# Running the Statistical Gauntlet in R

**Dr. Mark Williamson, PhD**

**Biostatistics, Epidemiology, and Research Design Core**

**DaCCoTA, University of North Dakota**

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

# Introduction

- Common for introductory statistics courses to use only one or two examples to explain a model or method
- Makes it difficult to adapt that example's particularities to your own work
- 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 introductory to intermediate learners

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

# Overview

- Today, we'll be using R
  - Access R software via https://cran.r-project.org/
  - Access R Studio software via https://rstudio.com/
  - Access R code at: https://med.und.edu/daccota/_files/docs/berdc_docs/model_gauntlet_rcode.txt
- All datasets used are available in R via packages or weblink
- 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

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

# Procedure

- Six examples per topic
- 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 R
    - 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


# One-sample t-test

UND UNIVERSITY OF NORTH DAKOTA

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

#1) Is the average miles per gallon (mpg) of 1975 Motor Trend vehicles less than 25?

**t = -4.61**, df = 31, **p-value = 3.293e-05**
95 percent confidence interval: -Inf 21.90
mean of x: 20.09

**Yes, MPG is less than 25.**

#2) Is the average quarter mile time in seconds (qsec) of 1975 Motor Trend vehicles greater than 17.5?

**t = 1.10**, df = 31, **p-value = 0.139**
95 percent confidence interval: 17.31 Inf
mean of x: 17.85

**No, quarter mile time is not greater than 17.5**

#3) Is the average weight in 1000 lbs (wt) of 1975 Motor Trend vehicles different than 3?

**t = 1.26**, df = 31, **p-value = 0.2185**
95 percent confidence interval: 2.86 3.57
mean of x: 3.22

**No, weight is not different than 3.**

#4) Is the log baseline number of seizures (lbase) for the placebo group of epileptics different than 0?

**t = -0.55**, df = 111, **p-value = 0.5849**
95 percent confidence interval: -0.190 0.11
mean of x: -0.04

**No seizure number not different from 0.**

#5) Is the average yield of barley for 1931 greater than 100?

**t = 1.73**, df = 29, **p-value = 0.04731**
95 percent confidence interval: 100.15 Inf
mean of x: 109.05

**Yes, yield is greater than 100.**

#6) Is the average yield of barely for 1932 different than 1931's (109)?

**t = -3.58**, df = 29, **p-value = 0.001235**
95 percent confidence interval: 84.07 102.20
mean of x: 93.13

**Yes, yield is different than 109.**

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

# 2. Two-sample t-test

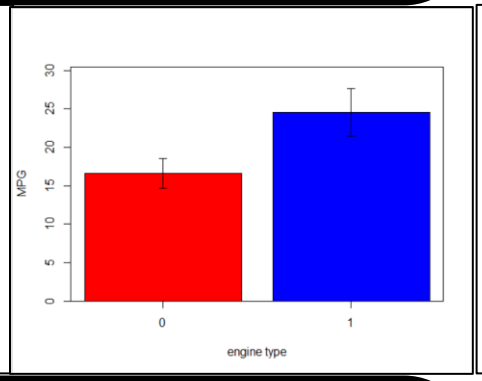
UNIVERSITY OF NORTH DAKOTA

❖ *Tests if the means of two different groups are different*

#1) Is the average miles per gallon (mpg) of 1975 Motor Trend vehicles different between v-shaped (0) and straight (1) engines (vs)?

**t = -4.67**
**p-value = 0.0001098**
**Group 0 mean = 16.62**
**Group 1 mean = 24.56**

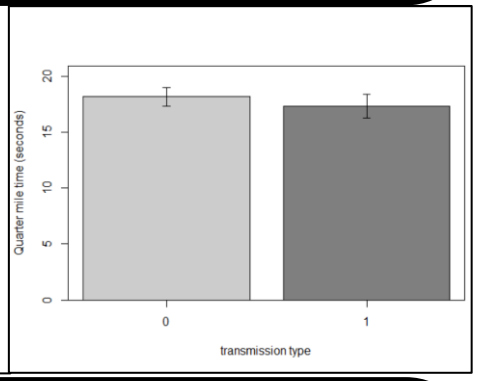
**Yes, straight engines (1) had higher MPG than v-shaped ones (0).**




#2) Is the average quarter mile time in seconds (qsec) of 1975 Motor Trend vehicles greater in automatic (0) compared to manual (1) transmissions (am)?

**t = 1.29**
**p-value = 0.1047**
**Group 0 mean = 18.18**
**Group 1 mean = 17.36**

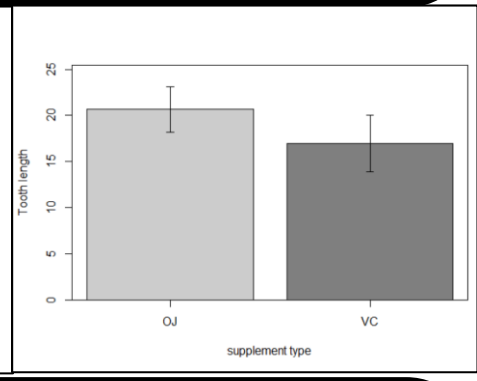
**No, automatic transmissions (0) did not have higher quarter mile times than manual (1).**




#3) Is the average tooth length of guinea pigs (len) different between orange juice (OJ) and ascorbic acid (VC) supplement of vitamin C (supp)?

**t = 1.92**
**p-value = 0.06063**
**OJ mean = 20.66**
**VC mean = 16.96**

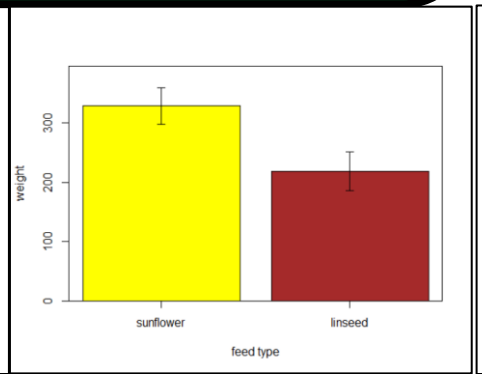
**No, OJ and VC supplemented guinea pigs did not have different tooth lengths.**




#4) Is the average weight (weight) lower in linseed fed chicks compared to sunflower fed ones (feed)?

**t = -5.34**
**p-value = 1.187e-05**
**Sunflower mean = 218.75**
**Linseed mean = 328.92**

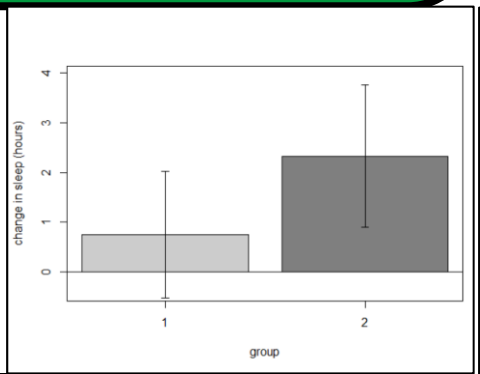
**Yes, linseed-fed chicks had lower weight than sunflower-fed chicks.**




#5) Is the change in hours of sleep (extra) different between two groups of students (group)?

**t = -1.86**
**p-value = 0.07939**
**Group 1 mean = 0.75**
**Group 2 mean = 2.33**

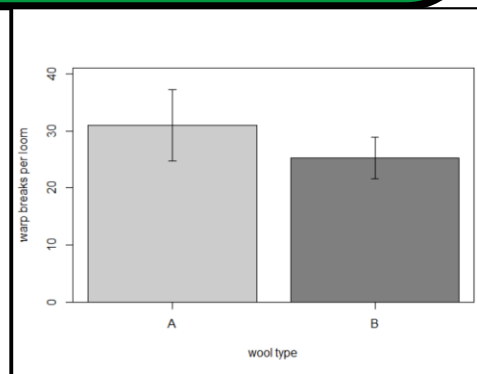
**No, group 1 and 2 did not have different changes in hours of sleep.**




#6) Is the number of warp breaks per loom (breaks) different between two types of wool (wool)?

**t = 1.63**
**p-value = 0.1098**
**Group A mean = 31.04**
**Group B mean = 25.26**

**No, type A and B wool did not have different number of warp breaks.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

3.

# Paired t-test

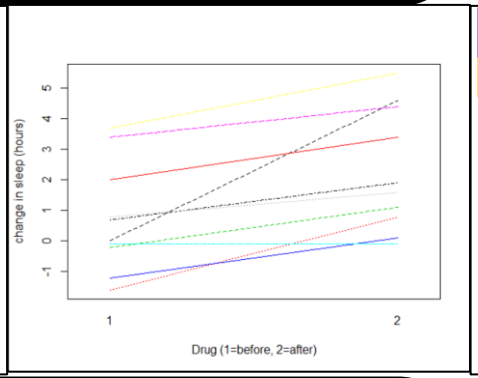
UND UNIVERSITY OF NORTH DAKOTA.

❖ *Tests if the means of two different paired groups are different*

#1) Is the change in hours of sleep (extra) different between students before (1) and after (2) a drug (group)?

**t = -4.06**
**p-value = 0.002833**
**Mean difference = -1.58**

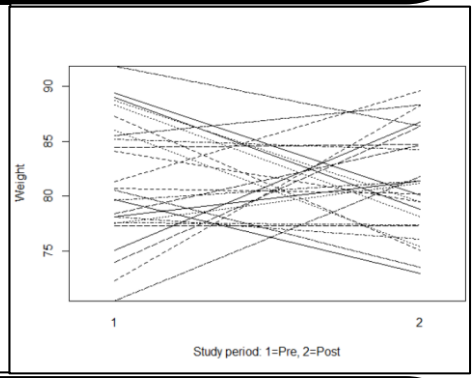
**Yes, students had a higher change in sleep after the drug compared to before.**




#2) Is the average weight (Weight) of anorexia females greater in the post study period compared to the pre study period (Period) for the Control group?

**t = -0.29**
**p-value = 0.6118**
**Mean difference = -0.45**

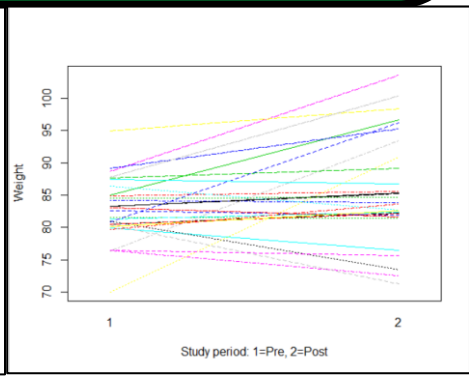
**No, female weights were not higher in the post-study period compared to the pre-study period for the Control group.**




#3) Is the average weight (Weight) of anorexia females greater in the post study period compared to the pre study period (Period) for the the Cognitive Behavioral Treatment group?

**t = 2.22**
**p-value = 0.01751**
**Mean difference = 3.01**

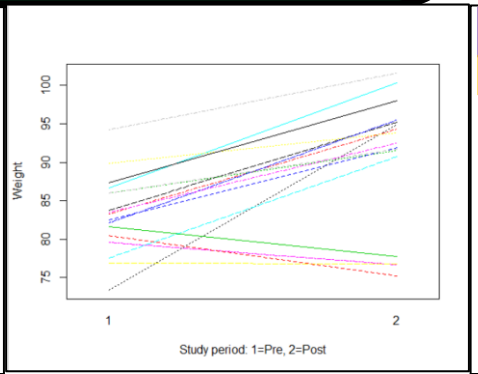
**Yes, female weights were higher in the post-study period compared to the pre-study period for the CBT group.**




#4) Is the average weight (Weight) of anorexia females greater in the post study period compared to the pre study period (Period) for the Family Treatment group?

**t = 4.1849**
**p-value = 0.0003501**
**Mean difference = 7.26**

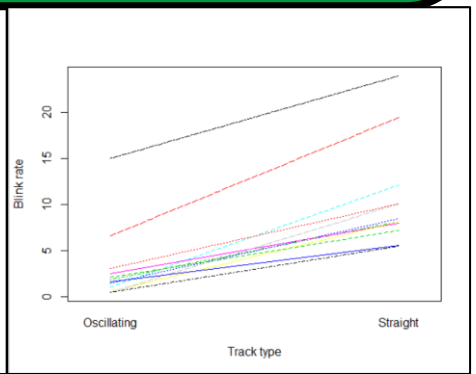
**Yes, female weights were higher in the post-study period compared to the pre-study period for the FT group.**




#5) Is the average blink rate (Rate) different between participants steering a pencil along a straight or oscillating track (Track)?

**t = -9.74**
**p-value = 9.621e-07**
**Mean difference = -7.5**

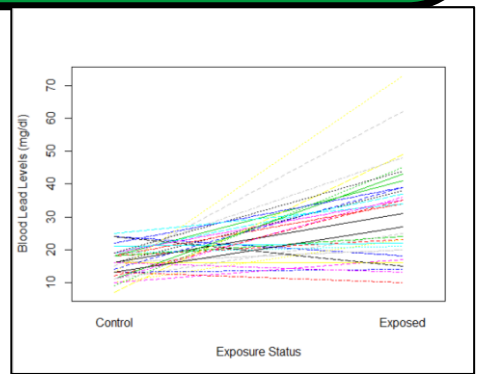
**Yes, blink rates were higher in the straight track compared to the oscillating track.**




#6) Is the average blood lead level (Levels) lower in control children compared to exposed children (Status)?

**t = -5.78**
**p-value = 1.018e-06**
**Mean difference = -15.97**

**Yes, blood lead levels were lower in the control children compared to the exposed children.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

# 4. One-way ANOVA

UNIVERSITY OF NORTH DAKOTA

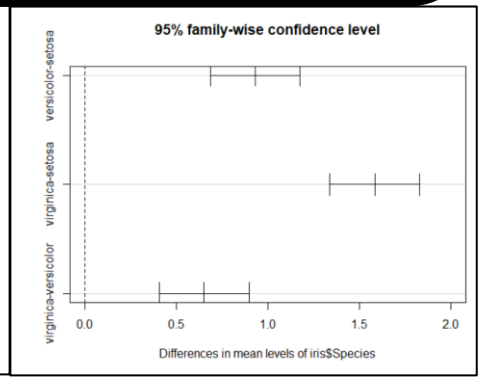
❖ *Tests if a variable's mean is different between a category with three or more groups*

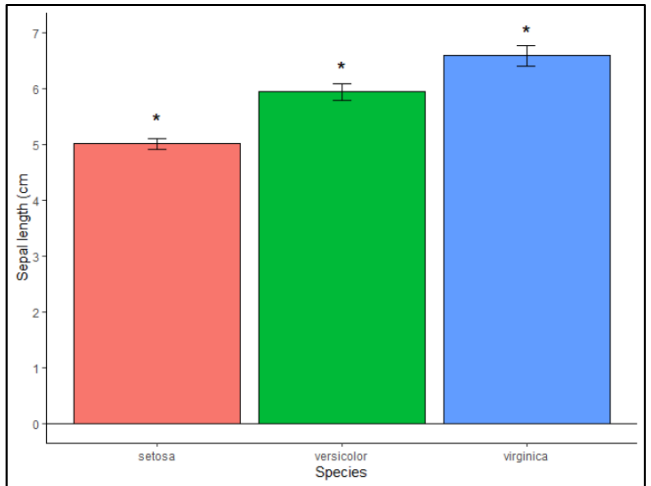
## #1) Is the average sepal length different across three species of iris (Species)?

**F = 119.3**

**p-value = <2e-16**

**Yes, versicolor had longer sepals than setosa and virginica had longer sepals than both.**



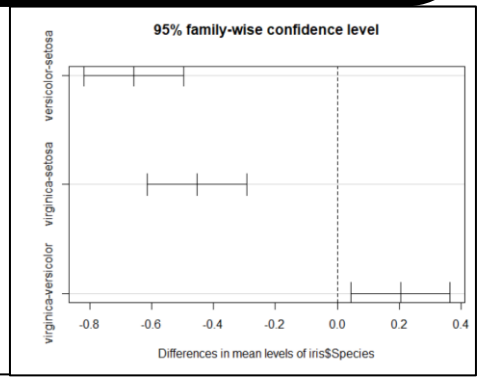



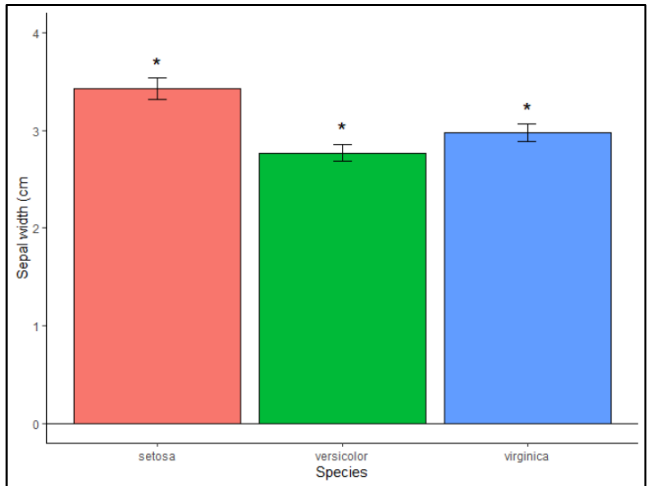

## #2) Is the average sepal width different across three species of iris (Species)?

**F = 49.16**

**p-value = <2e-16**

**Yes, virginica had wider sepals than versicolor and setosa had wider sepals than both.**



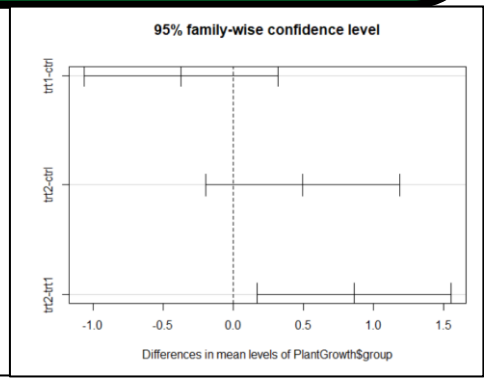



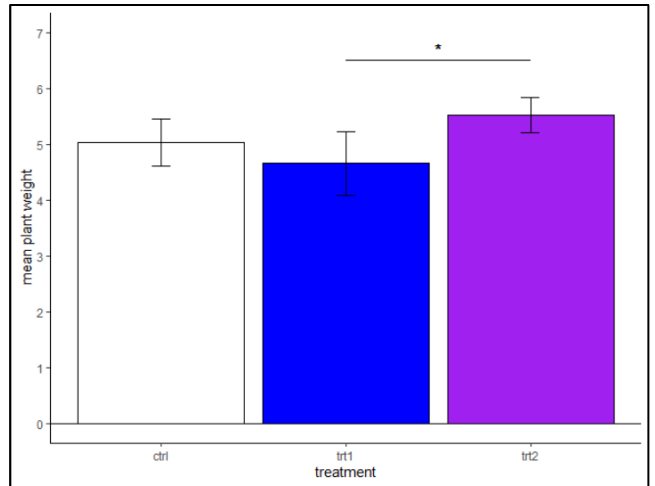

## #3) Is the average plant growth (weight) different across treatments (group)?

**F = 4.85**

**p-value = 0.0159**

**Yes, Treatment 2 had higher weights than Treatment 1.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

4.

# One-way ANOVA cont.

UND UNIVERSITY OF NORTH DAKOTA.

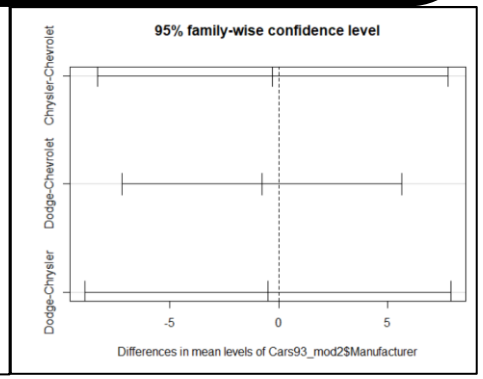
❖*Tests if a variable's mean is different between a category with three or more groups*

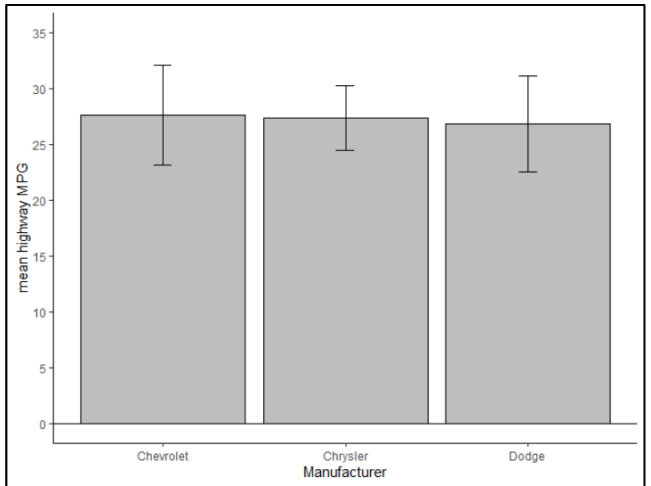
#4) Is the average highway MPG (MPG.highway) different across three car manufacturers (Manufacturer)?

**F = 0.052**

**p-value = 0.949**

**No, MPG is not different across the three manufacturers.**



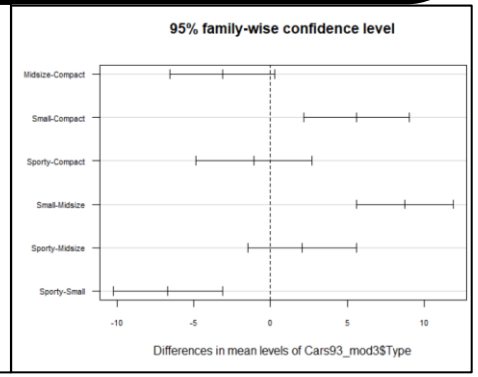



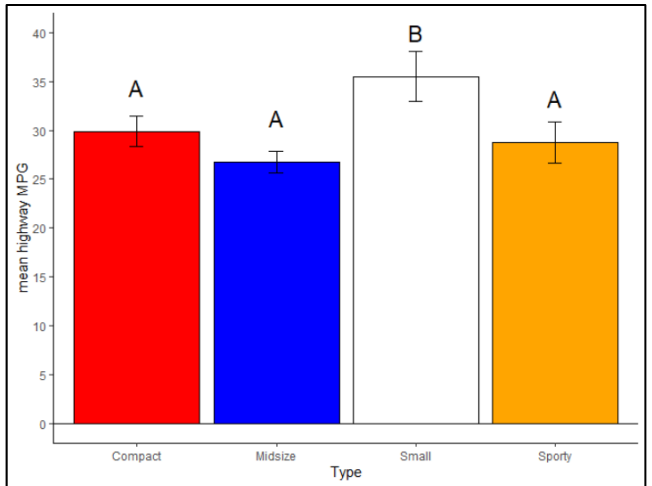

#5) Is the average highway MPG (MPG.highway) different across four types of cars (Type)?

**F = 18.99**

**p-value = 4.37e-09**

**Yes, Small cars had higher MPG than the other three types (Compact, Midsize, and Sporty).**



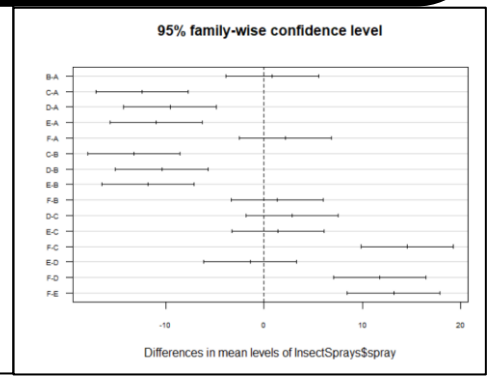



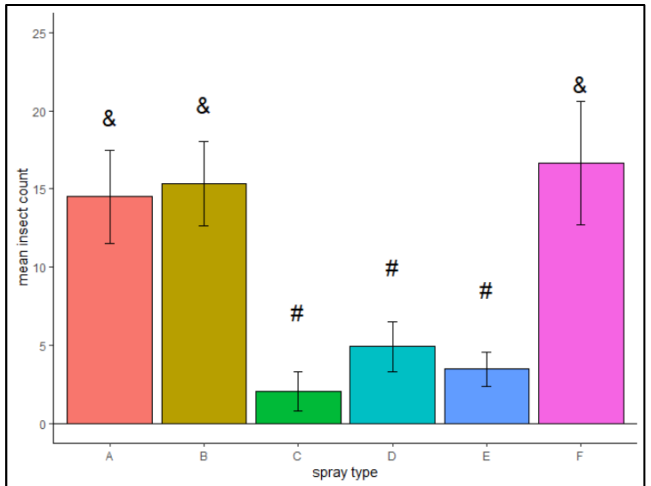

#6) Is the average number of insects (count) after spraying different across six types of spray (spray)?

**F = 34.7**

**p-value = <2e-16**

**Yes, spray types C, D, and E, had lower insect counts after spraying than spray types A, B, and F.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

5.

# Two-way ANOVA

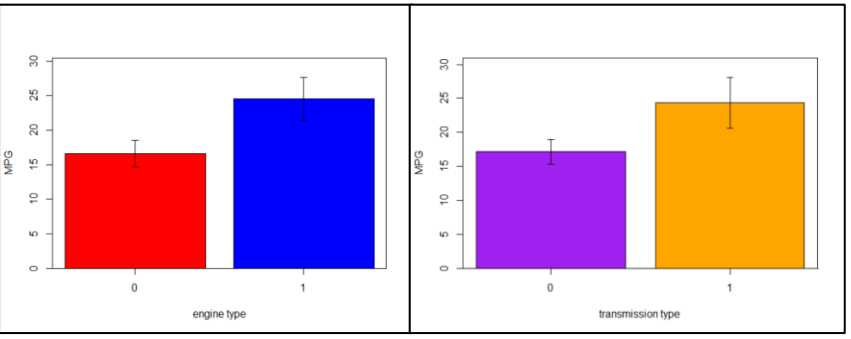
UND UNIVERSITY OF NORTH DAKOTA.

❖ *Tests if a variable's mean is different between two categories*

#1) Is the average miles per gallon (mpg) of 1975 Motor Trend vehicles different across the engine (vs) and transmission (am) type?

| Variable | F-value | p-value |
|---|---|---|
| vs | 41.20 | 5.98e-07 |
| am | 22.90 | 4.98e-05 |
| vs:am | 1.33 | 0.259 |

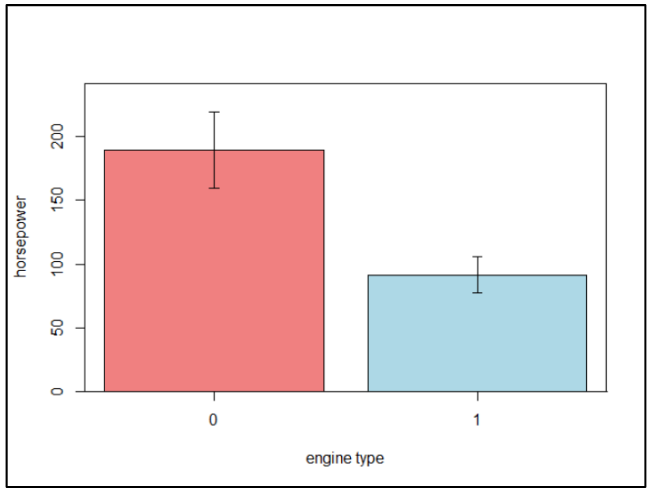
**Yes, MPG was higher in cars with straight engines (1) and manual transmission (1), but the interaction was not significant.**




#2) Is the average gross horsepower (hp) of 1975 Motor Trend vehicles different across the engine (vs) and transmission (am) type?

| Variable | F-value | p-value |
|---|---|---|
| vs | 31.75 | 4.91e-06 |
| am | 0.92 | 0.345 |
| vs:am | 0.05 | 0.820 |

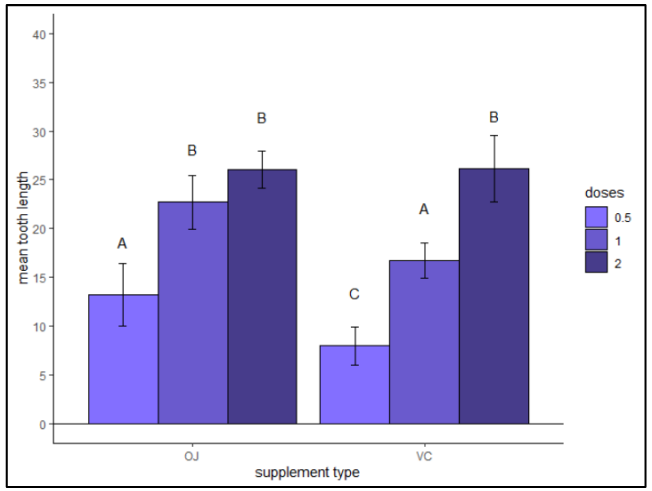
**Yes, horsepower was higher in cars with v-shaped engines (0), but transmission and the interaction were not significant.**




#3) Is the average tooth length of guinea pigs (len) different across supplement (supp) and dose (dose)?

| Variable | F-value | p-value |
|---|---|---|
| supp | 15.57 | 0.000231 |
| dose | 92.00 | < 2e-16 |
| supp:dose | 4.11 | 0.021860 |

**Yes, tooth length was significant across both variables and the interaction. Length increased with increasing dosage and differed between OJ and VC across dose categories.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

5.

# Two-way ANOVA cont.

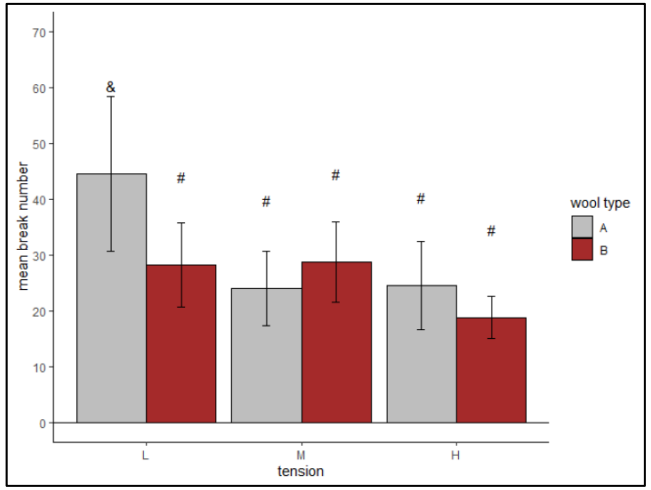
UND UNIVERSITY OF NORTH DAKOTA.

❖ *Tests if a variable's mean is different between two categories*

## #4) Is the number of warp breaks per loom (breaks) different across wool (wool) and tension (tension)?

| Variable | F-value | p-value |
|---|---|---|
| wool | 3.77 | 0.058213 |
| tension | 8.50 | 0.000693 |
| wool:tension | 4.12 | 0.021044 |

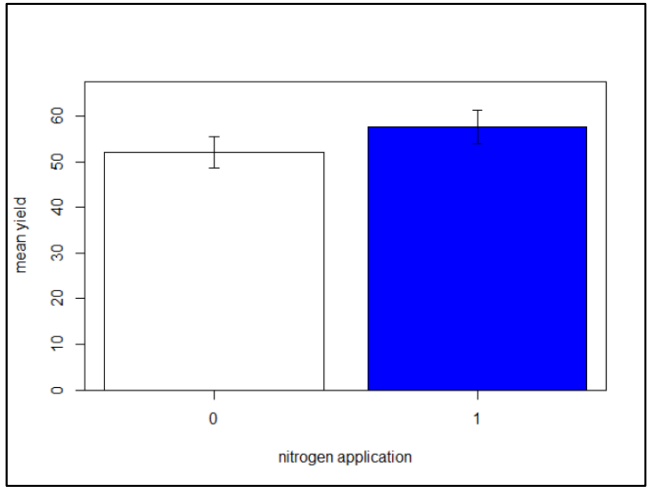
**Yes, the number of warp breaks was significant across tension and the interaction. Type A wool on low tension (&) had higher breaks than the other groups (#).**




## #5) Is the average yield for peas different across nitrogen (N) and phosphorus (P) application?

| Variable | F-value | p-value |
|---|---|---|
| N | 5.76 | 0.0263 |
| P | 0.26 | 0.6187 |
| N:P | 0.65 | 0.4305 |

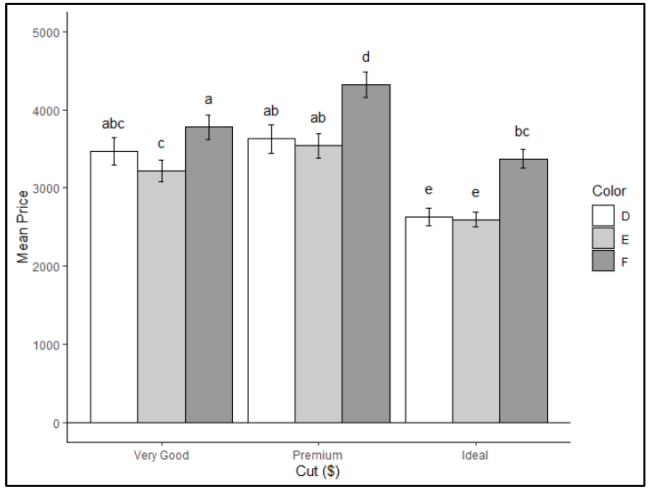
**Yes, yield was higher in Nitrogen (N) plots, but Phosphorus (N) and the interaction were not significant.**




## #6) Is the average price of diamonds different across cut (cut) and color (color)?

| Variable | F-value | p-value |
|---|---|---|
| cut | 157.62 | < 2e-16 |
| color | 100.41 | < 2e-16 |
| cut:color | 2.42 | 0.0462 |

**Yes, price was significant across cut, color, and interaction. Premium, F diamonds had a higher price than other groups, while ideal D-E diamonds lower prices than other groups.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

6.

# Blocked/Nested ANOVA

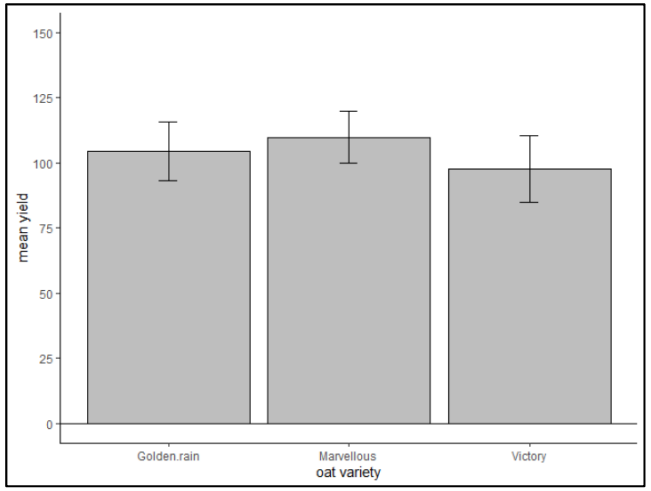
UND UNIVERSITY OF NORTH DAKOTA.

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

**#1) Is the average yield (Y) of oats per sub-plot different across 3 varieties (V), while controlling for block (B)?**

| Variable | F-value | p-value |
|---|---|---|
| V | 1.67 | 0.197 |

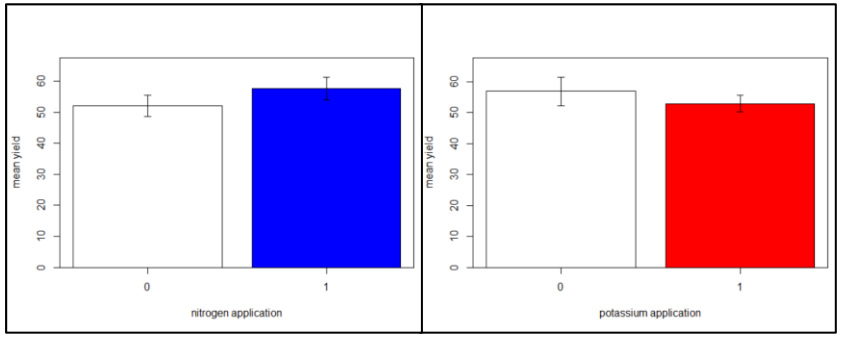
**No, even after controlling for block, the average yield was not different across the three varieties.**




**#2) Is the average yield for peas different across nitrogen (N) and potassium (K) application, while controlling for block?**

| Variable | F-value | p-value |
|---|---|---|
| N | 13.18 | 0.00247 |
| K | 6.63 | 0.02114 |
| N:K | 2.31 | 0.14959 |

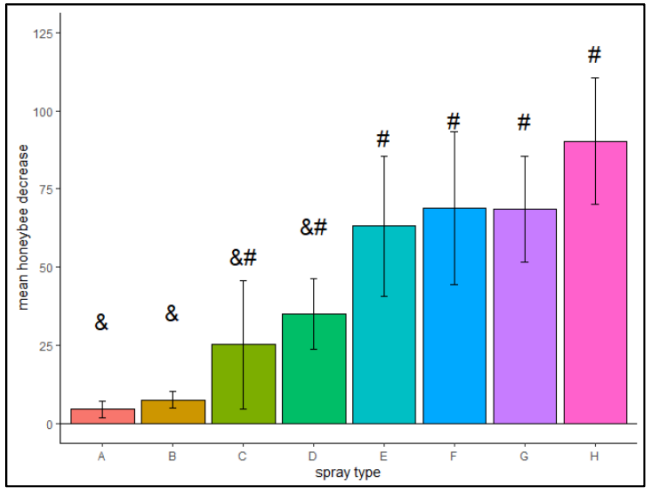
**Yes, yield was higher in plots with nitrogen applied and with no potassium applied. The interaction was not significant even controlling for block.**




**#3) Is the decrease in honeybees (decrease) different across spray type (treatments), while controlling for row (rowpos) and column (colpos)?**

| Variable | F-value | p-value |
|---|---|---|
| Treatment | 20.59 | 3.57e-13 |

**Yes, the decrease in honeybees was lower in sprays A-B, while it was higher in sprays E-H, when controlling for row and column.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

6.

# Blocked/Nested ANOVA cont.

UND UNIVERSITY OF NORTH DAKOTA.

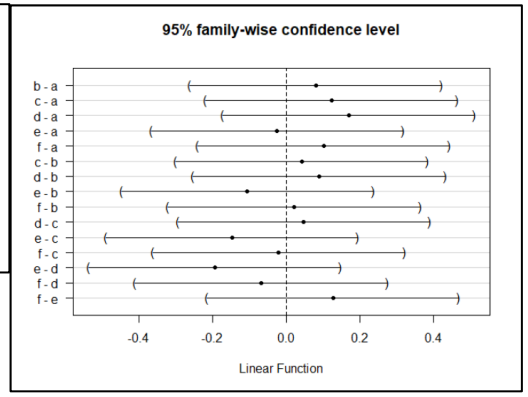
❖ *Tests if a variable's mean is different across categories while accounting for blocking/nesting*

#4) Is the log-optical density (logDens) different across samples (samples) when nested in block (Block)?

| Variable | F-value | p-value |
|---|---|---|
| sample | 11.21 | 0.00952 |

**No, while the overall model appeared significant using aov(), post hoc tests using results from using lme() found no differences in density across samples, even when nesting in block.**

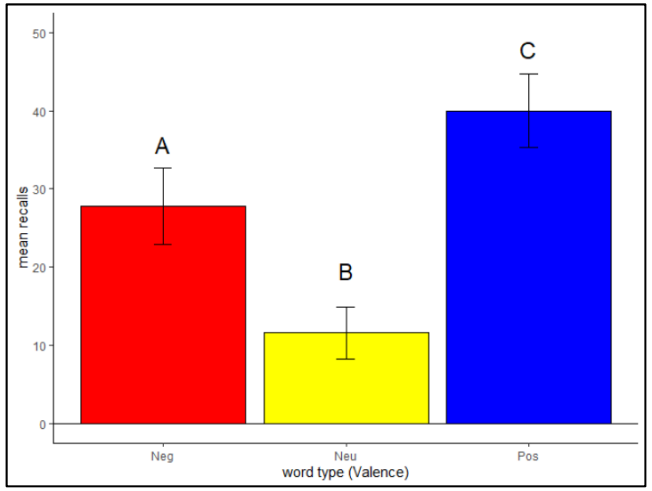
| Comp. | t-value | p-value |
|---|---|---|
| A:B | 0.67 | 0.5315 |
| A:C | 1.02 | 0.3536 |
| A:D | 1.42 | 0.2157 |
| A:E | -0.21 | 0.8436 |
| A:F | 0.85 | 0.4347 |




#5) Is the number of words recollected (Recall) different across word type (Valance) when nested in subject?

| Variable | F-value | p-value |
|---|---|---|
| Valence | 189.1 | 1.84e-07 |

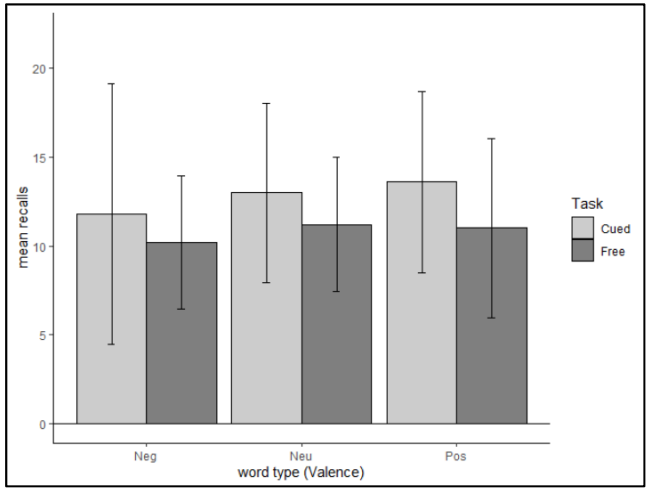
**Yes, word recall was highest for positive words, lower for negative words, and the lowest for neutral words, when accounting for nesting in subject.**




#6) Is the number of words recollected (Recall) different across word type (Valance) and (Task) when nested in subject?

| Variable | F-value | p-value |
|---|---|---|
| Task | 7.35 | 0.0535 |
| Valence | 1.46 | 0.288 |
| Task:Valence | 0.291 | 0.755 |

**No, word recall was not different across task, valence, or the interacting, even when nesting in subject.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

# 7. Simple Linear Regression

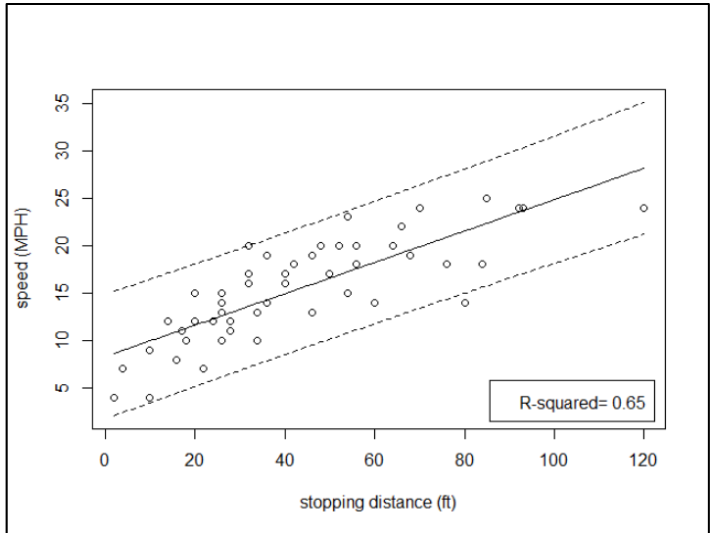
UND UNIVERSITY OF NORTH DAKOTA

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

## #1) Is there a relationship between car speed (speed) and stopping distance (dist)?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 8.28 | 9.47 | 1.44e-12 |
| dist | 0.17 | 9.46 | 1.49e-12 |

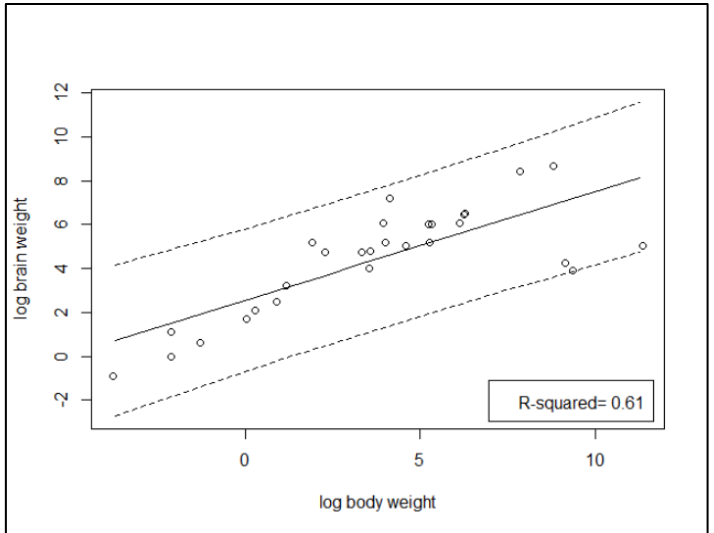
**Yes, there was a positive relationship. Car speeds increased as stopping distance increased.**




## #2) Is there a relationship between log animal brain weight (brain) and log body weight (body)?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 2.55 | 6.18 | 1.53e-06 |
| log_body | 0.50 | 6.35 | 1.02e-06 |

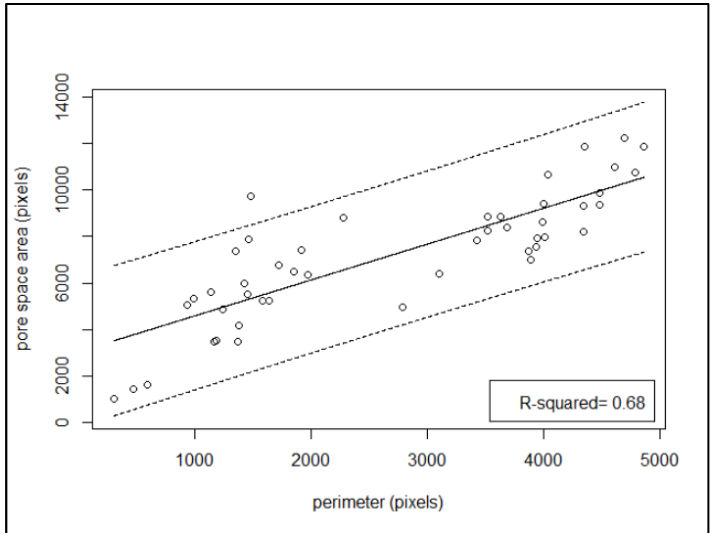
**Yes, there was a positive relationship. Brain weight in animals increased as body weight increased.**




## #3) Is there a relationship between the area of pore space (area) and perimeter (peri) in petroleum rock samples?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 3052.02 | 6.40 | 7.26e-08 |
| peri | 1.54 | 9.91 | 7.51e-13 |

**Yes, there was a positive relationship. Pore space area in rocks increased as perimeter increased.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

# 7. Simple Linear Regression cont.

UNIVERSITY OF NORTH DAKOTA

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

#4) Is there a relationship between the area of pore space (area) and permeability (perm) in petroleum rock samples?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 8197.86 | 16.46 | <2e-16 |
| perm | -2.43 | -2.93 | 0.00525 |

**Yes, there was a negative relationship. Pore space area in rocks decreased as permeability increased.**

#5) Is there a relationship between fertility rates (Fertility) and the percent of education beyond primary school for draftees (Education) in Swiss provinces?

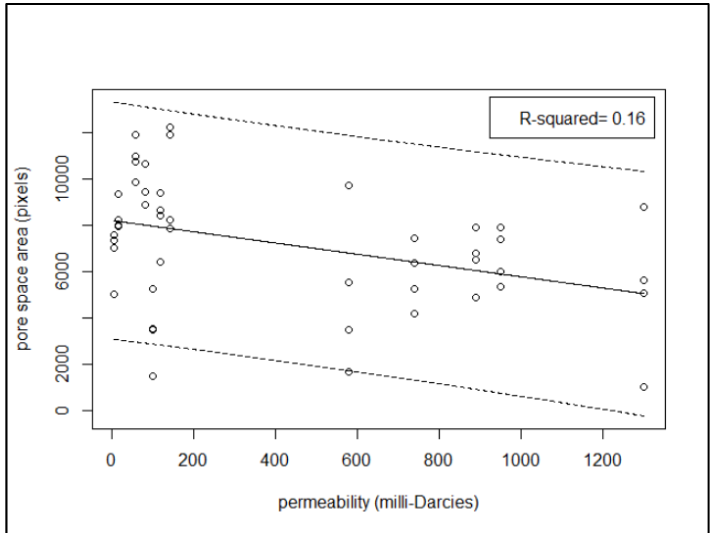
| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 79.61 | 37.84 | < 2e-16 |
| Education | -0.86 | -5.95 | 3.66e-07 |

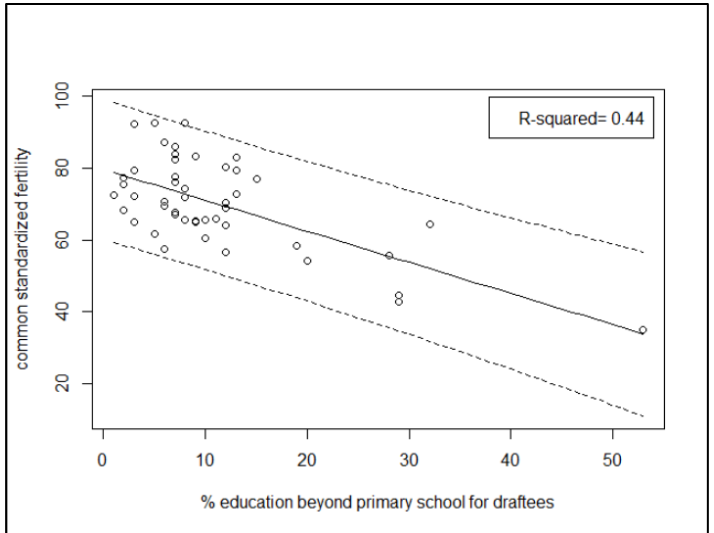
**Yes, there was a negative relationship. Fertility rates in Swiss provinces decreased as education percentage increased.**

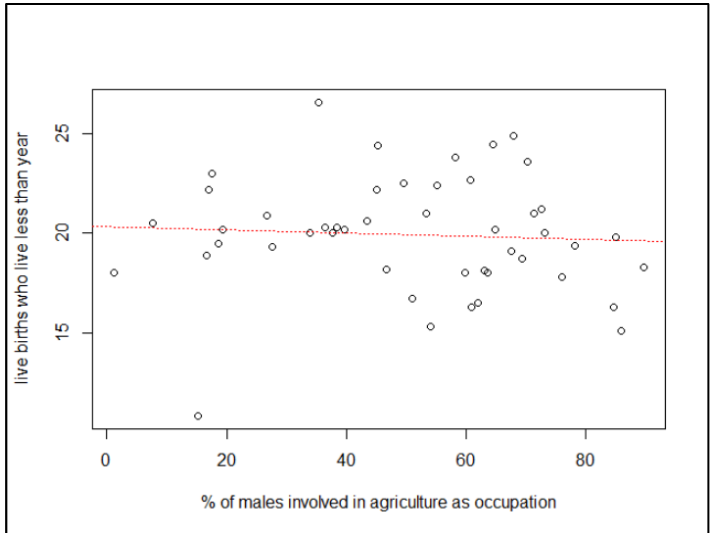
#6) Is there a relationship between infant mortality rates (Infant.Mortality) and the percent of males in agriculture (Agriculture) in Swiss provinces?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 20.34 | 19.23 | <2e-16 |
| Agriculture | -0.008 | -0.41 | 0.684 |

**No, there was no relationship between infant mortality rates and agriculture percentages in Swiss provinces.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

8.

# Multiple Linear Regression

UND UNIVERSITY OF NORTH DAKOTA

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

## #1) Can the area of pore space (area) in petroleum rock samples be predicted by perimeter (peri) and permeability (perm)?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 144.56 | 0.187 | 0.852 |
| peri | 2.19 | 11.10 | 1.81e-14 |
| perm | 2.85 | 4.42 | 6.14e-05 |

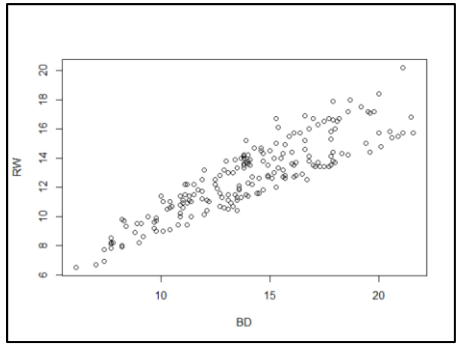
| Model | AIC |
|---|---|
| peri only | 844.96 |
| perm only | 890.91 |
| peri + perm | 829.64 |

Yes, both perimeter and permeability predicted area of pore space in rocks. The full model had the best fit. Pore space increased as perimeter and permeability increased.

## #2) Can frontal lobe size (FL) in crabs be predicted by rear width (RW) and body depth (BD)?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 1.44 | 8.77 | 8.15e-16 |
| BD | 1.01 | 88.62 | < 2e-16 |

| Model | AIC |
|---|---|
| RW only | 727.24 |
| BD only | 332.09 |
| RW + BD | 298.98 |

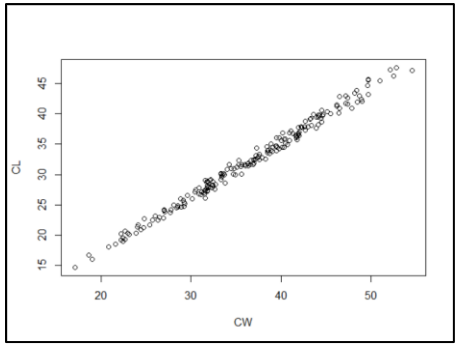



Yes, both rear width and body depth predicted frontal lobe size in crabs. The full model had the best fit, but since rear width and body depth were highly correlated, only the best fit single variable model (body depth) was retained. Frontal lobe size increased as body depth increased.

## #3) Can frontal lobe size (FL) in crabs be predicted by carapace length (CL), width (CW), and their interaction?

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 0.15 | 0.65 | 0.52 |
| CL | 0.48 | 67.31 | <2e-16 |

| Model | AIC |
|---|---|
| CL only | 438.50 |
| CW only | 538.01 |
| CL * CW | 398.90 |




Yes, both carapace length and carapace width predicted frontal lobe size in crabs. The full model had the best fit, but since carapace length and width were highly correlated, only the best fit single variable model (length) was retained. Frontal lobe size increased as carapace length increased.

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

8.

# Multiple Linear Regression cont.

UND UNIVERSITY OF NORTH DAKOTA.

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

#4) Can infant mortality rates (Infant.Mortality) in Swiss provinces be predicted by fertility rates (Fertility), % of males in agriculture (Agriculture), % Catholic (Catholic), and % education beyond primary school (Education)?

***Full Model***

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 9.61 | 2.00 | 0.05188 |
| Fertility | 0.15 | 2.82 | 0.00722 |
| Agriculture | -0.01 | -0.56 | 0.59047 |
| Catholic | -0.002 | -0.19 | 0.84997 |
| Education | 0.07 | 0.95 | 0.34924 |

***Simple Model***

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 13.13 | 5.83 | 5.51e-07 |
| Fertility | 0.10 | 3.07 | 0.00359 |

| Model | AIC |
|---|---|
| Full | 231.89 |
| Simple | 229.91 |

Yes, fertility rate predicted infant mortality rates in Swiss provinces. The rest of the variables were not significant. The simple model (Fertility only) compared to the full model had a better fit, though by less than two. The full model could be retained to show the other variables' non-effect.

#5) Can calories per portion (calories) in US cereals be predicted by protein (protein), fat (fat), carbohydrates (carbo), or their interactions (sqrt-transformed data)?

**Model selection**

Full Model: sqrt_calories ~ sqrt_protein * sqrt_fat * sqrt_carbo (AIC=7.58)

Final Model: sqrt_calories ~ sqrt_protein + sqrt_fat + sqrt_carbo + sqrt_protein:sqrt_carbo (AIC=2.02)

***Final Model***

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 8.65 | 4.60 | 2.20e-05 |
| Sqrt_protein | -1.69 | -2.20 | 0.031374 |
| Sqrt_fat | 1.19 | 6.97 | 2.85e-09 |
| Sqrt_carbo | 0.06 | 0.13 | 0.896086 |
| Sqrt_protein: sqrt_carbo | 0.61 | 3.60 | 0.000644 |

Yes, after model selection and checking correlation of covariates, protein, fat, and the interaction of protein and carbohydrates predicted the calories per portion of US cereals. Calories decreased as protein increased, while calories increased as fat and the interaction term of protein and carbohydrates increased.

#6) Can miles per gallon (mpg) be predicted by gross horsepower (hp), weight (wt), and quarter mile time (qsec), or their interactions?

**Model selection**

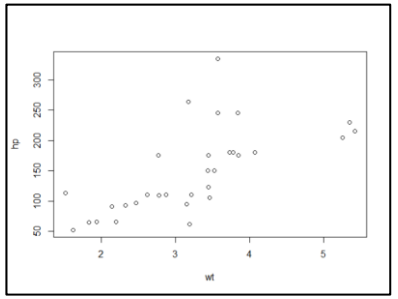
Full Model: mpg ~ hp * wt * qsec (AIC=56.83)

Red. Model: mpg ~ hp + wt + qsec + hp:wt (AIC=52.57)

***Final Model***

| Variable | Estimate | t-value | p-value |
|---|---|---|---|
| Intercept | 19.75 | 3.76 | 0.000765 |
| wt | -5.05 | -10.43 | 2.52e-11 |
| qsec | 0.93 | 3.51 | 0.001500 |

| Model | AIC |
|---|---|
| hp * wt * qsec | 149.64 |
| hp + wt + qsec + hp:wt | 145.38 |
| wt + qsect | 156.72 |




Yes, after model selection and removing horsepower because it was highly correlated with weight, weight and quarter mile time predicted miles per gallon for cars. MPG decreased as weight increased, while MPG increased as quarter mile time increased.

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*

#1) Can midwest counties' metropolitan status (inmetro=Yes vs. No) be predicted by the log population density(log_popdensity)?

| Variable | Estimate | z-value | p-value |
|---|---|---|---|
| Intercept | -21.81 | -10.26 | <2e-16 |
| log_popdensity | 2.88 | 10.03 | <2e-16 |

**Yes, as the log population density increases, the probability of the county having a metropolitan status increases.**

#2) Can midwest counties' metropolitan status (inmetro=Yes vs. No) be predicted by the percentage of college educated adults (percollege)?

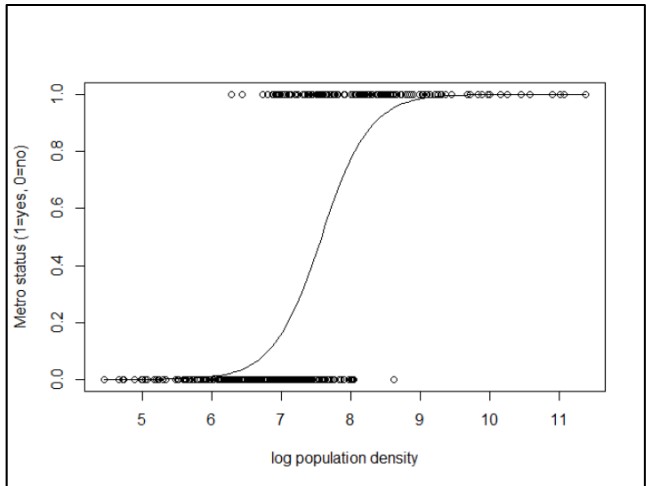
| Variable | Estimate | z-value | p-value |
|---|---|---|---|
| Intercept | -4.56 | -9.61 | <2e-16 |
| percollege | 0.21 | 8.44 | <2e-16 |

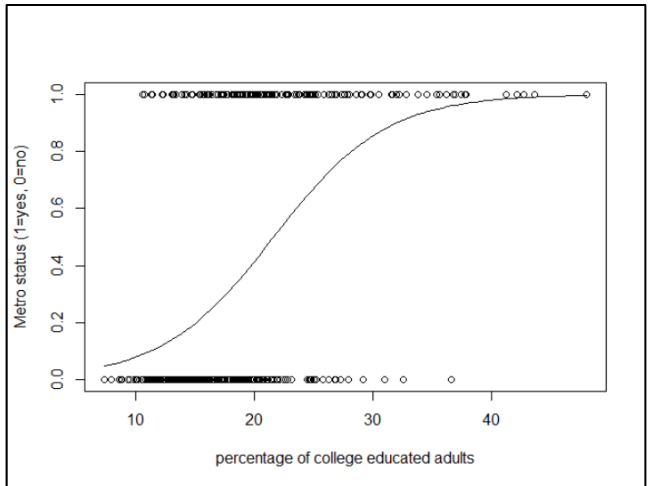
**Yes, as the percentage of college educated adults increases, the probability of the county having a metropolitan status increases.**

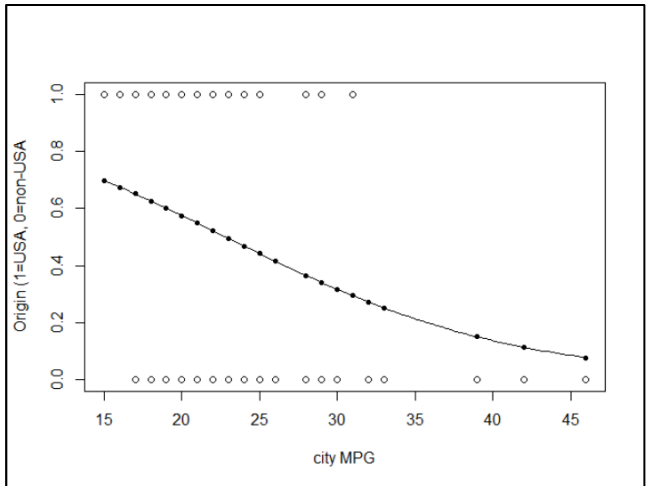
#3) Can a car's origin (Origin2: USA vs. non-USA) be predicted by city miles per gallon (MPG.city)?

| Variable | Estimate | z-value | p-value |
|---|---|---|---|
| Intercept | 2.45 | 2.40 | 0.0165 |
| MPG.city | -0.11 | -2.36 | 0.0183 |

**Yes, as city MPG increases, the probability that the car is of USA origin decreases.**



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

9.

# Logistic Regression cont.

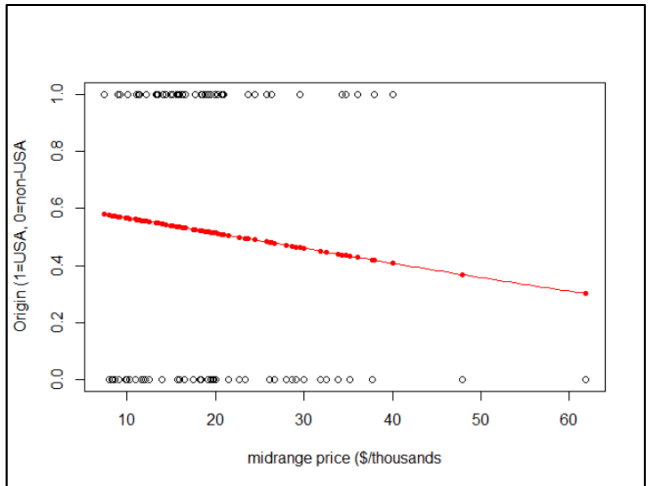
UND UNIVERSITY OF NORTH DAKOTA

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

#4) Can a car's origin (Origin2: USA vs. non-USA) be predicted by midrange price (Price)?

| Variable | Estimate | z-value | p-value |
|---|---|---|---|
| Intercept | 0.48 | 1.00 | 0.316 |
| Price | -0.02 | -0.96 | 0.336 |

**No, midrange price did not significantly predict origin status.**




#5) Can the presence of a bacteria species (y: yes vs. no) in children be predicted by drug status (ap: placebo/active) and compliance (hilo: low/high)?

| Variable | Estimate | z-value | p-value |
|---|---|---|---|
| Intercept | 1.35 | 4.74 | 2.11e-06 |
| spp | 0.79 | 2.12 | 0.0343 |
| hilolo | -0.48 | -1.38 | 0.1664 |

**Yes, drug status predicted the presence of bacteria in children, but compliance did not. Children with the placebo (app) were over twice as likely to have the bacteria compared to those with an active drug status (OR=2.21).**

***Odds Ratios (OR)***

| Variable | OR | 2.5% | 97.5% |
|---|---|---|---|
| app | 2.21 | 1.08 | 4.76 |
| hilolo | 0.62 | 0.31 | 1.22 |

#6) Can the presence of a tumor (status: tumor vs. none) be predicted by sex (sex: male/female) and treatment (rex: control/drug)?

| Variable | Estimate | z-value | p-value |
|---|---|---|---|
| Intercept | -1.38 | -5.61 | 2.07e-08 |
| rx | 0.96 | 2.61 | 0.00909 |
| sexm | -3.35 | -4.54 | 5.71e-06 |

**Yes, treatment and sex predicted the presence of tumors in rats. Rats given the treatment were over twice as likely to have a tumor compared to the control (OR=2.62), and male mice were much less likely to have a tumor compared to females (OR=0.035)**

***Odds Ratios (OR)***

| Variable | OR | 2.5% | 97.5% |
|---|---|---|---|
| rx | 2.62 | 1.27 | 5.43 |
| sexm | 0.035 | 0.006 | 0.118 |

# Acknowledgements

## References

**Title image**: https://commons.wikimedia.org/wiki/File:Spiessgasse_Frundsberger_Kriegsbuch_Jost_Ammann_1525.JPG

**Examples 5 and 6 for Blocked/Nested ANOVA**: https://personality-project.org/r/r.guide/r.anova.html

## 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)".***