Running the Statistical Gauntlet in SAS

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

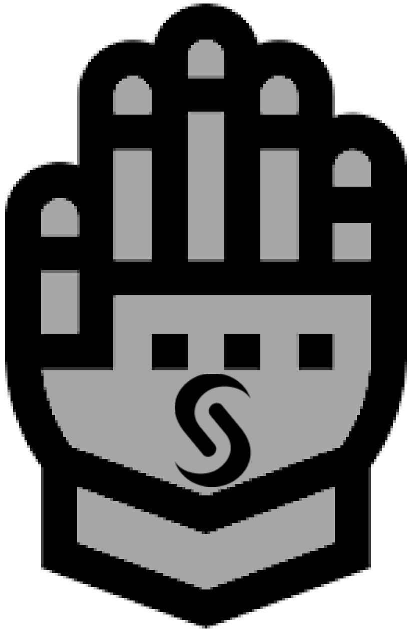
DaCCoTA, University of North Dakota

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Introduction

- Often, in an introduction to statistics, a single example is used to display a model or technique
- This can lead to difficulty in adapting that example's particularities to your own work
- It also fails to train your eye in reading and understanding patterns across examples
- Here, we aim to remedy that by providing, exhaustive, back-to-back examples
- Aimed at intermediate learners
- Get ready for a gauntlet, I hope it will serve you well



Assessment

- Before continuing, please take the pre-test

  Pre-Test: https://und.qualtrics.com/jfe/form/SV_cUqTGbRuYEDcRxA

- After finishing, please take the post-test and survey

  Post-Test: https://und.qualtrics.com/jfe/form/SV_0OqJTs8htwruJsa

  Survey: https://und.qualtrics.com/jfe/form/SV_56JT2olUAQBEpxk

Overview

- Today, we'll be using SAS Studio
- Access SAS Studio via https://www.sas.com/en_us/software/studio.html
- Access SAS code at https://med.und.edu/daccota/_files/docs/berdc_docs/model_gauntlet_sascode.txt
- Topics Covered
  - T-tests
    1) One-sample t-test
    2) Two-sample t-test
    3) Paired t-test
  - ANOVA
    4) One-way ANOVA
    5) Two-way ANOVA
    6) Blocked/Nested ANOVA
  - Regression
    7) Simple Linear Regression
    8) Multiple Linear Regression
    9) Logistic Regression

Procedure

- Six examples per topic
- Ignoring most assumptions condensing output for brevity
- The test statistic, p-value, and where appropriate, model fit will be outlined by color
- Each example includes:
  - Research question in the form of a sentence
  - Relevant statistical results from SAS
    - most values will be rounded to two decimal places
    - p-values will not be modified
  - Written answer to research question
  - Figure or table when appropriate
    - Some graphs will be of null results for clarity (greyscale or red)
    - Typically, only significant results are graphed
- Get ready to run the gauntlet!

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

1. One-sample t-test

UNIVERSITY OF NORTH DAKOTA

Tests if a variable's mean is different from a set value

#1) Is the average birth weight of White infants greater than 3200?

Mean
3411.2

DF	t Value	Pr > t
41857	78.92	<.0001

Yes, birth weight was significantly greater than 3200.

#2) Is the average birth weight of Black infant less than 3200?

Mean
3162.7

DF	t Value	Pr > t
8141	-5.49	<.0001

Yes, birth weight was significantly less than 3200.

#3) Is the average birth weight of Black infants different than the mean weight of White infants (3411.2)?

Mean
3162.7

DF	t Value	Pr > t
8141	-36.54	<.0001

Yes, birth weight was significantly different than 3411.2

#4) Is the average number of at bats for baseball players different than 400?

Mean
390.1

DF	t Value	Pr > t
321	-1.24	0.2158

No, number of at bats was not significantly different than 400.

#5) Is the log salary for baseball players less than 6?

Mean
5.9272

DF	t Value	Pr > t
262	-1.33	0.0928

No, the log salary was not significantly less than 6.

#6) Is the average number of home runs for baseball players greater than Barry Bonds (16)?

Mean
11.1025

DF	t Value	Pr > t
321	-10.10	1.0000

No, the number of home runs was not significantly greater than 16.

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

2. Two-sample t-test

- *Tests if the mean of two different groups is different*

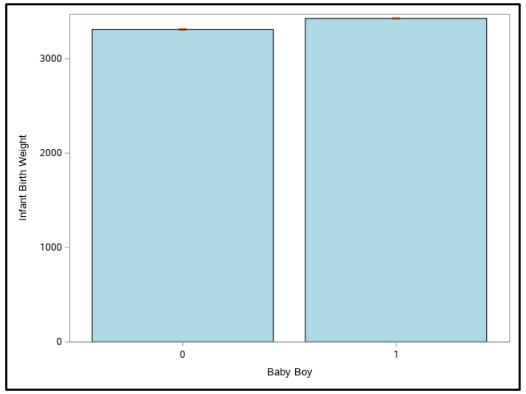
#1) Is the average birth weight of infants greater for boys compared to girls?

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	49998	-23.15	<.0001
Satterthwaite	Unequal	49993	-23.18	<.0001

Boy	Mean
0	3310.6
1	3427.3

Equality of Variances	
F Value	Pr > F
1.11	<.0001

Yes, birth weight was significantly greater for boys.



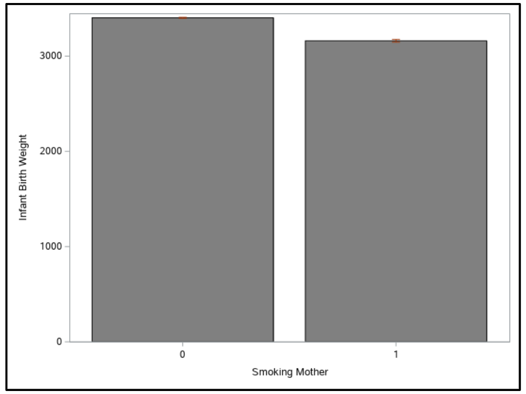
#2) Is the average birth weight of infants lower for smoking vs. non-smoking mothers?

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	49998	32.46	<.0001
Satterthwaite	Unequal	8474.1	31.68	<.0001

MomSmoke	Mean
0	3402.3
1	3160.9

Equality of Variances	
F Value	Pr > F
1.07	0.0004

Yes, birth weight was significantly lower for smoking mothers.



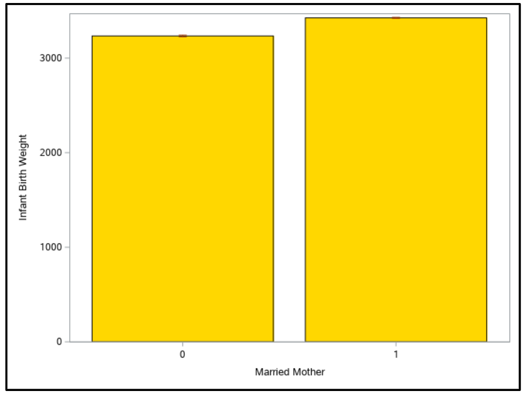
#3) Is the average birth weight of infants different between married and non-married mothers?

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	49998	-34.58	<.0001
Satterthwaite	Unequal	25443	-33.88	<.0001

Married	Mean
0	3234.4
1	3425.7

Equality of Variances	
F Value	Pr > F
1.10	<.0001

Yes, birth weight was significantly greater for married mothers.



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

2. Two-sample t-test

Tests if the mean of two different groups is different

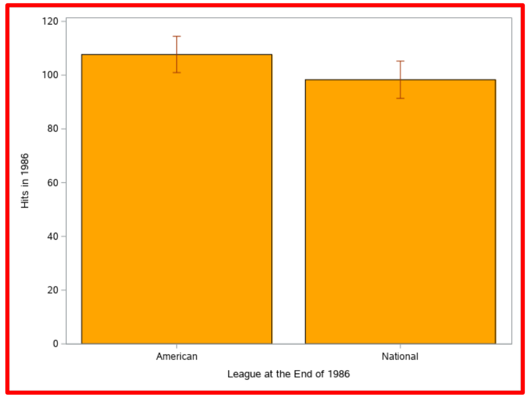
#4) Is the average number of hits for baseball players different across league?

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	320	1.91	0.0573
Satterthwaite	Unequal	315.99	1.92	0.0559

League	Mean
American	107.7
National	98.29

Equality of Variances	
F Value	Pr > F
1.13	0.4356

No, number of hits was not significantly different across league.



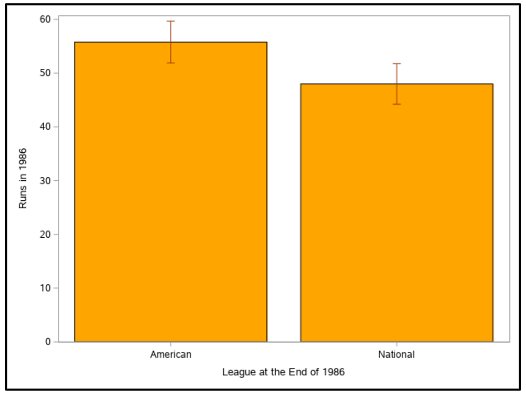
#5) Is the average number of runs for baseball players different across league?

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	320	2.81	0.0052
Satterthwaite	Unequal	319.05	2.84	0.0048

League	Mean
American	55.78
National	47.98

Equality of Variances	
F Value	Pr > F
1.27	0.1326

Yes, the number of runs was significantly greater in the American vs. the National League.



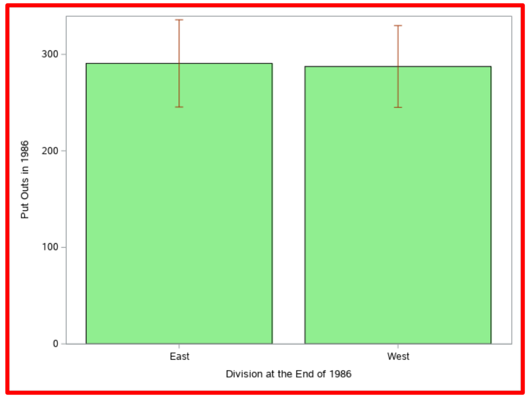
#6) Is the average number of outs for baseball players different across division?

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	320	0.10	0.9198
Satterthwaite	Unequal	317.48	0.10	0.9199

Division	Mean
East	290.6
West	287.5

Equality of Variances	
F Value	Pr > F
1.08	0.6181

No, the number of outs was not significantly different across division.



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

3. Paired t-test

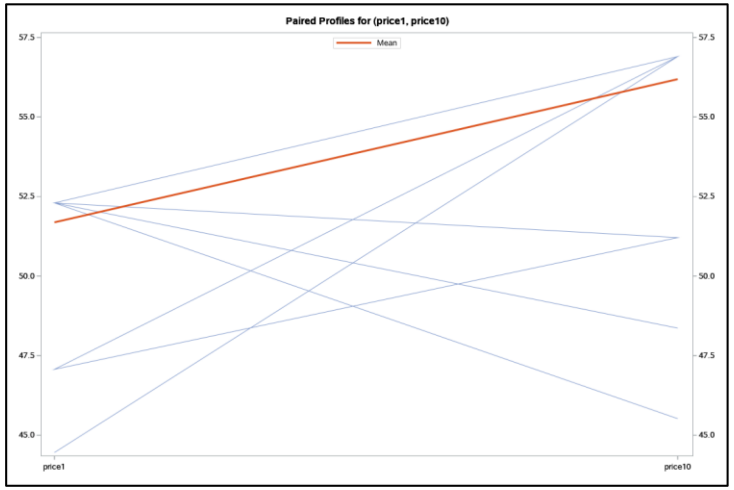
Tests if the means of two different paired groups are different

#1) Is the average unit price different across time for Product 1 and 10?

Mean
-4.50

DF	t Value	Pr > t
1019	-42.99	<.0001

Yes, unit price was significantly higher for Product 10.

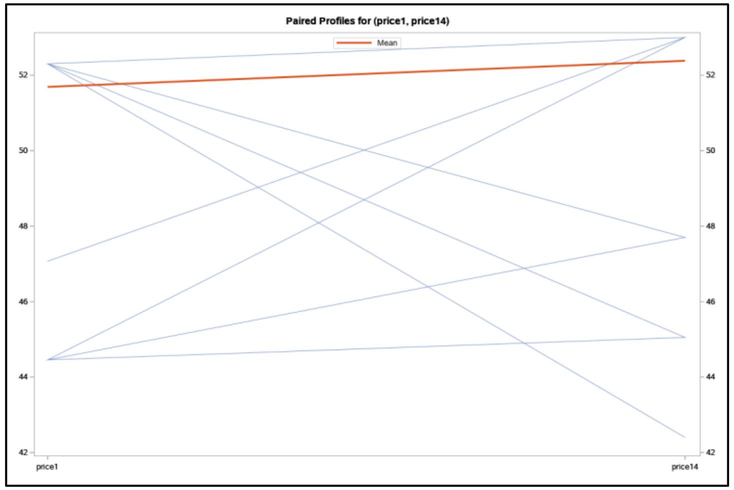


#2) Is the average unit price different across time for Product 1 and 14?

Mean
-0.69

DF	t Value	Pr > t
1019	-7.77	<.0001

Yes, unit price was significantly higher for Product 14.

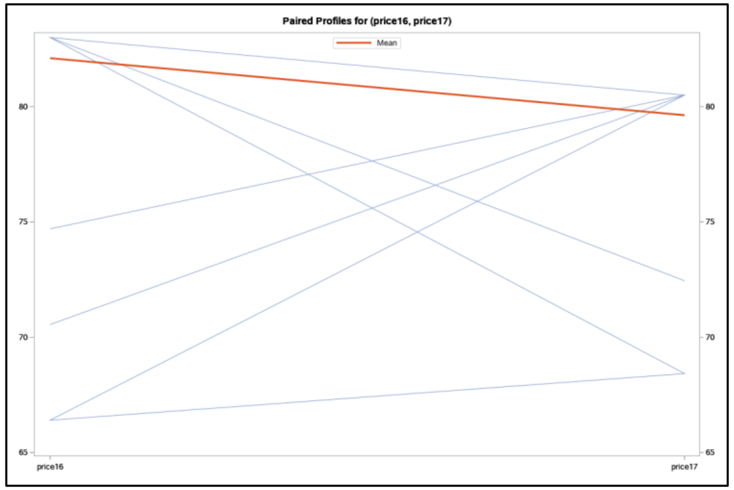


#3) Is the average unit price different across time for Product 16 and 17?

Mean
2.47

DF	t Value	Pr > t
1019	21.57	<.0001

Yes, unit price was significantly higher for Product 16.



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

3. Paired t-test

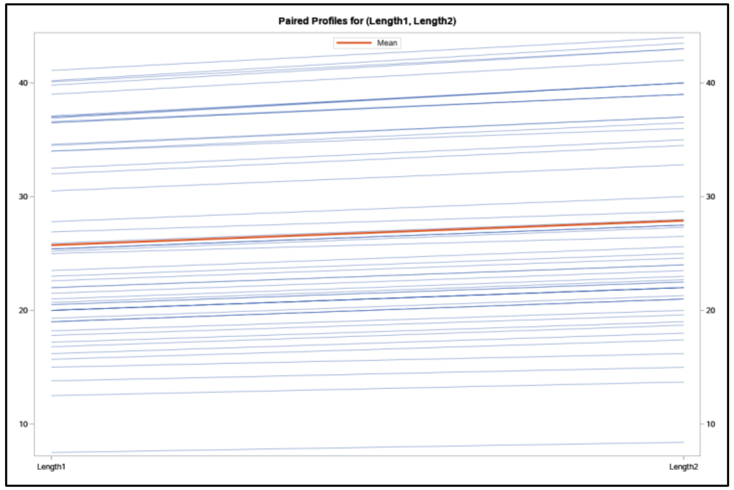
Tests if the means of two different paired groups are different

#4) Is the average length of perch different between measurement 1 and 2?

Mean
-2.16

DF	t Value	Pr > t
55	-31.91	<.0001

Yes, length was significantly higher for measurement 2.

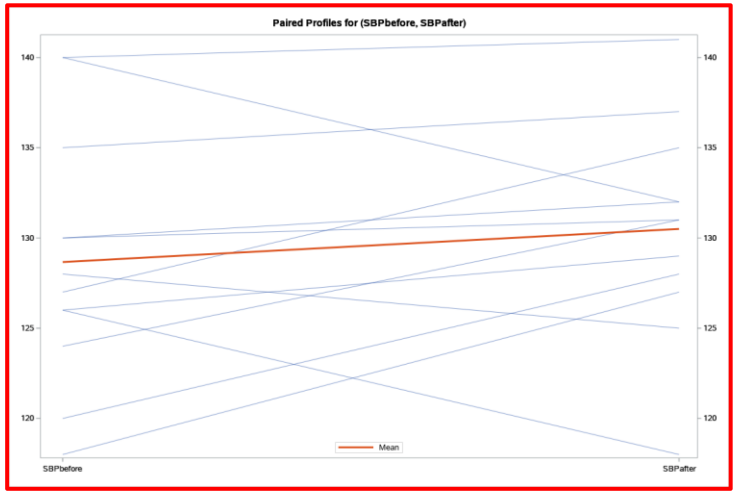


#5) Is average blood pressure different before versus after a stimulus?

Mean
-1.93

DF	t Value	Pr > t
11	-1.09	0.2992

No, blood pressure was not significantly different before and after a stimulus.

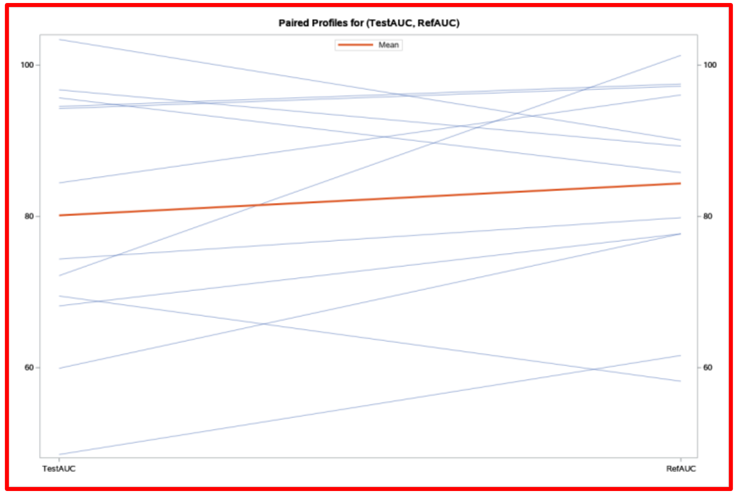


#6) Is the average AUC (area under serum-concentration curve) different between a test and reference drug?

Mean
-4.23

DF	t Value	Pr > t
11	-1.13	0.2834

No, AUC was not significantly different between drugs.



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

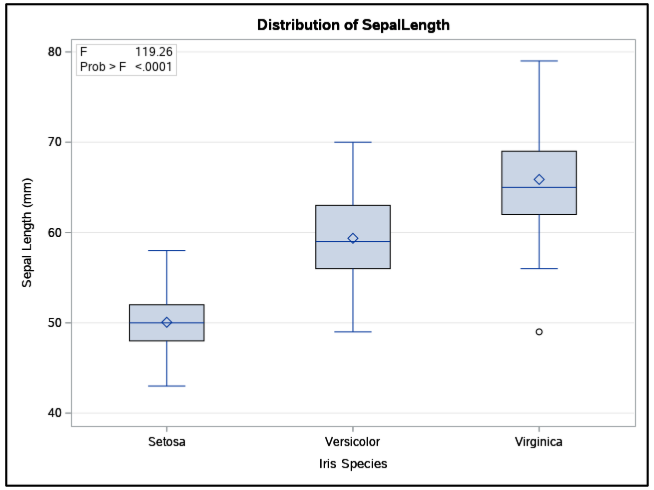
4. One-way ANOVA

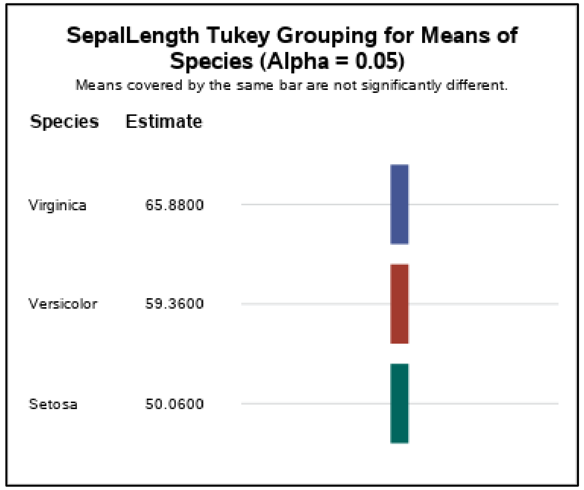
Tests if a variable's mean is different between a category with three or more groups

#1) Is the average sepal length different across iris species?

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Species	2	6321.21	3160.61	119.26	<.0001

Yes, sepal length was significantly different across species.

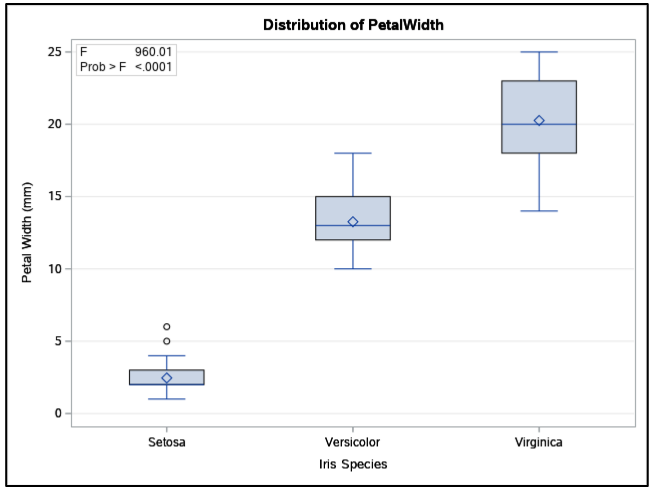


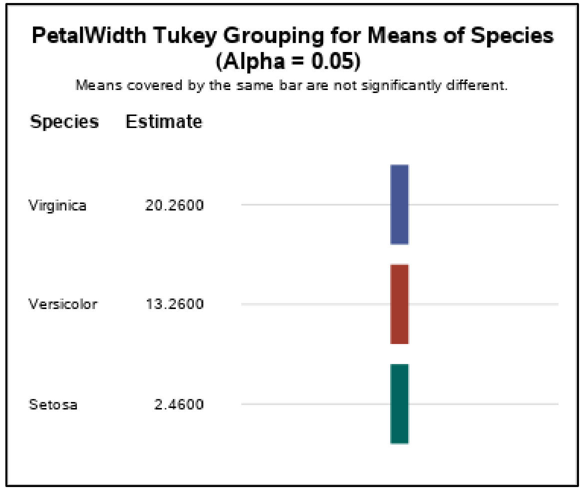


#2) Is the average petal width different across iris species?

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Species	2	8041.33	4020.67	960.01	<.0001

Yes, petal width was significantly different across species.





DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

4. One-way ANOVA

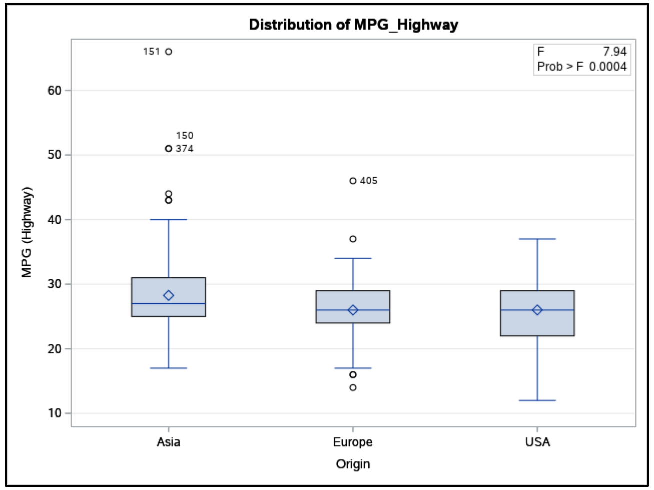
UNIVERSITY OF NORTH DAKOTA

Tests if a variable's mean is different between a category with three or more groups

#3) Is the average highway MPG different across car origin?

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Origin	2	506.71	253.36	7.94	0.0004

Yes, highway MPG was significantly different across origins.

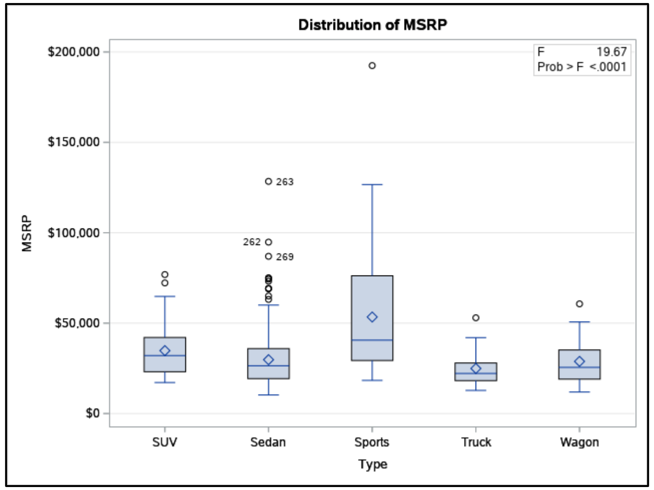


Comparisons significant at the 0.05 level are indicated by ***.				
Origin Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Asia - USA	2.2522	0.7294	3.7750	***
Asia - Europe	2.2577	0.6598	3.8556	***
USA - Asia	-2.2522	-3.7750	-0.7294	***
USA - Europe	0.0055	-1.6184	1.6294	
Europe - Asia	-2.2577	-3.8556	-0.6598	***
Europe - USA	-0.0055	-1.6294	1.6184	

#4) Is the average suggested retail price different across car type?

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Type	4	25355531765	6338882941	19.67	<.0001

Yes, suggested retail price was significantly different across type.



Comparisons significant at the 0.05 level are indicated by ***.				
Type Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Sports - SUV	18597	9126	28068	***
Sports - Sedan	23613	15958	31269	***
Sports - Wagon	24547	13144	35949	***
Sports - Truck	28446	16191	40700	***
SUV - Sports	-18597	-28068	-9126	***
SUV - Sedan	5017	-2023	12056	
SUV - Wagon	5950	-5049	16948	
SUV - Truck	9849	-2031	21728	
Sedan - Sports	-23613	-31269	-15958	***
Sedan - SUV	-5017	-12056	2023	
Sedan - Wagon	933	-8547	10413	
Sedan - Truck	4832	-5658	15322	
Wagon - Sports	-24547	-35949	-13144	***
Wagon - SUV	-5950	-16948	5049	
Wagon - Sedan	-933	-10413	8547	
Wagon - Truck	3899	-9571	17369	
Truck - Sports	-28446	-40700	-16191	***
Truck - SUV	-9849	-21728	2031	
Truck - Sedan	-4832	-15322	5658	
Truck - Wagon	-3899	-17369	9571	

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

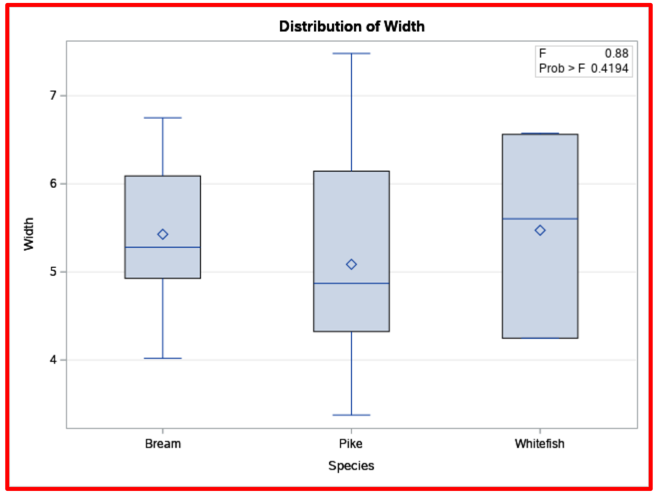
4. One-way ANOVA

- *Tests if a variable's mean is different between a category with three or more groups*

#5) Is the average width different across 3 fish species?

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Species	2	1.46	0.73	0.88	0.4194

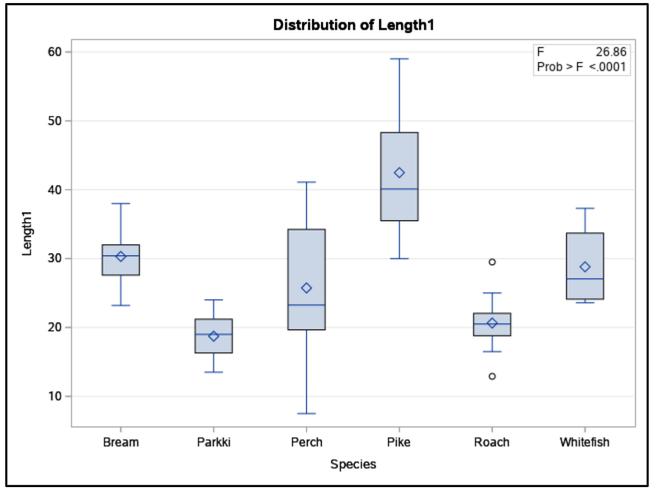
No, width was not different across species.



#6) Is the average length different across 6 fish species?

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Species	5	6053.886251	1210.777250	26.86	<.0001

Yes, length was different across species.



Comparisons significant at the 0.05 level are indicated by ***.

Species Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Pike - Bream	12.171	6.435	17.907	***
Pike - Whitefish	13.676	4.463	22.890	***
Pike - Perch	16.741	11.368	22.114	***
Pike - Roach	21.831	15.431	28.232	***
Pike - Parkki	23.749	16.241	31.257	***
Bream - Pike	-12.171	-17.907	-6.435	***
Bream - Whitefish	1.506	-7.068	10.079	
Bream - Perch	4.570	0.389	8.751	***
Bream - Roach	9.661	4.222	15.100	***
Bream - Parkki	11.578	4.872	18.285	***
Whitefish - Pike	-13.676	-22.890	-4.463	***
Whitefish - Bream	-1.506	-10.079	7.068	
Whitefish - Perch	3.064	-5.271	11.399	
Whitefish - Roach	8.155	-0.877	17.187	
Whitefish - Parkki	10.073	0.225	19.920	***
Perch - Pike	-16.741	-22.114	-11.368	***
Perch - Bream	-4.570	-8.751	-0.389	***
Perch - Whitefish	-3.064	-11.399	5.271	
Perch - Roach	5.091	0.036	10.145	***
Perch - Parkki	7.008	0.609	13.408	***
Roach - Pike	-21.831	-28.232	-15.431	***
Roach - Bream	-9.661	-15.100	-4.222	***
Roach - Whitefish	-8.155	-17.187	0.877	
Roach - Perch	-5.091	-10.145	-0.036	***
Roach - Parkki	1.918	-5.366	9.201	
Parkki - Pike	-23.749	-31.257	-16.241	***
Parkki - Bream	-11.578	-18.285	-4.872	***
Parkki - Whitefish	-10.073	-19.920	-0.225	***
Parkki - Perch	-7.008	-13.408	-0.609	***
Parkki - Roach	-1.918	-9.201	5.366	

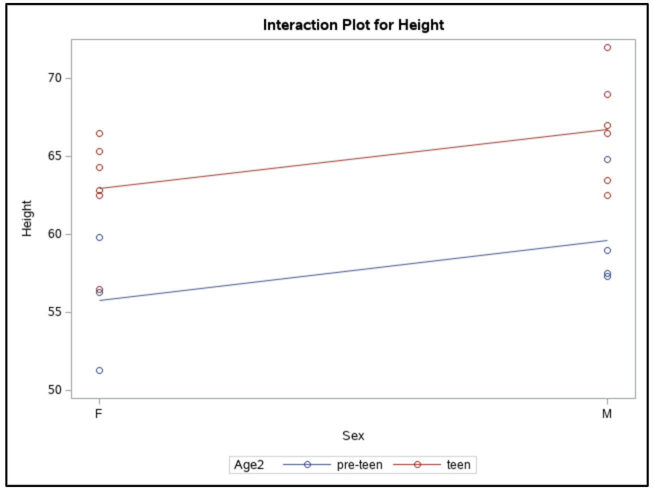
5. Two-way ANOVA

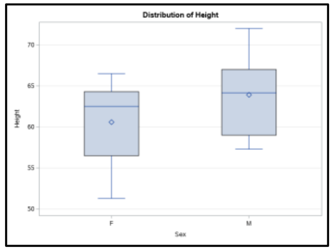
- *Tests if a variable's mean is different between a two categories with multiple groups each*

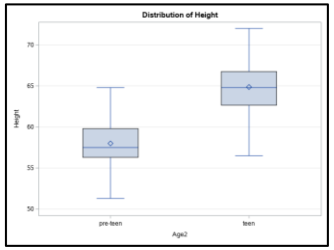
#1) Is average height different across age or sex for children?

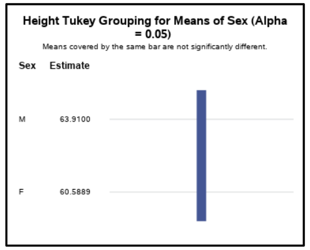
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	1	63.2875758	63.2875758	4.83	0.0442
Age2	1	222.5603030	222.5603030	16.97	0.0009
Sex*Age2	1	0.0075758	0.0075758	0.00	0.9811

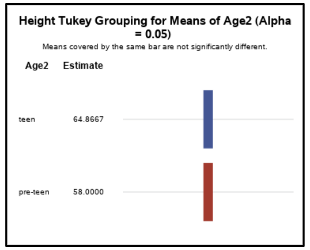
Yes, height was significantly different across age, but not for sex or the interaction between sex and age.







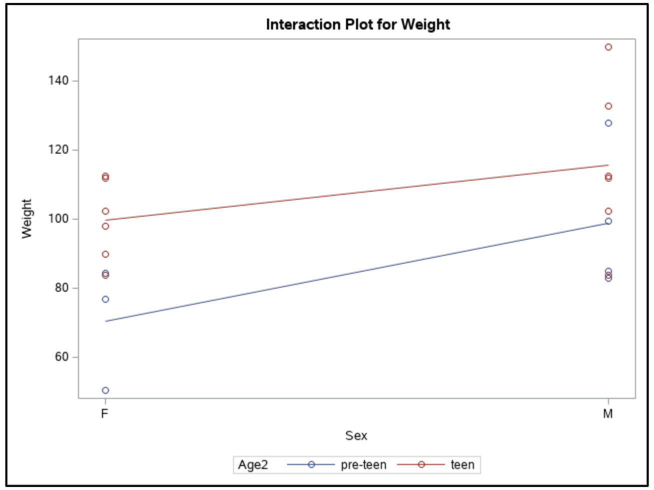


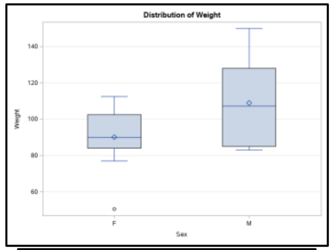


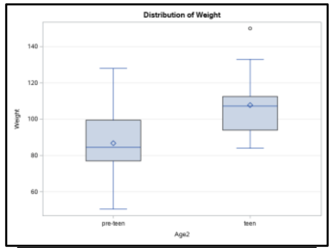
#2) Is average weight different across age or sex for children?

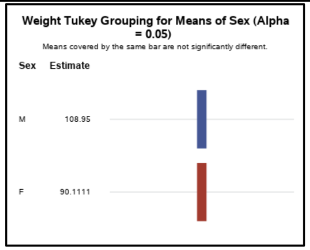
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	1	2116.001894	2116.001894	6.02	0.0269
Age2	1	2304.183712	2304.183712	6.55	0.0218
Sex*Age2	1	167.062500	167.062500	0.47	0.5013

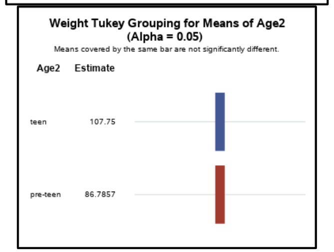
Yes, weight was significantly different across age and sex, but not for the interaction between sex and age.











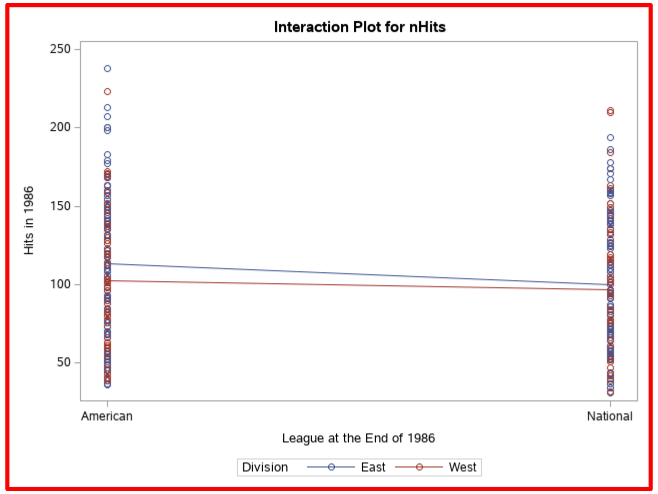
5. Two-way ANOVA

- *Tests if a variable's mean is different between a two categories with multiple groups each*

#3) Is the average number of hits for baseball players different across league or division?

Source	DF	Type III SS	Mean Square	F Value	Pr > F
League	1	7235.796171	7235.796171	3.75	0.0538
Division	1	3878.539028	3878.539028	2.01	0.1573
League* Division	1	1280.147494	1280.147494	0.66	0.4161

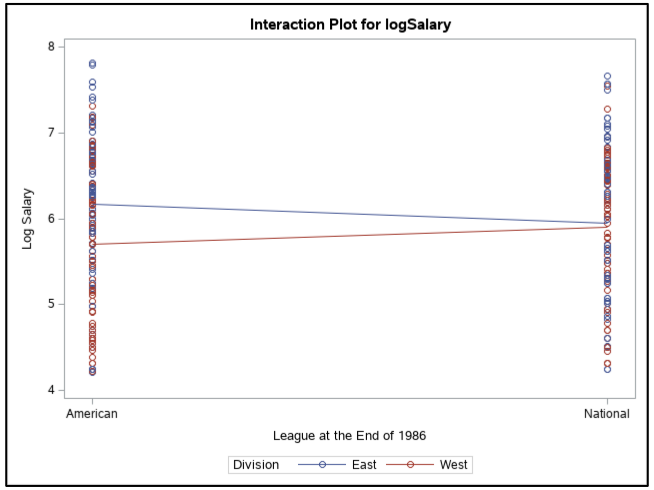
No, number of hits was not significantly different across league, division, or the interaction.

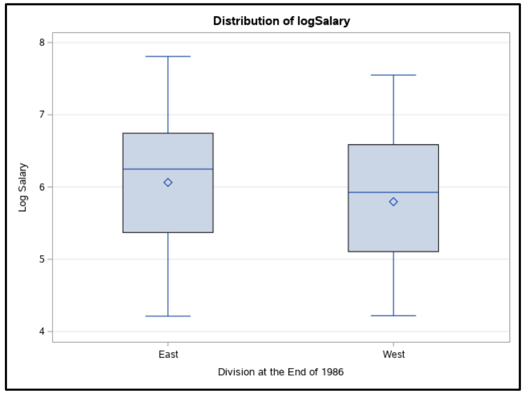


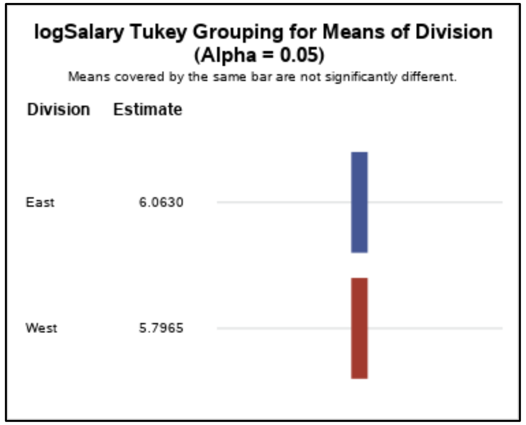
#4) Is the average log salary for baseball players different across league or division?

Source	DF	Type III SS	Mean Square	F Value	Pr > F
League	1	0.01679721	0.01679721	0.02	0.8828
Division	1	4.25477377	4.25477377	5.52	0.0196
League* Division	1	2.74858672	2.74858672	3.56	0.0602

Yes, log salary was significantly different across division, but not league or the interaction.







DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

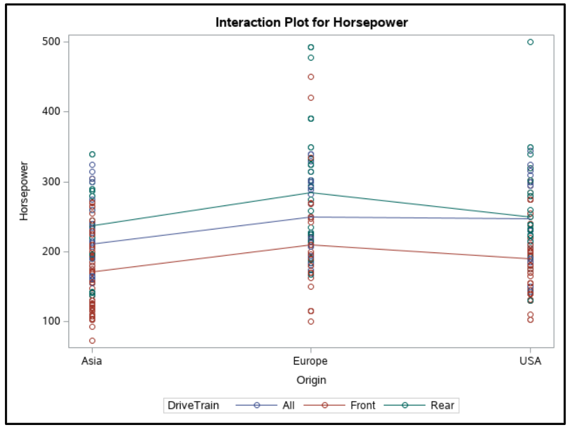
5. Two-way ANOVA

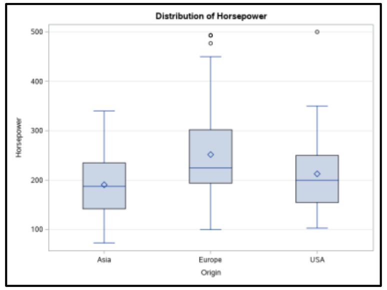
- *Tests if a variable's mean is different between a two categories with multiple groups each*

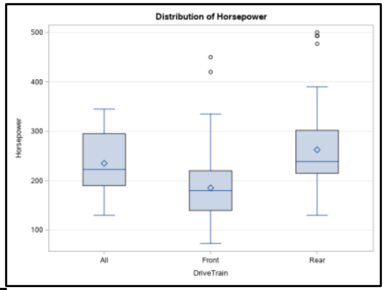
#5) Is the average horsepower for cars different across origin or drive train?

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origin	2	98610.3193	49305.1597	12.91	<.0001
DriveTrain	2	323276.2312	161638.1156	42.32	<.0001
Origin* DriveTrain	4	9396.6601	2349.1650	0.62	0.6520

Yes, horsepower was significantly different across origin and drive chain, but not the interaction.







Comparisons significant at the 0.05 level are indicated by ***.

Origin Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Europe - USA	39.071	21.308	56.834	***
Europe - Asia	61.192	43.713	78.671	***
USA - Europe	-39.071	-56.834	-21.308	***
USA - Asia	22.121	5.463	38.778	***
Asia - Europe	-61.192	-78.671	-43.713	***
Asia - USA	-22.121	-38.778	-5.463	***

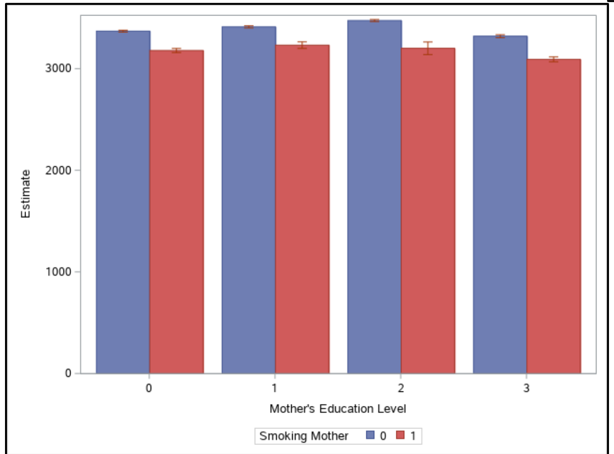
Comparisons significant at the 0.05 level are indicated by ***.

DriveTrain Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Rear - All	27.475	6.938	48.012	***
Rear - Front	77.232	60.333	94.131	***
All - Rear	-27.475	-48.012	-6.938	***
All - Front	49.757	31.780	67.734	***
Front - Rear	-77.232	-94.131	-60.333	***
Front - All	-49.757	-67.734	-31.780	***

#6) Is infant birth weight different across maternal education level or smoking status?

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MomEdLevel	3	42006795.9	14002265.3	44.95	<.0001
MomSmoke	1	143245320.3	143245320.3	459.85	<.0001
MomEdLevel* MomSmoke	3	3360388.6	1120129.5	3.60	0.0129

Yes, birth weight is different across maternal education level, smoking status, and the interation.



6. Blocked/Nested ANOVA

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

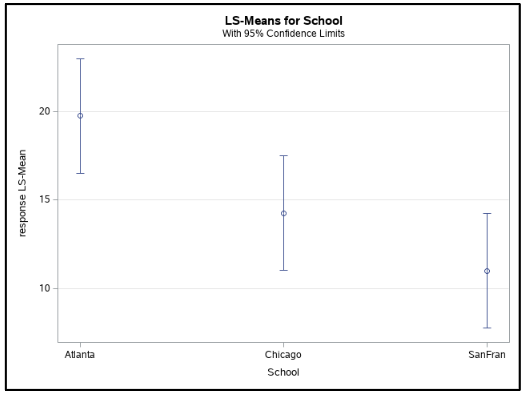
UNIVERSITY OF NORTH DAKOTA

§ Tests if a variable's mean is different across categories while accounting for blocking/nesting

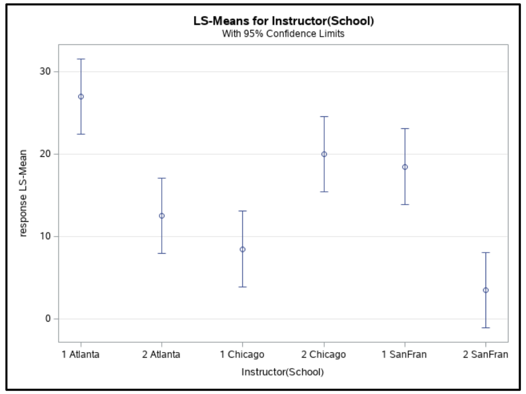
#1) Are average responses different across school and instructor, where instructor is nested in school?

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
School	2	6	11.18	0.0095
Instructor(School)	3	6	27.02	0.0007

Yes, responses were significantly different for both school and instructor.



Tukey Grouping for School Least Squares Means (Alpha=0.05)			
LS-means with the same letter are not significantly different.			
School	**Estimate**		
Atlanta	19.7500		A
			A
Chicago	14.2500	B	A
		B	
SanFran	11.0000	B	

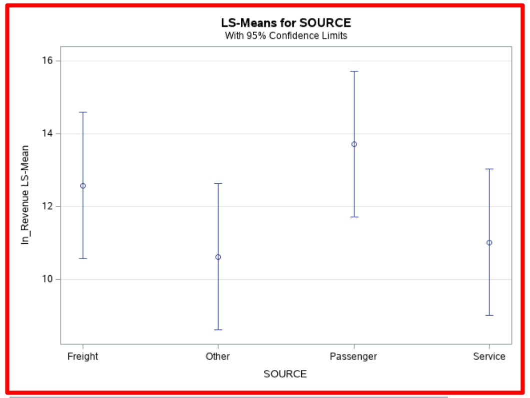


Tukey Grouping for Instructor(School) Least Squares Means (Alpha=0.05)					
LS-means with the same letter are not significantly different.					
School	**Instructor**	**Estimate**			
Atlanta	1	27.0000		A	
				A	
Chicago	2	20.0000	B	A	
			B	A	
SanFran	1	18.5000	B	A	C
			B		C
Atlanta	2	12.5000	B	D	C
				D	C
Chicago	1	8.5000		D	C
				D	
SanFran	2	3.5000		D	

#2) Is average log revenue for airlines different across flight type, where flight type is nested in flight source?

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
SOURCE	3	8	2.70	0.1161

No, log revenue was not different across flight source.



Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
TYPE(SOURCE)	2.1938	1.1349
Residual	0.4410	0.08120

6. Blocked/Nested ANOVA

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

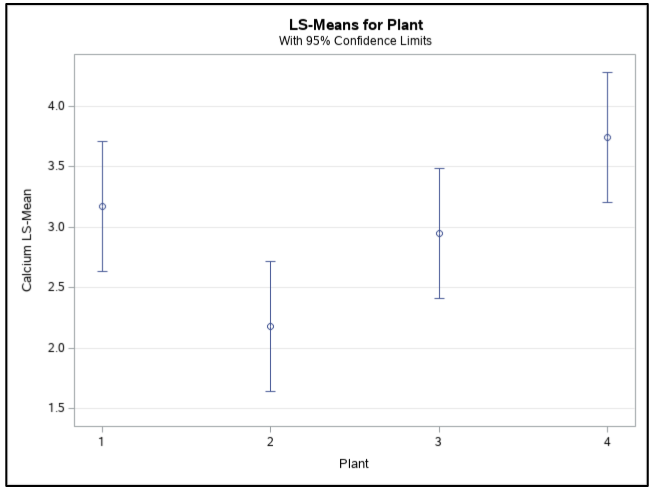
UNIVERSITY OF NORTH DAKOTA

§ *Tests if a variable's mean is different across categories while accounting for blocking/nesting*

#3) Are average calcium levels different across turnip plants, where samples are nested in leaves and plants?

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Plant	3	8	7.67	0.0097

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
Leaf(Plant)	0.1611	0.08220
Sample(Plant*Leaf)	0.000951	0.002717
Residual	0.005703	.



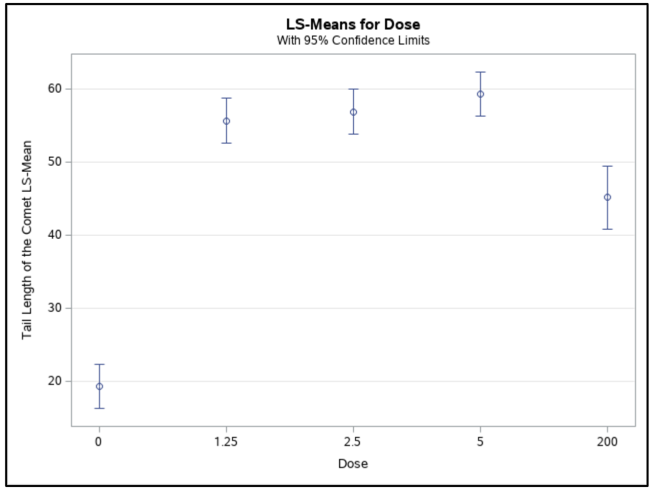
Tukey Grouping for Plant Least Squares Means (Alpha=0.05)			
LS-means with the same letter are not significantly different.			
Plant	**Estimate**		
4	3.7433		A
			A
1	3.1750	B	A
		B	A
3	2.9517	B	A
		B	
2	2.1783	B	

Yes, calcium levels were significantly different across plants.

#4) Is average cell DNA damage different across drug dose, when controlling for rat?

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Dose	4	4023	112.53	<.0001

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
Rat	13.9414	4.3715
Residual	83.5834	1.8636



Tukey-Kramer Grouping for Dose Least Squares Means (Alpha=0.05)		
LS-means with the same letter are not significantly different.		
1,2 Dimethylhydrazine dihydrochloride Dose Level	**Estimate**	
5	59.2416	A
		A
2.5	56.8405	A
		A
1.25	55.6127	A
200	45.1176	B
0	19.3232	C

Yes, cell damage was different across drug dose.

Blocked/Nested ANOVA

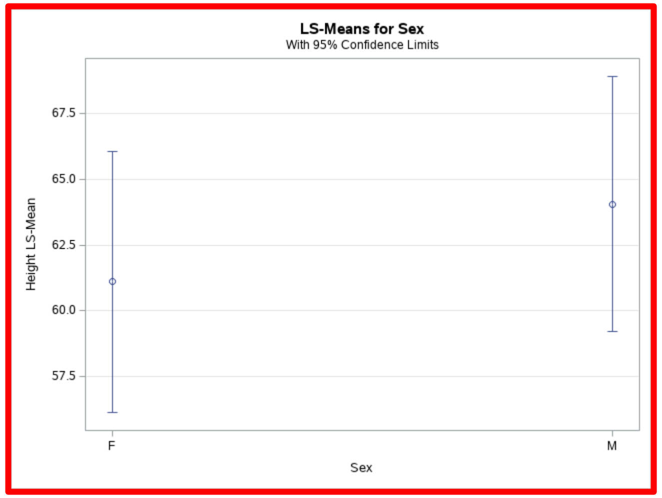
§ Tests if a variable's mean is different across categories while accounting for blocking/nesting

#5) Is average height different across sex in children when controlling for age?

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Sex	1	12	4.22	0.0624

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
Age	23.7445	18.8052
Residual	9.2149	3.8705

No, height was not significantly different across sex, even when controlling for age.



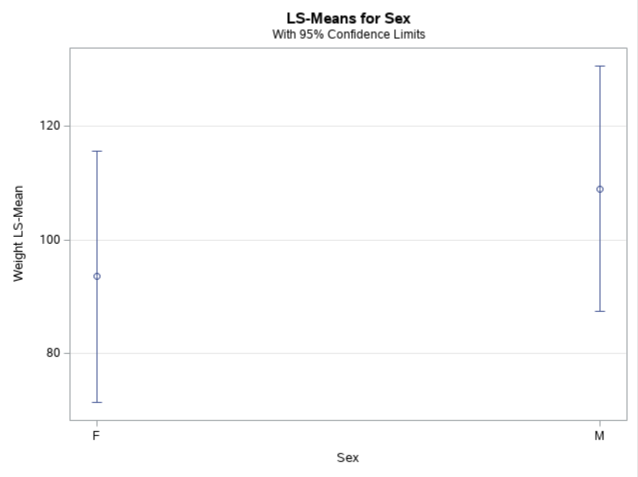
Tukey-Kramer Grouping for Sex Least Squares Means (Alpha=0.05)		
LS-means with the same letter are not significantly different.		
Sex	Estimate	
M	64.0582	A
		A
F	61.0993	A

#6) Is average weight different across sex in children when controlling for age?

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Sex	1	12	5.74	0.0338

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
Age	464.53	383.92
Residual	185.01	78.9816

Yes, weight was significantly different across sex when controlling for age.



Tukey-Kramer Grouping for Sex Least Squares Means (Alpha=0.05)		
LS-means with the same letter are not significantly different.		
Sex	Estimate	
M	108.98	A
F	93.5252	B

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

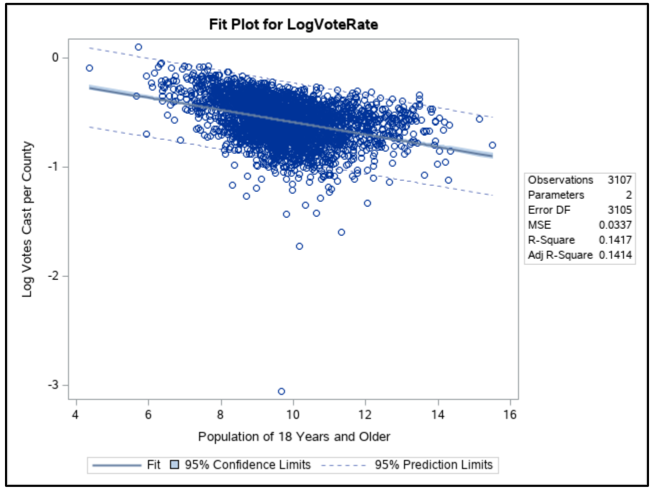
7. Simple Linear Regression

- *Tests if there is a relationship between a numerical response variable and one numerical predictor variable*

#1) Can the log number of votes be predicted by population in US counties?

Variable	Parameter Estimate	t Value	Pr > |t|
Intercept	-0.025	-1.04	0.3003
Pop	-0.056	-22.64	<.0001

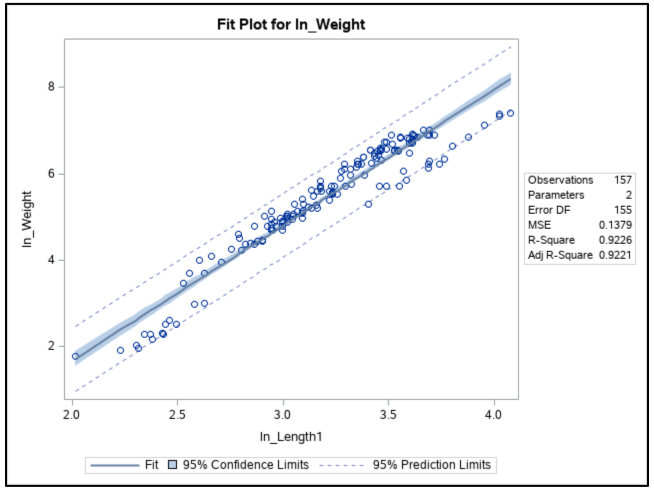
Yes, there was a significant negative relationship. As population increased, log voting rate decreased.



#2) Can log weight be predicted by log length for fish?

Variable	Parameter Estimate	t Value	Pr > |t|
Intercept	-4.63	-19.67	<.0001
Ln_length1	3.15	42.97	<.0001

Yes, there was a significant positive relationship. As log length increased, log weight increased.



#3) Can log weight be predicted by log width for fish?

Variable	Parameter Estimate	t Value	Pr > |t|
Intercept	1.54	23.23	<.0001
Ln_width	2.77	61.38	<.0001

Yes, there was a significant positive relationship. As log width increased, log weight increased.

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

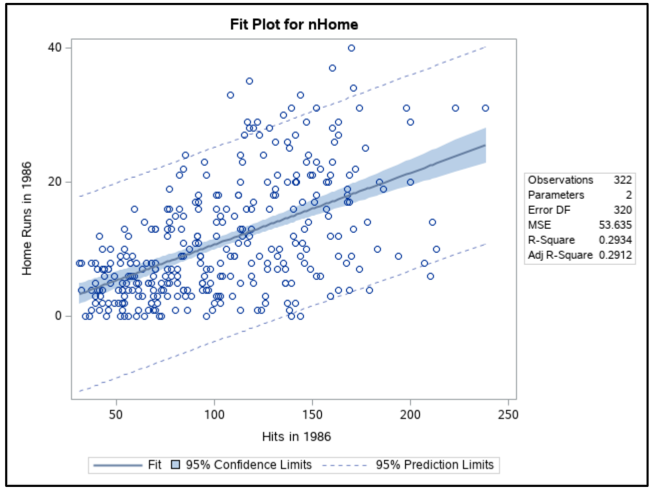
7. Simple Linear Regression

Tests if there is a relationship between a numerical response variable and one numerical predictor variable

#4) Can the number of home runs be predicted by the number of hits for baseball players?

Variable	Parameter Estimate	t Value	Pr > |t|
Intercept	0.076	0.07	0.9424
nHits	0.107	11.53	<.0001

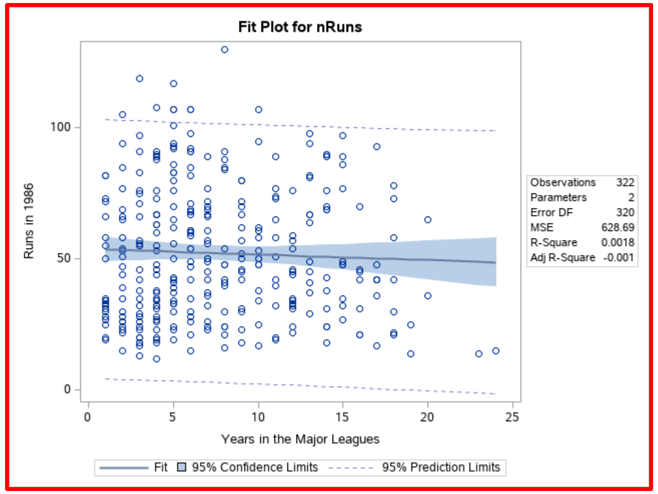
Yes, there was a significant positive relationship. As the number of hits increased, the number of home runs increased.



#5) Can the number of runs be predicted by the number of years in the major leagues for baseball players?

Variable	Parameter Estimate	t Value	Pr > |t|
Intercept	53.87	20.92	<.0001
YrMajor	-0.22	-0.76	0.4454

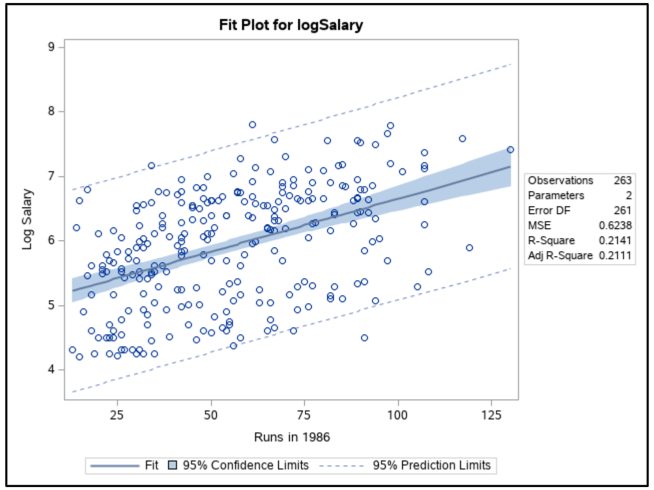
No, there was no significant relationship between the number of runs and the number of years in the major leagues.



#6) Can log salary be predicted by the number of runs for baseball players?

Variable	Parameter Estimate	t Value	Pr > |t|
Intercept	5.017	42.35	<.0001
nRuns	0.016	8.43	<.0001

Yes, there was a significant positive relationship. As the number of runs increased, log salary increased.



8. Multiple Linear Regression

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Tests if there is a relationship between a numerical response variable and multiple numerical predictor variables

#1) Can the log number of votes be predicted by population, education, and housing in US counties?

Yes, there was a significant negative relationship with population, and significant positive relationships with education and houses. The log number of votes increased as population decreased, education increased, and houses increased.

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	AICC
0	Intercept		1	-6951.7424
1	Pop		2	-7424.3175
2	Edu		3	-8672.4932
3	Houses		4	-9201.1182*

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.68664	0.02358	29.12	<.0001
Pop	Population of 18 Years and Older	1	-0.92521	0.01938	-47.74	<.0001
Edu	Population with 12th Grade and Higher	1	0.44113	0.01155	38.21	<.0001
Houses	Number of Owned Housing Units	1	0.43312	0.01802	24.04	<.0001

#2) Can the log number of votes be predicted by population, education, housing, and all interactions in US counties?

Yes, there was a significant relationship for all variables and several interactions.

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	AICC
0	Intercept		1	-6951.7424
1	Pop		2	-7424.3175
2	Edu*Houses		3	-8801.8095
3	Pop*Edu*Houses		4	-9368.6768
4	Houses		5	-9400.4895
5	Pop*Houses		6	-9503.4026
6	Pop*Edu		7	-9526.6255
7		Edu*Houses	6	-9528.2158
8	Edu		7	-9552.5803
9	Edu*Houses		8	-9563.3115*

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.486964421	0.33037095	7.53	<.0001
Pop	-0.920060663	0.11198424	-8.22	<.0001
Edu	-0.647090478	0.10383141	-6.23	<.0001
Pop*Edu	0.080618456	0.01054684	7.64	<.0001
Houses	1.079449517	0.08153133	13.24	<.0001
Pop*Houses	-0.087593996	0.01041664	-8.41	<.0001
Edu*Houses	0.042278596	0.01184683	3.57	0.0004
Pop*Edu*Houses	-0.000872167	0.00036969	-2.36	0.0184

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

8. Multiple Linear Regression

- *Tests if there is a relationship between a numerical response variable and multiple numerical predictor variables*

#3) Can the log weight be predicted by log length1, log length2, log length3,log height, and log width for fish?

Yes, there was a significant positive relationship. log weight increased as log width increased.

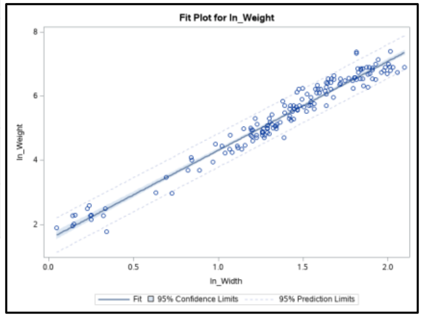
Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	AICC
0	Intercept		1	249.6171
1	ln_Width		2	-255.5954
2	ln_Length3		3	-438.7209
3	ln_Height		4	-561.0471
4	ln_Length2		5	-585.0382
5		ln_Length3	4	-586.0893*

Pearson Correlation Coefficients

	ln_Weight	ln_Width	ln_Height	ln_Length2
ln_Weight	1.00000	0.98004	0.92079	0.96605
ln_Width	0.98004	1.00000	0.90156	0.93023
ln_Height	0.92079	0.90156	1.00000	0.81081
ln_Length2	0.96605	0.93023	0.81081	1.00000

Parameter	Estimate	t Value	Pr > |t|
Intercept	1.54	23.23	<.0001
ln_Width	2.78	61.38	<.0001



#4) Can log salary be predicted by the number of hits, home runs, and runs for baseball players?

Yes, there was a significant positive relationship. Log salary increased as the number of hits and home runs increased.

Pearson Correlation Coefficients

	logSalary	nHits	nHome	nRuns
logSalary	1.00000	0.49233	0.37124	0.46268
nHits	0.49233	1.00000	0.54165	0.91167
nHome	0.37124	0.54165	1.00000	0.63965
nRuns	0.46268	0.91167	0.63965	1.00000

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.834927492	0.12687258	38.11	<.0001
nHits	0.008294892	0.00126440	6.56	<.0001
nHome	0.015799126	0.00630231	2.51	0.0128

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

8. Multiple Linear Regression

- *Tests if there is a relationship between a numerical response variable and multiple numerical predictor variables*

#5) Can log salary be predicted by the number of hits, home runs, outs, assists, and years in the major league?

Yes, there were significant relationships. Log salary increased as the number of hits and years in the major leagues increased.

Pearson Correlation Coefficients

	logSalary	nHits	nHome	nOuts	nAssts	YrMajor
logSalary	1.00000	0.49233	0.37124	0.22448	0.04997	0.56436
nHits	0.49233	1.00000	0.54165	0.32743	0.32131	-0.00803
nHome	0.37124	0.54165	1.00000	0.27319	-0.11134	0.09768
nOuts	0.22448	0.32743	0.27319	1.00000	-0.02520	-0.00995
nAssts	0.04997	0.32131	-0.11134	-0.02520	1.00000	-0.09730
YrMajor	0.56436	-0.00803	0.09768	-0.00995	-0.09730	1.00000

Parameter	Estimate	t Value	Pr > |t|
Intercept	4.053421841	35.87	<.0001
nHits	0.009264021	8.20	<.0001
nHome	0.004112093	0.78	0.4363
nOuts	0.000261019	1.89	0.0598
nAssts	-0.000237545	-0.82	0.4108
YrMajor	0.103663918	13.55	<.0001

#6) Can log salary be predicted by the number of at bats, hits, runs, home runs, walks, outs, assists, and years in the major league?

Yes, there were significant relationships. Log salary increased as the number of hits, walks, and years in the major leagues increased.

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	AICC
0	Intercept		1	204.2699
1	YrMajor		2	105.4641
2	nHits		3	-8.3967
3	nBB		4	-18.8356
4	nOuts		5	-19.7284
5	nAtBat		6	-20.6135*

Pearson Correlation Coefficients

	logSalary	nAtBat	nHits	nBB	nOuts	YrMajor
logSalary	1.00000	0.46183	0.49233	0.46920	0.22448	0.56436
nAtBat	0.46183	1.00000	0.96447	0.63578	0.34395	-0.00848
nHits	0.49233	0.96447	1.00000	0.60620	0.32743	-0.00803
nBB	0.46920	0.63578	0.60620	1.00000	0.30121	0.10870
nOuts	0.22448	0.34395	0.32743	0.30121	1.00000	-0.00995
YrMajor	0.56436	-0.00848	-0.00803	0.10870	-0.00995	1.00000

Parameter	Estimate	t Value	Pr > |t|
Intercept	3.997650575	35.92	<.0001
nHits	0.007609097	7.56	<.0001
nBB	0.006798852	3.30	0.0011
nOuts	0.000231664	1.72	0.0872
YrMajor	0.101189024	13.44	<.0001

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

9. Logistic Regression

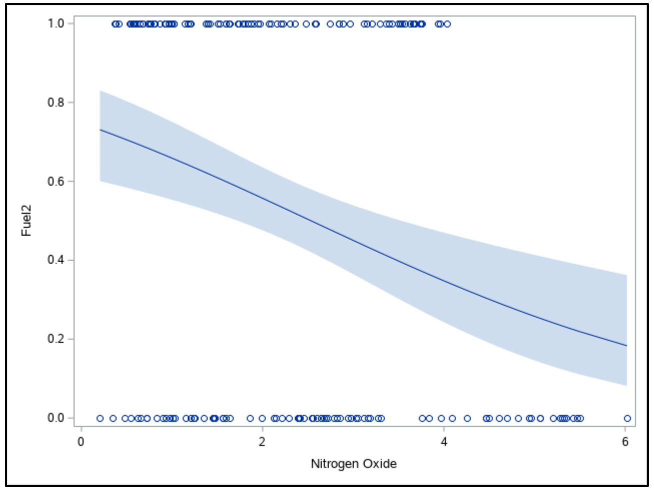
- *Tests if there is a relationship between a binary response variable and one or more predictor variables*

#1) Can fuel status (1=ethanol, 0=non-ethanol) be predicted by nitrogen oxide emission?

Type III Tests of Fixed Effects		
Effect	**F Value**	**Pr > F**
NOx	12.87	0.0004

Pearson Chi-Square / DF	1.00

Yes, there was a significant negative relationship. As nitrogen oxide emission increased, the probability of being ethanol decreased.

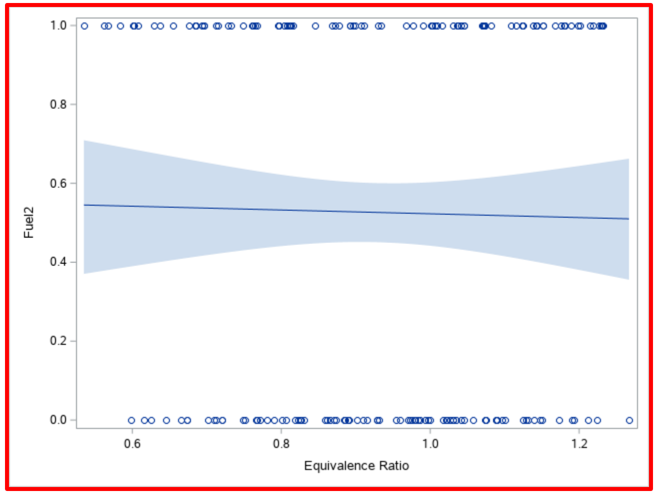


#2) Can fuel status (1=ethanol, 0=non-ethanol) be predicted by Equivalence Ratio?

Type III Tests of Fixed Effects		
Effect	**F Value**	**Pr > F**
EqRatio	0.05	0.81724

Pearson Chi-Square / DF	1.01

No, there was no significant relationship.

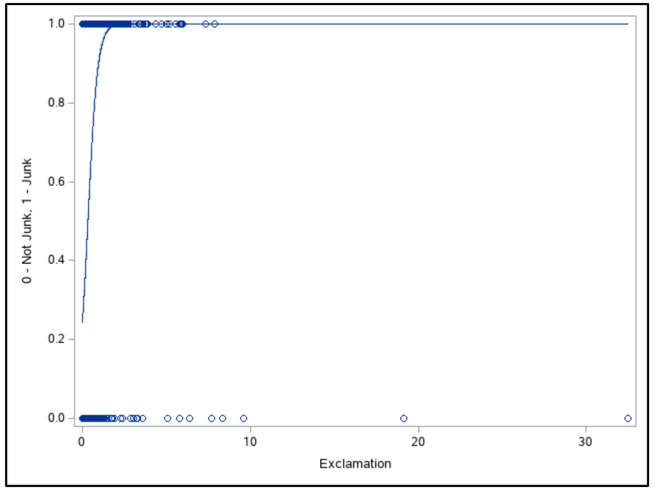


#3) Can Junk mail status (1=junk, 2=non-junk) be predicted by the frequency of exclamation marks?

Type III Tests of Fixed Effects		
Effect	**F Value**	**Pr > F**
Exclamation	578.85	<.0001

Pearson Chi-Square / DF	1.540E10

Yes, there was a significant positive relationship. As the frequency of exclamations increased, the probability of being junk mail increased. However, the model had poor fit.



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

9. Logistic Regression

UNIVERSITY OF NORTH DAKOTA

Tests if there is a relationship between a binary response variable and one or more predictor variables

#4) Can Junk mail status (1=junk, 2=non-junk) be predicted by the frequency several words and symbols?

Type III Tests of Fixed Effects		
Effect	**F Value**	**Pr > F**
Address	0.57	0.4493
Receive	49.28	<.0001
Report	0.20	0.6516
Free	148.03	<.0001
Credit	33.31	<.0001
Money	61.55	<.0001
Exclamation	160.40	<.0001
Dollar	278.37	<.0001

Pearson Chi-Square / DF	8.0357E8

Yes, there were significant relationships. As the frequency of the words 'receive', 'free', 'credit', 'money', and the symbols '!', and '$' increased, the probability of being junk mail increased. However, the model had poor fit.

#5) Can death status (1=dead, 0=censored) be predicted by risk category for post- bone marrow transplant leukemia patients?

Type III Tests of Fixed Effects		
Effect	**F Value**	**Pr > F**
Group	4.31	0.0154

Pearson Chi-Square / DF	1.02

Parameter Estimates				
Effect	**Disease Group**	**Estimate**	**t Value**	**Pr > |t|**
Group	AML-High Risk	0.5895	1.22	0.2246
Group	AML-Low Risk	-0.6874	-1.59	0.1148
Group	ALL	0	.	.

Odds Ratio Estimates				
Disease Group	**Disease Group**	**Estimate**	**95% Confidence Limits**	
AML-High Risk	ALL	1.803	0.693	4.688
AML-Low Risk	ALL	0.503	0.214	1.184

No, the AML-High Risk group was more likely to have died than ALL, while the AML-Low Risk was less likely, but the Odds Ratios were not significant.

#6) Can car type (1=sedan, 0=other) be predicted by origin, drive train, or cylinders?

Type III Tests of Fixed Effects		
Effect	**F Value**	**Pr > F**
Origin	3.37	0.0353
DriveTrain	28.58	<.0001
Cylinders	0.71	0.6450

Pearson Chi-Square / DF	1.01

Odds Ratio Estimates			
Comparison	**Estimate**	**95% Confidence Limits**	
Origin			
Asia vs. USA	0.898	0.518	1.558
Europe vs. USA	1.905	1.052	3.451
DriveTrain			
All vs. Front	0.096	0.051	0.178
Rear vs. Front	0.249	0.139	0.446

Yes, origin and drive train predicted car type, while cylinders did not. European cars were more likely to be sedans vs. US cars. All- and Rear-wheel-drive cars were less likely to be sedans vs. Front-wheel drive cars.

Acknowledgements

References

SAS-code:

- https://med.und.edu/daccota/_files/docs/berdc_docs/model_gauntlet_sascode.txt

Title image:

- https://commons.wikimedia.org/wiki/File:Spiessgasse_Frundsberger_Kriegsbuch_Jost_Ammann_1525.JPG

Selected Examples:

- https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_ttest_sect011.htm
- https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_ttest_sect013.htm
- https://online.stat.psu.edu/stat502/lesson/4/4.2/4.2.1
- https://support.sas.com/documentation/onlinedoc/stat/132/nested.pdf

DaCCoTA

The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729. For projects that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this presentation, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY