

Statistics for the Sciences

Qualitative Predictors and Interaction Effects

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Qualitative Predictors
- Interactions in Regression Models
- Lab 1
- Lab 2

Qualitative Predictors

- Qualitative or categorical predictor variables can be used in regression models. Many predictor variables of interest in business, economics, and the social and biological sciences are categorical. Examples of categorical predictor variables are gender (male, female), purchase status (purchase, no purchase), and disability status (not disabled, partly disabled, fully disabled).
- **Example.** Suppose we want to model Y (person's weight) as a function of X_1 (person's height) and a dummy variable X_2 (Gender), where

$$X_2 = \begin{cases} 1 & \text{Male} \\ 0 & \text{Female} \end{cases}$$

Consider the model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Qualitative Predictors

For males, it becomes

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 x_1.$$

For females, it becomes

$$E(Y) = \beta_0 + \beta_1 x_1.$$

- Why do we combine the data among men and women? Why do not we just model them separately?

Suppose we observe n_1 males and n_2 females.

males SSE has $df = n_1 - 2$

females SSE has $df = n_2 - 2$

But when we combine men and women using the model with interaction effect,

SSE has $df = n_1 + n_2 - 3$.

The larger df is an advantage as long as $\sigma_M^2 = \sigma_F^2$.

Qualitative Predictors

- With more than two levels, we create additional **dummy variables**. If there are c categories, we need $c - 1$ dummy variables.

$$Z_1 = \begin{cases} 1 & \text{category level 1} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{category level 2} \\ 0 & \text{otherwise} \end{cases}$$

\vdots

$$Z_{c-1} = \begin{cases} 1 & \text{category level } c - 1 \\ 0 & \text{otherwise} \end{cases}$$

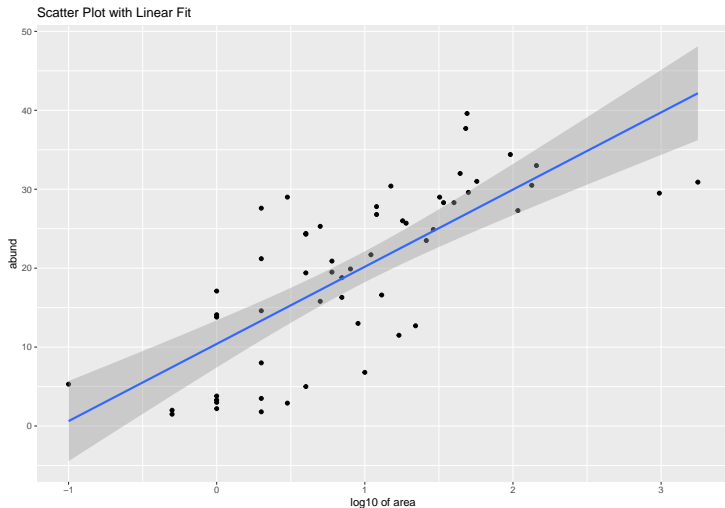
Qualitative Predictors

- Example (loyncat.csv): We re-analysed the data from Loyn (1987) by fitting a simpler model that just included grazing and log patch area.

##	abund	area	graze	grazecat	logarea
## 1	5.3	0.1	2	low	-1.00000
## 2	2.0	0.5	5	intense	-0.30103
## 3	1.5	0.5	5	intense	-0.30103
## 4	17.1	1.0	3	medium	0.00000
## 5	13.8	1.0	5	intense	0.00000
## 6	14.1	1.0	3	medium	0.00000
## 7	3.8	1.0	5	intense	0.00000
## 8	2.2	1.0	5	intense	0.00000
## 9	3.3	1.0	4	high	0.00000
## 10	3.0	1.0	5	intense	0.00000

Qualitative Predictors

- graze was treated as a numerical predictor
- The following is a SLR fit



Qualitative Predictors

- The levels of the categorical predictor grazecat

```
## [1] "zero"      "high"      "intense" "low"       "medium"
```

- We fit the MLR model `abund~logarea+grazecat` with 95% confidence intervals

```
## # A tibble: 6 x 7
##   term          estimate std.error statistic    p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)    15.7       2.77      5.68  0.000000687    10.2    21.3
## 2 logarea        7.25       1.26      5.77  0.000000490     4.73     9.77
## 3 grazecathigh   -1.59       2.98     -0.535  0.595        -7.57     4.39
## 4 grazecatintense -11.9       2.93     -4.06  0.000174       -17.8    -6.01
## 5 grazecatlow     0.383      2.91      0.131  0.896        -5.47     6.23
## 6 grazecatmedium -0.189      2.55     -0.0742 0.941        -5.31     4.93
```


Qualitative Predictors

- In the above model, 4 Qualitative Predictors are added so the MLR model is actually $\text{abund} \sim \log \text{area} + Z_1 + Z_2 + Z_3 + Z_4$

$$Z_1 = \begin{cases} 1 & \text{graze level} = \text{high} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{graze level} = \text{intense} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{graze level} = \text{low} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_4 = \begin{cases} 1 & \text{graze level} = \text{medium} \\ 0 & \text{otherwise} \end{cases}$$

Qualitative Predictors

- After we fit the model, when `graze level = zero`,

$$E(abund) = 15.7 + 7.25 \times \logarea$$

- When `graze level = high`,

$$E(abund) = 15.7 + 7.25 \times \logarea - 1.59$$

Interactions in Regression Models

- Example (`paruelo.csv`): Paruelo and Lauenroth (1996) analyzed the geographic distribution and the effects of climate variables on the relative abundance of a number of plant functional types (PFTs) including shrubs, forbs, succulents (e.g. cacti), C3 grasses and C4 grasses. There were 73 sites across North America. The response variable we will focus on is the relative abundance of C3 plants and there were six potential predictors: the latitude in centesimal degrees (LAT), the longitude in centesimal degrees (LONG), the mean annual precipitation in mm (MAP), the mean annual temperature in °C (MAT), the proportion of MAP that fell in June, July and August (JJAMAP) and the proportion of MAP that fell in December, January and February (DJFMAP).

##		c3	lat	long	map	mat	jjamap	djfmmap
## 1		0.65	46.40	119.55	199	12.4	0.12	0.45
## 2		0.65	47.32	114.27	469	7.5	0.24	0.29
## 3		0.76	45.78	110.78	536	7.2	0.24	0.20
## 4		0.75	43.95	101.87	476	8.2	0.35	0.15
## 5		0.33	46.90	102.82	484	4.8	0.40	0.14
## 6		0.03	38.87	99.38	623	12.0	0.40	0.11
## 7		0.00	32.62	106.75	259	14.5	0.47	0.17
## 8		0.02	36.95	96.55	969	15.3	0.30	0.14
## 9		0.05	35.30	101.53	542	13.9	0.44	0.13
## 10		0.05	40.82	104.60	421	8.5	0.31	0.14

Interactions in Regression Models

- We standardized the variables lat and long and consider these three variables only
 - ▶ See <https://mjkeough.github.io/examples/paruelo.nb.html> check why we standardize the variables

##		c3	slat	slong
## 1	0.65	1.1872051	2.04335832	
## 2	0.65	1.3606917	1.22289867	
## 3	0.76	1.0702902	0.68058728	
## 4	0.75	0.7252028	-0.70393837	
## 5	0.33	1.2814913	-0.55631779	
## 6	0.03	-0.2327448	-1.09085967	
## 7	0.00	-1.4113220	0.05436524	
## 8	0.02	-0.5948037	-1.53061361	
## 9	0.05	-0.9059481	-0.75677099	
## 10	0.05	0.1349713	-0.27972344	

Interactions in Regression Models

- Now consider the MLR model $c3 \sim \text{slat} + \text{slong}$

```
##  
## Call:  
## lm(formula = c3 ~ slat + slong, data = paruelo)  
##  
## Coefficients:  
## (Intercept)          slat          slong  
##    0.271370    0.174743   -0.006027
```

- Note that the average effect on $c3$ of a one-unit increase in slat is always 0.174743, regardless of the value of slong .

Interactions in Regression Models

- But suppose that increasing *slong* actually increases the effectiveness of *slat*, so that the slope term for *slat* should increase as *slong* increases.
- Model takes the form

$$\begin{aligned}c_3 &= \beta_0 + \beta_1 \textit{slat} + \beta_2 \textit{slong} + \beta_3 (\textit{slat} \times \textit{slong}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \textit{slong}) \times \textit{slat} + \beta_2 \textit{slong} + \varepsilon\end{aligned}$$

```
##  
## Call:  
## lm(formula = c3 ~ slat + slong + slat * slong, data = paruelo)  
##  
## Coefficients:  
## (Intercept)          slat          slong    slat:slong  
##    0.2652689    0.2011131    0.0001836    0.0640657
```

Interactions in Regression Models

```
##
## Call:
## lm(formula = c3 ~ slat + slong + slat * slong, data = paruelo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39563 -0.14722 -0.01491  0.11837  0.40268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2652689   0.0222747   11.909 < 2e-16 ***
## slat         0.2011131   0.0245323    8.198 8.69e-12 ***
## slong        0.0001836   0.0225357    0.008  0.9935
## slat:slong    0.0640657   0.0242342    2.644  0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1893 on 69 degrees of freedom
## Multiple R-squared:  0.4964, Adjusted R-squared:  0.4745
## F-statistic: 22.67 on 3 and 69 DF, p-value: 2.525e-10
```

Interactions in Regression Models

Interpretation

- The results suggests that interactions are important.
- The p-value for the interaction term $slat \times slong$ is low (0.01), indicating that there is strong evidence for $H_a : \beta_3 \neq 0$.
- Adjusted R^2 for model without interaction is

```
## [1] 0.4295385
```

- Adjusted R^2 for model with interaction is

```
## [1] 0.4744966
```


Interactions in Regression Models

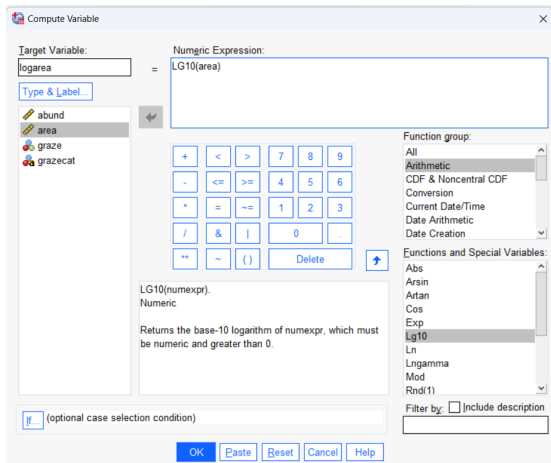
- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, `slat` and `slong`) do not.
- The **hierarchy principle**: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
 - ▶ The rationale for this principle is that interactions are hard to interpret in a model without main effects - their meaning is changed.

Interactions in Regression Models

- The concept of interactions applies just as well to qualitative variables, or to a combination of quantitative and qualitative variables.
- Especially if there is an interaction between a qualitative predictor and a quantitative predictor, we can interpret the interaction effects following the interpretation of the qualitative predictors. interpretation.

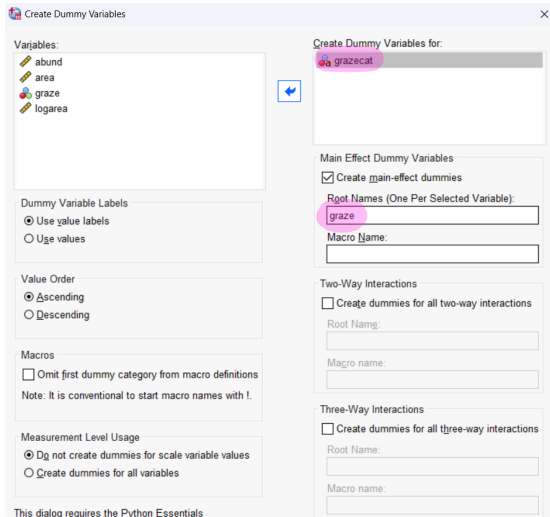
Lab 1

- Example (loyncat.csv): We re-analysed the data from Loyn (1987) by fitting a simpler model that just included grazing and log patch area.
- After importing the data to SPSS, we first add the new variable logarea, log10 transformation of area



Lab 1

- In SPSS when we fit a regression model, the independent variables must be numerical (scale). So we need to create dummy variables for a factor first.
- Click on Transform → Create Dummy Variables



The image shows the 'Create Dummy Variables' dialog box in SPSS. The 'Variables' list on the left contains 'abund', 'area', 'graze', and 'logarea'. The 'graze' variable is selected and moved to the 'Create Dummy Variables for:' list on the right. Under 'Main Effect Dummy Variables', the checkbox 'Create main-effect dummies' is checked, and the 'Root Names (One Per Selected Variable):' field contains 'graze'. Under 'Two-Way Interactions', the checkbox 'Create dummies for all two-way interactions' is unchecked. Under 'Three-Way Interactions', the checkbox 'Create dummies for all three-way interactions' is unchecked. The 'Macros' section has the checkbox 'Omit first dummy category from macro definitions' unchecked. The 'Measurement Level Usage' section has the radio button 'Do not create dummies for scale variable values' selected. A note at the bottom states 'This dialog requires the Python Essentials'.

Variables:

- abund
- area
- graze
- logarea

Dummy Variable Labels

- ☒ Use value labels
- ☐ Use values

Value Order

- ☒ Ascending
- ☐ Descending

Macros

- ☐ Omit first dummy category from macro definitions

Note: It is conventional to start macro names with !.

Measurement Level Usage

- ☒ Do not create dummies for scale variable values
- ☐ Create dummies for all variables

Create Dummy Variables for:

- grazecat

Main Effect Dummy Variables

- ☒ Create main-effect dummies

Root Names (One Per Selected Variable):

graze

Macro Name:

Two-Way Interactions

- ☐ Create dummies for all two-way interactions

Root Name:

Macro name:

Three-Way Interactions

- ☐ Create dummies for all three-way interactions

Root Name:

Macro name:

This dialog requires the Python Essentials

Lab 1

- From our discussions, we only need the first 4 dummy variables.

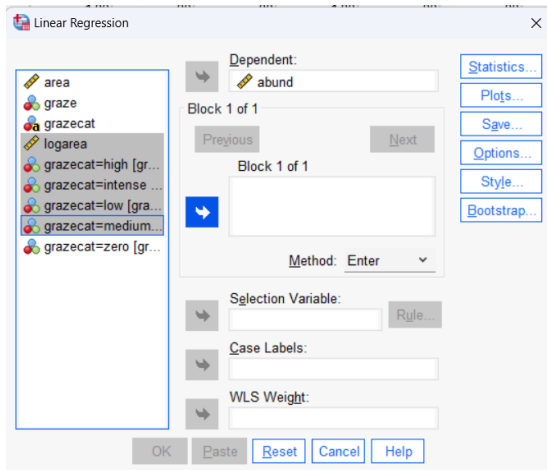
➔ **Create dummy variables**

Variable Creation

	Label
graze_1	grazecat=high
graze_2	grazecat=intense
graze_3	grazecat=low
graze_4	grazecat=medium
graze_5	grazecat=zero

Lab 1

- We now fit the regression model with logarea and the four dummy variables as independent variables.



Lab 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.853 ^a	.727	.700	5.8829

a. Predictors: (Constant), graze_4 grazecat=medium, logarea, graze_1 grazecat=high, graze_3 grazecat=low, graze_2 grazecat=intense

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4607.530	5	921.506	26.627	<.001 ^b
	Residual	1730.399	50	34.608		
	Total	6337.929	55			

a. Dependent Variable: abund

b. Predictors: (Constant), graze_4 grazecat=medium, logarea, graze_1 grazecat=high, graze_3 grazecat=low, graze_2 grazecat=intense

Lab 1

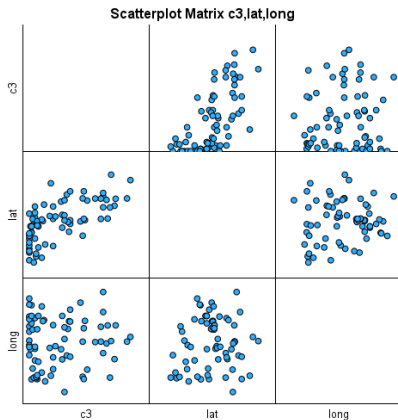
Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15.716	2.767		5.679	<.001
	logarea	7.247	1.255	.548	5.774	<.001
	graze_1 grazecat=high	-1.592	2.976	-.049	-.535	.595
	graze_2 grazecat=intense	-11.894	2.931	-.472	-4.058	<.001
	graze_3 grazecat=low	.383	2.912	.013	.131	.896
	graze_4 grazecat=medium	-.189	2.550	-.008	-.074	.941

a. Dependent Variable: abund

Lab 2

- Consider data `paruelo.csv`
- Import data, then create a scatter plot matrix of the variables `c3`, `lat` and `long`



Lab 2

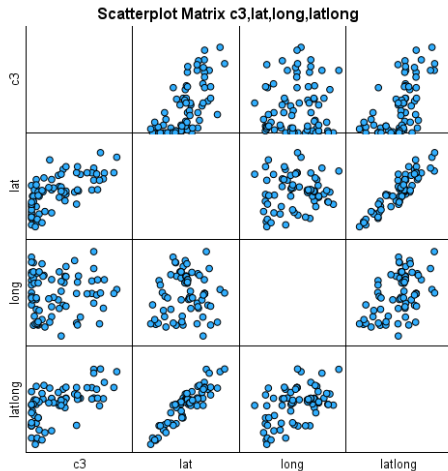
- Add a new column latlong which is lat*long

The image shows the SPSS 'Compute Variable' dialog box. On the left, a list of variables includes 'lat' and 'long'. The 'Target Variable' field contains 'latlong'. The 'Numeric Expression' field contains 'lat * long'. A list of functions is visible on the right, including 'All', 'Arithmetic', 'CDF & Noncentral CDF', 'Conversion', 'Current Date/Time', 'Date Arithmetic', and 'Date Creation'. The 'Functions and Special Variables' list includes '\$Casenum', '\$Date', '\$Date11', '\$JDate', '\$Sysmis', '\$Time', 'Abs', 'Any', 'Applymodel', and 'Arsin'. The 'Filter by' checkbox is unchecked, and the 'Include description' checkbox is checked. The 'OK' button is highlighted.

lat	long
46.40	119.55
47.32	114.27
45.78	110.78
43.95	101.87
46.90	102.82
38.87	99.38
32.62	106.75
36.95	96.55
35.30	101.53
40.82	104.60
47.75	102.50
33.48	105.55
33.33	99.23
45.33	99.10
39.10	96.60
41.55	101.80
43.73	112.67
44.25	112.15
29.58	102.92
43.53	103.45
31.60	110.50
50.70	107.72
32.97	101.18
36.05	97.23
41.87	113.08
32.00	109.12
43.75	102.33
38.17	113.25
37.10	111.87

Lab 2

- Scatter plot matrix again



Lab 2

- And correlation matrix as well

→ Correlations

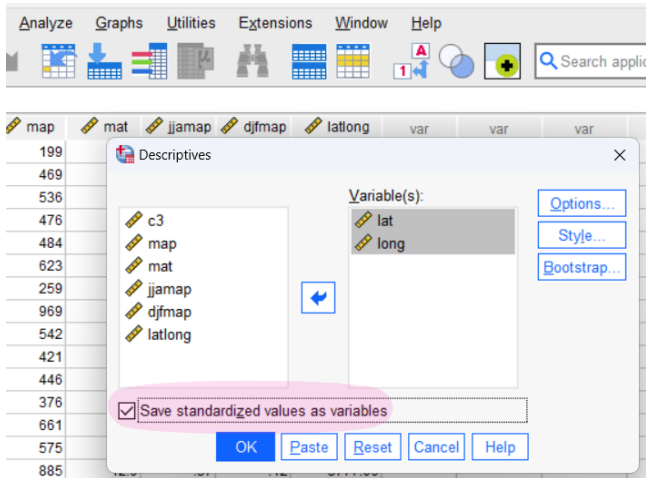
Correlations

		c3	lat	long	latlong
c3	Pearson Correlation	1	.667**	.042	.611**
	Sig. (2-tailed)		<.001	.727	<.001
	N	73	73	73	73
lat	Pearson Correlation	.667**	1	.097	.914**
	Sig. (2-tailed)	<.001		.416	<.001
	N	73	73	73	73
long	Pearson Correlation	.042	.097	1	.489**
	Sig. (2-tailed)	.727	.416		<.001
	N	73	73	73	73
latlong	Pearson Correlation	.611**	.914**	.489**	1
	Sig. (2-tailed)	<.001	<.001	<.001	
	N	73	73	73	73

** . Correlation is significant at the 0.01 level (2-tailed).

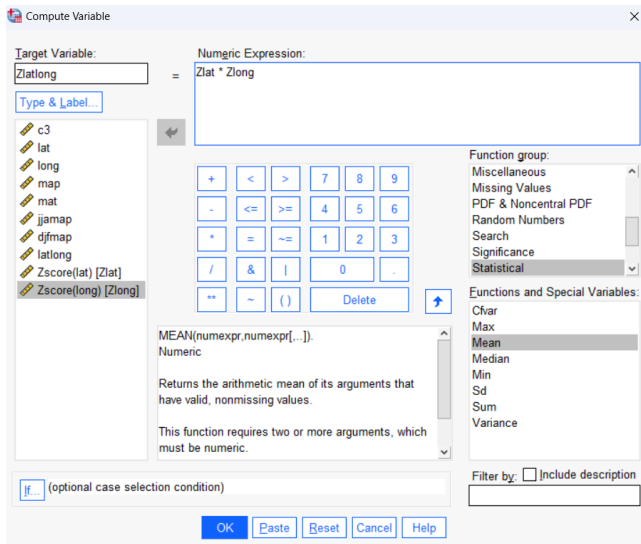
Lab 2

- Let's standardise the predictors to Zlat and Zlong by clicking on Analyze → Descriptive Statistics → Descriptives...



Lab 2

- Then add a new column `zlatlong` which is $Zlat * Zlong$



Lab 2

- Check the correlation matrix

Correlations

<i>Correlations</i>					
		c3	Zlat Zscore(lat)	Zlong Zscore(long)	Zlatlong
c3	Pearson Correlation	1	.667**	.042	-.069
	Sig. (2-tailed)		<.001	.727	.559
	N	73	73	73	73
Zlat Zscore(lat)	Pearson Correlation	.667**	1	.097	-.414**
	Sig. (2-tailed)	<.001		.416	<.001
	N	73	73	73	73
Zlong Zscore(long)	Pearson Correlation	.042	.097	1	-.134
	Sig. (2-tailed)	.727	.416		.257
	N	73	73	73	73
Zlatlong	Pearson Correlation	-.069	-.414**	-.134	1
	Sig. (2-tailed)	.559	<.001	.257	
	N	73	73	73	73

** . Correlation is significant at the 0.01 level (2-tailed).

Lab 2

- Finally, we fit the MLR model with interaction

Linear Regression

Dependent: c3

Block 1 of 1

Previous Next

Independent(s):

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

OK Paste Reset Cancel Help

Statistics... Plots... Save... Options... Style... Bootstrap...

lat long map mat jjamap djfmap latlong Zscore(lat) [Zlat] Zscore(long) [Zlong] Zlatlong

Lab 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.705 ^a	.496	.474	.18929

a. Predictors: (Constant), Zlatlong, Zlong Zscore(long), Zlat Zscore(lat)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.437	3	.812	22.670	<.001 ^b
	Residual	2.472	69	.036		
	Total	4.909	72			

a. Dependent Variable: c3

b. Predictors: (Constant), Zlatlong, Zlong Zscore(long), Zlat Zscore(lat)

Lab 2

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.265	.022		11.909	<.001
	Zlat Zscore(lat)	.201	.025	.770	8.198	<.001
	Zlong Zscore(long)	.000	.023	.001	.008	.994
	Zlatlong	.064	.024	.249	2.644	.010

a. Dependent Variable: c3

License



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#).