randomly selected and independent meeting the random condition. Distribution is not normal however the sample size is 40 which is >30 meeting the sample size condition. The normal plot is not linear, and the residual plot does not have homoscedasticity, so the population variances are not equal. This condition is not met. Considering that only 2 of the 3 assumptions are met it may not be appropriate for the ANOVA to be used.

Research Question 5: Which fields have evidence of a difference in population mean textbook costs, and how different are they?

n. Follow the JMP instructions "*Multiple Comparisons with One-way ANOVA using Fit Y by X*" in the JMP Guide to get your 95% Tukey multiple comparisons.

JMP Output:



Oneway Analysis of Cost By Field

Oneway Anova

Summary of Fit

Rsquare	0.252557
Adj Rsquare	0.19027
Root Mean Square Error	50.35822
Mean of Response	126
Observations (or Sum Wgts)	40

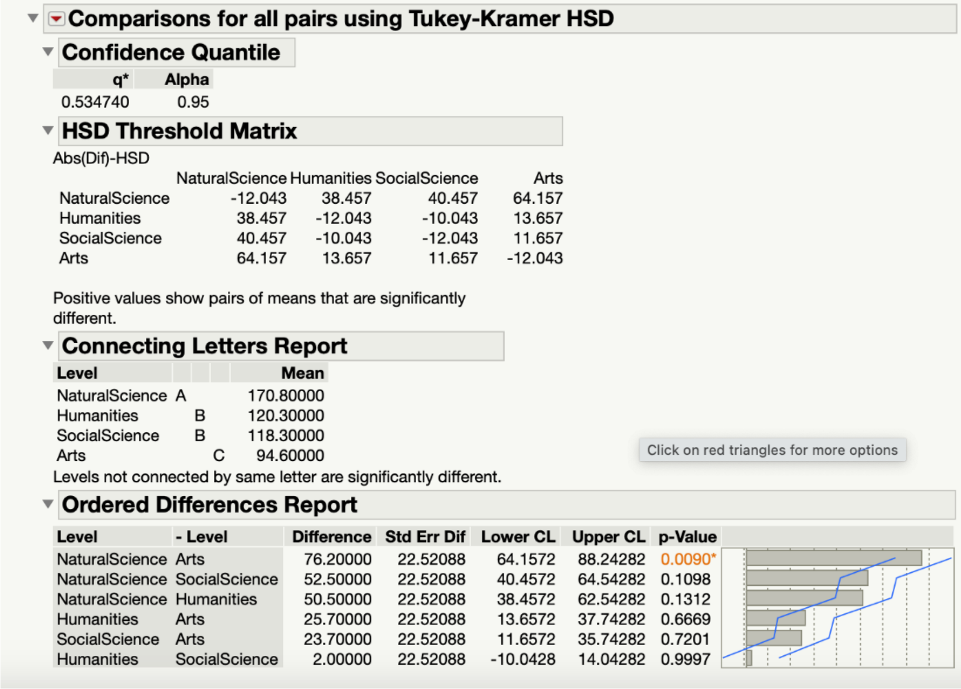
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Field	3	30847.80	10282.6	4.0547	0.0140*
Error	36	91294.20	2535.9		
C. Total	39	122142.00			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 5%	Upper 5%
Arts	10	94.600	15.925	93.59	95.61
Humanities	10	120.300	15.925	119.29	121.31
NaturalScience	10	170.800	15.925	169.79	171.81
SocialScience	10	118.300	15.925	117.29	119.31

Std Error uses a pooled estimate of error variance



Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
0.534740	0.95

HSD Threshold Matrix

Abs(Dif)-HSD

	NaturalScience	Humanities	SocialScience	Arts
NaturalScience	-12.043	38.457	40.457	64.157
Humanities	38.457	-12.043	-10.043	13.657
SocialScience	40.457	-10.043	-12.043	11.657
Arts	64.157	13.657	11.657	-12.043

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level				Mean
NaturalScience	A			170.80000
Humanities		B		120.30000
SocialScience		B		118.30000
Arts			C	94.60000

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
NaturalScience	Arts	76.20000	22.52088	64.1572	88.24282	0.0090*
NaturalScience	SocialScience	52.50000	22.52088	40.4572	64.54282	0.1098
NaturalScience	Humanities	50.50000	22.52088	38.4572	62.54282	0.1312
Humanities	Arts	25.70000	22.52088	13.6572	37.74282	0.6669
SocialScience	Arts	23.70000	22.52088	11.6572	35.74282	0.7201
Humanities	SocialScience	2.00000	22.52088	-10.0428	14.04282	0.9997

o. Answer Research Question 5 using the Tukey comparisons p-values and confidence intervals.

Type answer here: The groups Natural Science and Arts have a significant difference based on the ordered differences report with a p-value of 0.0090. The other relationships between the groups do not have significant p-values, so there is no significant difference for any other pairing.

We are 95% confident that the population mean textbook costs for Natural Sciences and Arts is between 64.1572 and 88.24282 dollars.